Computational Argumentation for the Automatic Analysis of Argumentative Discourse and Human Persuasion

RAMON RUIZ-DOLZ

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

VALENCIAN RESEARCH INSTITUTE FOR ARTIFICIAL
INTELLIGENCE (VRAIN)

A thesis submitted for the degree of
Doctor of Philosophy in Computer
Science

Supervised by
Dr. Stella Heras
Dr. Ana García-Fornes

- April 2023 -

Supervisors

Dr. Stella Heras
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (Spain)

Dr. Ana García-Fornes
Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (Spain)

External Reviewers

Dr. Carlos Ivan Chesñevar
Universidad Nacional del Sur, Argentina

Dr. Judith Masthoff
Utrecht University, Netherlands

Dr. Chris Reed
University of Dundee, United Kingdom

Thesis Defense Committee Members

Dr. Floriana Grasso
University of Liverpool, United Kingdom

Dr. Vicente Javier Julian Inglada
Universitat Politècnica de València, Spain

Dr. Chris Reed
University of Dundee, United Kingdom

Abstract

Computational argumentation is the area of research that studies and analyses the use of different techniques and algorithms that approximate human argumentative reasoning from a computational viewpoint. In this doctoral thesis we study the use of different techniques proposed under the framework of computational argumentation to perform an automatic analysis of argumentative discourse, and to develop argument-based computational persuasion techniques. With these objectives in mind, we first present a complete review of the state of the art and propose a classification of existing works in the area of computational argumentation. This review allows us to contextualise and understand the previous research more clearly from the human perspective of argumentative reasoning, and to identify the main limitations and future trends of the research done in computational argumentation. Secondly, to overcome some of these limitations, we create and describe a new corpus that allows us to address new challenges and investigate on previously unexplored problems (e.g., automatic evaluation of spoken debates). In conjunction with this data, a new system for argument mining is proposed and a comparative analysis of different techniques for this same task is carried out. In addition, we propose a new algorithm for the automatic evaluation of argumentative debates and we evaluate it with real human debates. Thirdly, a series of studies and proposals are presented to improve the persuasiveness of computational argumentation systems in the interaction with human users. In this way, this thesis presents advances in each of the main parts of the computational argumentation process (i.e., argument mining, argument-based knowledge representation and reasoning, and argument-based human-computer interaction), and proposes some of the essential foundations for the complete automatic analysis of natural language argumentative discourses.

Resumen

La argumentación computacional es el área de investigación que estudia y analiza el uso de distintas técnicas y algoritmos que aproximan el razonamiento argumentativo humano desde un punto de vista computacional. En esta tesis doctoral se estudia el uso de distintas técnicas propuestas bajo el marco de la argumentación computacional para realizar un análisis automático del discurso argumentativo, y para desarrollar técnicas de persuasión computacional basadas en argumentos. Con estos objetivos, en primer lugar se presenta una completa revisión del estado del arte y se propone una clasificación de los trabajos existentes en el área de la argumentación computacional. Esta revisión nos permite contextualizar y entender la investigación previa de forma más clara desde la perspectiva humana del razonamiento argumentativo, así como identificar las principales limitaciones y futuras tendencias de la investigación realizada en argumentación computacional. En segundo lugar, con el objetivo de solucionar algunas de estas limitaciones, se ha creado y descrito un nuevo conjunto de datos que permite abordar nuevos retos y investigar problemas previamente inabordables (e.g., evaluación automática de debates orales). Conjuntamente con estos datos, se propone un nuevo sistema para la extracción automática de argumentos y se realiza el análisis comparativo de distintas técnicas para esta misma tarea. Además, se propone un nuevo algoritmo para la evaluación automática de debates argumentativos y se prueba con debates humanos reales. Finalmente, en tercer lugar se presentan una serie de estudios y propuestas para mejorar la capacidad persuasiva de sistemas de argumentación computacionales en la interacción con usuarios humanos. De esta forma, en esta tesis se presentan avances en cada una de las partes principales del proceso de argumentación computacional (i.e., extracción automática de argumentos, representación del conocimiento y razonamiento basados en argumentos, e interacción humano-computador basada en argumentos), así como se proponen algunos de los cimientos esenciales para el análisis automático completo de discursos argumentativos en lenguaje natural.

Resum

L'argumentació computacional és l'àrea de recerca que estudia i analitza l'ús de distintes tècniques i algoritmes que aproximen el raonament argumentatiu humà des d'un punt de vista computacional. En aquesta tesi doctoral s'estudia l'ús de distintes tècniques proposades sota el marc de l'argumentació computacional per a realitzar una anàlisi automàtic del discurs argumentatiu, i per a desenvolupar tècniques de persuasió computacional basades en arguments. Amb aquestos objectius, en primer lloc es presenta una completa revisió de l'estat de l'art i es proposa una classificació dels treballs existents en l'àrea de l'argumentació computacional. Aquesta revisió permet contextualitzar i entendre la investigació previa de forma més clara des de la perspectiva humana del raonament argumentatiu, així com identificar les principals limitacions i futures tendències de la investigació realitzada en argumentació computacional. En segon lloc, amb l'objectiu de sol·lucionar algunes d'aquestes limitacions, hem creat i descrit un nou conjunt de dades que ens permet abordar nous reptes i investigar problemes prèviament inabordables (e.g., avaluació automàtica de debats orals). Conjuntament amb aquestes dades, es proposa un nou sistema per a l'extracció d'arguments i es realitza l'anàlisi comparativa de distintes tècniques per a aquesta mateixa tasca. A més a més, es proposa un nou algoritme per a l'avaluació automàtica de debats argumentatius i es prova amb debats humans reals. Finalment, en tercer lloc es presenten una sèrie d'estudis i propostes per a millorar la capacitat persuasiva de sistemes d'argumentació computacionals en la interacció amb usuaris humans. D'aquesta forma, en aquesta tesi es presenten avanços en cada una de les parts principals del procés d'argumentació computacional (i.e., l'extracció automàtica d'arguments, la representació del coneixement i raonament basats en arguments, i la interacció humà-computador basada en arguments), així com es proposen alguns dels fonaments essencials per a l'anàlisi automàtica completa de discursos argumentatius en llenguatge natural.

El desenvolupament d'aquesta tesi ha sigut possible gràcies al suport d'un ampli grup de persones les quals han estat al meu costat per a recolzar-me i guiar-me en els moments més complicats.

En primer lloc vull agrair als meus pares, els quals han estat al meu costat al llarg de tot el meu procés de desenvolupament com a la persona que sóc avui en dia. El seu paper ha sigut essencial per a arribar fins a aquest punt, i per suposat, per a donar els meus primers passos en el fascinant món de la recerca acadèmica.

En segon lloc m'agradaria expressar els meus agraïments a la meua parella per haver-me donat el seu suport incondicional en els moments més necessaris. Aquesta tesi representa el començament d'un projecte molt més gran al teu costat.

En tercer lloc vull agrair també als meus companys de laboratori. L'ambient agradable i col·laboratiu que he pogut experimentar al VRAIN m'ha permés disfrutar de tots i cada un dels anys en els quals he format part del laboratori i del grup d'investigació.

Finalment, m'agradaria acabar expressant els meus agraïments a les dues persones fonamentals en la definició i consolidació de la meua carrera acadèmica, les meues tutores. Des que vaig començar a treballar al VRAIN, el suport rebut per part d'Ana i de Stella ha sigut incommensurable. Sota la seua tutela excepcional ha sigut possible desenvolupar totes les idees i propostes presents en aquesta tesi doctoral.

Moltes gràcies a totes i tots.

Ramon

Contents

IV Argument-based Computational Persuasion   183

8 A Qualitative Analysis of the Persuasive Properties of Argumentation Schemes   185

9 Toward the Prevention of Privacy Threats: How Can We Persuade Our Social Network Platform Users?   209

10 Persuasion-enhanced Computational Argumentative Reasoning   235

V   Discussion   267

11 Discussion   269

VI Conclusion and Future Work   273

12 Conclusion and Future Work   275

Bibliography   279

List of Figures

List of Tables

Part I

Introduction and Objectives

Introduction and Objectives

"No! Try not. Do. Or do not. There is no try." – Yoda, Jedi Master.

One of the most popular questions that has arisen in the field of computer science in recent years is: "*Can computer systems be intelligent?*". However, before giving an answer to this question we should think about the own definition of intelligence. Does it consist of being able to automatically detect patterns and structures without explicit instructions on how to do it? Does it consist of processing information and reasoning about it in order to make a judgement? Or does it consist of interacting with human users in a way that they can understand, trust, and believe in what the computer system is saying? Probably, there is no unique answer to these questions, and each of them plays an important role in the definition of intelligence. In this PhD thesis, we approach some of these questions through the use and proposal of new computational argumentation techniques, and how can they be developed, implemented, and improved for the automatic analysis of argumentative discourse, and for engaging with human users in persuasive interactions. Thus, the research described in this thesis belongs to the intersection of argumentation theory and artificial intelligence, which we briefly define and contextualise below.

First, argumentation is an essential part of human communication, where the natural language and logics converge. Through arguments and argumentation, humans are able to communicate their logical reasoning in a coherent and understandable way to others. This way, argumentation can be seen as the process conducted by humans that involves the logical reasoning carried out inside our minds and the natural language that allows to communicate and express our thoughts with others (e.g., using arguments in a debate or writing an essay). Thus, it is common to partially dissociate these two main aspects of argumentation when doing research on the effect of arguments used by humans. Some typical aspects related to the natural language part of argumentation that have been investigated are: the language used, the expressiveness, the intonation, or the emotional response [254, 376, 385]. Conversely, the logical reasoning aspect of argumentation studies the different patterns that are used by humans in argumentative reasoning, the use of fallacies, enthymemes, and other structures commonly used in human discourses [286, 377, 388, 399]. Furthermore, argumentation plays a major role in different types of human dialogue such as negotiation, persuasion, and eristic dialogues among others, where the use of arguments allow to improve the human interaction experience [231]. Therefore, argumentation can be considered one of

the most significant manifestations of human intelligence.

On the other hand, Artificial Intelligence (AI) is the area of research that investigates the possibilities of computer systems approaching problems that involve reasoning, optimisation, and pattern recognition among others [330]. However, researchers have not yet reached a solid agreement on a general definition of AI. To simplify the understanding of AI and its definitions, in this thesis we will always refer to specific tasks that have been researched under the umbrella of AI such as Natural Language Processing (NLP), Knowledge Representation and Reasoning (KRR), and Human-Computer Interaction (HCI). In fact, we might be far from developing an Artificial General Intelligence (AGI), but where state-of-the-art AI systems excel is in approaching these specific tasks that were previously unapproachable. Recent advances in Deep Learning (DL) represent a leap forward in AI research, and made possible to significantly improve the results obtained in these tasks by previous approaches (e.g., symbolic reasoning) [210]. However, DL has its own limitations. Even though DL algorithms are able to learn very good implicit representations of data hard to explicitly encode by humans (e.g., text, audio, or images), the outputs provided by these algorithms are sometimes hard to explain and understand, and have a strong dependency on the distributions present in the data used to train them [408]. Moreover, DL algorithms provide outstanding results on specific tasks and problems, but are not usually designed to perform, learn, and understand a heterogeneous set of tasks.

Given the relevance of argumentation in human reasoning and communication, and the latest advances and limitations in AI related tasks such as NLP, KRR, or HCI, computational argumentation has emerged again as one of the most promising multidisciplinary areas of research underlying AI. Computational argumentation is a relatively new field of study that was born in the intersection of AI, graph theory and formal logic inspired by concepts studied in philosophy, linguistics and dialectics. Originally, most of the research was carried out following a formal approach, considering the logical elements and structures that underlie argumentation. However, recent results obtained in different NLP tasks by DL algorithms have sparked interest in an informal approach to computational argumentation research. Such is the case of all the research done in argument mining [207], assessment of natural language arguments [343], or natural language argument generation [173] tasks. The previously described dissociation between aspects of human

argumentation can also be observed in the computational argumentation area of research. A rich variety of disciplines converge in this topic in order to study argumentation from the computational viewpoint: NLP, formal logic, HCI, behavioural studies, psychology, etc. Therefore, when analysing research in computational argumentation, it is possible to observe a collection of work in different directions that are sometimes difficult to relate to each other.

With the research framed into this PhD thesis, our main objective is to bridge the gap between the heterogeneous advances and proposals in computational argumentation research, and to leverage the existing synergies between different approaches and paradigms. For that purpose, we have firstly conducted a thorough review of the state of the art in computational argumentation, and proposed a classification of the reviewed works from the human reasoning viewpoint. This classification has allowed us to identify the main limitations and needs in the computational argumentation area of research which have served as the main motivation of the present PhD thesis. Furthermore, we have been able to hypothesise the potential synergies and common points between different approaches in the literature that we have developed and proposed in this thesis. Therefore, in the works included in this compendium we have contributed to some of the major limitations in argument mining, argument-based KRR, and argument-based HCI, the three pillars of computational argumentation. This way, it has been our objective to investigate if it is possible to improve the performance of argumentation systems and to explore new aspects of argumentation by combining argumentation theory concepts, NLP algorithms, and behavioural studies, rather than relying on a particular concept exclusively.

1.1 Motivation

Research in computational argumentation focuses on approaching the different phases of human argumentative reasoning from the computational viewpoint. These phases are: (i) the *identification* of new arguments, (ii) the *analysis* of the identified arguments, (iii) the *evaluation* of the argumentative structures, and (iv) the *invention* of new arguments [395]. Through the *identification*, humans detect argumentative elements in natural language sources (e.g., when participating a debate,

or when reading an essay). During the *analysis*, we find relations between the argumentative elements and provide an argumentative structure to the previously identified elements. In the *evaluation*, humans assign different strength to the arguments depending on a varied set of factors such as their personal preferences, coherence, culture, beliefs, etc. Finally, the *invention* part of the argumentative process consists of using all the available information to create new arguments and structure them to support some specific conclusion and elaborate a consistent discourse. Therefore, it is possible to observe strong influences of these phases of human argumentation on how computational argumentation research has been structured and carried out. For example, between research in argument mining and the *identification/analysis* of arguments; between argument-based KRR and the *analysis/evaluation* of arguments; and between the research in the automatic generation of natural language arguments and argument-based HCI, and the *invention* phase of human argumentation. The identification of these influences and relations between areas such as argument-based NLP and KRR or between argument-based KRR and HCI is, in fact, one of the objectives of this thesis, which we developed and analysed in-depth.

It is common in AI research (in particular DL) to focus on a specific problem or task, and trying to improve state-of-the-art results in this specific aspect. However, having a broader viewpoint of an heterogeneous area of research such as computational argumentation, helps to understand what is done and what is not, and what are the most important challenges and limitations. It is important to note that not every phase of human argumentation needs to be approached when doing research in computational argumentation. In some situations, identifying arguments and structuring them is enough if we only want to approach the automatic analysis of argumentation. In some other cases, the evaluation of natural language argumentative features could also be enough if we only want to automatically assess arguments. Considering our analogy with human argumentation, we do not always invent or evaluate arguments after identifying and analysing them. Sometimes we just want to retrieve argumentative information and structure it in our minds, and sometimes we draw conclusions by evaluating these arguments. However, at the end, a relation between the different phases of argumentation should exist if we want our argumentative reasoning to be consistent and coherent.

In the computational argumentation area of research, this typical practise of DL

research approaching very specific tasks and problems can also be observed, since most of the latest DL-based contributions are aimed at addressing a specific task and improving the previous results achieved on it [314]. It is important to improve state-of-the-art results in specific argumentative tasks, since this will make possible to extract richer information, to have better data representations, and to improve the interaction with human users. However, overlooking the whole argumentative process and the possible existing synergies between contributions in different disciplines could cause a bottleneck in the advances achieved by research aimed at approaching the human argumentative reasoning from a computational viewpoint. A great example of the relevance of exploring synergies can be observed in the recent research framed in the *Project Debater*, carried out by IBM [345], were researchers and engineers present an autonomous debating system designed to address the entire argumentative process. This system was able to achieve competitive results in a live debate tournament. We have observed in this project a strong focus on natural language aspects of arguments and a concerning overlook of explicit argumentative underlying logic and human reasoning patterns. This can be an important issue, specifically in a context where data biases are directly reflected in the outputs of the algorithms. For example, when determining the *strength* of an argument looking exclusively at the natural language of the argument and disregarding the logical structure and validity of it explicitly. If only the language probabilistic distribution is brought into consideration to model arguments, it is possible that the strongest argument is the most frequent one, or the strongest for some specific "*dominant*" audience in the available corpora, but not because of its logical soundness.

The research conducted in this thesis has been carried out to overcome some of these relevant needs and limitations in computational argumentation research. First, we create and release new natural language text and speech corpora for doing research in the automatic analysis of human argumentation in professional debates. Second, we improve the understanding and performance of algorithms approaching natural language argumentative tasks. Third, we address new computational argumentation problems that have not been researched yet, such as the automatic evaluation of professional spoken argumentative debates. And fourth, we conduct a transversal investigation of the common points between some of the different disciplines underlying computational argumentation research, (i.e., deep

learning, formal argumentation theory, and the psychology of persuasion), and the benefits of considering them as a whole instead focusing exclusively on a specific paradigm.

We propose, therefore, a more cohesive perspective in the area of computational argumentation, where NLP algorithms, formal argumentation approaches, and human behavioural studies can be understood as parts of a whole rather than as mutually exclusive alternatives. For that purpose, all the experimentation has been contextualised and supported by the research projects: *Intelligent Agents for Privacy Advice in Social Networks* (TIN2017-89156-R), *Technologies for Human Emotional Organisations* (PROMETEO/2018/002), *Affective Intelligent Agents for Persuading Civic Behaviour in Virtual Environments* (ID2020-113416RB- I00), and a 6-month internship on an argument mining project at the National Institute of Informatics in Japan. This support and project involvement has motivated our research in the domains of professional debates, political argumentation, and privacy magament in OSNs. Thus, we explored and proposed the use of new algorithms for the analysis of natural language human argumentative discourses in debate tournaments, the automatic extraction of arguments in political debates, and we analysed and improved the computational argumentative reasoning for HCI aimed at educating in privacy management and preventing privacy violations in OSNs. The consideration of these specific use cases has been of utmost importance to clearly observe the power of computational argumentation and its relevance for human-oriented AI research.

1.2   Objectives

Taking into account the motivation presented above, we proposed three major objectives for this PhD thesis. With the work framed in this thesis, we want to do significant advances in two important aspects of computational argumentation research: (i) the automatic analysis of human argumentative discourse, and (ii) the improvement of computational persuasion through the use of arguments when interacting with human users. Therefore, we define our objectives as follows:

1. Review the state of the art in computational argumentation from the human reasoning perspective.

    1.1 Identify and analyse the main contributions to computational argumentation.

    1.2 Classify these contributions in an easy-to-understand structure from the human argumentative reasoning viewpoint.

    1.3 Identify the actual challenges and limitations in computational argumentation research.

    1.4 Analyse the future of computational argumentation research.

2. Propose new techniques for the automatic analysis of human argumentative discourses.

    2.1 Create a new corpus of natural language argumentation that enables the automatic analysis of complete structured argumentative debates.

    2.2 Analyse, and evaluate new algorithms for argument mining in different domains and languages.

    2.3 Propose a new transversal method for approaching the automatic evaluation of argumentative debates that combines concepts from argumentation theory and NLP.

3. Improve the persuasive human-computer interactions through the use of arguments and computational argumentation reasoning.

    3.1 Analyse the persuasive properties of human argumentative reasoning.

    3.2 Study and evaluate the persuasive power of arguments when used with different user models.

    3.3 Integrate argumentation theory and computational persuasion concepts into a framework that enables a user-tailored interaction with different arguments aimed at improving their persuasiveness.

1.3   Structure of the Thesis

This PhD thesis is structured into six parts that have been designed and organised as follows:

- **Part I: Introduction and Objectives**. The first part of this thesis presents the introduction and the motivation of the research work carried out. Furthermore, we provide a detailed list with the research objectives defined at the beginning, a list of the academic publications produced, and research projects that supported this PhD thesis.

- **Part II: Preliminaries and Literature Review.** In this part, the first objective is addressed. We present a complete analysis of the research in computational argumentation from a human reasoning perspective. All the basic concepts in computational argumentation are summarised, and a general classification aimed at identifying the main contributions, limitations, and needs in computational argumentation research is provided.

- **Part III: Automatic Analysis of Argumentative Discourse.** In this part, the second objective is addressed. The most significant contributions resulting of the research conducted in this thesis aimed at analysing the human argumentative discourse are described. A brief description of the chapters included in this part is presented below:

  Chapter 3. *VivesDebate*, our annotated multilingual corpus of argumentation in a debate tournament and its annotation process are thoroughly described in this chapter. A comparison and analysis of existing argumentative corpora is also carried out to contextualise and understand the advantages that presents the *VivesDebate*.

  Chapter 4. An original algorithm for segmenting and classifying arguments is described in this chapter. The algorithm was submitted to an international shared-task achieving competitive results.

  Chapter 5. A complete analysis of Transformer-based architectures for automatically identifying argumentative relations in multiple domains is presented in this chapter. State-of-the-art results are reported using these architectures and an interesting degree of robustness to domain variance is also observed in the experiments.

  Chapter 6. A hybrid algorithm for evaluating argumentative debates is described in this chapter. The proposed algorithm combines concepts from

argumentation theory and NLP, and presents promising results in a rather under-researched task such as the automatic assessment of human debates.

Chapter 7. *VivesDebate-Speech*, an extension of the original *VivesDebate* corpus is proposed. Considering all the observed experimental limitations, we release an speech extended version of the original corpus. We also conduct first-of-its-kind experiments approaching the argument segmentation combining audio and text features.

- **Part IV: Argument-based Computational Persuasion.** In this part, the third and last objective is addressed. The most significant contributions resulting from the research conducted in this thesis aimed at understanding and improving computational persuasion techniques through the use of arguments are described. A brief description of the chapters included in this part is presented below:

  Chapter 8. An analysis of the persuasive principles behind some of the most common patterns of human argumentative reasoning is presented in this chapter. From the results of this study it is possible to understand the persuasive properties that are implicit to some argumentative reasoning structures.

  Chapter 9. A study and evaluation of the persuasive power of arguments when used to educate teenagers in managing their privacy in online social networks is described in this chapter. From the results of this study, it is possible to observe which arguments in terms of their structure and content are more persuasive, and how variations in user models correlate with variations in the persuasiveness of arguments.

  Chapter 10. A new theoretical framework for human persuasion that relies in concepts from argumentation theory is proposed in this chapter. The framework is evaluated and validated through a study with human participants that shows an improved persuasiveness thanks to the user modelling combined with computational argumentative reasoning.

- **Part V: Discussion.** In this part, a discussion of the observed results and its implications is conducted.

- **Part VI: Conclusion and Future Work.** Finally, this part summarises the most important conclusions of our research. Furthermore, it also describes the open challenges that remain unexplored at the end of this thesis.

1.4 List of Publications

During the development of this thesis, several academic contributions have been published. In the following lists, all the produced papers are grouped in the categories JCR academic journals, GII-GRIN-SCIE international conferences, and other international conferences and workshops. The publications marked with an asterisk (∗) have been included as a chapter in this PhD thesis.

- Journals listed in the JCR:

    - (∗)**Ramon Ruiz-Dolz**, J. Alemany, S. Heras, and A. García-Fornes. Toward the Prevention of Privacy Threats: How Can We Persuade Our Social Network Platform Users? . Human-centric Computing and Information Sciences (HCIS), Accepted, 2022. *Impact Factor: 6.558, Q1*

    - (∗)**Ramon Ruiz-Dolz**, J. Alemany, S. Heras, and A. García-Fornes. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. IEEE Intelligent Systems, 36(6), (pp. 62-70), 2021. *Impact Factor: 6.744, Q1*

    - (∗)**Ramon Ruiz-Dolz**, M. Nofre, M. Taulé, S. Heras, and A. García-Fornes. VivesDebate: A New Annotated Multilingual Corpus of Argumentation in a Debate Tournament. Applied Sciences, 11(15), 7160, 2021. *Impact Factor: 2.838, Q2*

- International conferences listed in the GII-GRIN-SCIE:

    - (∗)**Ramon Ruiz-Dolz**, J. Taverner, S. Heras, A. Garcia-Fornes, and Vicente Botti. A Qualitative Analysis of the Persuasive Properties of Argumentation Schemes. In Proceedings of the *30th ACM Conference on User Modeling, Adaptation and Personalization*, (pp. 1-11), 2022. *Class: 3 (CORE B)*

- A. Brännström, T. Kampik, **Ramon Ruiz-Dolz**, and J. Taverner. A Formal Framework for Designing Boundedly Rational Agents. In Proceedings of the *International Conference on Agents and Artificial Intelligence (3)*, (pp. 705-714), 2022. *Class: 3 (CORE B)*

- **Ramon Ruiz-Dolz**. Towards an artificial argumentation system. In Proceedings of the *29th International Conference on International Joint Conferences on Artificial Intelligence - Doctoral Consortium Track*, (pp. 5206-5207), 2021. *Class: 1 (CORE A++)*

- Other international conferences:

  - (∗)**Ramon Ruiz-Dolz**. A Cascade Model for Argument Mining in Japanese Political Discussions: the QA Lab-PoliInfo-3 Case Study. In Proceedings of the *16th NTCIR Conference on Evaluation of Information Access Technologies*, (pp. 175-180), 2022.

  - **Ramon Ruiz-Dolz**, J. Alemany, S. Heras, and A. García-Fornes. Automatic Generation of Explanations to Prevent Privacy Violations. In Proceedings of the *2nd EXplainable AI in Law Workshop*, 2019.

  - **Ramon Ruiz-Dolz**, S. Heras, J. Alemany, and A. García-Fornes. Towards an Argumentation System for Assisting Users with Privacy Management in Online Social Networks. In Proceedings of the *19th Workshop on Computational Models of Natural Argument*, (pp. 17-28), 2019.

1.5 List of Research Projects

The research carried out in this thesis is framed within the following academic projects, which have served as the main source of financial support for this work:

- Intelligent Agents for Privacy Advice in Social Networks

  - *Funder:* Ministerio de Economia y Empresa (TIN2017-89156-R).

  - *Lead Applicant:* E. Argente, and A. García-Fornes.

  – *Years:* 2018 - 2021.

- Technologies for Human Emotional Organisations

  – *Funder:* Generalitat Valenciana (PROMETEO/2018/002).
  – *Lead Applicant:* V. Botti.
  – *Years:* 2018 - 2021.

- Affective Intelligent Agents for Persuading Civic Behaviour in Virtual Environments

  – *Funder:* Ministerio de Economia y Empresa (PID2020-113416RB-I00).
  – *Lead Applicant:* E. Argente, and A. Espinosa.
  – *Years:* 2021 - 2023.

1.5. LIST OF RESEARCH PROJECTS

Part II

Preliminaries and Literature Review

A Structured Review and Outlook of Computational Argumentation Research from a Human Reasoning Perspective

RAMON RUIZ-DOLZ, STELLA HERAS AND ANA GARCÍA-FORNES

Abstract

Computational Argumentation studies how human argumentative reasoning can be approached from a computational viewpoint. Human argumentation is a complex process that has been studied from different perspectives (e.g., philosophical or linguistic) and that involves many different aspects beyond pure reasoning. The heterogeneity of human argumentation is present in Computational Argumentation research, in the form of various tasks that approach the main phases of argumentation individually. With the increasing interest of researchers in Artificial Intelligence, we consider that is of great importance to provide coherence to the Computational Argumentation research area. Thus, in this paper, we present a general viewpoint of Computational Argumentation, from the perspective of how human argumentation has been approached by computer systems. For that purpose, the following contributions are produced: (i) a solid structure for Computational Argumentation research mapped with the human argumentation process; (ii) a collective understanding of the tasks approached by Computational Argumentation and their synergies; (iii) a thorough review of the most important advances in each of these tasks; and (iv) an analysis and a classification of the future trends in Computational Argumentation research and the most relevant open challenges in the area.

2.1 Introduction

Man is, by nature, a social animal. As pointed out by classic philosophers such as Seneca [341] or Aristotle [25], social interaction with other beings is an intrinsic feature of humans. These social interactions can be manifested in many different ways such as cooperation, competition, or conflict, among others. However, one of the most important commonalities of social interaction manifestations is the human capacity of speech. Human beings use speech as a means of social interaction, since it represents an effective way of communication and thereby, interaction. Argumentation is the process by which humans structure the discourse in a discussion or a debate. With the use of argumentation, debaters can elaborate their reasons and supports towards a specific idea, and draw appropriate conclusions on some topic. Therefore, argumentation can be considered one of the structural pillars of human speech.

Argumentative discourse has been studied and analysed for a long time, from Ancient Greek philosophers such as Aristotle to the present by scholars, philosophers, linguists and psychologists. The multiple perspectives on the analysis of arguments are reflected in the variety of definitions and interpretations of argument models and rhetoric (e.g., Aristotelian, Rogerian, Toulmin, etc.) [221, 44, 372]. Computational Argumentation is where all of these concepts merge with the theory of computation. Computational models of argument [28, 320] are argumentation models that are created and designed to be "*understood*" and interpreted by computers. Thus, Computational Argumentation is the branch of Artificial Intelligence where computational models of arguments are used to approximate human argumentative reasoning from the computational viewpoint [50]. However, the different interpretations of arguments and human argumentation are reflected in an heterogeneous set of research approaching the most important aspects of Computational Argumentation from different perspectives. We find it important to provide coherence and a structure to all of these research works, in a context in which interest in Artificial Intelligence and its underlying areas is experiencing strong growth.

CHAPTER 2. COMPUTATIONAL ARGUMENTATION FROM A HUMAN REASONING PERSPECTIVE

In this paper, we present an analysis of Computational Argumentation research from the perspective of how computers can approach the different phases of human argumentation through the most relevant contributions and advances. Previous review work in this area has been focused either on a specific application domain (e.g., legal, decision making) [291, 262], on a particular task of the Computational Argumentation process (e.g., argument mining, argument-based knowledge representation and solving) [274, 207], or on practical aspects of Computational Argumentation (e.g., argumentation systems, machine learning in argumentation) [101, 84, 98]. However, none of them provides a general structure of the whole Computational Argumentation research area. Our work differs from others since it analyses Computational Argumentation from a general perspective, inspired by the different phases that shape human argumentation, and provides a global structure to the whole Computational Argumentation process.

Therefore, with this work we present a structured review and outlook for the main research work conducted in each task of Computational Argumentation. In this paper, an attempt has been made to follow a line of argument that captures the different views and conceptions of the terms in the Computational Argumentation area of research by the different research communities involved in it. As contributions, the objectives of this article are to do the following:

- Structure Computational Argumentation research into the three main tasks of Argument Mining, argument-based Knowledge Representation and Reasoning, and argument-based Human-Computer Interaction.

- Propose a mapping between the three main tasks of Computational Argumentation and the structure of human argumentative reasoning.

- Provide a holistic viewpoint of Computational Argumentation research, and the existing synergies among its different disciplines.

- Review and classify the most significant breakthroughs in the Computational Argumentation area of research.

- Present the recent trends in Computational Argumentation research and the most relevant open challenges in the area.

Thus, our review work serves as a guideline to the complex, heterogeneous, and multidisciplinary research area of Computational Argumentation. With the proposed perspective based on the human argumentative reasoning process, we create a simple but effective scheme of Computational Argumentation research that is easy to follow and understand. We also make it easier to contextualise the diverse research conducted in the different disciplines framed into Computational Argumentation for researchers that might not be familiar with computer science. This is of utmost importance in a multidisciplinary field such as Computational Argumentation, since researchers from different disciplines (e.g., philosophy, linguistics, logic) converge, having human argumentation as the context in common. Furthermore, with our work, we reference a large set of task-specific survey papers so that the reader can easily delve into any of the main tasks of Computational Argumentation.

The rest of the article is structured as follows. Section 2.2 depicts the whole Computational Argumentation process and provides a schematic representation of all the phases involved in it. Section 2.3 presents the Argument Mining task and reviews the most important published research. Section 2.4 focuses on the argument representation and solving concepts of Computational Argumentation. The main approaches proposed in the literature to provide argument knowledge representations are reviewed, together with the concept of argument acceptability and its underlying argumentation semantics. Section 2.5 reviews the Human Computer-Interaction side of Computational Argumentation, specifically focusing on the automatic generation of arguments and the argument-based computational persuasion. Section 2.6 presents some future research directions that have recently raised their interest in the community, or that have not been thoroughly researched yet, and analyses the remaining open challenges of the reviewed work. Finally, Section 2.7 summarises the most important ideas reached after an extensive analysis of the main tasks of Computational Argumentation research.

2.2   The Computational Argumentation Process

Computational Argumentation studies the integration of the argumentative hu-
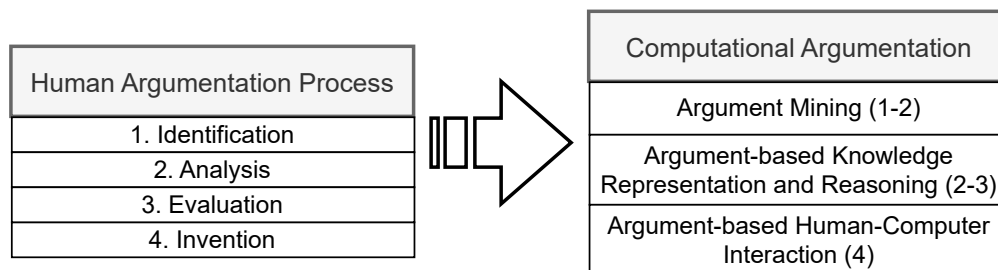
Figure 2.1: Relation between the phases of human argumentative reasoning and
the Computational Argumentation research.

man reasoning process into the knowledge extraction, representation and process-
ing of computational intelligent systems. The process by which humans apply
argumentative reasoning is divided into four different phases: identification, anal-
ysis, evaluation, and invention [395]. The first phase is the *identification* of argu-
ments and argumentative elements in the discourse. When debating or dialoguing,
in order to elaborate a coherent rebuttal, humans try to identify the main conclu-
sions and the premises that lead to them. Once these argumentative elements are
identified, humans *analyse* the following: the underlying argumentative patterns
and structures; how the identified premises and conclusions are related; or even
whether there are implicit premises that could be relevant for understanding the
argumentative reasoning (i.e., enthymemes). The *evaluation* of the identified argu-
mentative elements and its structures is the next phase in the human argumentative
reasoning process. Humans measure the strength of arguments based on many
different factors, such as their knowledge of some specific topic, their personal
preferences (i.e., values) due to belonging to a specific audience, or the coherence
of the argumentative discourse, among others. Finally, humans *invent* new argu-
ments and structure them in order to support and prove some specific conclusion,
which can be either a rebuttal to a previous conclusion in a debate, or a new idea
to be developed.

The first step in this paper is to investigate how these phases of human argu-
mentative reasoning are related to Computational Argumentation research. Sev-
eral lines of research have appeared along with Computational Argumentation,
approaching the different phases of human argumentative reasoning. We have clas-
sified the different phases of the human argumentation process as they have been

translated into Computational Argumentation, identifying three main groups of related tasks. First, **argument mining** research explores how a computer program can automatically identify and extract argumentative elements and their relations from a piece of text or speech (e.g., newspapers, political debates, legal documents, etc.) [265]. Second, we identify an important amount of work on **argument-based knowledge representation and reasoning**. This line of research is mainly focused on investigating how we can provide intelligent systems with frameworks, easing the computational representation of arguments and their evaluation. These representation frameworks can be either abstract [124] or structured [63] depending on how the arguments are instantiated. Furthermore, argumentation frameworks allow us to perform a logical and dialectical evaluation of arguments with the use of semantics [38]. This evaluation of the represented knowledge (i.e., arguments) is an approximation of the rational reasoning carried out by humans in the argumentative process. Finally, the third major line of research in Computational Argumentation is the **argument-based human-computer interaction**. Most of the research grouped here focuses on the automatic generation of arguments and their persuasive properties. The generation of arguments, either template-based or natural language generated, allows an argumentative intelligent system to automatically create new arguments from its knowledge on some specific topic and display them in a way that is understandable to humans (e.g., written or spoken language). The study of the persuasive properties of arguments within argument-based computational persuasion covers an important amount of the contributions considering the human-computer interaction part of argumentation research. The research on the persuasive properties of arguments sheds light on which argument or reasoning pattern is the most suitable in order to persuade people, taking into account a specific topic or domain. Thus, it is possible to improve the trustworthiness of computational intelligent systems. Figure 2.1 depicts the proposed mapping between the human argumentative reasoning and the Computational Argumentation processes.

We can observe how all of the fundamental elements of human argumentative reasoning have been approached by, at least, one of the identified tasks underlying Computational Argumentation. The identification phase of the argumentative process is completely tackled by Argument Mining. Therefore, every Computational Argumentation system that is designed to automatically retrieve or analyse argu-

ments will compulsorily carry out the Argument Mining task. The analysis phase
is usually approached by both Argument Mining and argument-based knowledge
representation techniques. This phase involves arguments and their structures, so
the automatic retrieval of argumentative relations and how argumentative struc-
tures are computationally represented are framed into the analysis of arguments.
The evaluation of arguments is usually approached by the works proposed in the
context of argument-based knowledge representation and reasoning. The study
of argument representations and the definition of argument sets (e.g., acceptable
arguments) in these representations allow the computational evaluation of argu-
ments. Finally, the creation of new human language shaped arguments from the
previous computational representations is approached by the works grouped under
the argument-based human-computer interaction. Research on this topic explores
how arguments can be automatically generated and how persuasive the arguments
can be in different domains and for different audiences.

The evaluation of arguments has been approached in the literature from differ-
ent perspectives, as analysed in [390], where multiple dimensions of the quality of
computational arguments are proposed. The argument evaluation research grouped
under the argument-based knowledge representation and reasoning mainly focuses
on logical and dialectical (i.e., rational) aspects of argumentation. The computa-
tional persuasion aspect of argumentation could also be classified into the assess-
ment of the rhetorical effectiveness of arguments, considering non-rational aspects
such as human emotions. Therefore, in this survey, we have considered the analy-
sis of argument-based computational persuasion as a cornerstone of the invention
phase of human argumentation. This decision has been made because of the rele-
vance of human features in the reviewed research (i.e., personality, emotions, etc.)
and due to the importance of human persuasion when defining new dialogue strate-
gies and coordinating the generation of new arguments. Furthermore, there is also
a big conceptual gap between the research on the logical and dialectical evaluation
of arguments (i.e., argument-based knowledge reasoning) and the research on the
persuasive analysis of arguments (i.e., argument-based computational persuasion),
which leads to the literature classification proposed in this survey.

Depending on its domain, its requirements, or its purposes an argumentation
system may not deal with the complete process, but only some specific underlying
tasks. For instance, an argumentation system aimed at assisting in the analysis of

argumentative text may only require the Argument Mining task of Computational Argumentation. In this case, the only phases of human argumentative reasoning needed to achieve the purposes are identification and analysis. On the other hand, a complete argumentation system aimed at debating complex topics with humans such as the IBM project debater[1] [345] must effectively approach the complete Computational Argumentation process. In this case, the system needs to be able to automatically identify, analyse, evaluate, and invent arguments in order to effectively carry out a complete debate. Let us illustrate the complete process with a simple example. Consider the four arguments depicted in Table 2.1. An artificial argumentation system must be able to automatically identify the four arguments A1, A2, A3, and A4. Furthermore, the attacks between A1 and A2 with each other, the attack from A3 to A2, and the attack from A4 to A3 must also be automatically detected. Thus, a simple graph representation consisting of four nodes (i.e., arguments) and three edges (i.e., attack relations) is internally generated in order to ease the evaluation of the detected arguments. When presenting this information to a human, the argumentation system should be able to either depict the structure of the previously analysed arguments or to generate a new argument such as "*We should pay our taxes since it is our main way to contribute to society's needs. Furthermore, we can democratically choose how the taxes will be managed every four years.*", which is created using the most solid conclusions from the previously identified arguments. Some aspects of this new piece of text, such as the structure or its own content need to be considered in the process of argument generation in order to improve its persuasive capacities when used to convince different audiences. With this example, we have briefly depicted how all tasks of the Computational Argumentation process would be carried out in a simple situation. However, this is a highly complex process, whose specifics are reviewed in the following sections.

2.3   Argument Mining

Argument Mining is the automatic identification and retrieval of arguments in

---

[1]https://www.research.ibm.com/artificial-intelligence/
project-debater/

Argument ID	Natural Language Argument
A1	*"We should pay our taxes since it is an important way to contribute to the improvement of society's needs such as public health or education."*
A2	*"We should not pay any taxes as politicians do not manage them properly."*
A3	*"This is not a problem caused by the taxes themselves, but by the people who manage them. Every four years citizens can vote and choose different politicians if they are not satisfied."*
A4	*"The inefficient management of resources does not depend on the political party. We will not be able to overcome this problem by only voting for different politicians."*

Table 2.1: Four natural language arguments used to illustrate the different tasks of Computational Argumentation.
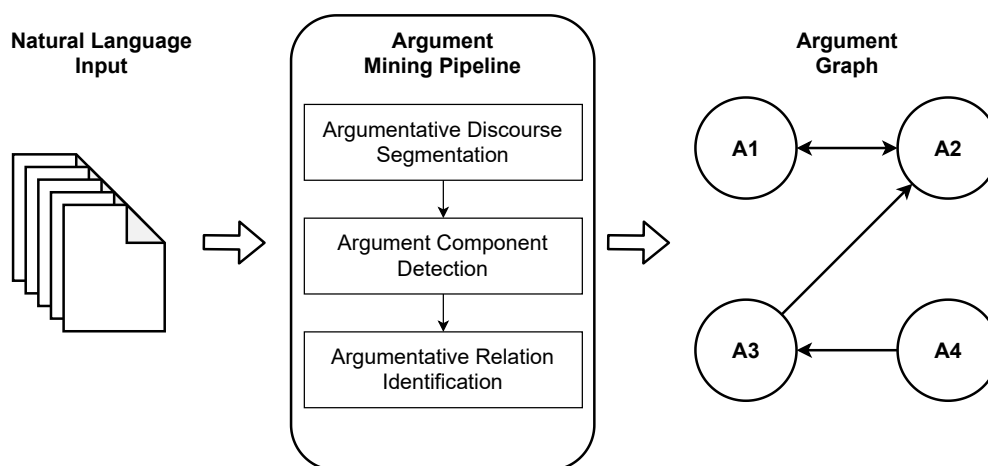
Figure 2.2: Structure of the complete Argument Mining pipeline.

any type of natural language source, and their structuring by approximating human reasoning [265]. Argument Mining research focuses on the definition and evaluation of new algorithms which allow an intelligent system to tackle the identification and the structural analysis phases of the human argumentative reasoning process. In this section, we focus on providing a clear and schematic perspective of Argument Mining rather than presenting a thorough review of the area (for which we refer the reader to the most recent literature reviews of this area [355, 71, 207]).

Research on Argument Mining can be broken down into three main sub-tasks: argumentative discourse segmentation, argument component detection, and automatic identification of argumentative relations between propositions. Thus, the

aim of the complete Argument Mining pipeline is to take a natural language input and generate an argument graph as output (see Figure 2.2). Let us retake the arguments defined in Table 2.1 to illustrate these Argument Mining sub-tasks. Consider as input the following natural language text obtained from the transcriptions of a debate about the tax payment:

"*We should pay our taxes since it is an important way to contribute to the improvement of the society's needs such as public health or education. ... We should not pay any taxes as politicians do not manage them properly. ... This is not a problem caused by the taxes themselves, but by the people who manage them. Every four years citizens can vote and choose different politicians if they are not satisfied. ... The inefficient management of resources does not depend on the political party. We will not be able to overcome this problem by only voting for different politicians.*"

In the first step of the Argument Mining pipeline (i.e., argumentative discourse segmentation), the complete text needs to be divided into the relevant units of text containing argumentative information. Thus, an Argument Mining system would produce the eight argumentative segments depicted in Table 2.2. In the second step of the pipeline (i.e., argument component detection), the argumentative role of each one of the identified elements has to be determined. These roles may differ from one work to another. In our example, we will consider the typical instance of this problem where the main objective is to identify and distinguish between the premises and claims that make up the arguments. Thus, E1, E3, E5, and E8 would be classified as claims and E2, E4, E6, and E7 as premises. Finally, the third step of the pipeline (i.e., argumentative relation identification) aims at automatically extracting the argumentative relations existing between the previously detected propositions. In our example, support relations would be detected between E2 and E1 (making up argument A1), between E4 and E3 (making up argument A2), between E6 and E5 (making up argument A3), and between E7 and E8 (making up argument A4). Furthermore, attack relations would be detected between arguments A1-A2, A2-A1, A3-A2, and A4-A3. Thus, after processing the initial natural language input text the result of the Argument Mining pipeline would be a graph-like data structure containing the most relevant argumentative information. With this example, we have presented a simple instance in which the most important aspects of the pipeline can be observed. However, the

Segment ID	Natural Language Segment
E1	*"we should pay our taxes"*
E2	*"since it is an important way to contribute to the improvement of society's needs such as public health or education"*
E3	*"we should not pay any taxes"*
E4	*"as politicians do not manage them properly"*
E5	*"this is not a problem caused by the taxes themselves, but by the people who manage them"*
E6	*"every four years, citizens can vote and choose different politicians if they are not satisfied"*
E7	*"the inefficient management of resources does not depend on the political party"*
E8	*"we will not be able to overcome this problem by only voting for different politicians"*

Table 2.2: Natural language argumentative segments extracted during the first step of Argument Mining.

complexity of these sub-tasks depends completely on how the input is annotated and how fine-grained these annotations are (e.g., from attack/support [99] to argumentation schemes [388, 396]). Currently, there is no standard annotation for Argument Mining. However, researchers have proposed methods such as the Argument Interchange Format (AIF [204]) to provide a framework to standardise the argumentative annotations and the computational representation of arguments. The following sections review each sub-task of the Argument Mining pipeline, presenting how different algorithms have been proposed to tackle the presented sub-tasks on an heterogeneous set of corpora with different annotations and, consequently, complexity.

Argumentative Discourse Segmentation

Argumentative discourse segmentation is the first task undertaken within the Argument Mining pipeline. This task consists of automatically splitting the natural language input into atomic discourse elements and detecting whether these elements belong to an argumentative structure or contain argumentative information. Previous work on discourse analysis has identified the existence of these atomic discourse elements as Elementary Discourse Units (EDUs). EDUs may appear in different forms in the discourse represented by clauses [150, 159], by sentences [280], or by turns to talk [332], among others. In [274], Argumentative Discourse

Units (ADUs) are introduced as an argumentative representation of the EDUs. An ADU is defined as the "*minimal unit of argumentative analysis*", meaning that it can be either an individual EDU containing argumentative information or a group of related EDUs forming an argumentative structure. Research on argumentative discourse segmentation has been strongly influenced by these two concepts of EDUs and ADUs. Initial research approached the segmentation of argumentative discourses by splitting a natural language input into sentences and classifying them as either *Argument* or *Non-Argument* [242, 265, 266]. The ideas presented in these works were further developed in [156] where new features were used, and in [336] where Conditional Random Fields (CRF) were proposed for the classification, but the initial segmentation was still based on the sentence structuring of the input. However, as pointed out in [334, 77], context is very relevant when automatically detecting arguments from natural language. Depending on the context, a sentence may or may not contain argumentative information. Thus, approaching the argumentative segmentation by splitting the input into sentences may not be enough to robustly identify arguments.

The automatic identification of EDUs has been approached in many different ways and resulted in a better general segmentation than previous approaches. The perceptron algorithm is explored in [141] as the main approach to learn to identify the words in the last segment of an EDU (i.e., its boundary). In [357], a Multi-Layer Perceptron (MLP) is used to identify EDU boundaries in a similar way. Work by Joty et al. [188] presents a new approach where EDU segmentation is considered as a sequence labelling problem. CRFs achieved state-of-the-art results [140] on this task by using this new approach.

More ambitious work directly tackles ADU identification. This task is significantly more complex than EDU identification since the argumentative function of each unit must also be considered to define boundaries between propositions. In [208], a two level separation of argumentative discourse segments is proposed: 1) the argumentative text, which considers propositions such as pieces of evidence or claims; 2) the discursive language used to create argumentative structures using the previous propositions (i.e., argument/non-argument). In [224], an alternative to ADU segmentation is proposed. CRFs are used to detect argumentative propositions in natural language text. In this approach, the natural language input is split into words, and ten different word features are used to structure the input of the

model. Work by Levy et al. [212] presents a formalisation of the claim detection
task in a specific context. An original cascade architecture is proposed in order
to automatically detect claims from natural language inputs. For segment bound-
ary detection, a combination of a maximum likelihood estimation (MLE) proba-
bilistic model and a logistic regression (LR) classifier is used. Another different
approach is presented in [4], where the context surrounding words is considered
as an additional feature for providing further information to the model. A neural
network architecture is used in that work to model the probability distribution of
ADU boundaries. Finally, in [213], an unsupervised approach based on typical
argumentative structures is proposed as a way to automatically retrieve arguments
from any natural language input. Although it has its limitations, that work presents
an inexpensive and effective method for identifying argumentative propositions in
large corpora.

The complexity of the automatic segmentation of argumentative discourse is
reflected in assumptions and oversimplifications in the literature of the field. Such
is the case of contextual information and external knowledge assumptions, that are
commonly used by debaters as a resource to avoid redundancy in the argumentative
discourse. Therefore, a robust algorithm for argumentative discourse segmentation
needs to deal with these complications of the discourse in some way. Recent work
by Jo et al. [186] proposes a cascade model to complete argumentative proposi-
tions with omitted information. This is an important contribution towards more
robust argumentative discourse segmentation approaches, which will lead to better
performing Argument Mining models.

Argument Component Detection

Once the natural language input is segmented, the argumentative analysis contin-
ues with the automatic identification of argumentative roles within these segments.
The argument component detection part of the Argument Mining pipeline focuses
on approaching this problem. The argumentative roles being identified may vary
significantly from one work to another. It depends on how the argument compo-
nent detection problem is instantiated. In general, most of the reviewed research
approaches this problem from two different perspectives: evidence identification
and premise/claim classification [207].

The former focuses on a lower level analysis of argument components. Evidence can be used to support either a premise or a claim in argumentation. Therefore, the automatic detection of pieces of evidence in argumentative debate makes it possible to have a more accurate perspective on the validity of premises and claims [268]. Furthermore, automatic evidence identification has been proven to be a very important step when developing tools for the legal domain [394] or when analysing clinical trials in the medical domain [227]. Different types of evidence and classifications have been identified in the literature. For instance, in [138], an evidence is categorised into fact, definition, cause, value, and action. On the other hand in [310], pieces of evidence are classified into statistical, testimonial, anecdotal, and analogical evidence types. This heterogeneity is reflected in the different instances of the evidence identification problem. In [268], online user reviews are analysed to determine the nature of the observed propositions. They address the task as a three-class classification problem, where propositions can belong to the *unverifiable*, *verifiable non-experiential*, and *verifiable experiential* classes. These three classes are interpreted as non-evidence, objective evidence, and subjective evidence, respectively. With this approach, it is possible to have a further understanding of online reviews and even detect weakly supported propositions. From the results of that work, it is observed that Support Vector Machines (SVM) provide promising results for this task. The same authors proposed a revision of this task in [269], where a new corpus is presented with more fine-grained proposition classes. Here, the authors propose up to five different classes: *fact* (i.e., non-experiential fact), *testimony* (i.e., experiential fact), *value* (i.e., value judgements), *policy* (i.e., action policy proposition), and *reference* (i.e., reference to a specific resource). Work by Niculae et al. [256] provides promising results on the automatic evidence detection task by using the previous corpus [269] with a feature-based SVM and a Recurrent Neural Network (RNN) approach. A final modification of the classes proposed in the works above is done in [134]. The *reference* class is substituted by a *rhetorical statement* where figurative phrases, emotions, or rhetorical questions are taken into account. In that work, the authors present a new annotated corpus considering this new class from an online opinion forum. However, empirical results are not provided in this new instance of the task.

Online Social Network message analysis is also another important domain where the evidence detection and classification problem has been studied. In [3],

evidence detection is again instantiated as a three-class classification problem. In this approach, Twitter publications are classified depending on their source: *news* for news media, *blog* standing for personal blog posts, and *no evidence* if no evidence supporting the claim is detected. SVMs again represent the best approach to automatically detect such pieces of evidence. Following this line, [132] presents another instance of the problem where tweets are classified into *facts* and *opinion* classes. This is an interesting domain where the automatic evidence detection can be a task of utmost importance, specially nowadays with the important number of fake news and non-valid reasoning claims spreading in Online Social Networks.

Finally, another important domain where the automatic detection of evidence has been researched is the political domain. In [270, 184], pieces of evidence and claims from political debates are automatically detected and ranked depending on their authenticity and check-worthiness. In [253], claims stated in parliamentary sessions are classified into the *true*, *false*, *stretch*, and *dodge* labels. SVMs also provided the best results on this instance of the task. Automatic evidence detection and classification in the political domain helps to evaluate the quality of the argumentative reasoning presented by politicians, and even detect an invalid reasoning (e.g., fallacies). However, as mentioned above, in some situations it can be hard to determine the nature of an argumentative sentence without knowing its context. Work by Rinott et al. [311] addresses this issue and proposes a new method for evidence-mining from Wikipedia articles. In that work, a new linguistic feature-based architecture is proposed to undertake the task. The promising results obtained in their experiments paved the way towards less oversimplified instances of this task.

The second main group of argument component detection research focuses on the classification of text segments into premises, claims, and conclusions. This task was first introduced in [265], where argumentative propositions were classified into *premise* and *conclusion* classes. In [352], this idea was further developed and the authors proposed a classification task considering the *major claim*, *claim*, *premise*, and *none* labels. In that work, the propositions are represented using various sets of linguistic features. Furthermore, a comparison between different classification algorithms using these sets of features is provided. SVMs perform the best on this instance of the task. Another instance of the problem is presented in [273], where a corpus of argument annotated *microtexts* is presented. In this approach, linguistic

features are also considered and six different classifiers are compared. The results show that SVMs performed better when detecting less frequent classes. The naive Bayes classifier obtained the best results on more common labels. To sum up, in [5], the authors present a complete analysis of classical machine learning approaches and the relevance of different linguistic features used to train these models for Argument Mining tasks.

Work by Persing and Ng [279] introduces a new concept to the argument component detection task. An end-to-end architecture based on rules and classical machine learning algorithms is presented to tackle the whole Argument Mining process. With the significant improvements in other Natural Language Processing tasks, such as machine translation, neural architectures also caught the attention of Argument Mining researchers. Neural network architectures and deep learning implied an important improvement from previous approaches since no more hand-crafted linguistic features were required. In [135], the authors take advantage of these improvements to propose a new end-to-end Argument Mining pipeline based on a Long Short-Term Memory (LSTM) neural architecture. This new approach does not require manually generating the feature representation of the input structures. This idea is further developed in [245], where an original neural pointer-based architecture is proposed to automatically mine claims and premises from online debate forums. A recent work by Morio and Fujita [246] presents state-of-the-art results in argument component detection using a Graph Convolutional Network (GCN) architecture to detect both argument boundaries and their components using the corpus published in [353]. An interesting analysis is presented in [162], where a comparison of classic machine learning (i.e., SVMs), basic neural networks (i.e., Feed-Forward Neural Networks), and recurrent neural networks (i.e., Long Short-Term Memory) is done. The results demonstrate that with enough data, more complex neural architectures learn how to perform this task better.

In addition to the task of argument component detection itself, there are other aspects of argumentation, such as the language or the domain, that need to be addressed in order to robustly approach this task. Most of the publicly available Argument Mining corpora are in English. In [136], the argument component detection task is approached from a cross-lingual perspective. The authors consider three different components (i.e., *major claim*, *claim*, *premise*) in six different languages. From the results, it is observed that, by combining machine translation

with embedding projection, it is possible to obtain almost the same results as training an in-language model for Argument Mining. The second major aspect to be considered when addressing an Argument Mining task is the domain. Furthermore, good performing domain-specific Argument Mining systems may suffer from loss of robustness in domains other than the one/s used to train the system, mainly due to the learning techniques used, which in many situations is undesirable. Work by Stab et al. [354] addresses this problem and presents a new attention-based architecture that is able to generalise better than previous approaches. This aspect is also pointed out in [37], where the authors present new algorithms for expanding argumentative topics when mining arguments for a debate. With this approach, it is possible to extend the information range to retrieve useful arguments for a specific purpose, regardless of the inclusion of the topic in the argument itself.

Automatic Identification of Argumentative Relations

Argument relation mining is the last sub-task framed within Argument Mining research. This task is focused on the automatic extraction of argumentative structures existing in a given natural language input. Although it is not easy to determine a complexity ranking of the Argument Mining sub-tasks, mainly due to the many different instances of each problem and how each corpus is annotated, argument relation mining is usually considered the most complex task of all of them [28]. The critical points in argument mining identified in the sections above, such as the linguistic nuances and the context, play a major role in the argument relation identification sub-task.

Many different approaches to argument relation mining can be identified in the literature depending on the methods used to approach the problem (i.e., Parsing algorithms, Textual Entailment Suites, LR, SVMs, and Neural Networks), the available annotated corpora, and their annotations. Preliminary work on the identification of argumentative relations is presented in [265] where rule-based Context-Free Grammars are used to predict the relations of argument components and its subsequent structure. In [383, 276], parsing algorithms are used in order to determine argumentative relations between two propositions. A binary classification approach of the problem is presented here, with only *supports* and *attacks* being considered. Argumentation theory concepts such as abstract argumentation frame-

works [124] (i.e., argument graphs) have also been used in [70] to complement a Textual Entailment Suite (TES), and in [77] together with Random Forests (RF). Both works present a binary instance of the task, where only support and attack relations are taken into account. A different approach based on machine learning techniques is presented in [61], where arguments obtained from an online debate forum are used to train SVM classifiers. In this approach, three linguistic features are used: textual entailment to determine if the proposition entails the hypothesis of the argument, semantic textual similarity to measure the semantic equivalence between two texts, and stance alignment to represent the alignment of a given argument. Similar features are considered to train SVMs in [252], where argument relation mining is carried out in a corpus of political speeches. In [279], an end-to-end argument mining approach is presented where argument relation mining is tackled using a Maximum Entropy (ME) classifier. The argument relation classification is also instantiated as a binary classification problem in that work. Student essays are used to train the models to detect attack and support relations between argumentative propositions. Work by Stab and Gurevych [353] presents another end-to-end pipeline where binary argument relations are identified using integer linear programming. In this approach, arguments are represented using linguistic features that are fed to the models. A different perspective of the argument relation identification problem is presented in [234], where the authors consider monologue political speeches. A new corpus is created from non-debate political speeches taking into account three different relation classes: *attack*, *support*, and *no-relation*. In [255], an argument mining pipeline is used to assist in the automatic scoring of argumentative essays. The presented approach automatically detects binary relations between *major claims*, *claims*, and *premises*. The resulting argument structure is used to determine the scores of the essays.

Neural network architectures were introduced aiming at argument relation mining in [99, 171, 135]. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) architecture are proposed to approach the automatic identification of attack/support argument relations between text propositions. These architectures provide better representations than the hand-crafted linguistic features, and, therefore, obtain better and more robust results. In [245], argument relations between forum posts are detected using a new proposed pointer neural architecture. This architecture allows identifying binary relations inside a post (claim/premise

argument structure) and between different posts (attack/support relations).

In [325], the authors present state-of-the-art results in an instance of the automatic identification of argumentative relations where four different classes are considered: *Inference*, *Conflict*, *Rephrase*, and *No relation* (AIF standard relation classes). The authors present an approach based on the Transformer neural architecture which makes it possible to capture longer-term linguistic dependencies, and generates better internal representations of the arguments. The presented results are evaluated using six corpora from different domains, in an attempt to provide a less domain dependant perspective of the work. However, an interesting idea is raised in [256], where the difficulty of comparing Argument Mining works is pointed out. This difficulty is mainly due to the wide variety of corpora, domains, and instances of each Argument Mining problem. In [94], the authors propose a set of dataset independent baselines to overcome the presented problem between different argument relation identification works. Furthermore, similar to previous Argument Mining tasks, as pointed out in [271], context and background knowledge are of utmost importance to robustly determine argumentative relations in many situations. An approach based on SVMs and Bidirectional LSTM representations is proposed to improve argument relation identification using background knowledge. Finally, in [32], the authors propose a transition-based model built upon a Transformer architecture that incrementally captures argumentative components and their relations in a natural language input. This efficient method combines contextual information to achieve state-of-the-art results in two different corpora. They also place emphasis on the efficiency problems that *standard* models may have in the identification of argumentative relations. This is due to the exponential growing of argument pairs with respect to the document length, and it is a topic that deserves more attention if we want to have efficient Argument Mining systems to use *in the wild*.

Cross-lingual and Cross-domain Argument Mining

From the previous task-specific reviewed work, we can see that there is great heterogeneity in the data. The argumentative classes and their distribution significantly differ from one work to another, and the argumentative concepts are understood and instantiated in different ways (e.g., what is a premise and a claim,

which kind of argumentative relations exist, etc.). These differences have a great impact on the natural language arguments used in each experiment and make the results specific for each work. Therefore, it is very hard to generalise their findings and have a robust model that is capable of approaching Argument Mining in different domains. This is an intrinsic challenge of human argumentation since there is no unique and valid definition and different types of analyses have been conducted over the years. Furthermore, we can observe a clear dominance of the English language in the existing research, which makes it more difficult to draw general conclusions and apply the findings in under-researched languages such as Chinese, French, Spanish, Catalan, or Japanese, among others. This challenge has been partially addressed in recent work in cross-lingual and cross-domain Argument Mining, providing a more generalised approach to language and domain.

First, regarding cross-lingual argument mining, it was initially researched in [136], where the authors approach the problems of argument segmentation and classification considering corpora in three different languages (i.e., English, German, and Chinese). The authors proposed two different strategies to approach cross-lingual argument mining using the machine translation and projection techniques. Multilingual word embeddings are combined with an LSTM architecture to approach argument segmentation and classification (considering the *claim*, *premise*, and *major-claim* classes) in an experimental setup with three different languages. From their findings, it is possible to observe how languages belonging to distant language families (i.e., Chinese) are more difficult to deal with than languages belonging to closer language families (i.e., English and German). This work was extended in [313] and [348], where they explored the whole argument mining pipeline in the English and Portuguese languages. For that purpose, they extend the previously mentioned techniques of machine translation and projection with BiLSTM, attention, and BERT architectures. In parallel with these works, a machine translation-based approach was proposed in [370] where the authors take five European languages into consideration (i.e., Spanish, French, Italian, German, and Dutch). Multilingual BERT architectures are trained following two different strategies: first, including all of the languages in the training set; and second, grouping them into Roman and Germanic languages. Again, they could observe how it is possible to achieve improvements by considering only languages that are closer, but this improvement drops significantly when the heterogeneous set

of languages is considered as a whole. Exploring effective cross-lingual Argument Mining strategies for distant languages is a challenge that remains unsolved, and this problem can significantly slow down the development of under-researched languages.

Second, research in cross-domain Argument Mining has also been addressed in recent years. Initially researched in [4] where the sub-task of argument segmentation was considered, the domain challenge was addressed by combining corpora belonging to different domains during the training process. During the evaluation, both in-domain and cross-domain results are provided, emphasising the importance of conducting cross-domain research in argument mining to produce robust models and solutions. In [65], the authors approach the problem of classifying argumentative components in a cross-domain setup. For that purpose, they make use of convolutional neural networks and handcrafted features to improve the results obtained by classical machine learning algorithms. More recently, in [325], a broad set of Transformer-based architectures was compared in the sub-task of identification of argumentative relations. For that purpose, the authors trained the models on a larger and more generic corpus and evaluated them using five smaller domain-specific corpora. From the results, it was possible to observe a good performance in general, and interesting capabilities of generalisation without the need for handcrafted features. Following the research on Transformer-based architectures for argument mining, Alhamzeh et al. [17] tackle the segmentation problem considering two different corpora and a transfer learning strategy. They extended this work in [18], where transfer learning was combined with a model ensemble to improve their performance on the same problem.

In the case of cross-domain research, it is possible to observe an improved capability of generalisation with the combination of the most recent Natural Language Processing (NLP) algorithms and the appropriate learning strategies. The cross-lingual challenge is, however, more problematic since a vast majority of the research is conducted in English and in language families that are closer to the Germanic and Romanic languages (i.e., European). Therefore, to prevent the neglect of minority languages in computational argumentation research, it is advisable to look more closely at aspects such as cross-lingual approaches.

Summary

The research items reviewed in this section have been summarised in Table 2.3. Even though it is hard to compare argument mining research due to the existing heterogeneity in different corpora and task instances, with this table, we highlight the most important aspects which may influence the performance of the experiments carried out in each work. In addition to the proposed approach and the obtained results, we summarise the main features of the corpora and the chosen instance of each task tackled by the authors in each work. We consider the corpus availability (i.e., Public or Private); the corpus size represented by the number of sentences (S) and/or documents (D); a descriptor of the task instance; and (if it is a classification instance of the task) the class balance (i.e., equivalent proportion of samples among the different classes) and the number of classes (N_c) in the presented experiments. A balanced class corpus is used in the experiments when the number of samples between the different classes is the same or very close. The class balance helps to improve the results of the experiments when testing the models. However, in real situations, it is hard to find perfectly balanced distributions between argumentative classes (e.g., supports or inferences are more common in a discourse than attacks or conflicts).

Finally, some of the challenges addressed by Argument Mining research in recent years are worth mentioning. With all of the research carried out in Argument Mining over the previous years and the consolidation of this research task within Computational Argumentation, more complex aspects have caught the attention of researchers. Having knowledge of the best performing algorithms and techniques and having established an heterogeneous but somehow standard way to approach Argument Mining, we could observe that recent research is generally focused on exploring beyond the individual argument mining sub-tasks themselves.

The first of these trends is presented with end-to-end Argument Mining research. The most recent end-to-end model proposals are designed to undertake the complete Argument Mining pipeline, instead of focusing on a very specific sub-task. A good example of this line of research is presented in [248], where the authors propose Multi-Task Argument Mining (MT-AM). MT-AM is an end-to-end cross-corpus training model that relies on both a source and a target corpus and allows sharing knowledge across the different Argument Mining tasks and

Research	Approach	Task Instance	Data Availability	Corpus Size	Class Balance (N_c)	Results
[242]	Naive Bayes	Arg./No-Arg.	Public	3789 (S)	Yes (2)	73.75 (Acc)
[156]	Linear Regression	Arg./No-Arg.	Private	16000 (S)	No (2)	77.10 (F1)
[336]	Word2Vec + CRF	Arg./No-Arg.	Public	300 (D)	No (2)	32.21 (F1)
[17]	SVM + DistilBERT	Arg./No-Arg.	Public	402/340 (D)	No (2)	79.21 (F1)
[141]	Perceptron	Discourse Segmentation	Public	385 (D)	-	90.5 (F1)
[357]	MLP	Discourse Segmentation	Public	7123 (S)	-	86.07 (F1)
[140]	CRF	Discourse Segmentation	Public	8447 (S)	-	92.6 (F1)
[224]	CRF	Argument Segmentation	Private	200 (D)	-	60.70 (F1)
[4]	BiLSTM	Argument Segmentation	Public	402/300/340 (D)	-	88.54 (F1)
[212]	MLE+LR	Claim Identification	Public	326 (D)	-	12 (Pre)
[213]	Unsupervised	Claim Identification	Public	83M (S)	-	34 (Pre)
[268]	SVM	Evidence Classif.	Private	1047 (D)	No (3)	68.99 (F1)
[256]	SVM + RNN	Evidence Classif.	Public	3800/7100 (S)	No (5/3)	73.5/79.3 (F1)
[3]	SVM	Evidence Classif.	Private	531k (S)	No (3)	78.6 (F1)
[132]	SVM	Evidence Classif.	Public	1459 (S)	No (2)	80 (F1)
[253]	SVM	Evidence Classif.	Public	478 (D)	No (4)	55.10 (F1)
[370]	mBERT	Evidence Classif.	Public	6735 (S)	No (2)	80 (F1)
[352]	SVM	Argument Component Classif.	Public	1879 (S)	No (4)	72.6 (F1)
[135]	LSTM	Argument Component Classif.	Public	402 (D)	No (4)	34.35 (F1)
[136]	LSTM	Argument Component Classif.	Public	829 (D)	No (4)	70.71 (F1)
[348]	mBERT + CRF	Argument Component Classif.	Public	402 (D)	No (4)	75.74 (F1)
[246]	BiLSTM + GCN + CRF	Argument Component Classif.	Public	402 (D)	No (4)	67.55 (F1)
[32]	Transition BERT-based	Argument Component Classif.	Public	402/731 (D)	No (3/4)	88.4/82.5 (F1)
[18]	SVM + DistilBERT	Argument Component Classif.	Public	6817/1436/2683 (S)	No (3)	66.1 (F1)
[248]	Longformer + Span-biaffine	Argument Component Classif.	Public	1833/112/731/500/60 (D)	No (3)	77.41 (F1)
[33]	Generative Transformer	Argument Component Classif.	Public	402/731 (D)	No (3)	75.94/57.72 (F1)
[276]	MST parser	Support/Attack	Public	112 (D)	No (2)	71 (F1)
[70]	TES + Arg. Graph	Entail/Attack	Public	200 (S)	Yes (2)	67 (Acc)
[77]	RF + Arg. Graph	Support/Attack/None	Public	854 (S)	*	77.5 (Acc)
[61]	SVM	Support/Attack/None	Public	1013/1285 (S)	No (3)	70.5/81.1 (F1)
[252]	SVM	Support/Attack	Public	265 (S)	No (2)	72.4 (Acc)
[279]	ME classifier	Support/Attack	Private	1473 (S)	No (2)	20.4 (F1)
[234]	SVM	Support/Attack	Public	1462 (S)	No (2)	77 (F1)
[99]	LSTM	Support/Attack/None	Public	*	Yes (3)	89 (F1)
[171]	Logistic Regression	Support/Attack	Public	66542 (S)	Yes (2)	65.4 (F1)
[313]	Inner-Attention	Support/Attack/None	Public	477 (D)	No (3)	53.4 (F1)
[325]	Transformer	Inference/Conflict/Rephrase/None	Public	12392 (S)	No (4)	70 (F1)
[32]	Transition BERT-based	Relation/No-Relation	Public	402/731 (D)	No (2)	82.5/67.8 (F1)
[248]	Longformer + Span-biaffine	Support/Attack/None	Public	1833/112/731/500/60 (D)	No (3)	45.97 (F1)
[33]	Generative Transformer	Support/Attack	Public	402/731 (D)	No (2)	50.08/16.57 (F1)

Table 2.3: Comparison and summary of the reviewed AM research. The table is divided into three main blocks, each containing research focused on the AM sub-tasks: (i) argumentative discourse segmentation; (ii) argument component detection; and (iii) argument relation mining, respectively. The Corpus Size is represented with the number of sentences (S) or documents (D). N_c defines the number of classes in each corpus. The fields with (∗) indicate an aspect of a corpus which is not described in the original publication.

corpora. The proposed model is compared with the single-task learning model that relies on Longformer (i.e., a Transformer-based architecture specifically designed for long input sequences) for learning the natural language representations, and a span-biaffine attention architecture for dealing with the classification problem.

From their findings, it is possible to observe that the MT-AM model effectively shares knowledge across the different corpora and easily outperforms the single-task baselines. In addition to the end-to-end research trend, this work is also an example of the relevance of cross-corpora (e.g., cross-domain and cross-lingual) learning research, since it makes it possible to leverage the limited publicly available resources for Argument Mining. Another recent work in end-to-end Argument Mining with an original perspective is presented in [33]. The authors propose a Transformer-based generative model that takes a string of text as the input and produces a stream of labels as the output. Instead of a classification problem, this approach models the complete Argument Mining pipeline in a way similar to how machine translation models operate, and it is able to outperform the previous work baselines in several Argument Mining benchmarks.

Furthermore, we can observe a significant increase of the research interest in the proposal of new Argument Mining algorithms to undertake complex natural language analyses in different application domains. Such is the case of the medical domain, where Argument Mining has played an important role on automatising the analysis of text-based health reports. In both [228] and [356], Transformer-based models are used to automatically extract claims and premises and to support clinicians in the decision making process for their patients. In general, Argument Mining has proved to be useful in natural language-based decision making problems where an elaborated reasoning can be helpful in making the final decision. In [86], the authors demonstrate the utility of including argumentative features into the assessment of crowdsourced reviews and their relevance for new potential customers. The proposed method combines the benefits of argument mined features with formal argumentation theory (reviewed in Section 2.4) to estimate the helpfulness of a given review. In this same direction, works [326] and [172] exploit argumentative features that are mined from text to build an argumentative graph and estimate the winner of a debate, which is a helpful task for supporting the jury and critical thinking students.

In addition to decision support problems, Argument Mining has also recently played an important role in natural language analysis systems. We observed an incremental use of Argument Mining techniques for analytic purposes such as the detection of fallacious reasoning [152], promoting mutual understanding in polarised debates [403], analysing scientific literature [215], analysing argumentation

in dialogues [333], and finding reasoning structures in text essays [93]. all of these
works show the impact of Argument Mining in many different use cases and rep-
resent how this area of research is opening up a wide set of exciting applications
of Argument Mining and Computational Argumentation for the coming years.

2.4   Argument-based Knowledge Representation and Reasoning

With the objective of structuring and processing argumentative information
from a computational viewpoint, researchers have focused on two aspects of Com-
putational Argumentation: argument representation and argumentation solving.
The former studies the computational representation of arguments. Some works
refer to our definition of argument representation as the syntactic part of argu-
mentation theory. The latter studies the different argumentation semantics which
provide formal definitions for argument properties such as acceptability or defeat.
With the argumentation semantics, it is possible to make an evaluation of the previ-
ously represented arguments. In this section, we analyse the fundamental concepts
and proposals related to formal argumentation research. However, it is an extensive
topic and some advanced aspects have not been reviewed due to the introductory
purpose of the present work. Furthermore, a comprehensive overview of the re-
search on formal argumentation on which most of the reviewed research has been
based can be found in [289].

Argument Representation

First of all, we will focus on reviewing the existing work in argument representa-
tion. Depending on how the arguments are represented, the works in the literature
can be divided into the two main groups of abstract and structured argument repre-
sentations. The differences between the two representations do not necessarily im-
ply exclusion. In fact, structured argument representation approaches are usually
based on abstract argumentation, but they provide a finer-grained level of detail
when defining the internal elements of an argument. Therefore, the choice of a

computational representation of arguments for a specific problem will depend on its domain or the users' needs.

Abstract Argument Representations

The research done by Dung in [124] is considered to be one of the initial contributions that bridge the gap between argumentation theory and Artificial Intelligence. In that work, Dung introduced the concept of an argumentation framework, which serves to make a graph-based representation of argumentation. In this approach, arguments are considered as abstract entities rather than linguistic propositions. Thus, an abstract argumentation framework is defined as a pair of a set of arguments and a binary relation representing attack relationships between pairs of arguments.

This definition laid the foundations over which multiple variations of the original framework have been proposed. These variations add new dimensions to the computational representation of argumentation. In [49], Value-based Argumentation Frameworks (VAF) are formally defined. A VAF is a variation of Dung's AF where the values defended or promoted by the arguments are also considered in the abstract representation of arguments. Thus, a VAF is defined as a 5-tuple where the set of values, the relation between arguments and values, and the preferences over these values are taken into account. In human argumentation, it is very common to find different arguments with opposing claims attacking each other from different perspectives. These perspectives can be motivated by more complex aspects such as ideology or interests. In these situations, two opposed arguments may be acceptable even if their claim proposes confronted ideas. The VAF allows the representation of these situations with the integration of the new value property of arguments.

Human preferences were further developed in [23] and [24], where the authors point out a logical limitation of the previous VAFs regarding the representation of attack relations. To overcome this limitation, Preference-based Argumentation Frameworks (PAF) are proposed. A PAF is represented as a tuple of three elements considering the set of arguments, the attack relations, and a partial or total preordering among the set of arguments. With this preordering, it is possible to encode the human preferences over the arguments included in the framework.

Differently, Modgil [240] proposed an approach to represent preferences, but
aimed at preserving the general properties of the original argumentation frame-
work. In the proposed Extended Argumentation Frameworks (EAF), preferences
are not defined by some externally given preference ordering, but are themselves
claimed by arguments. For that purpose, EAFs are defined as a 3-tuple where in
addition to the set of arguments and attack relations, a new type of relation between
the set of arguments and the set of attack relations is included. This new relation
allows claimed preferences to be represented explicitly during the argumentation
process.

Another important aspect of human argumentation is the indirectness of the ar-
gumentative relations. This is addressed in [79], where the authors propose Bipolar
Argumentation Frameworks (BAF). This framework provides the original abstract
AF with support relations in addition to attacks. The support relation between ar-
guments enables the representation of indirect conflicts in the resulting argument
graph. In the original AF, it was only possible to represent direct attack or defeat
relations between arguments. However, an argument that does not explicitly attack
another argument may be supporting a third argument that does, resulting in an
indirect conflict between arguments. These situations can be perfectly represented
with a BAF. The framework has the required properties to consider finer-grained
relations in an argumentation dialogue. A deeper study of BAFs properties is pre-
sented in [22]. Furthermore, an in-depth survey on the different interpretations of
the argumentative support relation in argumentation systems is presented in [101].

In some situations where online social interaction may occur (e.g., e-commerce
sites), it can be interesting to associate a numerical value to each argument repre-
senting its popularity or strength based on previous voting done by the community
[137]. In [43], Baroni et al. formally define the Quantitative Argumentation De-
bate framework (QuAD), which extends the original AF with a score function.
Through this score function it is possible to incorporate external information, such
as voting or an expert assessment, into the representation of arguments.

The proposed QuAD is framed into the issue-based information system (IBIS),
which is a decision-making support system where decisions are made after a com-
putational debate is solved considering several aspects such as economical, tech-
nical, or environmental, among others. A deep analysis of the properties of a
generalised version of this framework (i.e., Quantitative Bipolar Argumentation

Framework) is done in [42]. An alternative approach for this problem is presented in [128], where weighted argumentation frameworks (WAF) are formally defined. Under this paradigm, numerical values (i.e., weights) are assigned to each attack relation between arguments rather than to the arguments themselves. A complete study on the properties of these frameworks is presented in [58], where the authors report the well-foundedness of the WAFs. Thanks to the quantification, a deeper analysis of argumentation can be done and solutions can be found in situations where abstract frameworks have none.

In argumentative debates, two related concepts are of utmost importance: uncertainty and anticipation. When arguing, in order to present a consistent line of reasoning, it is important to prevent attacks from opponents' arguments. However, the presence of these arguments may not be known. Thus, estimating potential attacks incoming from arguments which have not yet been uttered, but may appear during an argumentative debate, is also an important aspect of argumentation. The inclusion of this stochastic aspect of argumentation was first introduced in [127]. In [214], Li, Oren and Norman formally define the Probabilistic Argumentation Frameworks (PrAF), a variation of the original Dung's AF considering uncertainty in addition to its previous properties. For that purpose, a probability distribution of the arguments and their relations is included in the definition of the PrAFs.

In the constellation approach, probability functions map every argument and argumentative relation to their likelihood values. A PrAF reflects the uncertain representation of arguments and relations in its topology. Therefore, multiple AFs may be instantiated from a specific PrAF. The likelihood values assigned to each argument and relation will determine the whole set of AFs that can be induced (i.e., generated) from a PrAF. Only in the case that all of the likelihood values assigned are the maximum, a unique Dung's AF will be induced. A deep analysis of the probabilistic properties of PrAFs is conducted in [176]. Work by Hunter [178] further develops the PrAF concepts and presents an approach for generating argument graphs from a given probability distribution. This proposed approach can be seen as an intermediate step between the estimation of argument probability distributions and the final computational representation of argumentation.

An alternative probabilistic approach for argumentation is presented in [362]. The author proposes an epistemic approach for the PrAFs, where the probability functions are related to the credibility of an argument rather than the PrAFs

topology. Therefore, the topology is fixed, but the credibility of the existing arguments is modelled with a probability distribution. The epistemic approach is further developed in [181], where a complete modelling and a formal definition of this approach is done. Similar to the quantitative (QuAD, WAF) approaches, epistemic argument representations provide the tools for a quantitative analysis of argumentation, based on the credibility probabilities assigned to each argument. A complete work on the graphical (computational) representations and reasoning of the epistemic approach for probabilistic argumentation is done in [180].

In addition to the PrAFs, uncertainty in argumentation has also been studied under the Incomplete Argumentation Frameworks (IAF) [46], which are an extension of the Dung AF. In this approach, uncertainty is introduced from two different perspectives: attack incompleteness and argument incompleteness. In both perspectives, a set of uncertain attacks or arguments is defined when representing the argumentation framework. Then, all of the possible completions of the IAF are considered when defining the properties of arguments (e.g., acceptability), enabling the analysis of argumentation in situations where the existence of an argument or a conflicting relation between arguments is unknown.

There is a final aspect of argumentation theory that has also been approached in abstract argument representation research: argumentation dynamics. From the human perspective of argumentation, it is very common to have variations in the argumentative discourse (e.g., when a debater presents a new argument). However, the previously presented frameworks are not capable of dealing with such dynamics. The concepts of argument abstraction (removal) and argument refinement (addition) were first introduced in [59, 57] and [60, 78], respectively.

These concepts define the Dynamic Argumentation Frameworks (D-AF) and represent the update of abstract argument representations considering any possible variation in the argumentative discourse. This alternative paradigm for computational argumentative reasoning has been recently surveyed in [121], where the authors provide a detailed analysis of the research carried out in dynamic computational representations of arguments.

Finally, Control Argumentation Frameworks (CAF) are proposed in a recent work by Dimopoulos, Mailly and Moraitis [116]. The main purpose of CAFs is to combine the concepts of uncertainty and dynamics in a generalised representation of argumentation. This way, CAFs are defined as 3-tuples consisting of a fixed

2.4. ARGUMENT-BASED KNOWLEDGE REPRESENTATION AND REASONING
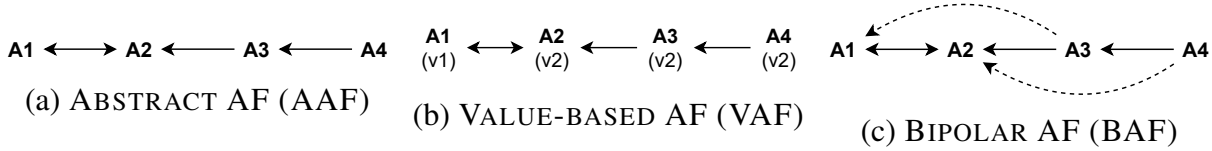
Figure 2.3: Abstract argumentation frameworks instantiated in an example with real arguments.

part, an uncertain part, and a control part. Each of these elements is represented by a partial argumentation framework, depicting different types of information. The fixed part framework represents the known arguments and attack relations. The uncertain part framework represents the arguments and attack relations that can exist in the future (argumentation dynamics). And finally, the most interesting aspect of the framework is that, in the control part, it includes a set of arguments and relations that can be added or not to the global framework by the argumentative agent itself. This way, CAFs provide argumentative systems with a certain degree of control and robustness in dynamic environments where we might want to be prepared to counter arguments that might be uttered in the future. The control part is important to *defend* one's goals and beliefs from arguments that can be added in a dynamic environment (i.e., uncertain part).

Probabilistic Control Argumentation Frameworks (PCAF) were defined in [145], extending the CAFs with probabilistic distributions. For that purpose, the authors introduce the concept of controlling power, a probability that a specific control configuration will make it possible to *defend* one's goals and beliefs from uncertain arguments. With this approach, the robustness of the CAFs is improved, and the computational cost reduced.

We will now retake our ongoing example in Figure 2.3 in order to illustrate how these abstract representations can be used in situations with real arguments. It is important to pay attention to the fluidity of the concept of abstraction in different representations. The most important idea of abstract argument representations is that the linguistic notion of an argument and any sense of structure disappear. Thus, having identified the complete four arguments A1, A2, A3, and A4 depicted in Table 2.1 and then analysing the eight argumentative segments described in Table 2.2, it is possible to generate an abstract representation of these four arguments using Dung's AF [124].

First, the set of arguments A should be instantiated considering all of the available arguments in this situation, $A = \{A1, A2, A3, A4\}$. Second, the set of binary attacks between arguments should be defined according to the nature of the identified arguments, $R = \{(A1, A2), (A2, A1), (A3, A2), (A4, A3)\}$ (Figure 2.3a). This is the *purest* abstract representation of the arguments presented above. However, it is possible to make alternative abstract representations of this situation using different frameworks based on our needs. For example, a VAF [49] could be instantiated in this situation by considering the two major values promoted by these arguments: social good ($v1$) and resource management ($v2$). Thus, in addition to A and R, the following can also be included in the definition of the framework (Figure 2.3b): the set of values $V = \{v1, v2\}$; a mapping function val, such as val(A1) $= v1$, val(A2) $= v2$, val(A3) $= v2$, val(A4) $= v2$; and a preference relation over the values $valpref = v1 > v2$. Furthermore, it is also possible to make an abstract representation of this situation using a BAF [79] by considering the addition of the support relations $S = \{(A3, A1), (A4, A2)\}$ to the previously defined sets A and R (Figure 2.3c).

Structured Argument Representations

As an intermediate step between the pure abstract representation of arguments presented in the subsection above, and the complete linguistically structured arguments which can be found in human argumentation, researchers explored the representation of structured arguments. This representation of arguments lays its foundation in formal logic concepts, considering an argument as a pair $< \Phi, \alpha >$, where Φ is the support and α is the claim of the argument [54]. For example, in the argument A1 ("*We should pay our taxes since it is an important way to contribute to the improvement of the society's needs such as public health or education.*"), Φ_1 would contain the support of the argument (i.e., "*it is an important way to contribute to the improvement of the society's needs such as public health or education*"), and α_1 would represent its claim (i.e., "*we should pay our taxes*"). This same structure can be identified in the remaining arguments from our running example: A2 $= < \Phi_2$ ("*politicians do not manage them properly*"), α_2 ("*we should not pay our taxes*") $>$; A3 $= < \Phi_3$ ("*this is not a problem caused by the taxes themselves, but by the people who manage them*"), α_3 ("*every four years cit-*

izens can vote and choose different politicians if they are not satisfied") $>$; and A4
$= \; < \; \Phi_4$ ("*the inefficient management of resources does not depend on the political party*"), α_4 ("*we will not be able to overcome this problem by only voting for different politicians*") $>$.

Given this formal definition of an argument, it is possible to provide more fine-grained representations of the existing conflict relations among them: defeater arguments, and the characteristic assumption attacks of undercuts and rebuttals [259, 292, 344]. A defeater argument is defined as an argument whose conclusion proves the refutation of the support of the defeated argument, and, therefore, its claim. This situation can be observed when two different types of assumption attacks occur in an argumentation.

First, an undercut happens when an argument directly attacks the inference from the premise of another argument to its conclusion. An undercut can be observed in our example between arguments A3 and A2, since A3's claim α_3 ("*every four years citizens can vote and choose different politicians if they are not satisfied*") is in conflict with A2's support Φ_2 ("*politicians do not manage them properly*"). The weakening of the inference created from the support of an argument towards its claim (i.e., "we should not pay our taxes") in a debate undermines the credibility of the whole argument. However, the most natural way to attack another argument is to directly oppose its claim. This situation is studied in the literature as a rebuttal. An example of a rebuttal can be observed between arguments A2 and A1, since α_2 ("*we should not pay our taxes*") is attacking α_1 ("*we should pay our taxes*"), which is the claim of A1, and vice-versa. Similar to the attacks, structured argumentation make possible to be more specific on the supports between arguments. In [102], the authors characterise four different types of support in structured argumentation: the conclusion support, the premise support, the intermediate support, and the sub-argument support. These logical definitions of arguments and their relations made possible a computational representation of argumentation using logic programming languages. Defeasible Logic Programming (DeLP) [146] is a paradigm of the combination of logic structured argumentation and a logic programming language based on Prolog syntax.

Another approach to structured argument representation is the assumption-based argumentation (ABA) [63]. This approach presents a computational framework based on Dung's abstract argumentation and incorporates the logical notion

of argument structures. Essentially, assumption-based argumentation provides abstract frameworks with a set of defeasible assumptions. These assumptions are true until the contrary is proved by another argument attacking its veracity. However, the semantics of assumption-based argumentation (i.e., the notions of acceptability) are not different from the abstract argumentation approach. Furthermore, ABA differs from its logic-based counterparts in two major aspects. First, the explicit definition of rebuttal presented above does not happen in ABA, but rebuttals can be induced from the generated representations [290]. Second, in ABA, arguments are not required to have supports in order to exist. Original assumption-based argument representations rely on tree data structures. However, research done by Craven and Toni [109] restates this concept and provides a graph-based alternative which makes possible to consider further aspects of argumentation than the original approach. In fact, graph-based ABA can be seen as a finer-grained argument representation of the abstract argumentation graphs. A complete tutorial on ABA is presented in [371].

Finally, work by Prakken [288] merges the abstract concepts of argumentation (i.e., abstract argument graphs) with the logical notions of argument structures and their tree-based representations. ASPIC(+) provides a framework for the definition of argumentation systems relying on these two ideas. A fully detailed tutorial on the definitions of ASPIC(+) and its implications is carried out in [241]. Except for the traditional logic-based approaches, the other reviewed structured representations rely on the definition of rules. However, the rule-based mechanism used in each representation is not the same. On the one hand, DeLP and ASPIC(+) have defeasible rules. On the other hand, ABA only allows deductive rules in their argument representations. The differences on the rule treatment by each approach can be a determining factor for defining the structure of the arguments, and even for their evaluation.

In some domains and situations (e.g., legal reasoning), it was shown that even logic-based structured representations of arguments were too abstract [286]. The need for a more specific structured representation of different argumentation patterns leads to the definition of argumentation schemes. An argumentation scheme formally defines the inference structure of an argument. These definitions reflect the most common reasoning patterns that can be identified within human argumentation. Thus, argumentation schemes provide the more specific structures for the

definition and representation of arguments. An argumentation scheme is defined by a set of premises and a conclusion. A very common example of an argumentation scheme, which can be usually found in debates, is the argument from expert opinion. The major premise is that the source E is an expert in the domain S which contains the proposition A. The minor premise is that the expert E asserts that proposition A is true (false). Finally, its conclusion is that A is true (false). Every argumentation scheme also has a set of critical questions, which may weaken the credibility of an argument. A typical critical question for the expert opinion arguments would be the expertise question (i.e., How credible is E as an expert source?) or the field question (i.e., Is E really an expert in the field that A is in?). The work in [400] makes the most complete compilation of argumentation schemes is done. More than sixty different schemes and their critical questions are grouped into three major classes (i.e., Source-based, Practical reasoning, and Causal reasoning) depending on the nature of the underlying reasoning pattern. This classification has been refined in subsequent research [398].

A dialogue-based computational representation of arguments resulting from the combination of the Argument Interchange Format (AIF) [204] and Inference Anchoring Theory (IAT) [68] has been proposed to bridge the gap between natural language argumentation and the computational representation of arguments in argumentative dialogues. Therefore, this computational representation of arguments can be understood as being both an annotation guideline for natural language argumentation for Argument Mining and NLP tasks and a structured representation of arguments that are compatible with other frameworks such as ASPIC(+). With AIF, it is possible to computationally encode argumentative propositions and their relations, together with the utterances in the natural language dialogue. Additionally, IAT serves as the *glue* that links the flow of dialogue with the argumentative structures derived from such dialogue.

The different argument representation approaches reviewed in this work are represented in Table 2.4. We have summarised the most relevant attributes (columns) that define each framework (rows). We have also sorted the different representations depending on their level of abstraction, with the Abstract AF (AAF) being the most abstract representation of arguments and the argumentation schemes being the linguistically richest representation. Furthermore, some frameworks such as the probabilistic or the dynamic, can be viewed as meta-frameworks

Framework	Attacks	Supports	Features	Uncertainty	Dynamics	Structure
AAF [124]	✓	×	×	×	×	
BAF [79]	✓	✓	×	×	×	
VAF [49]	✓	×	✓ (Value)	×	×	
PAF [23]	✓	×	✓ (Preference order)	×	×	
EAF [240]	✓	×	✓ ($A \times R$ relation)	×	×	
QuAD [43]	✓	✓	✓ (Score)	×	×	Abstract
WAF [128]	✓	✓	✓ (Weight)	×	×	
PrAF [214]	○	○	○	✓	×	
IAF [46]	✓	×	×	✓	×	
D-AF [59, 60]	○	○	○	×	✓	
CAF [116]	✓	×	×	✓	✓	
PCAF [145]	✓	×	✓ (Probability)	✓	✓	
DeLP [146]	✓	✓	×	×	×	Logic
ABA [63]	✓	✓	✓ (Assumptions)	×	×	Logic + Inference
ASPIC(+) [288]	✓	✓	○	○	○	AF + Logic
Arg. Schemes [400]	✓	✓	×	×	×	Reasoning Pattern
AIF [204]	✓	✓	✓ (Text and Reasoning)	×	×	Natural Language

Table 2.4: Comparison of the most relevant argument representation approaches.
Note: ✓ and × indicate whether a framework has a specific attribute or not; ○
indicates if the framework is compatible with some attribute, even if it was not
considered in its original definition.

that provide basic frameworks with new functions (i.e., uncertainty and dynamics).

Argumentation Solving

Once the argumentative reasoning is computationally represented by an argumen-
tation framework, the next step is to understand and evaluate the contributions of
each argument to the presented reasoning. We call this task argumentation solving,
i.e., understanding the solving idea as the identification of acceptable and defeated
arguments. The acceptability concept of an argument may vary depending on its
definition, which can be different from one argumentation domain or framework
to another. Along with the definition of Dung's abstract argument representations,
the concept of argumentation semantics was proposed [124]. Argumentation se-
mantics are the rules on which each argument evaluation method relies. Usually,
argumentation semantics rule the definition of the extensions of an argumentation
framework. These extensions are a subset of the complete set of arguments, repre-
senting the idea of collective acceptance (i.e., arguments belonging to an extension

can not conflict with each other).

The argumentation semantics have been proposed mainly around two concepts that are defined considering the graph structures underlying the argumentation frameworks. The first of these concepts is conflict-freedom. We can say that a set of arguments is conflict free if there are no relations of conflict between any argument included in this set. The second of these concepts is admissibility. It is possible to affirm that an extension is admissible if in addition to being conflict-free, the arguments belonging to the extension are able to defend themselves (i.e., that there can be no external attack on any one of the arguments that are included in the set that is not responded to with a counter attack by the arguments in the set) [124].

Four different extension-based semantics were proposed originally in [124]: complete, grounded, stable, and preferred. Depending on which semantics are considered, the argumentation solving problem will output different sets of acceptable arguments (i.e., extensions). The complete semantics define the complete extensions of an argumentation framework as the set of arguments capable of completely defending themselves and which must include every argument that the set defends. Retaking our example, in our abstract AF = $<\{$A1, A2, A3, A4$\}$, $\{($A1, A2$)$, $($A2, A1$)$, $($A3, A2$)$, $($A4, A3$)>$ (Figure 2.3a), the set of complete extensions would be E_{CO}(AF) = $\{\{$A4$\}$, $\{\{$A1, A4$\}$, $\{$A2, A4$\}\}$. The grounded extension is the minimal complete extension (wrt. set inclusion). To compute the grounded extension, every initially unattacked argument is added to the set. Then, all of the arguments attacked by these are removed from the framework. These two steps are repeated until no attacked arguments exist in the set. In our example, the grounded extension would be E_{GR}(AF) = $\{\{$A4$\}\}$. The stable semantics define the stable extensions of an argumentation framework since every conflict-free set of arguments is capable of attacking all of the arguments that are not included in it. The existence of stable extensions of any cyclic framework with three or more arguments is not guaranteed. Considering our example, the stable extension would be E_{ST}(AF) = $\{\{$A1, A4$\}$, $\{$A2, A4$\}\}$. Finally, the preferred semantics define the preferred extensions as the subset-maximal of admissible sets, with the only requirement being that the arguments included in that sets must be able to defend themselves from any existing attack. Thus, preferred semantics can be seen in the middle of complete and stable semantics, since every stable extension is preferred and every

preferred extension is complete. In our example, the preferred extension would be $E_{PR}(\text{AF}) = \{\{A1, A4\}, \{A2, A4\}\}$.

In addition to these four original semantics, researchers have proposed different alternatives to overcome their limitations and to improve some undesired behavioural aspects in specific contexts: stage [381], semi-stable [74], ideal [126], CF2 [40], and prudent [105] semantics. We will not delve further into the formal definition and analysis of all of the existing semantics since it is not the main objective of this work and the basic notions of argumentation semantics have already been described. The formal definition of all of the mentioned semantics and their differences can be found in [39]. Furthermore, a complete introduction to the abstract argumentation semantics is presented in [38], where the authors describe in detail the most important underlying concepts of argumentation semantics and formally define their most common classes.

We have seen many different alternatives to the Dung's abstract AF. Therefore, in some cases, these semantics will need to be adapted to the variations of each framework. Such is the case of the work in [23], where the proposed PAF leverages the preference-based preordering of arguments to overcome the possible violation of the conflict-freedom property by the acceptable extensions produced by VAFs. Work by Amgoud and Naim [21] explores how the argumentation semantics need to be restated in order to evaluate weighted bipolar argumentation graphs. The authors propose new semantics which are compliant with the properties of such frameworks. Furthermore, the same authors propose the concept of ranking-based semantics in [20]. In this approach, arguments are ranked during the evaluation based on their degree of acceptability using a score function. A complete review and analysis of these semantics is done in [114], where the most important ranking-based and scoring semantics are compiled (e.g., categoriser-based ranking semantics [53], burden-based semantics, or discussion-based semantics [20]). Another work that is framed in the bipolar argumentation frameworks domain is [118], where Doder et al. present a set of postulates and formalities for the integration of ranking-based semantics into weighted bipolar argumentation-based systems. This new idea of semantics makes possible to do a gradual evaluation where arguments are not completely defeated but just weakened.

Furthermore, the particularities of the quantitative argument representations make it harder to effectively use algorithms based on purely abstract semantics for

the evaluation arguments. In the paper where the QuAD frameworks are proposed, an algorithm for evaluating arguments based on the aggregation of the values assigned to the arguments is defined [43]. However, in [299], the authors identify some of its limitations and propose an improved version of the algorithm to do the evaluation of such frameworks. Similarly to the ranking-based approach, there is not a unique extension or labelling of acceptable arguments. The main difference is that the scores in ranking-based semantics must be calculated directly from the argumentation framework, without considering external sources.

It is also worth to mention that dynamic argumentation allows dealing with argument graphs that may vary with the addition or removal of existing arguments over time. In this situation, specific algorithms for argumentation semantics need to be specified. In [216], the authors propose an approach based on partitioning the original graph into modified subgraphs, which improves the efficiency of calculating the extensions of the complete, preferred, ideal, and grounded semantics. This approach is corrected and further polished in [41], where the authors provide an analysis of graph topology-related properties of argumentation semantics. Work by Alfano, Greco and Parisi [15] presents an incremental algorithm to calculate the extensions of an argumentation framework considering one update at a time. The algorithm recalculates extensions on a significantly reduced version of the previous partitions making an important contribution to the efficiency of argumentation solving in dynamic environments.

Argumentation solving is, in fact, a hard problem regarding its computational complexity. As pointed out in [28], the computational complexity of calculating any of the previously mentioned extensions from a given argument graph goes beyond P and NP complexity classes. The implications of this affirmation mean that any deterministic algorithm for argumentation solving will have an exponential (worst-case) cost of execution under the exponential time hypothesis. This complexity issue has been analysed in the literature for most of the existing argument representations: for the abstract (Dung's), value-based, and assumption-based argumentation frameworks [129]; for the probabilistic (bipolar) argumentation frameworks [139]; and for dynamic argumentation solving [257]. The International Competition on Computational Models of Argumentation[2] (ICCMA)

---

[2]http://argumentationcompetition.org/

[364, 201] is proposed as a way to develop efficient algorithms for argumentation solving. The participant algorithms [144] are awarded based on their computational efficiency. In this way, it is possible to have frequently updated baselines regarding argumentation solving computational costs. To overcome the complexity issue of argumentation solving, recent research has focused on doing a probabilistic modelling of the problem using Markov networks [284] and deep neural networks [108]. Promising results can be observed considering the grounded, preferred, stable, and complete semantics. However, dynamic argumentation remains as future work for these probabilistic approaches. Also, in [133], the authors propose a complete solver based on the answer-set programming paradigm that incorporates the concepts of credulous and skeptical acceptance and calculates the extensions for a given framework. Thanks to the ICCMA, significant advances have been made in the implementation of efficient and accurate argumentation solvers in recent years, details of which can be found in [202]. In [80], Cerutti et al. present a thorough review of the existing implementations for formal argumentation models and provide a comparison of the most important solvers designed for abstract and structured argument representations. Furthermore, the authors also identify the complexity problem related to computational argumentation reasoning (i.e., solving) and summarise other existing approaches to overcome this limitation.

Some recent research has focused on understanding how these formal proposals fit into the human reasoning process. First approached in [81], the authors present an empirical evaluation of the performance of different formal argument models when applying human reasoning instead of argumentation semantics. The findings show that human reasoning takes the domain much more into account than argumentation semantics, which (almost) completely disregard the context of argumentation. Similarly, in [281], an evaluation of the probabilistic models (both constellation and epistemic) and the bipolar models is conducted through an empirical evaluation with humans. From the results, the authors state that handling uncertainty through probabilistic models is a critical aspect from the human reasoning perspective and that an abstract model with standard semantics is not enough to approximate human reasoning. Finally, it is also worth mentioning the work by Rahwan et al. [300], where the combination of semantics and behavioural-based models of human reasoning is proposed to overcome acceptability conflicts. This way the behavioural-based model of human reasoning is used to solve any

potential conflict between the predictions of different semantics.

2.5 Argument-based Human Computer Interaction

The Computational Argumentation line of research that deals with the invention phase of human argumentative reasoning can be framed in the argument-based human-computer interaction. An argumentative intelligent system that interacts with human users must be able to do the following: 1) "*translate*" computational representations of arguments to a human understandable representation (e.g., readable or spoken natural language); and 2) define dialogue strategies that remain trustworthy and persuasive for every human user interacting with the system. Figure 2.4 provides a graphical representation of this last area of Computational Argumentation research which has been identified and reviewed in this paper. Retaking our example, after having defined the set of acceptable arguments in their computational representations (i.e., A1 and A4 considering the AF in Figure 2.3a and admissible semantics), generate natural language arguments must be generated before engaging in direct interaction with human users. Furthermore, different arguments may be more or less persuasive for different human users depending on several aspects (e.g., personality, preferences, etc.). Thus, argument generation research deals with the transition from computational representations of abstractly represented arguments to their complete natural language versions. On the other hand, argumentative persuasion research focuses on an efficient use of arguments to persuade human users. For example, the use of the first generated argument (i.e., Figure 2.4, "Argument 1") should be prioritised to persuade a user who values social benefit the most (i.e., Figure 2.4, User 1). However, our second generated argument (i.e., Figure 2.4, "Argument 4") will be more persuasive for a user who is concerned with the problem of political resource management (i.e., Figure 2.4, User 2). In this section, we review the most important research that focuses on making advances in argument generation and argument-based computational per-
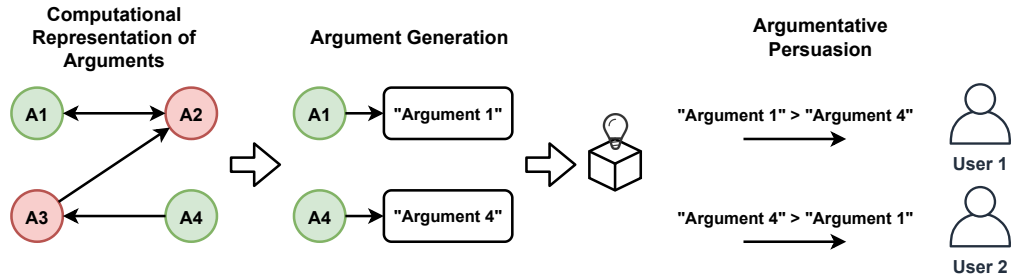
Figure 2.4: Main steps in argument-based human computer interaction: (i) The acceptable set of arguments (i.e., A1 and A4) is taken from the computational representation. (ii) Natural language arguments (i.e., "Argument 1" and "Argument 4") are generated from their abstract/structured representations. (iii) Arguments are used to convince human users following different strategies.

suasion.

Argument Generation

Argument generation is the first task to be tackled towards a complete argument-based human-computer interaction. Research on argument generation focuses on exploring methods to "*translate*" computer systems' knowledge on some specific topic or domain into arguments that are new and understandable to humans. In [305], a new rule-based framework is defined to deal with argument generation for the first time. The ideas proposed along with this framework are further developed in [304], where argument components (i.e., logical propositions) are ordered following new specific strategies with the objective of improving the coherence and persuasiveness of the generated arguments. In parallel, work by Zukerman, McConachy and Korb [415] approaches this task by using Bayesian reasoning and proposes a new method to automatically generate arguments from an argument graph given a topic and a context. Similar to previous research in [305] and [304], in [414], the authors improve the Bayesian reasoning argument generation approach with different argument strategies. These strategies make it possible to refine of the process by considering different argument components (i.e., goals, supports, claims, premises) to generate the final arguments depending on the strategy chosen (i.e., *reductio ad absurdum*, inference to the best explanation, reasoning by cases, and premise to goal). Work by Carenini and Moore [75] laid the foundations of modern argument generation research. The authors presented

Research	Approach	Arg. Control	Arg. Repetition	Arg. Originality
[55], [56], [322], [324], [365]	Template-based Generation	Max.	High	Min.
[158], [295] [296], [309], [337], [350], [392], [397]	Argument Database Information Retrieval	High	Low	Low
[30], [157], [169], [260], [339]	Natural Language Generation	Min.	Min.	Max.

Table 2.5: Comparison and summary of the reviewed research in automatic argument generation. Three major features are considered in our comparison: the user control over the argument generation process (*Arg. Control*); the amount of repeated arguments generated by each approach (*Arg. Repetition*); and the originality of the new generated arguments (*Arg. Originality*).

an original two-sided approach, relying on argumentation theory concepts for the argument component selection and structure definition, and on computational linguistics for the natural language generation. Three main branches of the recent automatic argument generation research can be identified: template-based argument generation, argument information retrieval, and approaches based on natural language generation and computational linguistics.

Table 2.5 depicts the most relevant research in this area, grouped into these three major approaches for the automatic generation of arguments. Three fundamental features that are intrinsic to the process of generating new arguments are used to compare these three approaches. The first one is the user control over the creation of new arguments (*Arg. Control*). This feature indicates how controllable the argument generation process is, which can be of utmost importance in specific domains, such as the education domain or a privacy sensitive domain. The sec-

ond one is the number of repeated arguments that will be created at the end of the
process (*Arg. Repetition*). This is a relevant feature when looking for naturalness
in dialogues. It is important to minimise the number of repeated utterances, and
even internal structures in different arguments. The third one is the originality of
the created arguments (*Arg. Originality*). In some situations where *creativity* is
a major aspect when analysing new arguments, it might be desirable to generate
original arguments rather than reusing arguments used in previous debates.

First, template-based argument generation is a straightforward, but effective,
way to automatically generate arguments in controlled environments. Work by
Bilu et al. [56] defines a set of templates matching previously detected premises in
a given topic. Argument Mining techniques are used to undertake the automatic de-
tection of claims and premises. Then, the argument generation is done by combin-
ing these previously "*mined*" argument components with the use of templates. The
ideas presented in that work are further developed in [55] where machine learning
and neural network architectures are used to match "first principles" (i.e., relevant
arguments on a wide range of topics) and predefined classes of arguments. These
"first principles" are used to represent the main ideas of every argument generated
in some specific domain. A different template-based approach is presented in [365]
and [322]. Work by Thomas, Oren and Masthoff [365] proposes a set of templates
based on argumentation schemes (i.e., stereotyped patterns of human reasoning
[400]), which are used to automatically generate arguments following the different
reasoning patterns. The proposed system asks the user for a minimal input (e.g., a
goal, a commitment, a belief, or some liking) and generates persuasive arguments
aimed at changing user behaviour in the healthy eating domain. On the other hand,
work by Ruiz-Dolz et al. [322] presents a set of argument templates designed to
prevent privacy violations in online social networks. The proposed argumentation
system [324] automatically retrieves all of the needed user features from the net-
work and analyses every piece of content shared in it. When a potential privacy
violation is detected, the system uses the features together with the templates to
create persuasive arguments. Four different types of arguments are proposed to
persuade social network users: Privacy, Risk, Trust, and Content. Each type of ar-
gument deals with the potential privacy violation from a different perspective: the
privacy configuration, the scope of the publication, the trust of the users involved
towards the author and the sensitive content detected in it.

An important contribution to modern argument generation research is "*Carneades*" [397], a complete argumentation system. "*Carneades*" is capable of generating new arguments from a previously loaded argument component database. This tool allows generating complex reasoning patterns such as argumentation schemes, but the argumentative database (i.e., argumentation frameworks) must be previously structured and annotated. Argument information retrieval techniques can consider a wider range of cases than the template-based approaches. However, their scope is usually bounded to the quality of the underlying argument knowledge database. Work by Sato et al. [337] presents an argument generation system for debates. This approach is based on a sentence retrieval algorithm. Each sentence is previously analysed to obtain key concepts such as the topic or the polarity. Then, a predefined set of rules is used to score and rank the available arguments in order to be able to generate (i.e., select) the most suitable one for the ongoing debate. In [309], the authors propose an argument generation system that automatically generates Toulmin arguments (i.e., claim, data, and warrant) for a given topic. The method presented in that work is based on the initial generation of a knowledge database from which the claims, data and warrants are retrieved. These elements are retrieved using a score that is based on contextual similarity aimed at generating the most coherent possible arguments. In the work carried out by Wachsmuth et al. [392] a standard structured pipeline to undertake the argument generation task is proposed. The authors present a "*rhetorical strategy*" that can be used for duly synthesising arguments and a new dataset to validate the proposed strategy. This strategy also relies on having an available argument knowledge database and is divided into three steps: content selection (i.e., Argumentative Discourse Units), argumentative component structuring, and phrasing the style of the generated argument.

Neural network architectures are introduced to perform the argument information retrieval in [350], where an argument generation system that is capable of retrieving arguments from an heterogeneous set of topics is presented. The system architecture uses a pre-processed argument knowledge database. A topic is given by the user and the system automatically retrieves the top related arguments from the previously indexed database. A Bidirectional Long-Short Term Memory (BiLSTM) neural network is used to automatically extract argumentative features to improve the quality ranking of argumentative sentences. A large amount of data

is usually required in order to completely take advantage of these models. The largest dataset for argument quality ranking is presented in [158]. The authors also present the first approach for ranking arguments depending on their quality in a specific context, based on the state-of-the-art natural language processing Transformer architecture (i.e., BERT). An important aspect of argument generation is the evaluation of the quality of the generated arguments, not only regarding their semantics but also their coherence in the context in which they are generated by an argumentation system. Work by Rach et al. [295] introduces a new framework to evaluate argument search techniques. The framework provides an interface with a chatbot that allows human users to rate and assign different categories to the arguments displayed in each dialogue step. These categories are assigned based on how suitable the argument generated in the conversation is, and thus are used to calculate the quality of the different argument search approaches. That work compares the argument search engines presented in [391] and [350]. Finally, arguments are generated using a modification of the template-based approach of [296].

Recent research on argument generation has set its sights on the natural language generation area of Natural Language Processing. The latest advances in this area have made it possible to automatically generate rich text without the need to constantly search large databases. Work by El Baff et al. [30] presents the task of automatically synthesising arguments as a language modelling task. In this approach, the authors consider the ADUs as the "*words*" and the arguments as the "*sentences*" of a language model. This approach uses the "*rhetorical strategy*" proposed in [392] in order to define the sub-tasks carried out by the model (i.e., *select*, *arrange* and *phrase*). In [169], the authors tackle the automatic generation of counterarguments. For that purpose, they create a corpus from an online forum with conflicting pairs of comments (i.e., claims). The proposed method first identifies the words in the original comment that should be removed or replaced and then generates the needed modifications. A neural encoder-decoder gated recurrent unit (GRU) model with attention is used in that work to automatically generate counterarguments. In [339], the authors present an approach that is based on controllable language models to automatically learn to generate arguments. The presented pipeline is divided into three different steps. First, arguments are classified and analysed from large data sources. Then, the language model is fine-tuned on these new processed data sources. Finally, the model is able to infer new ar-

guments using specific control codes that are specified in the first step. With this approach, complete new arguments can be generated from only a few conceptual words used as control codes. Similarly, in [157], the authors present a claim generation pipeline based on the GPT-2 model architecture. The model is fine-tuned on a large collection of argumentative text. In that work, claims are generated from a topic input. This is a less restrictive approach than the previous reviewed work and allows to generating new arguments considering only its topic. Finally, [260] describes the creation of a corpus of argumentative rebuttals. This corpus is aimed at overcoming the topic (i.e., domain) limitation of previous argument generation research and finding more natural continuations for argumentative debates.

Work by Le et al. [209] compares the argument retrieval approach with the natural language generation approach. Neural architectures are used in both cases, but better results are achieved using the argument retrieval system. However, a very limited corpus is used, which could be the cause of misleading results. Conversely, [174] and [173] propose a combination of neural architectures for natural language generation with argument retrieval techniques to improve the quality of the generated arguments. Promising results are presented, which may imply that a combination the two approaches may provide the system with richer information resulting in better generated arguments, especially with the current scarcity of available corpora. In conclusion, we have been able to observe how each approach presents different strengths and weaknesses (see Table 2.5). Template-based approaches maximise the control over the process of generating new arguments, but thus sacrifice the naturalness of the dialogue and the originality of the created arguments. Other approaches (i.e., argument retrieval and natural language generation) partially sacrifice the user control in the generation process, but they can be more useful in creating natural argumentation dialogues or original arguments on a specific debate topic. Thus, a different approach for generating arguments can be the optimal one based on the needs and the application domain.

Argument-based Computational Persuasion

The final step when designing the human-computer interaction of an argumentation system focuses on analysing how arguments must be used in a direct dialogue with human users. An argument is a piece of reasoning that supports, rebuts, or

justifies some specific claim. In an argumentative dialogue, arguments are used to strengthen one's own claims and to weaken others' claims. Thus, in the end, any computational system that directly dialogues with human users using arguments, regardless of the final goal of the system (i.e., reach an agreement, decision making, recommendation, conflict resolution, etc.), needs to be persuasive in order to be considered effective. This problem has been approached from a formal viewpoint, benefiting from prior research in argument representation and reasoning [189] (see Section 2.4). A complete review of this perspective on argument-based computational persuasion is done in [287]. However, in this work, we focus on the most recent approaches proposed in the literature where more informal aspects are brought into consideration (e.g., natural language, user models). In this last task of the Computational Argumentation process, most of the research aims at analysing the persuasive aspect of arguments. However, argument persuasiveness is hard to model and understand. The argument-based persuasive dialogue can be structured into four major sub-tasks: the modelling of persuasive aspects of the user, the automatic estimation of argument persuasiveness, the definition of dialogue strategies, and the direct interaction with human users. Table 2.6 classifies the identified work related to argument-based computational persuasion and provides a general structure for research in this area.

Empirical Studies for Argument-based Persuasion

The audience is of utmost importance to accurately model the persuasiveness of arguments. Thus, an important part of the research in argument persuasion has focused on studying how different human features can relate to the perception of persuasion. In [367], the authors carried out a study of how human personality could be relevant in creating persuasive messages to convince humans in the healthy eating domain. For that purpose, personality is represented with the Big Five personality dimensions [317], and messages are grouped into the six persuasive principles of Cialdini [91]. Through these principles, it is possible to classify the persuasive intentions of a given message into six categories: reciprocity, commitment and consistency, consensus or social proof, authority, liking, and scarcity. These principles of persuasion can be found every day in different situations such as marketing and advertisements. In the case of argumentation, we need to pay at-

| Research | Approach | Argument-based Persuasive Dialogue | | | |
		User Modelling	Persuasiveness Estimation	Dialogue Strategy	User Interaction
[367], [92], [368]	Study	✓	×	×	×
[278], [413], [151], [31]	Machine Learning	×	✓	×	×
[220], [131], [191], [119]	Machine Learning	✓	✓	×	×
[148], [267], [244], [8], [9], [272]	Reinforcement Learning	×	×	✓	×
[316], [83], [82]	Chatbot	×	×	✓	✓
[164]	Study + Machine Learning	✓	✓	×	×
[369]	Study + Chatbot	✓	×	×	✓
[315]	Machine Learning + Chatbot	×	✓	✓	✓
[179]	Framework	✓	×	✓	×

Table 2.6: Classification of the identified research in argument-based computational persuasion. Note: ✓ and × indicate whether a research work approaches an argument-based persuasive dialogue sub-task or not.

tention to the usage of these principles since an argument whose persuasive intent
can be classified under one of these principles, but whose premises are not provid-
ing a solid enough justification for its claims, can be considered a fallacious argu-
ment and thus be used for manipulative purposes. This problem has been studied
from the argumentative viewpoint with the argumentation schemes and their crit-
ical questions. For an argument defined under a specific reasoning scheme, these
questions allow us to verify if its rational elements (i.e., premises and claims) are
well justified, and the applied reasoning is valid. With these critical questions,
it is possible to easily detect fallacies and invalid arguments that without further
analysis could be accepted. Recently, a study to find the existing relations between
argumentation schemes with the persuasive principles of Cialdini [328] was con-
ducted. The observed relations can be used to analyse manipulative persuasion
in argumentation through the concept of the previously defined critical questions.
The initial research relating personality, arguments, and persuasion was continued
in both [92] and [369] works. In [92], the authors explore the existing correlations
between user descriptive features such as personality, gender, or age with the per-
suasive principles. The results depict a set of significant correlations found in their
study, which can be used to define user-tailored persuasive strategies. In [369],
persuasive arguments are generated automatically by an argumentation system in
the healthy eating and email security domains. The authors present the results of a
human evaluation of the automatically generated persuasive messages.

From the results, it is possible to conclude that persuasive properties such as
the six persuasive principles or the reasoning structure of the arguments (i.e., ar-
gumentation schemes [400]) have a strong impact on the human interpretation of
messages and arguments. Work by Thomas, Masthoff and Oren [368] delves fur-
ther into the meaning of this conclusion. A complete study is presented to validate
a new scale to measure the persuasiveness of messages and arguments perceived
by human users. The strong domain dependency of persuasion related research
is highlighted at the end of that work. Furthermore, from the results of the study
presented, it is possible to observe that the reasoning structure of the arguments
embeds better message persuasiveness than only considering Cialdini's principles.

A similar approach is presented in [323], where the persuasive power of an
argument is defined as a ranking of human preferences on different arguments,
and a complete study of the effect of user descriptive features (i.e., personality and

social interaction statistics) on the persuasive power of arguments is carried out. In this work, in addition to content-based argument types (i.e., privacy, trust, risk, and content), the persuasive power of argumentation schemes is also investigated. This research has recently been extended in [328], where the authors present a qualitative analysis, instead of a quantification of the persuasive power, based on user modelling features. For that purpose, the authors study the existing relations between the reasoning structures of arguments and the six principles of persuasion described above. From their findings, it is possible to observe that some of the most common patterns used in human argumentative reasoning have an underlying persuasive purpose.

A different approach is presented in [164], where user features are represented as preferences for different domain-specific concerns. The empirical results obtained in this study validate the human user preference model proposed. Furthermore, the authors use machine learning techniques to predict these preferences, which are used to improve the persuasiveness of an argument-based dialogue system. A more thorough review of previous similar work exploring argument-based computational persuasion in a variety of domains using different user features (i.e., models) is carried out in [179]. The authors propose a consistent framework for computational persuasion. The proposed framework is made up of three components: the domain model, the user model, and the dialogue engine, which are needed to generate user-tailored and domain-specific persuasive strategies.

Machine Learning for Estimating the Persuasiveness of Arguments

The impact of the same argument used in different contexts or presented to different audiences can substantially differ. Therefore, the automatic estimation of argument persuasiveness is a complex task that can be instantiated in a wide range of application domains. It is commonly approached as a mainstream machine learning task, but the input representation usually varies a lot from each research work mainly depending on its aim and domain. The first annotated corpus for argument strength prediction is presented in [278]. Every argument is annotated with a score reflecting its strength in a given context. Furthermore, the authors use linguistic features of the arguments to approach the prediction of their strength values using the SVM regression. Work by Zhang et al. [413] analyses the flow of content in a

debate in order to estimate the strength of the arguments used. The presented approach uses handcrafted argument descriptive features to predict which side (i.e., for or against) wins the debate.

An new perspective is presented in [315], where the authors present an agent-based approach to assist users with the choice of best arguments in a discussion. The agents use different policies based on the predictions of the most appropriate argument in a given context. SVMs, Decision Trees, and Multi-Layer Neural Networks are used to model human discussions and to estimate which argument is the best in a given context. The results are validated with human participants, concluding that the use of machine learning techniques to estimate the best (i.e., persuasive) argument can significantly improve the natural interaction of argumentation systems with human users.

Complementing all of these research works, Lukin et al. [220] carried out an analysis on how argument persuasiveness can vary depending on the audience. Users with different personalities may find emotional or rational reasoning more or less persuasive, pointing out the importance of considering audience features when trying to estimate the persuasiveness of arguments in different domains. Work by Gleize et al. [151] presents a different task instance for the automatic estimation of persuasion. This instance of the task relies on evidence pairs that are compared considering their persuasiveness in a specific domain. A new annotated corpus of argument pairs is presented and a Siamese Neural Network is trained on this corpus to predict which evidence is more convincing from every pair. With this approach, it is possible to determine the most convincing evidence, regardless of the audience or the context. Diversely, in [191], user tailored argument persuasion is predicted, considering three different personal features. These features are the users' interests, their prior beliefs towards some specific topic (previously analysed in [131]), and the users' personality. A new dataset is created from an online debate forum, and linguistic features from the written messages are used together with the previously introduced features to perform the prediction using a logistic regression classifier. On the other hand, in [401], the authors introduce the importance of the style complementing the content of the arguments in order to achieve better persuasion. In [31], the authors explore the impact of the written style of news editorials on the persuasiveness of arguments. Style features are modelled using psychological, emotional, and argumentative features. A bipolar instance of the

political spectrum is considered in the experiments (i.e., conservative and liberal). A comparison between content and style-based SVM classifiers is done, which is aimed at finding out the relevance of the style on the persuasiveness of arguments used in a newspaper. From the results of the experiments, the authors conclude that the written style has a significant influence on how an argument can influence the reader.

One of the main challenges of the machine learning approaches is to define the user-specific utility functions that make it possible for the models to learn how human persuasion works. This challenge has recently been addressed by Donadello et al. [119], where the authors explore two methods for utility prediction in combination of decision trees for modelling persuasion in dialogues. The first of these methods is called Evidence as Amount of Information, which consists of using a Support Vector Regression model to estimate the utility of an argument based on a set of evidence present in other different arguments. The second of the proposed methods is called Evidence as Depth of Searching, which is focused on clustering the user utility vectors. With this clustering, the model can group users based on similar utility features. Their findings show how machine learning is useful for problems of this kind, and the two explored methods make it possible to learn better utility functions for different user models.

Reinforcement Learning for Modelling Argumentative Dialogues

An alternative approach to the estimation of the arguments' persuasiveness aimed at finding dialogue strategies has recently been explored using Reinforcement Learning [359]. This is a machine learning approach that relies on a proper modelling of the environment (i.e., the dialogue) and the definition of a set of actions (e.g., argument utterance) and rewards (e.g., persuaded or not). Work by Georgila and Traum [148] brings these concepts to the Computational Argumentation domain. A simple negotiation case is presented, where a florist and a grocer who share a retail space try to convince each other on its configuration. The uses of different arguments are the actions available to each participant in this negotiation. The reward used by the Reinforcement Learning algorithm depends on the final agreement reached by both parties. An extension of that work is proposed in [267], where the presented approach considers the possibility of negotiating on

more than one issue.

Extending Reinforcement Learning research for argument-based persuasive dialogues, Monteserin and Amandi [244] present an instance of the Reinforcement Learning framework in an argument-based negotiation game. In this approach, argumentative agents learn the argument selection policy from the response (reward) of another agent while having a discussion. A simulated multi-agent system is used to validate the proposal in a laboratory experiment rather than a real situation.

A more complete dialogue model is explored in [8] where the authors present a Reinforcement Learning agent that is integrated into the "*DE*" model. The "*DE*" model has a set of five different moves available for the dialogue agents: assertion, question, challenge, withdrawal, and resolution demand. These moves are the actions that are available for the Reinforcement Learning argumentation agent. The rewards are awarded at the end of the negotiation, i.e., a positive one for the winning agent and a negative one for the losing one. Dialogue coherence and relevance are added to the previous approach in [9], which provide the Reinforcement Learning agent with additional dimensions to improve the naturalness of the learnt strategies. Recent work by Pecune and Marsella [272] explores how user conversational goals can impact a previously learnt dialogue policy. For that purpose, a new framework is presented where the Reinforcement Learning agent is integrated. Two versions of this agent are used in the experiments, one which takes the conversational goals into account in the policy learning phase and another which does not. The presented results shed light on a new important aspect that may help to improve the previous contributions in Reinforcement Learning dialogue policy learning, which are the users' conversational goals.

Argument-based Persuasive Chatbots

Finally, Computational Argumentation research has also focused on the development of trustworthy, persuasive chatbots, in order to find the most natural way to interact with human beings. The initial concept of an argumentative agent that is capable of having a dialogue with a human user was presented in [316]. In that work, the authors combine concepts of argumentation theory and argument-based knowledge representation with machine learning techniques on human dialogues and the optimisation of a Markov Decision Process. The resulting agent is limited

to persuading people in two different domains. Furthermore, it does not consider any model of the dialogue counterpart, which can be an important drawback considering previous research [163], where modelling opponents allow to increase the persuasiveness of the system. Work by Chalanguine et al. [83] explores the relevance of different types of arguments and user concerns when interacting with an argumentative chatbot. In that work, the authors analyse human reaction when using different types of arguments on them aimed at modifying their behaviour. Furthermore, the authors also point out an observed improvement in the interaction with human users when arguments regarding important concerns are used. These observations are integrated into a real argumentative chatbot and validated in two different experiments with real human-computer interaction. However, the interaction with this chatbot is limited due to the current restrictions of machine natural language understanding. In [82], the authors further develop the previous research limitations, proposing an argumentative chatbot that is complemented with a crowd-sourced argument graph. The argument graph is used by the chatbot as a knowledge representation database, which makes it possible to select the best argument or counterargument for a given situation in a coherent way. This method is validated with an experiment consisting of direct human interaction with the chatbot, and the results show that the argument graph allows the chatbot to increase its human persuasiveness.

To conclude this section, we consider that it is important to provide some discussion with respect to a concerning aspect of human persuasion i.e., manipulation. The development of computational persuasive systems is typically associated with the good will of assisting human users and helping them to achieve specific goals (e.g., doing more sport, having a healthy diet, preventing online privacy threats). However, when the conducted studies and experiments take into consideration features outside the purely rational nature of arguments (e.g., user personality, emotions), the findings may be used to manipulate human users. As we observe in the work reviewed in this section, user modelling features play a major role in argumentative persuasion. This is mainly because of the subjective nature of persuasion. Humans are not purely rational, and, therefore, the inclusion of emotional features (e.g., the personality) in the user models not only helps to increase the persuasiveness of the argumentative systems, but it also makes the interaction more natural for humans. Some works propose a complete user model consisting of both

rational-based features (i.e., prior beliefs) and emotional-based features (i.e., personality traits) [191]. Some other works include behavioural aspects in the user model (e.g., online social interaction record) [323] in addition to the personality traits so that the arguments can be generated taking their personal preferences into account [324].

However, we have not been able to identify any "purely rational" argument-based Human-Computer interactive system. This is entirely understandable, as the computer itself is a barrier for humans to feel such interactions natural, so the research aims to make the interactions as natural as possible. Therefore, we consider that the transparency of the research works on their proposed algorithms, the datasets used, and conducting a discussion about the ethical concerns of each work is more relevant than this duality of emotional and rational argument-based persuasion. Interestingly, we can relate this potential issue with the first of the three tasks of Computational Argumentation proposed in this paper. In fact, to prevent these malicious manipulative intentions in argument-based human-computer interactive systems, it can be helpful to have accurate and robust Argument Mining algorithms for detecting fallacious reasoning in natural language inputs. This is also a good example of how the Computational Argumentation process resembles the human argumentative reasoning process, and how after the invention of an argument (argument-based Human-Computer Interaction), it is advisable to identify, analyse, and evaluate the argument (Argument Mining & argument-based Knowledge Representation and Reasoning) before accepting its claims.

2.6 The Future of Computational Argumentation

In this section, we present our perspective on the future of Computational Argumentation research. We have divided this section into two parts. First, we review and analyse relevant contributions that have been made in yet unexplored aspects of Computational Argumentation or that propose a new research aspect themselves. Second, we present our general perspective on the open challenges in all of the analysed aspects of Computational Argumentation. This way, we seek to provide the reader not only with what has already been widely researched, but

also with what still needs to be explored in order to contribute to the development of Computational Argumentation.

Future Research Directions

Some specific aspects of different Computational Argumentation problems have not been reviewed in their corresponding sections. This is mainly due to the lack of research on these aspects or because of their recent appearance in the literature. We present a review of all of these research topics which either have been already presented but not researched in depth, or have recently been proposed for the first time.

Unsupervised Argument Mining

One of the major drawbacks of the supervised learning techniques used in Argument Mining is the huge dependency on large corpora. Every state-of-the-art neural architecture proposed for Natural Language Processing has more parameters than the previous ones. This large number of parameters can only be learnt with inputs consisting of thousands of millions of words. Other tasks such as language modelling rely on a training method that does not need annotations. However, that is not possible with Argument Mining, and argumentative corpora is hard and expensive to annotate. Recently, unsupervised Argument Mining has caught the interest of researchers. Unsupervised machine learning techniques (i.e., clustering) have been used in [373, 62] to identify major argumentative topics to ease the task of mining opinions or arguments in a natural language corpus. In [238], the authors present a new annotated corpus that is designed to calculate argument similarity on different topics. An approach based on a combination of hand-engineered features (i.e., n-gram cosine, Rouge, semantic textual similarity, Word2Vec, and Linguistics Inquiry Word Count) and clustering is proposed aimed at scoring this similarity between arguments. A different perspective on this topic is presented in [213], where the authors propose a new technique to automatically extract arguments from a large natural language corpus based on the repetition of a linguistic structure. Work by Reimers et al. [308] presents improvements to previous work on argumentative topic grouping. The authors evaluate unsupervised methods considering classical text features such as tf-idf (i.e., term frequency - inverse docu-

ment frequency), and the latest neural language models embeddings (i.e., ELMo
and BERT). An unsupervised feature approach for the automatic retrieval of best
counterarguments is presented in [393]. A new corpus consisting of argument-
counterargument pairs is presented. The counterargument retrieval can be seen as
the inverse approach of argument topic grouping. The authors approach this task
considering the aspect and stance of every available argument.

Argument Graph Mining

A very recent line of research is the automatic identification of argument graphs
(i.e., argumentation frameworks or argumentative structures). A preliminary prob-
abilistic approach for abstract argumentation was presented in [312], where the
authors research the inference of argumentative graphs (i.e., argument/attack iden-
tification) from the probabilistic abstract argumentation viewpoint. In this ap-
proach, the labelling of the arguments is used to infer the probabilistic argumenta-
tion framework structure. Following this trend, in [258], the authors do a thorough
analysis of the argumentation framework (i.e., argument graph) synthesis prob-
lem from the abstract viewpoint. The complexity, the basic properties, and sev-
eral algorithms are investigated and proposed for the synthesis of argumentation
frameworks. However, both works analyse the automatic generation of argument
graphs by considering argumentation semantics (i,e,. acceptability/defeat) as in-
put instead of natural language text. Recent research [147] reintroduces this idea
by combining previous argument mining techniques and decomposing arguments
into smaller functional elements (i.e., the target concept, the aspect, the opinion
on the target concept, and the opinion on the aspect). With these functional ele-
ments, the authors propose an approach to generate argument graphs based on the
detection of links between claim/premises and the relational types of inferences/-
conflicts in natural language text. The automatic generation of internal argument
graph structures is also tackled in [247]. The authors propose a new model archi-
tecture which is split into two parts. First, the model classifies the text spans into
different argumentative classes and generates embeddings (i.e., vector representa-
tions) of these. Second, the model predicts the edges between propositions using
an attention module that computes the scores of all of the proposition pairs. The
output of the presented approach is an argument graph of the internal elements of

an argument. Another perspective of this problem is presented in [178], where the authors present a theoretical framework for the automatic generation of argument graphs. This approach takes probabilistic information that can be acquired using machine learning models and creates a representative argumentation framework. Finally, work by Lenz et al. [211] presents a complete argument mining pipeline to automatically generate argumentation graphs from natural language text.

Deep Learning for Computational Argumentation Solving

One of the major issues related to the reasoning aspect of computational argumentation (i.e., solving) is the wall imposed by the complexity of this problem under the current paradigm of computing. Recent research has focused on overcoming this limitation with the use of Deep Learning approaches [199, 107] instead of using symbolic algorithms. The existing research on this topic is not very extensive, but similar results to classical approaches have been reported with significantly lower computational cost [108]. The recent interest on this topic has been manifested with the new Approximate Track introduced in the last edition of the International Competition on Computational Models of Argumentation (ICCMA'21), where the evaluation of the submitted algorithm takes into account their runtime in addition to their performance when defining argument's acceptability [225, 363].

Argument Summarisation

The automatic generation of text summaries has been an important part of the most recent NLP research. However, it is not possible to identify much work on the argumentative domain. First introduced in [36], argument summarisation is presented as the task of identifying small sets of talking points from a large set of arguments. A new corpus is created that is aimed at approaching the task of identifying these *key points* and generating summaries containing the main ideas stated in the arguments. The automatic generation of news summaries is a common application domain of the research in text summarisation. However, even if arguments are usually found in news, only a recent work has focused on this aspect. In [361], the authors present an approach for the automatic summarisation of news editorials from an argumentative viewpoint. A new corpus and annota-

tion guidelines for argument summarisation are proposed, and some experiments
using classical models for text summarisation are carried out. Finally, in [318],
the authors present a new corpus that is aimed at argument mining and argument
summarisation created from samples of competitive formal debates. The authors
provide a complete description of the created corpus and some preliminary results
in the argument summarisation task.

Argument-based Explainable AI

Computational Argumentation provides a set of techniques to extract, represent,
and use arguments when interacting with human users. These techniques are of ut-
most utility when designing new explainable Artificial Intelligence systems. Work
by Zheng et al. [412] proposes an argumentation-based approach that combines
the use of black-box models to learn arguments, and an argumentation system to
provide users with explanations on the final decision. In that work, a medical ex-
ample considering the detection of dementia is presented. Similarly, in [95], a new
approach for binary classification relying on artificial neural networks and abstract
argumentation frameworks is presented. The proposed approach uses neural net-
works for selecting the set of features and an abstract argumentation framework
to do the binary classification. The argumentation framework allows explanations
for each classification to be easily generated. A different approach for explaining
Artificial Intelligence inspired by case-based reasoning and abstract argumentation
is introduced in [110]. This approach relies on the definition of a previously de-
fined complex setting of cases. Abstract argumentation frameworks are generated
from the defined cases (treated as arguments). The outcomes (and explanations)
of the system are provided by such frameworks as an arbitrated argumentative
dispute. Explainable recommender systems using argumentation are presented in
[96]. The authors propose an approach that combines NLP techniques to extract
product reviews and their assigned votes. Then, a quantitative bipolar argumen-
tation framework is used to make a graph-based representation of the previously
extracted reviews (treated as arguments). Finally, a score is assigned to each ar-
gument, and gradual semantics are used to evaluate the argumentation framework.
The acceptable set of arguments is provided to the human user as the explanations
on a given product (i.e., a movie). all of the previous ideas presented in the explain-

able AI reviewed papers converge in [97]. The authors present a new paradigm for generating dialectical explainable predictions: Data-Empowered Argumentation (DEAr). DEAr is defined as a transparent prediction method from which dialectical explanations can be drawn in a natural way (contrary to the well-known black box models). This approach relies on the idea of extrapolating data into arguments and generating argumentative debates from data. The dialectical outcomes of such debates are usually called predictions in other paradigms, but they are generated following a transparent, explainable method.

Emotions in Argumentation

The final aspect we are reviewing in this section and which has caught the interest of researchers in the previous years is the emotional part of argumentation. Although it is not a new line of research, there is not much work focusing on it, mainly due to the high complexity of researching this topic. Doing experiments in this area requires human participants and specific hardware for emotion recognition, which significantly limits the possibilities for researchers. The research in this area can be classified into two main groups: work researching emotional reasoning and work researching the emotional impact of argument-based reasoning. On the one hand, to understand what emotional reasoning is, it is important to distinguish between rational and emotional reasoning. Whilst the first relies on logic, emotional reasoning appeals to human emotions to convince or persuade other humans [236]. A modelling of natural argumentation using intelligent agents that combine emotional reasoning with its logical counterpart is presented in [76]. A study on argumentative persuasion showed that human users, in general, found it more natural to combine both aspects than to use purely rational persuasive strategies [229, 230]. Further research on this topic raised the idea of how emotions may directly influence the evaluation of arguments [254], and, therefore, should be taken into account when dealing with emotional entities. Following this idea, in [111], the authors formally define an emotional argumentation framework. This framework relies on the basic concepts of Dung's abstract argumentation framework and includes the emotional factor as an additional representative element of arguments. Relatedly, in [165], a framework to emotionally model participants in a dialogue is presented. Finally, in [219], the authors propose a new class of Ar-

gumentation Schemes that consider emotions as *objects* in argumentation. These patterns are called emotional argumentation schemes.

On the other hand, we have identified a recent line of research that aims at investigating if rational reasoning can trigger human emotions, and understand their implications. For that purpose, in [51], the authors present a preliminary experiment where the participants were debating while wearing specific hardware for detecting their emotions. The observed results indicate clear trends that can be discerned from an emotional analysis while having argumentative dialogues. These experiments were followed by further research with a similar study where persuasive aspects of argumentation were also taken into account [52, 385]. The three Aristotelian argumentative strategies (i.e., Ethos, Logos, and Pathos) are taken into account in the experiments. The results of this second experiment show that Logos strategies require a greater effort from the participants, whilst Pathos and Ethos induce more engagement. Furthermore, Pathos strategies (i.e., emotional appealing strategies) have shown to be the most persuasive strategy of the Aristotelian approach. A third dimension is added in [386], where the participants' personality is taken into account. The results show how, in some situations, participants' personality traits may be closely related to the brain activity of debaters and the triggered emotions.

Open Challenges

Argumentation theory has been studied since a long time ago. However, Computational Argumentation is a (relatively) new area of research in Artificial Intelligence. Furthermore, not all of the aspects of Computational Argumentation have been researched at the same pace. The structure of Computational Argumentation research presented in this paper is aimed at emphasising the existing differences among these aspects. While Argument Mining is closely related to the advances presented in Natural Language Processing research, Argument-based Knowledge Representation and Reasoning mainly relies on formal logic and graph theory, and argument-based human-computer interaction focuses on the understanding of human behaviour and social interactions. Although the reviewed research on these three aspects presents promising results, there is still a long way to go before considering the problem solved. In the following sections, we analyse the most promi-

nent open challenges in each one of the three aspects of Computational Argumentation.

Argument Mining

This line of research is directly related to advances in Natural Language Processing. This relation, however, is double-edged. On one hand, Argument Mining-related research can benefit from advances of other non-related research framed in the NLP context. However, NLP research is in constant evolution and still has many limitations. Thus, exclusively relying on these advances may have a significant drawback effect on the progress of Argument Mining research. Furthermore, state-of-the-art Argument Mining depends on data, but argumentative data is hard to obtain and expensive to annotate. Thus, this constraint may also be hampering research on this topic. Exploring the behaviour of less data-dependent techniques in this area, and creating larger corpora are two important contributions that can have fruitful results in Argument Mining.

Furthermore, argumentative discourse is usually highly dependant on the context. In spoken conversations, it is very common to try to avoid redundancy and repetitions (e.g., with the use of pronouns). A similar resource is used in argumentation, an enthymeme is an argumentative construction where either a claim or a premise has been omitted, assuming its knowledge from the other part of the interaction. The correct automatic detection of such elements and their interpretation is still an important challenge in Argument Mining research. Furthermore, Argument Mining has been mainly researched using text exclusively. For a complete analysis of spoken debates and conversations, the inclusion of acoustic features is an important extension to the existing systems.

The last aspect that we think deserves attention in Argument Mining research is the design of complete functional Argument Mining systems. Although full pipelines and end-to-end approaches have been researched in the literature, their evaluation is focused on their performance on the underlying tasks individually and not holistically. However, once the tasks of segmenting text and detecting argumentative components have been performed, relations between arguments need to be identified in order to provide a complete analysis of argumentative discourses. Argumentation is usually presented in the shape of long and complex reasoning.

Therefore, doing a complete automatic argumentative analysis of such inputs may
have severe runtime cost implications. Research on how to efficiently analyse argumentative inputs still remains a challenge. Furthermore, real-time Argument
Mining has not yet been explored, which would significantly benefit from this last
identified open challenge.

Argument-based Knowledge Representation and Reasoning

Computational representations of arguments are mainly divided into two different
approaches: abstract and structured. While abstract representations leave aside any
linguistic or structural aspect of arguments, structured approaches try to keep these
aspects when generating the computational representation of arguments. Both approaches are acceptable, depending on the nature and needs of each specific application domain. Computational representations of arguments play a major role
in the Computational Argumentation process since they define the needed frameworks to *embed* the argumentative information into computer systems. Thanks to
these *embeddings* (i.e., representations), it is possible to propose new algorithms
to automatically calculate and solve a specific argumentation framework.

The main limitation of argumentation solving are the high computational complexity of the problem and the lack of domain-specific information for approximating human reasoning. Defining the extensions of an argumentation framework
can be, in the worst cases, beyond the NP complexity class. It is here where the
most recent probabilistic approaches, based on machine learning and deep learning, may play a major role. An approximate algorithm track for argumentation
solving was included in the 2021 edition of the International Competition on Computational Models of Argumentation (ICCMA 2021) to explore and research this
new perspective of the task. Overcoming the limitations caused by the complexity
of this problem is still an important open challenge in the area. On the other hand,
more research needs to be done on new argument representations or *embeddings*
that are better able to capture the application domain or the context that formal
approaches need. This way, computational argumentation reasoning will provide
better approximations to human reasoning in real use cases rather than specifically
designed experimental cases.

Argument-based Human Computer Interaction

As reviewed in previous sections of this work, the automatic generation of arguments has been approached from different perspectives. Template-based and retrieval techniques are significantly limited by the defined templates or the available argument databases. Natural language argument generation has presented a more flexible approach to this problem. However, natural language generation is a Natural Language Processing area of research and has similar drawbacks to the ones presented in Section 2. In order to generate completely coherent arguments that are not limited to some specific domains, further research and more argumentative data will be needed. Therefore, it seems to be the best approach to the automatic generation of less constrained natural language arguments, but it is far from being solved. Different limitations can be observed in the computational persuasion aspect of Computational Argumentation. Human persuasion is hard to model and measure from the computational viewpoint. Although several models of argument-based computational persuasion have been proposed in the literature, most of them still need to be evaluated considering a larger set of human participants. Experiments are usually carried out in small populations due to the complexity of conducting large-scale experiments with human participants. Furthermore, a very limited amount of public data containing human aspects of argument persuasion is available to the research community.

Finally, the majority of Computational Argumentation research usually focuses on a single task, and uses the *standard* techniques applied in each of these tasks. After doing a complete review of the main tasks in Computational Argumentation, we have found a lack of works that bridge the gap among them. Thus, we have identified the gradual integration of the three main tasks underlying Computational Argumentation research analysed in this work as a long-term open challenge.

2.7 Conclusions

In the last few years, we have been experiencing a great era in the development of Artificial Intelligence. With the significant improvements in computing hardware, the creation of large datasets and corpora, and the exceptional empirical

results observed in research and industry experiments, AI has caught the interest
of researchers worldwide. Computational Argumentation was born from theoretical argumentation concepts and logic (e.g., nonmonotonic logic, epistemology,
legal reasoning, and philosophy among others). Recently, new paradigms of Computational Argumentation have been proposed as a way to integrate the latest AI
advances. The development of new algorithms has made it possible to undertake
tasks such as the automatic identification of argumentative structures or the automatic generation of natural language arguments. However, there is still a big gap
between the different research communities and their interpretations of Computational Argumentation. These differences are reflected in the heterogeneity that
can be found in the reviewed research, both among the research conducted inside
each task of Computational Argumentation, and between research conducted in
different tasks.

In this work, we have presented a complete structure for Computational Argumentation research from the perspective of human argumentation and reasoning.
In our analysis, we have put together different aspects (e.g., formal logic, natural language processing, and human behaviour analysis) and given a coherent
and consistent interpretation framed into the area of Computational Argumentation. We divided the reviewed work into three main tasks: Argument Mining,
Argument-based Knowledge Representation and Reasoning, and Argument-based
Human-computer Interaction. Thanks to the proposed structure, we have provided
a general understanding and a connected perspective of the most important advances for the different underlying tasks of Computational Argumentation. Furthermore, we have performed a thorough analysis of the most promising results
achieved by previous research in each of these tasks. However, Computational
Argumentation research still must consolidate and explore new angles of human
argumentation from the AI perspective. We have summarised the most important
future trends in Computational Argumentation and the main open challenges that
still require more attention by the research community.

Part III

Automatic Analysis of Argumentative Discourse

VivesDebate: A New Annotated Multilingual Corpus of Argumentation in a Debate Tournament

RAMON RUIZ-DOLZ, MONTSERRAT NOFRE, MARIONA TAULÉ, STELLA HERAS AND ANA GARCÍA-FORNES

Abstract

The application of the latest Natural Language Processing breakthroughs in computational argumentation has shown promising results, which have raised the interest in this area of research. However, the available corpora with argumentative annotations are often limited to a very specific purpose or are not of adequate size to take advantage of state-of-the-art deep learning techniques (e.g., deep neural networks). In this paper, we present VivesDebate, a large, richly annotated and versatile professional debate corpus for computational argumentation research. The corpus has been created from 29 transcripts of a debate tournament in Catalan and has been machine-translated into Spanish and English. The annotation contains argumentative propositions, argumentative relations, debate interactions and professional evaluations of the arguments and argumentation. The presented corpus can be useful for research on a heterogeneous set of computational argumentation underlying tasks such as Argument Mining, Argument Analysis, Argument Evaluation or Argument Generation, among others. All this makes VivesDebate a valuable resource for computational argumentation research within the context of massive corpora aimed at Natural Language Processing tasks.

3.1 Introduction

Argumentation is the process by which humans reason to support an idea, an action or a decision. During this process, arguments are used by humans to shape their reasoning using natural language. In an attempt to understand how really does human reasoning work, researchers have focused on analysing and modelling the use of arguments in argumentation. This problem has been usually approached by philosophers and linguists [26, 375, 400]. However, recent advances in Artificial Intelligence (AI) show promising results in Natural Language Processing (NLP) tasks that were previously unfeasible (e.g., machine translation, text summarisation, natural language generation), but now leave an open door to explore more complex aspects of human language and reasoning. Computational argumentation is the area of AI that aims at modelling the complete human argumentative process [301, 320], and encompasses different independent tasks that address each of the main aspects of the human argumentation process (Figure 3.1). First, argument mining [265, 207] focuses on the automatic identification of arguments and their argumentative relations from a given natural language input. Second, argument representation [124, 49, 125, 288] studies the best computational representations of argument structures and argumentative situations in different domains. Third, argument solving (or evaluation) [124, 38, 299, 108] researches on methods and algorithms to automatically determine the set of *acceptable* (i.e., winner) arguments from the complete set of computationally represented arguments. Finally, the argument generation [350, 55, 322, 30] task is mainly focused on the automatic creation of new arguments from a context and a set of known information regarding some specific topic.

Due to the huge heterogeneity of the tasks, each one of them requires different corpus structures and annotations to be approached from the computational viewpoint. Thus, depending on the data source, the annotations, and the task, a corpus might only be useful to approach a unique aspect of argumentation. For example, a simple corpus with small unrelated pieces of text annotated with *argument/non-argument* labels will only be useful for the argument mining task. In addition to this limitation, there is the great complexity underlying the human annotation of such corpora. This elevated complexity has a devastating impact on the versatility of the available corpora. The majority of the identified publicly
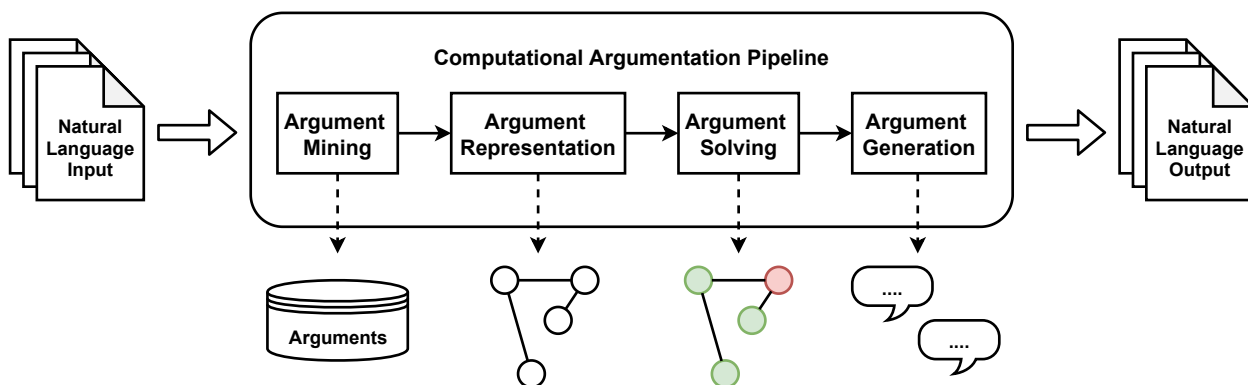
Figure 3.1: General pipeline for Computational Argumentation tasks.

available data for computational argumentation research is annotated either for a
very specific task of the complete argumentation process, and does only consider
its most superficial aspects (see Section 3.5). Furthermore, Deep Learning (DL)
[155] has recently shown outstanding results in many different AI areas (e.g., NLP
and Computer Vision among others). DL differs from the previous classic ma-
chine learning approaches in a major aspect, the data representations. While clas-
sic machine learning approaches usually required an important effort on feature
engineering the input for each specific task, DL algorithms model the represen-
tation of each input automatically during the training process. However, despite
presenting significantly superior results, DL approaches require a large amount of
data to observe this improvement in the experimentation. Computational argu-
mentation research has recently focused on the implementation of DL algorithms
to approach each of its underlying tasks. In argument mining, the Transformer
architecture [380] that presented outstanding results in the majority of NLP tasks
has become the focus of attention. Recent research compares and proposes new
Transformer-based neural architectures for both argument mining [411] and ar-
gument relation identification [325]. In argumentation solving, recent research
proposes a deep graph neural network to automatically infer the *acceptable* argu-
ments from an argumentation graph [108]. Finally, the latest argument generation
research proposals explore the use of DL architectures to automatically generate
natural language arguments [30, 339] rather than using templates or retrieving ar-
guments from a database. Thus, this trend of applying, adapting and proposing new
state-of-the-art approaches to the computational argumentation research makes the

creation of new high-quality large corpora a priority. From the publicly available resources for computational argumentation research it is possible to observe a strong trade-off between the size of the corpora and the *quality* of the annotations. We understand the term *quality* in this situation as the depth that annotations present from an argumentative viewpoint. The majority of the most extensive corpora available for computational argumentation research are usually focused on a very specific argumentative concept (e.g., segmentation, argument component identification, etc.), and only consider short pieces of argumentative text, which makes possible to simplify (or even automate) the annotation process. However, these simplifications imply a significant loss of context and information from the annotated arguments.

The main objective of this article is to present *VivesDebate*, a new annotated argumentative multilingual corpus from debate tournaments. For that purpose, the contribution of this paper is threefold: (i) the creation of a new resource for computational argumentation research; (ii) the description of the annotation guidelines followed in the creation of the corpus; and (iii) the comparison and review of ten of the most relevant corpora for computational argumentation research. The *VivesDebate* corpus has been created based on three main aspects that are of paramount importance in recent developments in AI and computational argumentation: the size, the quality, and the versatility provided by the corpus in its different possible uses. The *VivesDebate* corpus has a total of *139,756* words from 29 annotated debates from the 2019 university debate tournament organised by the *"Xarxa Vives d'universitats"*[1]. Each debate is annotated in its complete form, making it possible to keep the complete structure of the arguments raised in the course of the debate. Thus, the presented corpus is a relevant contribution for most of the main computational argumentation tasks such as argument mining, argument representation and analysis, argument solving, and argument generation and summarisation. Furthermore, the debates have been machine-translated from its original language (i.e., Catalan) to Spanish and English languages. The *VivesDebate* corpus is released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0) and can be freely downloaded from Zenodo.

The rest of the paper is structured as follows. Section 3.2 defines professional

---

[1]https://www.vives.org/programes/estudiants/lliga-debat-universitaria/

debate tournaments and their structure, from which our corpus has been created. Section 3.3 thoroughly describes the methodology followed during the annotation process. Section 3.4 analyses the *VivesDebate* corpus and presents its most relevant features. Section 3.5 analyses and compares the most important existing corpora for computational argumentation research. Finally, Section 3.6 highlights the main conclusions and the future research work that will take this corpus as its starting point.

3.2 Argumentation in Professional Debate Tournaments

Debate tournaments and competitions exist in different forms. Each type of debate has its own rules, structure, and aim. However, regardless of these differences, in every type of debate the winner is always the one presenting the best arguments and the most solid reasoning. Thus, given this condition, it is natural to think that one of the best sources to analyse the human argumentative discourse are debate competitions, mainly due to the higher quality of the reasoning presented by the participants. In this section we thoroughly describe the standard academic debate tournament, which has served as the source for the corpus presented in this paper. This type of debate presents one of the most popular structures and rules used in university debate tournaments. First, a controversial topic is chosen, and the debating question is proposed in a way that two conflicting stances are created (in favour or against). Each debate is divided into three main phases: the introduction, the argumentation and the conclusion. Each team, consisting of 3 to 5 debaters (university students), is randomly assigned a stance for the tournament topic at the beginning of each debate. The team opening the debate is also drawn before its start. In the subsequent description of the flow of the debate, we will assume that the proposing (in favour) team begins, and the opposing (against) team follows up. Thus, the proposing team opens the debate with a 4 minute introduction, where the main aspects that will be used to support their arguments are presented. Then, the opposing team is able to introduce their main ideas on the topic in another 4 minute introduction. Once the introduction phase concludes, each team has two rounds of 6 minutes to argue their stances by presenting new arguments or sup-

Introduction (INTRO) 8 min	Proponent team: Presentation of its main lines of argument	4 min
	Opponent team: Presentation of its main lines of argument	4 min
Argumentation (ARG1) 12 min	Proponent team: Development of its main arguments	6 min
	Opponent team: Development of its main arguments and rebuttal	6 min
Argumentation (ARG2) 12 min	Proponent team: Reinforcement of its main arguments and rebuttal	6 min
	Opponent team: Reinforcement of its main arguments and rebuttal	6 min
Conclusion (CONC) 8 min	Opponent team: Conclusion of the debate	4 min
	Proponent team: Conclusion of the debate	4 min

Figure 3.2: General structure for academic debate tournaments.

porting the previously introduced ones. Furthermore, in the argumentation phase, participants can also attack the arguments proposed by the other team. Finally, the debate is closed by each team's 4 minute conclusion. The order in which each team concludes its argumentation is inverted with respect to the previous phases of the debate. Thus, in our instance of a debate where the proposing team's introduction was the first phase, the opposing team will be the first to present its main conclusions, and the debate will be closed with the conclusion of the proposing team. Figure 3.2 summarises the presented structure of academic debates from which the *VivesDebate* corpus has been created.

The outcome of each debate is decided by a Jury that evaluates six different aspects of the debate weighted by their relevance. First, the Jury assesses how solid is each team's thesis and how has it been defended during the debate (22.5%). Second, the Jury evaluates essential aspects of the argumentation such as the relation of the arguments with the topic, the strength and originality of the presented arguments, and the coherence of the discourse and its structure (22.5%). Third, the Jury assesses how well each team has reacted and adapted to the adversary's attacks and arguments (20%). Fourth, the Jury evaluates the security in discourse and the capacity of finding weak spots in the adversary's argumentation (15%). Fifth, aspects such as the oral fluency, the semantic and grammatical correctness, the richness of the vocabulary used, and the non-verbal language are also assessed by the Jury (10%). Finally, the Jury considers positively the respectful attitude shown during the debate (10%).

A numerical score is assigned to each one of these aspects during the final deliberation, and a weighted sum of the six values indicates the score of each team. Furthermore, each team can be penalised if some specific conditions are met during the debate. Three different penalisation degrees are considered depending on their severity: warnings, minor faults, and serious faults. The warnings do not have a direct impact on the previously defined score, and may happen when the team members talk between them during the debate, and when the speakers do not comply with the assigned duration of each phase of the debate. A minor fault will reduce the score by 0.5, and happens with the accumulation of two warnings, with minor behavioural issues, and when a team uses fake news to support their arguments. Finally a serious fault reduces the score by 3 points, and will only happen if a team commits serious disrespectful acts (e.g., insults, racism, misogyny, etc.), or violates the rules of the tournament. Each Jury is specifically constituted for each debate and composed of, at least, three members that are assigned before starting the debate. Thus, the final score (*FS*, Equation 3.1) consists of a normalisation of the score (*S*) minus the penalisation (*P*) assessed by each member of the Jury ($\forall j \in J$):

$$(3.1) \qquad FS = \frac{\sum_{j}^{J} S_j - P_j}{|J|}$$

3.3 Annotation Methodology

In this section, we describe the annotation tagset used, the criteria applied and the annotation process carried out, including the Inter-Annotator Agreement tests conducted for the annotation of the *VivesDebate* corpus. The annotation task consists of three main subtasks: First, the annotators review and correct the transcriptions automatically obtained by the MLLP transcription system[2] [187], the IberSpeech-RTVE 2020 TV Speech-to-Text Challenge award winning transcription system developed by the Machine Learning and Language Processing (MLLP) research group of the VRAIN. Then, the Argumentative Discourse Units (ADUs) of each

---

[2]https://ttp.mllp.upv.es/

debate, which are the minimal units of analysis containing argumentative information, are identified and segmented. Finally, the different types of argumentative relationships between the previously identified ADUs are annotated. All these tasks were manually carried out by two different annotators and supervised by a third senior annotator. In the following, we describe in more detail each of these subtasks.

Revision and correction of automatic transcriptions

The first task we performed was the revision of the automatic transcriptions of each debate. The duration of each debate is approximately 50 minutes. The annotators reviewed whether these automatic transcriptions corresponded to the audio recorded in the videos. Due to the time required to completely correct all the transcriptions, we decided to focus on the quality of the transcriptions corresponding to the ADUs in order to ensure the comprehensibility of the arguments presented during the debate and avoid misunderstandings, as a general criterion. The remaining text, which will not be part of the ADUs, was checked for inconsistencies or glaringly obvious errors. Regarding the specific criteria established, we agreed:

1. To maintain the transcriptions of the different linguistic variants of the same language used in the original debates, Balearic, central and north-western Catalan and Valencian, as well as the use of words or expressions from other languages, but which are not normative, such as borrowings from Spanish and from English. It is worth noting that the eastern variants (Balearic and central Catalan) are the prevailing variants of the MLLP transcriber.

2. To correct spelling errors, such as '*autonoma' instead of 'autònoma/autonomous' (missing accent), '*penalitzar vos' instead of 'penalitzar-vos/penalize you' (missing hyphen).

3. To amend those words that were not correctly interpreted by the automatic transcriber, especially wrongly segmented words. For instance, '*debatre/to debate' instead of 'debatrà/he or she will debate', or '*desig de separa/the desire to (he/she) separates' instead of 'desig de ser pare/the desire to be

a father'. In the first example, the automatic transcription does not interpret correctly the tense and person of the verb, and in the second example it wrongly interprets as a single word ('separa') two different words 'ser pare', probably due to the elision of 'r' when we pronounce 'ser' and due to the confusion that can be caused by the Catalan unstressed vowels 'a' and 'e', which in some linguistic variants are pronounced the same way. Most of these errors are related to homophonous words or segments, which the automatic transcriber cannot distinguish correctly, and are also probably due to the different linguistic variants of the same language used in the debates (Balearic, central and north-western Catalan and Valencian).

4. To follow the criteria established by the linguistic portal of the Catalan Audiovisual Media Corporation [3] for spelling the names of persons, places, demonyms, etc.

5. To capture and write down the main ideas in those cases in which the quality of the audio does not allow us to understand part of the message conveyed.

6. Noisy sounds and hesitations (e.g. 'mmm', 'eeeeh'), self-corrections (e.g. 'mètode arlt alternatiu/arlt alternative method') and repetitions ('el que fa el que fa és ajudar/what it does it does is to help)' are not included because they do not provide relevant information for computational tasks focused on argumentation.

The data that is automatically transcribed and manually reviewed appears in plain text format without punctuation marks and without capital letters. The revision of the automatic transcriptions took us an average of two hours for each debate, even though we performed a superficial revision of those fragments of text in which ADUs were not present. Therefore, this type of revision is undoubtedly a time-consuming task.

---

[3] http://esadir.cat/

Segmentation of debates in Argumentative Discourse Units (ADUs)

The next task consisted of segmenting the text into ADUs (**ADU** tag) and annotating it, identifying: a) the participant who uttered the ADU and the part of the debate in which it was used (i.e. in the introduction, argumentation or conclusion) by means of the **PHASE** tag; b) their stance towards the topic of the debate (against or in favour) annotated with the **STANCE** tag; and the number of the argument presented during the argumentation part tagged as **ARGUMENT_NUMBER** (the identified arguments are labelled with a numerical value from 1 to the maximum number of arguments found in the debate) (see Section 3.2 and Figure 3.2 for more information). The preparatory part of the debate, in which the topic and the stance each team had to adopt were decided, and the conclusions were not segmented.

An Argumentative Discourse Unit (ADU) is defined as the minimal unit of analysis containing argumentative information [274]. Therefore, an argument can consist of one or more ADUs, each contributing a different (or complementary) argumentative function (e.g., premises, pieces of evidence, claims).

Next, we describe the criteria followed for the segmentation of ADUs and the tags assigned to each ADU for identifying them. This task was also manually performed by the same two annotators and reviewed by the senior annotator.

Segmentation criteria

Two general criteria were applied to the segmentation of ADUs: The first is that ADUs are created following the chronological order in which they appear in the discourse. Each ADU was assigned a unique **ID** tag for identifying them and showing their position in the chronological sequence. The second criterion is related to the quality of ADUs, which means that their content has to be clear, comprehensible and coherent. Therefore, this involves a further revision and, if necessary, a correction of the text. Each ADU corresponds to a transcribed text segment considered as a unit of argumentation and was included in the **ADU** tag. The ADUs are generally equivalent to a sentence or a dependent clause (for instance a subordinated or coordinated clause). It is worth noting that we also found ADUs which contained a subsegment that was, in turn, another ADU (for instance, rela-

tive clauses, as we will see below). In this case, the second ADU was also assigned its own ID and ADU tags.

The specific criteria followed for the segmentation of ADUs were the following:

- Punctuation marks must not be added to the content of the ADUs. The annotators could view the debate recording to solve ambiguous interpretations and avoid a misinterpretation of the message.

- Anaphoric references are left as they are, there is no need to reconstruct them, that is, the antecedent of the anaphora is not retrieved.

  (1a) [és una forma d'explotació de la **dona**]$_{ADU1}$
        [és una forma de cosificar-**la**]$_{ADU2}$
        [i és una forma que fa que vulneri la **seva** dignitat]$_{ADU3}$

  (1b) [it is a way of exploiting **women**]$_{ADU1}$
        [it is a way of objectifying **them**]$_{ADU2}$
        [and it is a way of violating **their** dignity]$_{ADU3}$

  In example (1), the text contains three different ADUs and two anaphoric elements appear in the second and third ADUs, '-la/her' in ADU2 and 'seva/her' in ADU3, which refer to the same entity ('dona/woman'), but we did not retrieve their antecedent in the corresponding ADUs.

- Discourse markers (2) must be removed from the content of the ADUs, but the discourse connectors (3) will be kept. Discourse connectors are relevant because they introduce propositions indicating cause, consequence, conditional relations, purpose, contrast, opposition, objection, etc., whereas discourse markers are used to introduce a topic to order, to emphasize, to exemplify, to conclude, etc. We followed the distinction between discourse connectors and markers established in the list provided by the Language and Services and Resources at UPC[4].

---

[4]`https://www.upc.edu/slt/ca/recursos-redaccio/criteris-linguistics/`
`frases-lexic-paragraf/marcadors-i-connectors`

(2) Examples of discourse markers: 'Respecte de/ regarding'; 'en primer lloc/first or firstly'; 'per exemple/for instance'; 'en d'altres paraules/in a nutshell'; 'per concloure/in conclusion'.

(3) Examples of discourse connectors?: 'per culpa de/due to'; 'a causa de/because of'; 'ja que/since'; 'en conseqüència/consequently'; 'per tant/therefore'; 'si/if'; 'per tal de/in order to'; 'tanmateix/however'; 'encara que/although'; 'a continuació/then'.

- Regarding coordination and juxtaposition, we segmented coordinated sentences differently from coordinated phrases and words: a) In coordinated sentences, each sentence was analysed as an independent ADU and the coordinating conjunction (e.g. copulative, disjunctive, adversative, distributive) was included at the beginning of the second sentence (4). The type of conjunction can be used to assign the argumentative relation in the following task; b) In coordinated phrases and words, each of the joined elements are included in the same single ADU (5).

(4a) [l'adopció s'està quedant obsoleta]$_{ADU1}$ [**i** per això hem de legislar]$_{ADU2}$

(4b) [adoption is becoming obsolete]$_{ADU1}$ [**and** that's why we have to legislate]$_{ADU2}$

(5a) [**justícia i gratuïtat** per evitar la desigualtat social]$_{ADU1}$

(5b) [**justice and gratuity** to avoid social inequality]$_{ADU1}$

- Regarding subordinated sentences, the subordinated (or dependent) clause is analysed as an ADU that is independent from the main (or independent) clause, and includes the subordinating conjunction (6). The type of subordinating conjunction (e.g. causal, conditional, temporal, etc.) can be used to assign the argumentative relation.

(6a) [**si** s'acaba el xou]$_{ADU1}$ [s'acaba la publicitat]$_{ADU2}$

(6b) [**if** the show is over]$_{ADU1}$ [advertising is over]$_{ADU2}$

In example 6, two different ADUs are created, in which the second clause (ADU2) will be then annotated as an inference argumentative relation from

the first clause (ADU1). In this way, if there later appears a proposition that is only related to one of the two previous clauses, this proposition can be related to the corresponding ADU.

- Regarding relative clauses, the relative clause is included in the same ADU as the main clause, because these clauses function syntactically as adjectives. However, they can be treated as a subsegment of the ADU if the relative clause acts as an argument (7).

(7a) [suposa un desig de les persones que fa perpetuar un rol històric de la dona]$_{ADU1}$ [que fa perpetuar un rol històric de la dona]$_{ADU2}$
[afirma i legitima que les dones han de patir]$_{ADU3}$

(7b) [this presupposes a desire by people to perpetuate a historical role for a woman]$_{ADU1}$
[to perpetuate a historical role for a woman]$_{ADU2}$
[this asserts and legitimatizes the idea that women must suffer]$_{ADU3}$

In (7), the relative clause is segmented as an independent ADU2, because it is the argument to which ADU3 refers to.

- In reported speech or epistemic expressions, we distinguish whether the epistemic expression is generated by one of the participants in the debate (8) or is generated by another (usually well-known or renowned) person (9). In the former, the subordinate clause is only analysed as an ADU while, in the latter, the whole sentence is included in the same ADU.

(8a) Jo pense que es deurien prohibir les festes amb bous ja que impliquen maltractament animal
[es deurien prohibir les festes amb bous]$_{ADU1}$
[ja que impliquen maltractament animal]$_{ADU2}$

(8b) I think bullfights should be banned as they involve animal abuse
[bullfights should be banned as they involve animal abuse]$_{ADU1}$
[as they involve animal abuse]$_{ADU2}$

(9a) [Descartes pensa que cos i ànima són dues entitats totalment separades]$_{ADU1}$

(9b) [Descartes thinks that body and soul are two totally separate entities]$_{ADU1}$

In example (8), the ADU already indicates which specific participant uttered this argument in the STANCE and ARGUMENT_NUMBER tags associated, as we describe in more detail below (Section 3.4). Therefore, including this information would be redundant. However, in example (9), it would not be redundant, and could be used in further proposals to identify, for instance, arguments from popular, well-known or expert opinion and arguments from witness testimony.

- With regard to interruptions within the argumentative speech produced by the same participant, the inserted text will be deleted (10), whereas if the interruption is made by a participant of the opposing group, it will be added to another ADU (11).

(10a) el que estan fent vostès, i aquest és l'últim punt, és culpabilitzar a la víctima
[el que estan fent vostès és culpabilitzar a la víctima]$_{ADU1}$

(10b) what you are doing, and this is my last point, is to blame the victim
[what you are doing is to blame the victim]$_{ADU1}$

(11a) centenars de dones han firmat un manifest per tal de garantir d'adherir-se a la seua voluntat de ser solidàries, *que passa si els pares d'intenció rebutgen el nen i on quedaria la protecció del menor en el seu model*, completament garantida per l'estat
[centenars de dones han firmat un manifest per tal de garantir d'adherir-se a la seua voluntat de ser solidàries completament garantida per l'estat]$_{ADU1}$
[que passa si els pares d'intenció rebutgen el nen i on quedaría la protecció del menor en el seu model]$_{ADU2}$

(11b) hundreds of women have signed a manifesto to ensure that they adhere to their willingness to be in solidarity, *what happens if the intended parents reject the child and where would the protection of the child be in your model*, fully guaranteed by the state

[hundreds of women have signed a manifesto to ensure that they adhere to the to their willingness to be in solidarity fully guaranteed by the state]$_{ADU1}$

[what happens if the intended parents reject the child and where would the protection of the child be in your model]$_{ADU2}$

In (11), the initial fragment of text is segmented into two different ADUs. The interruption (in italics) is segmented separately and tagged as ADU2. If an argumentative relation is observed in an interruption made by the same participant, this part of the text will be analysed as a new ADU and, therefore, will not be removed since it establishes the relationship between arguments.

- Interrogative sentences are analysed as ADUs, because they can be used to support an argument (12), except when they are generic questions (13).

(12a) [Això fa que no necessàriament ho valori econòmicament?]$_{ADU1}$ [No]$_{ADU2}$

(12b) [Does that necessarily mean that I do not value it economically?]$_{ADU1}$ [No]$_{ADU2}$

(13a) Què en penses d'això?

(13b) What do you think about that?

It should be noted that tag questions are not annotated as ADUs. In (14) 'oi?/right?' is not tagged as an ADU.

(14a) [Això fa que no necessàriament ho valori econòmicament]$_{ADU1}$ oi?

(14a) [Does that necessarily mean that I do not value it economically]$_{ADU1}$ right?

- In the case of emphatic expressions (15), only the main segment is included in the ADU.

(15a) sí que [hi ha la possibilitat]$_{ADU1}$

(15b) yes [there is the possibility]$_{ADU1}$

- Examples and metaphoric expressions are annotated as a single ADU, because the relationship with another ADU is usually established with the whole example or the whole metaphor (16). In cases in which the relationship with another ADU only occurs with a part of the metaphorical expression or example, a subsegment can be created with its corresponding identity ADU.

(16a) [aquest mateix any una dona es va haver de suïcidar just abans del seu desnonament o per exemple una mare va saltar per un pont amb el seu fill perquè no podia fer-se càrrec d'un crèdit bancari]$_{ADU1}$
[si la gent és capaç de suïcidar-se per l'opressió dels diners com no es vendran a la gestació subrogada]$_{ADU2}$

(16a) [this same year a woman had to commit suicide just before her eviction or for example a mother jumped off a bridge with her child because she could not pay back a bank loan]$_{ADU1}$
[if people are able to commit suicide because of the oppression of money why shouldn't they sell themselves in surrogacy]$_{ADU2}$

The ADU2, in (16), which includes an example, is related to the previous ADU1.

- Expressions including desideratum verbs (17) are not considered ADUs.

(17a) A mi m'agradaria anar a l'ONU i explicar els mateixos arguments per a que aquesta prohibició no sigui només a Espanya

(17b) I would like to go to the UN and present the same arguments so that this prohibition is not only in Spain

Annotation of argumentative relationships between ADUs

Once the ADUs are identified and segmented, the aim of the following task is to establish the argumentative relationships between ADUs and annotate the

type of relation held. We use the **RELATED_ID** (**REL_ID**) and **RELA-TION_TYPE** (**REL_TYPE**) tags for indicating these argumentative relationships. The REL_ID tag is used to indicate that an ADU2 holds an argumentative relationship with a previous ADU1 (18). The ID identifies the corresponding ADUs. It is worth noting that the relationships between ADUs almost always point to previous ADUs, following the logic of discourse, and that not all the ADUs have argumentative relations with other ADUs. There are cases in which several ADUs maintain a relationship with a single previous ADU, and all of them are indicated (see Table 3.2). An ADU may be related to more than one previous ADU, but the type of relationship with each of them is different. In these cases, we annotated the REL_ID and REL_TYPE for each different type of relationship generated from the same ADU. The annotation of the argumentative relations mainly occurs in the argumentation phase of the debate, but may also appear in the introduction phase.

Next, we describe the three type of argumentative relationships -inference, conflict and reformulation- which represent different semantic relations between two propositions. These relationships are annotated with the REL_TYPE tag and the corresponding values are **RA** for inference, **CA** for conflict and **MA** for reformulation. The notation used for the argumentative relations has been adopted from the Inference Anchoring Theory (IAT) [69] paradigm in order to provide a coherent labelling with previous corpora.

- Inference (RA) indicates that the meaning of an ADU can be inferred, entailed or deduced from a previous ADU (18). As already indicated, the direction of the inference almost always goes from one ADU to a previous ADU, but we have also found cases in which the direction is the opposite, that is, the inference goes from a previous ADU to a following one, although there are fewer cases (19). Therefore, inference is a meaning relation, in which the direction of the relationship between ADUs is relevant, and this direction is represented by the REL_ID tag.

(18a) [la gestació subrogada és una pràctica patriarcal]$_{ADU1}$
[ja que el major beneficiari d'aquesta pràctica és l'home]$_{ADU2}$ **REL_ID=1 REL_TYPE=RA**

(18b) [surrogacy is a patriarchal practice]$_{ADU1}$ [since the main beneficiary of this practice is the man]$_{ADU2}$ **REL_ID=1 REL_TYPE=RA**

(19a) [no tot progrés científic implica ...
... un progrés social]$_{ADU1}$ **REL_ID=2;3 REL_TYPE=RA**
[l'energia nuclear és la mare de la bomba atòmica]$_{ADU2}$
[Els pesticides que multiplicaven les collites han estat prohibits per convertir el aliments en insalubres]$_{ADU3}$

(19b) [not all scientific progress ...
... implies social progress]$_{ADU1}$ **REL_ID=2;3 REL_TYPE=RA**
[nuclear energy is the mother of the atomic bomb]$_{ADU2}$
[pesticides that multiplied crops have been banned
for making food unhealthy]$_{ADU3}$

In example (18), REL_TYPE=RA and REL_ID=1 indicate that ADU2 is an inference of ADU1, whereas in example (19) the REL_TYPE=RA and REL_ID=2;3 are annotated in ADU1 because it is an inference of ADU2 and ADU3, which appear in the original text below ADU1.

• Conflict is the argumentative relationship assigned when two ADUs present contradictory information or when these ADUs contain conflicting or divergent arguments (20). We consider that two ADUs are contradictory 'if they are extremely unlikely to be considered true simultaneously' [112].

(20a) [vol tenir és dret a formar una família]$_{ADU1}$
[formar famílies no és un dret]$_{ADU2}$ **REL_ID=1 REL_TYPE=CA**

(20b) [she wants to have the right to form a familiy]$_{ADU1}$
[to form families is not a righty]$_{ADU2}$ **REL_ID=1 REL_TYPE=CA**

• Reformulation is the argumentative relationship in which two ADUs have approximately the same or a similar meaning, that is, an ADU reformulates or paraphrases the same discourse argument as that of another ADU (21). The reformulation or paraphrase involves changes at different linguistic levels, for instance, morphological, lexical, syntactic and discourse-based changes [197].

(21a) [ja n'hi ha prou de paternalism]$_{ADU1}$ [ja n'hi ha prou que ens tracten com a xiquetes]$_{ADU2}$ **REL_ID=1 REL_TYPE=MA**

(21b) [enough of paternalism]$_{ADU1}$
[enough of treating us like children]$_{ADU2}$ **REL_ID=1 REL_TYPE=MA**

It should be noted that repetitions are not considered reformulations. A repetition contains a claim or statement with the same content as a previous one, i.e. the same argument. We consider an ADU to be a repetition only if it is exactly the same as a previous one and we do not therefore segment them and they are not annotated as ADUs.

It is worth noting that when a team mentions the opposing team's argument, that mention is not considered an argument. When their reasoning is referred to, the reference will be ascribed directly to the opposing team's argument.

Annotation process

The annotation of the *VivesDebate* corpus was manually performed by two different students of Linguistics specially trained for this task for three months and supervised by two expert annotators[5]. The annotation of the corpus was carried out in two main phases. The aim of the first phase was twofold: for the training of the annotators and for defining the annotation guidelines, that is, to establish the definitive tagset and criteria with which to annotate the debates. In this phase, we conducted different Inter-Annotator Agreement tests in order to validate the quality of the annotation of the different tasks involved, that is, the revision of automatic transcriptions, the segmentation of each debate into ADUs and the annotation of argumentative relations between ADUs. These tests allow us to evaluate the reliability of the data annotated, which basically means whether or not the annotators applied the same criteria for solving the same problem in a consistent way. These inter-annotator tests are also useful for evaluating the quality of the annotation guidelines, that is, to check whether the different types of phenomena to be treated are covered and the criteria are clearly explained, and to update the guide-

---

[5]The annotators are members of the Centre de Llenguatge i Computació (CLiC) research group `http://clic.ub.edu/`.

lines when necessary. In the second phase, after the training of the annotators, the remaining files in the corpus were annotated by each annotator independently.

Inter-Annotator Agreement tests

We carried out, first, a qualitative analysis in order to validate that the team of annotators was applying the same criteria in the revision of the automatic transcriptions of debates. This analysis consisted of the revision of three files (*Debate15.csv*, *Debate11.csv*, and a debate from the previous year's edition which was not included in our corpus) by the two annotators in parallel and the comparison of the results obtained by the senior annotators. The team met to discuss the problems arising from the comparison of the results in order to resolve doubts and inconsistencies. We devoted three sessions to this until we solved all disagreements and reached the same results in the revision of transcriptions, which explains why we revised the three files selected (one per session). The initial guidelines were updated with the new criteria established, such as following the criteria of the linguistic portal of the Catalan Audiovisual Media Corporation for spelling the named entities or writing down the main ideas in those cases in which the quality of the audio did not allow us to understand part of the message conveyed. In a nutshell, we ensure that the text of the ADUs was transcribed correctly, maintaining the linguistic variant originally used, whereas in the remaining text we applied a more superficial revision.

Once the transcription of texts obtained reliable results, we initiated the segmentation task, which was by far the most difficult task in the whole annotation process. We conducted three Inter-Annotator Agreement (IAA) tests until we reached an acceptable agreement for the segmentation of the transcribed texts into ADUs (see Table 3.1). We used the same file (*Debate6.csv*) in the two tests. We calculated the observed agreement and the Krippendorff's alpha [198]. The criteria followed for the evaluation of the Inter-Annotator Agreement test were the following:

- In the case of the PHASE, STANCE and argument REL_TYPE tags, we considered agreement to be reached when the annotators assigned the same value to each tag, while disagreement was considered to be when the value was different.

- In the case of the ADU tag, we considered agreement to exist when the span of the ADUs matched exactly, and disagreement to exist when the span did not match at all or coincided partially. We have also conducted a third Inter-Annotator Agreement test for evaluating the ADU tag considering partial agreement. In this case, we considered agreement to exist when the span of the ADUs coincided partially (22)-(25).

(22a) [la cosificació que s'està fent de la dona]$_{ADU1}$

(22b) [the objectivation of women]$_{ADU1}$

(23a) [**ens centrarem en** la cosificació que s'està fent de la dona]$_{ADU1}$

(23b) [**we will focus on** the objectivation of women]$_{ADU1}$

(24a) [el vincle que es genera entre ella i el nadó que porta al seu ventre]$_{ADU1}$ [**és trencat de manera miserable**]$_{ADU2}$

(24b) [the bond between her and the baby she carries in her womb]$_{ADU1}$ [**is broken in a miserable way**]$_{ADU2}$

(25a) [el vincle que es genera entre ella i el nadó que porta al seu ventre és trencat de manera miserable]$_{ADU1}$

(25b) [the bond between her and the baby she carries in her womb is broken in a miserable way]$_{ADU1}$

The disagreements found are basically of two types: a) the inclusion or omission of words at the beginning or at the end of the ADU (22) vs. (23); and b) the segmentation of the same text into two ADUs or a single ADU (24) vs. (25), the latter being stronger disagreement than the former. For instance, one of the annotators considered 'is broken in a miserable way' to be a different ADU (24), whereas the other annotator considered this segment part of the same ADU (25). Finally, we agreed that it should be annotated as a single ADU (25), because 'is broken' is the main verb of the sentence, and the argument is that what is broken is the bond between the mother and the baby.

As shown in Table 3.1, the results obtained in the Inter-Annotator Agreement tests for the PHASE, STANCE and REL_TYPE tags are almost perfect (above

Table 3.1: Results of the Inter-Annotator Agreement Tests.

Tag	Observed agreement %	Krippendorff's alpha
STANCE (AGAINST/FAVOUR)	99.05	0.979
PHASE (INTRO/ARG1,ARG2,ARG3/CONC)	94.60	0.925
REL_TYPE (RA/CA/MA)	86.00	0.913
ADU (1st IAA Test)	70.80	0,392
ADU (2nd IAA Test)	76.60	0,777
ADU (3rd IAA Test partial disagreements)	91.20	0.917

0.97), and acceptable for the ADU tags (0.77), which correspond to the segmentation into ADUs and the assignation of the type of argument, following Krippendorff [198] recommendations. The observed agreement (91.20%) and the corresponding alpha value (0.91) for the ADU tag increase when we consider there to be partial agreement ($\alpha \geq 0.80$ the customary requirement according to Krippendorff). The team of annotators met once a week to discuss problematic cases and resolve doubts in order to minimise inconsistencies and guarantee the quality of the final annotation. The results obtained are very good given the complexity of the task.

3.4 The VivesDebate Corpus

Data Collection

The *VivesDebate* corpus has been created from the transcripts of the 29 complete debates carried out in the framework of the 2019 *"Xarxa Vives d'universitats"* university debate tournament. During this competition, 16 different teams from universities belonging to the autonomous regions of Valencia, Catalonia and the Balearic islands debated in Catalan language on the topic *"Should surrogacy be legalised?"*. In addition to the original language of the annotated data, automatic translations to Spanish and English languages using the MLLP machine translation toolkit [143, 182] have also been included. The results and the evaluation of the debates were directly retrieved from the organisation, but post-processed by us in order to focus on the argumentative aspects of the debates and to preserve the anonymity of the jury and the participant teams. Furthermore, we would also like to remark that the data collected is part of a competition where the stances (i.e.,

favour or against) are assigned randomly. Thus, any argument or opinion existing
in the corpus is used to elaborate a logically solid reasoning, but it is not necessarily
supported by the participants.

Structure and Properties

The *VivesDebate* corpus is structured into 30 different CSV documents publicly
available to download from Zenodo. The first 29 documents correspond to a unique
debate each one, containing its three phases: introduction, argumentation, and con-
clusion. The structure of each document (Table 3.2) contains the identified ADUs
(rows) in Catalan (ADU_CAT), Spanish (ADU_ES), and English (ADU_EN),
and covers the six features that define every identified ADU (columns). First of
all, each ADU is assigned a unique ID created following the chronological order
(i.e., 1, 2, ..., N; where N is the total number of ADUs in a debate). This ID
allows for an intuitive representation of the flow of discourse in each debate. Sec-
ond, each ADU is classified into one of the three phases of competitive debate (i.e.,
INTRO, *ARG1*, *ARG2*, and *CONC*) depending on when has it been uttered. Third,
each ADU that forms part of one of the arguments put forward by the debaters is
assigned an argument number. This number allows to group every identified ADU
under the same claim. The same number used for the ADUs belonging to different
stances does not imply any type of relation between them, since their related claim
is different (i.e., argument number 1 in favour and argument number 1 against
stand for two different arguments). Fourth, each ADU is classified according to
the stance (i.e., in favour or against) for which it was used. Finally, the existing
argumentative relations between ADUs are identified and represented with the re-
lation type (i.e., Conflicts or CA, Inferences or RA, and Rephrases or MA), and
the ID(s) of the related ADU(s).

   Each document (i.e., debate) was annotated independently and in its entirety.
The conclusions of each team are considered as a unique ADU separately, since
they represent a good summary of the argumentative discourse from both teams'
perspectives. In addition to the 29 annotated debate documents, the corpus has
a supplementary evaluation file (*VivesDebate_eval.csv*). The jury's evaluation of
one debate (*Debate29.csv*) was not available for the creation of the evaluation file.
Thus, this file contains an anonymised version of the jury evaluation of the first 28

Table 3.2: Structure of the *VivesDebate* corpus CSV documents (*Debate7.csv*). (∗) An empty value in the *Arg. Number* column indicates that the ADU does not explicitly belong to any argument presented by the Favour or Against team to specifically support their stance.

ID	Phase	Arg. Number (∗)	Stance	ADU_CAT, ADU_ES, ADU_EN	Related ID	Relation Type
1	INTRO		FAVOUR	*quan mireu aquí què veieu* (*cuando miráis aquí qué veis*) (*when you look here what do you see*)		
2	INTRO		FAVOUR	*cinquanta euros* (*cincuenta euros*) (*fifty euros*)	1	RA
3	INTRO		FAVOUR	*el nostre nou déu* (*nuestro dios*) (*our new god*)	2	RA
. . .	. . .	. . .	. . .	. . .	. . .	. . .
43	ARG1	1	FAVOUR	*vivim en un món on ha guanyat els valors del neoliberalisme* (*vivimos en un mundo dominado por los valores del neoliberalismo*) (*we live in a world dominated by the values of neoliberalism*)		
44	ARG1	1	FAVOUR	*uns valors que ens diuen que si tenim diners som guanyadors* (*valores que dicen que si tenemos dinero somos ganadores*) (*values that say that if we have money we are winners*)	43	RA
45	ARG1	1	FAVOUR	*i si som guanyadors podem comprar tot allò que desitgem* (*y si somos ganadores podemos comprar lo que deseemos*) (*and if we are winners we can buy whatever we want*)	44	RA
46	ARG1	1	FAVOUR	*és el model que està imperant en la gestació subrogada* (*es el modelo que impera en la gestación subrogada*) (*is the prevailing surrogacy model*)	43;44;45	RA
. . .	. . .	. . .	. . .	. . .	. . .	. . .
144	ARG1	3	AGAINST	*per suposat que no* (*por supuesto que no*) (*of course not*)	143	CA
145	ARG1	3	AGAINST	*no és que vullguem que hi haja més xiquets* (*no es que queramos que haya más niños*) (*not that we want there to be more children*)	140	RA
146	ARG1	3	AGAINST	*sinó tot el contrari* (*sino todo lo contrario*) (*quite the contrary*)	145	MA
. . .	. . .	. . .	. . .	. . .	. . .	. . .

Table 3.3: Structure of the *VivesDebate* evaluation file.

Debate	Stance	Score	Thesis Solidity	Argument Quality	Adaptability
Debate1	Favour	3.32	3.25	3.37	3.33
Debate1	Against	3.29	3.33	3.18	3.38
Debate2	Favour	3.41	3.5	3.33	3.42
Debate2	Against	3.43	3.17	3.58	3.58
. . .	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
Debate28	Favour	2.75	3.25	2.75	2.17
Debate28	Against	2.36	2.77	2.25	2.00

debates (Table 3.3). However, since some aspects used by the judges to evaluate
the debating teams are not reflected in our corpus (e.g., oral fluency, grammatical
correctness, and non-verbal language), we have excluded them in our argumen-
tative evaluation file. A numerical score in the range of 0 to 5 (100%) which
combines the thesis solidity (35%), the argumentation quality (35%), and the ar-
gument adaptability (30%) is provided as the formal evaluation for each debate.
The presented structure makes the *VivesDebate* corpus a very versatile resource
for computational argumentation research. Not only because of its size, but due
to all the information it contains, it can be used for argument mining, argument
analysis and representation, argument evaluation, and argument summarising (i.e.,
generation) research tasks.

The resulting *VivesDebate* corpus comprises a total of *139,756* words (tokens).
Furthermore, the corpus presents an average of *4,819* words per document indepen-
dently annotated. Words are grouped into a total of *7,810* ADUs, with an average
of *269* ADUs per document. In our corpus, an argument is built from multiple
ADUs sharing argumentative relations. A total of *1,558* conflicts, *12,653* infer-
ences, and *747* rephrases between the identified ADUs have been annotated in
the *VivesDebate* corpus, with an average of *54* conflicts, *436* inferences, and *26*
rephrases per document. A summary of the structure and a breakdown for each
of the included debates is presented in Table 3.4. In addition to all these corpus
statistics, we retake the "argument density" metric proposed in [387]. This met-
ric computes the density of arguments in a corpus by normalising the number of

annotated inference relations to the total word count. The *VivesDebate* presents an "argument density" of 0.091, which is significantly higher compared to the densities of previously existing similar corpora such as the US2016 [387] (0.028 density) for *97,999* words, or the DMC [183] (0.033 density) for *39,694* words.

3.5 Related Work: Other Computational Argumentation Corpora

As noted in the introduction, the existing resources for computational argumentation present significant differences depending on their main purposes. Thus, we consider it important to contextualise our contribution to the computational argumentation research within the existing related work. For that purpose, we present a thorough comparison between the most prominent available resources for the computational argumentation research community. One of the first public corpora focused on the argument mining task was presented in [351], where the authors annotated 90 persuasive essays in English obtained from an online forum. In this corpus, two different aspects of arguments were annotated, the argument components (i.e., claim and premise) and the argumentative relations (i.e, attack and support). Another early resource to satisfy the needs of argument mining researchers was presented in [275]. The authors present a new corpus of 112 annotated "microtexts", short and dense written arguments in German which were also professionally translated into English. This corpus was annotated taking into account the argumentative structure of the text, where each argument has a central claim with an argumentative role (i.e., proponent and opponent), and several elements with different argumentative functions (i.e., support, attack, linked premises and central claims). These "microtexts" were generated in a controlled experiment where 23 participants were instructed to write argumentative text on a specific topic. A different approach was introduced in [183], where dialogue spoken argumentation samples were used to create the *Dispute Mediation Corpus* (DMC). Three different sources were considered to retrieve up to 129 mediation excerpts which were analysed by a unique professional annotator. The sources from which these excerpts were annotated are 58 transcripts found in academic papers, 29 online website mediation scripts, 14 scripts provided by professional mediators, and 28 analyses of

Table 3.4: Structure and properties of the *VivesDebate* corpus. Score F and A
represent the score assigned to the favour and against teams respectively according
to our processing of the original evaluation.

File	Words	ADUs	Conflicts	Inferences	Rephrases	Score F	Score A
Debate1.csv	3979	198	4	158	22	3.32	3.29
Debate2.csv	5178	371	60	310	32	3.41	3.43
Debate3.csv	4932	311	63	360	22	4.39	4.31
Debate4.csv	6243	308	8	229	37	4.01	4.15
Debate5.csv	5389	270	45	505	48	4.38	3.28
Debate6.csv	4387	324	45	219	42	2.94	3.02
Debate7.csv	4523	299	11	236	18	3.31	3.13
Debate8.csv	4933	220	5	185	15	3.31	3.92
Debate9.csv	5574	352	45	309	18	4.12	4.21
Debate10.csv	4284	279	12	207	39	4.39	3.46
Debate11.csv	5720	239	5	202	16	3.60	3.64
Debate12.csv	5305	283	83	477	49	4.12	4.36
Debate13.csv	3646	138	10	106	6	2.94	2.60
Debate14.csv	4790	302	74	400	47	3.83	3.80
Debate15.csv	4550	173	23	113	13	3.95	3.94
Debate16.csv	4887	288	94	639	53	3.33	3.39
Debate17.csv	3891	164	8	123	8	3.00	3.26
Debate18.csv	3701	166	6	149	4	2.80	2.77
Debate19.csv	4645	186	13	159	1	4.24	4.34
Debate20.csv	5484	306	33	1306	55	3.53	3.49
Debate21.csv	5064	278	102	1076	42	3.17	3.18
Debate22.csv	4669	330	16	408	1	4.40	4.22
Debate23.csv	4420	266	136	917	26	2.74	2.69
Debate24.csv	5139	267	380	1002	39	4.41	4.37
Debate25.csv	4828	321	7	337	0	4.09	3.88
Debate26.csv	4440	290	16	328	5	4.16	3.93
Debate27.csv	5012	234	106	645	24	3.49	2.33
Debate28.csv	4254	310	21	344	2	2.75	2.36
Debate29.csv	5889	337	51	1203	72	-	-
VivesDebate	139,756	7,810	1,558	12,653	747	-	-

meta-discourse elements in mediation interactions from a mixture of the previous sources. The DMC corpus was annotated using the Inference Anchoring Theory (IAT), containing up to eleven structural features of arguments useful for the argument mining task: locutions, assertions, assertive questions, pure questions, rhetorical questions, assertive challenges, pure challenges, popular concessions, inferences, conflicts, and rephrases. Furthermore, graphical representations of the complete structures useful for argument analysis can be loaded in the OVA+[6] tool. In addition to the argument mining and analysis tasks, the automatic evaluation of arguments is an important aspect in the analysis of argumentative discourses. Work [269] presents the *Consumer Debt Collection Practices* (CDCP) corpus, where 731 user comments from an online forum are annotated with their argumentative structures, and capturing the *strength* of the identified arguments. In the CDCP corpus, each comment is segmented into elementary units (i.e., Facts, Testimony, Value, Policy, and Reference). Support relations between these elementary units are annotated in order to provide structural information. The authors defined the evaluability of an argument for those cases in which all the propositions that make up this argument are supported by an explicit premise of the same type of elementary unit. The *strength* of an argument is measured by comparing the type of the elementary units that make it up. Another important part of computational argumentation which was not approachable from the reviewed corpora is the automatic generation of natural language arguments. This is a recent research topic which requires an important amount of data to achieve competitive results. In [260], the authors present a new annotated corpus aimed at approaching this task. The GPR-KB-55 contains 200 speeches from a debate competition that were analysed, each one debating one of the 50 different topics existing in these speeches. The resulting corpus consists of 12431 argument pairs containing a claim and its rebuttal with annotations regarding the relevance of a claim to its motion, the stance of the claim, and its appearance in a piece of speech (i.e., mentioned/not mentioned, explicit/implicit). Even though some linguistic annotations were done, the argument structure or the flow of discourse were not annotated in the GPR-KB-55 corpus. This corpus is part of the IBM Project Debater[7] which encompasses a large set of different

---

[6]http://ova.arg-tech.org/
[7]https://www.research.ibm.com/artificial-intelligence/
project-debater/

corpora, each one aimed at a specific task of the argumentative process. A different perspective on natural language argument generation is presented in [318], where the authors provide a new corpus aimed at the word-level summarisation of arguments. The DebateSum corpus consists of 187,386 debate summaries without any structural annotation, retrieved from the debate tournaments organised by the National Speech and Debate Association. The only annotation provided by this corpus is the segmentation of arguments-evidences-summary triplets extracted directly from the transcripts. Again, the usefulness of this resource remains strongly linked to the specific tasks of argument summarisation and language modelling. At this point, it is possible to observe the strong dependence between the analysed corpora and the different tasks of computational argumentation. The *US2016* debate corpus was presented in [387] as the largest argumentative corpus with a great versatility between different aspects of argumentation such as discourse analysis, argument mining, and automatic argument analysis. The *US2016* compiles the transcripts of the 2016 US presidential election TV debate and the subsequent online forum debate (i.e., Reddit). The text is analysed and divided into argument maps consisting of 500 to 1500 words. The annotation process is carried out independently for each argument map, where the text is segmented into ADUs and argumentative relations between ADUs are identified. The final corpus has 97,999 words, with a sub-corpus of 58,900 words from the TV debates and 39,099 words from the Reddit discussion. Despite the improvement achieved with this new corpus, we still identify two major issues which may hinder the performance of the trained models and the scope of the experiments: the quality of the uttered arguments, and the traceability of discourse. Electoral campaigns and debates are usually focused at reaching the majority of voters rather than properly using arguments, or having a rational debate. Furthermore, the arguments retrieved from an online forum might neither be of the ideal quality. Thus, the trained models using this data can be biased in a way that does not reflect the reality of a more rational and logical argumentation. Finally, the most recent argumentative corpus was presented in [123]. The authors present the ReCAP corpus of monologue argument graphs extracted from German education politics. More than 100 argument graphs are annotated from natural language text sources like party press releases and parliamentary motions. This corpus annotates the ADU segments identified in the text and relations between different ADUs (i.e., inferences). Furthermore, the

authors have also included annotations of the underlying reasoning pattern (i.e., argumentation schemes) of arguments. These are not the unique resources published in the literature for computational argumentation research. Many research done in argument mining includes new corpora, for the healthcare domain in [228], for legal argumentation in [409], and for online social network analysis in [132] among other different domains. However, for our comparison, we have focused on the most used corpora in computational argumentation research, and corpora created from a more generalist perspective.

A comparison of the previously analysed corpora is presented in Table 3.5. Furthermore, we have added the *VivesDebate* corpus to the comparison in order to provide a reference to understand the significance of our contribution. Seven different features that we consider indicators of the quality of a corpus have been analysed in our comparison. First, the format of the argumentative data indicates if the arguments are retrieved from a monologue (M) or a dialogue (D). Furthermore, it is also important to know the source of the arguments, if they come from a text source (T) or from a speech transcript (S). The domain indicates the context from which the corpus has been created (e.g., competitive debate, online forum, etc.). This is a key feature to determine the quality of the arguments contained in the resulting corpus, since major linguistic aspects such as the richness of vocabulary or the originality of arguments will be significantly different from one domain to another. The tasks feature indicates in which argumentative tasks a corpus can be useful: Argument Mining (AM), Argument Analysis (AA) and representation, Argument Evaluation (AE), Argument Generation (AG), and Argument Summarisation (AS). This feature is important to observe the versatility of each analysed corpus. The language indicates if a corpus is available in English (EN), German (DE), or Catalan (CAT). Finally, we have taken into account the size of each corpus in words (W) and/or sentences (S); and the annotation ratio, which indicates the average number of words (or sentences) per each independently annotated document *w/d* (*s/d*). This last feature can give us an idea of the contextual information preserved in the annotation process. For instance, it is not the same to annotate a complete debate (higher annotation ratio), than to split the debate into smaller argumentative structures to simplify the annotation process (lower annotation ratio).

Thus, it is possible to observe how, in addition to the quality improvement

Table 3.5: Comparison of computational argumentation corpora. (∗) Automatically translated languages.

Research	Identifier	Format	Source	Domain	Tasks	Language	Size	Annotation Ratio
[351]	Persuasive Essays	M	T	Online Forum	AM	EN	34,917 (W)	388 *w/d*
[275]	Microtexts	M	T	Controlled Experiment	AM	EN+DE	576 (S)	5 *s/d*
[183]	DMC	D	S	Academic+Online+Professional	AM+AA	EN	18,628 (W)	144 *w/d*
[269]	CDCP	D	T	Online Forum	AM+AE	EN	4,931 (S)	6.7 *s/d*
[260]	GPR-KB-55	D	S	Competitive Debate	AG	EN	12,431 (S)	41 *w/d*
[387]	US2016	D	T+S	Political+Online	AM+AA	EN	97,999 (W)	189 *w/d*
[387]	US2016TV	D	S	Political	AM+AA	EN	58,900 (W)	492 *w/d*
[387]	US2016Reddit	D	T	Online Forum	AM+AA	EN	39,099 (W)	137 *w/d*
[318]	DebateSum	D	S	Competitive Debate	AS	EN	101M (W)	520 *w/d*
[123]	ReCAP	M	T	Political	AM+AA	DE+*EN*(∗)	16,700 (W)	150 *w/d*
	VivesDebate	D	S	Competitive Debate	**AM+AA+AE+AG/AS**	CAT+*ES*(∗)+*EN*(∗)	139,756 (W)	**4819 *w/d***

mainly due to the source (i.e., competitive debate), our corpus can be useful in a wider variety of computational argumentation tasks. Furthermore, the *VivesDebate* corpus presents an annotation ratio significantly higher compared to the previous work. This approach makes it possible to improve the richness of the annotations by keeping longer-term argumentative relations and allowing to have a complete representation of the flow of the debate.

3.6   Conclusion

In this paper we describe *VivesDebate*, a new annotated multilingual corpus of argumentation created from debate tournament transcripts. This work represents a major step forward in publicly available resources for computational argumentation research. Next, we summarise the main improvements brought about by the creation of this corpus.

First, because of its size. The *VivesDebate* corpus is, to the best of our knowledge, one of the largest publicly available resources annotated with relevant argumentative propositions, and argumentative and dialogical relations. With a total of *139,756* words and an argument density of 0.091, in addition to its size, the *VivesDebate* corpus also improves the previously available argumentation corpora in terms of their density.

Second, because of the quality of the argumentative reasoning data. The *Vives-Debate* corpus has been created from the transcripts of 29 complete competitive debates. Annotating spoken argumentation is usually harder and more expensive

than textual argumentation, so the majority of the publicly available corpora for computational argumentation research are created from social network and online forum debates. Furthermore, most of the available spoken argumentation corpora are from the political debate domain, which does not have a solid structure and the quality of argumentation is harder to evaluate. By creating a new corpus from the transcripts of a debate tournament, the improvement of the argumentative quality compared to previously available corpora is threefold: (i) debate tournaments have a well-defined argumentation structure, which eases their modelling; (ii) the only motivation behind the debates is the argumentation itself, so that participants need to argue using the strongest arguments and present a coherent reasoning to win the debate; and (iii) the debates are objectively evaluated by an impartial jury, analysing parameters that are directly related to the quality of arguments and argumentation.

Third, because of its versatility. The size, the structure, the annotations, and the content of the *VivesDebate* corpus makes it useful for a wide range of argumentative tasks such as argumentative language modelling, the automatic identification of ADUs in an argumentative dialogue (i.e., Argument Mining), the elaboration and analysis of complex argument graphs (i.e., Argument Analysis), the automatic evaluation of arguments and argumentative reasoning (i.e., Argument Evaluation), and the automatic generation of argument summaries (i.e., Argument Generation/Argument Summarising). Furthermore, the corpus is available in its original version in Catalan, and in machine-translated versions to Spanish and English languages, leaving an open door to multilingual computational argumentation research.

Even though the *VivesDebate* corpus brings significant improvements over existing resources for computational argumentation research, it has its own limitations. The debates contained in the corpus belong to a unique tournament, which means that every annotated debate will have the same topic in common. This feature is directly related to the observable language distribution, which will be biased by the *"Should surrogacy be legalised?"* topic. However, since our corpus is aimed at computational argumentation research rather than language modelling, this should not be an important issue. Furthermore, this data bias can be easily amended with a topic extension of the *VivesDebate* corpus. The other main limitation of the corpus are the Spanish and English machine-translated versions, which

may not be as linguistically correct as the original version in Catalan.

As future work, we plan to overcome some of these limitations and to deepen on the argumentative analysis and annotation of the corpus. First, we plan to improve the Spanish and English machine-translated versions of the *VivesDebate* corpus with a professional translation. We also want to improve the argumentative annotations of the corpus by deepening on the logical and rational aspects of argumentation. On its current form, it is possible to do a general structural analysis of the arguments. With the identification and annotation of stereotyped patterns of human reasoning (i.e., argumentation schemes [400]), it will be possible to bring the automatic detection and analysis of arguments to a deeper level. However, this is a complex task, and it has only been superficially researched in the literature. Finally, we are also exploring the possibility of organising a new shared task focused on the argumentative analysis of natural language inputs.

A Cascade Model for Argument Mining in Japanese Political Discussions: the QA Lab-PoliInfo-3 Case Study

RAMON RUIZ-DOLZ

Abstract

The rVRAIN team tackled the Budget Argument Mining (BAM) task, consisting of a combination of classification and information retrieval sub-tasks. For the argument classification (AC), the team achieved its best performing results with a five-class BERT-based cascade model complemented with some handcrafted rules. The rules were used to determine if the expression was monetary or not. Then, each monetary expression was classified as a premise or as a conclusion in the first level of the cascade model. Finally, each premise was classified into the three premise classes, and each conclusion into the two conclusion classes. For the information retrieval (i.e., relation ID detection or RID), our best results were achieved by a combination of a BERT-based binary classifier, and the cosine similarity of pairs consisting of the monetary expression and budget dense embeddings.

4.1 Introduction

The automatic analysis of natural language arguments has made possible to improve computer systems for human assistance in the domains of medicine [226],

121

academic research [34], web discourse analysis [161], and autonomous debate [345] among others. The argument mining task is present in many different domains and instances [207]. However, due to its heterogeneity and complexity, it is considered as an important challenge in the Natural Language Processing (NLP) research community. The underlying linguistic structures in natural language argumentation present a great challenge for both, the human annotation of new corpora [327]; and the training/evaluation of new models for domain-independent argument mining [192, 325] or for different instances of the problem belonging to the same domain (e.g., legal) [384, 285]. Thus, advances in argument mining research will benefit from as many as different viewpoints (e.g., domains and/or task instances) approaching this task.

In this work, we describe the participation of our team *rVRAIN* to the Budget Argument Mining (BAM) task organised for the *QALab PoliInfo 3*[1] and the *NTCIR-16*[2]. The BAM is a combination of classification and information retrieval sub-tasks in the domain of political debate analysis. First, the argument classification sub-task is aimed at determining if a given monetary expression belongs to an argument, and which is its argumentative purpose (i.e., either claim or premise). Second, the relation ID detection sub-task is aimed at finding relations between monetary expressions uttered in an argumentative discourse and political budget items.

Our approach presents a BERT-based cascade model for argument mining in japanese political discussions. The model proposed in this paper for solving the BAM task tackles independently the argument classification and the relation ID detection tasks. For the former, we propose the use of handcrafted rules to determine if an expression is monetary or not. Then, a BERT-based cascade model is trained to classify each argumentative monetary expression into premise or claim, and their subsequent sub-classes (i.e., three premise and two claim sub-classes). For the latter, a BERT-based binary classifier is trained to identify possible relations between political budget items and monetary expressions. Each (possible) identified relation is then scored using the cosine similarity, and only the top relations are brought into consideration.

The rest of the paper is structured as follows. Section 4.2 reviews the related

---

[1]https://poliinfo3.net/
[2]http://research.nii.ac.jp/ntcir/ntcir-16/index.html

research and contextualises the contribution of this paper to the area of argument mining. Section 4.3 briefly defines the BAM task and the corpus used to carry out our experiments. Section 4.4 presents the architecture of the model proposed for solving the BAM task. Section 4.5 depicts the observed results in three different stages of the competition. Finally, Section 4.6 discusses the obtained results and analyses future lines of research and open challenges.

4.2   Related Work

Argument mining approaches the automatic identification, classification and structuring of argumentative natural language [265]. It has been typically decomposed into different sub-tasks in the literature [207, 32]: argumentative discourse segmentation, argument component detection, and argumentative relation identification. Each one tackles a different step belonging to the global goal of argument mining.

Argumentative discourse segmentation is the task of detecting argument spans in a given natural language input. For example, identifying where an argumentative component begins and ends throughout the full interventions of politicians in a discussion. A basic approach for this task was to consider complete sentences and classify them into *argument/non-argument* [265]. In [212], the authors propose an unsupervised approach for claim segmentation based on the appearance of common linguistic structures used for argumentation (e.g., "that"). However, recent research has emphasised the relevance of context for improving the segmentation of arguments and argumentative components in natural language inputs [4]. In spoken dialogue it is common to omit contextual information to ease its flow, aimed at overcoming this problem a cascade model for identifying argumentative propositions completing the missing context is proposed in [186].

The detection and classification of argument components is the argument mining task aimed at understanding the argumentative purpose of the previously segmented text. Research in this topic has usually focused on the identification of argumentative evidence and on the premise/claim classification [207]. We will focus on the latter since it has a direct relation with the BAM task and the model proposed in this work. The argument component detection is a very descriptive repre-

sentation of the previously mentioned existing heterogeneity in argument mining research. Initially introduced in [265], the task was instanced as a binary classification problem. The authors make use of classical machine learning algorithms to predict premise and claim classes for the argumentative expressions. Subsequent research focused on a linguistic enrichment of the task proposed a new instance where up to four classes (i.e., *major claim*, *claim*, *premise*, *none*) were considered [352]. Some of the latest research in this topic has explored the use of end-to-end neural network-based architectures [135, 245], graph convolutional networks [246], and attention-based architectures [354] to improve previous experimental results.

Finally, the argumentative relation identification task focuses on detecting argumentative structures between the argumentation components (e.g., premises or claims). This task has been classically considered one of the most complex tasks in argument mining, and approached as a sentence pair classification problem with two classes (i.e., attack and support) [99, 135, 171]. Recent research has investigated the behaviour of state-of-the-art NLP techniques when approaching a cross-domain multi-class instance of this task [325]. However, since the BAM task and our proposed model does not approach this sub-task of argument mining, we will not go any further into this aspect.

4.3 Budget Argument Mining

This work approaches the Budget Argument Mining (BAM) instance of the argument mining task. The BAM is aimed at improving the automatic argumentative analysis of political discussion transcripts through the use of NLP techniques. It includes the argumentative discourse segmentation and the argument component detection sub-tasks. For that purpose, monetary expressions are detected in the transcripts, and it must be determined if an expression belongs to an argument or not, and which is its argumentative role in the discussion. Furthermore, the required analysis is enriched with the relation of each monetary expression with a political budget item. This way, the resulting analysis will provide a set of argumentative components and their type detected in the transcripts of a discussion, and a set of relations between the arguments and budget items.
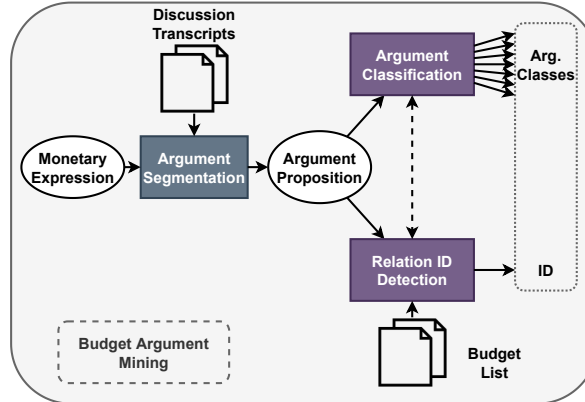
Figure 4.1: Budget Argument Mining task diagram.

Therefore, the BAM consists of two different sub-tasks: the argument classification (AC) and the relation ID detection (RID) (see Figure 4.1). A complete description of the whole task can be found in its overview [194]. However, the basic ideas of BAM are presented in the following sections in order to make this paper self-contained.

Argument Classification (AC)

The AC sub-task is aimed at covering the two first parts of argument mining: segmentation and classification. Thus, for a given monetary expression appearing in an utterance, we need to analyse if it belongs to an argumentative proposition, and which is its role in argumentation. First, the argumentative propositions containing the monetary expressions need to be segmented from the natural language transcripts. Second, these segments must be classified into seven different argumentative classes: (i) Premise: Past and Decisions; (ii) Premise: Current and Future; (iii) Premise: Other; (iv) Claim: Opinions, suggestions and questions; (v) Claim: Other; (vi) Not monetary expression; and (vii) Other.

Relation ID Detection (RID)

The RID sub-task is aimed at determining if the monetary expressions uttered in the discussion are related to a specific item in the budget list. For this purpose,

Table 4.1: Class distribution of the BAM training data.

		Premise			Claim			
	Past	Future	Other	Opinions	Other	**Non monetary**	**Other**
N	260	622	212	98	23	27	6

each argument containing any monetary expression must be segmented. Then, a relation between the segmented text and the budget items must be established.

Data

The data released for the BAM task is structured into three different documents: the budget data (*PoliInfo3_BAM-budget.json*), the training data (*PoliInfo3_BAM-minutes-training.json*), and the test data (*PoliInfo3_BAM-minutes-test.json*). Each document contains information from the Japanese national diet, and from three different local circumscriptions (i.e., Otaru, Ibaraki, and Fukuoka).

The budget document is a list consisting of 768 different budget items. Each budget item has eleven descriptive features: an *identifier*, a *title*, a *url*, an *item*, the *budget amount*, a list of *categories*, the *types of account*, the *department*, *last year's budget*, a *description*, and a *budget difference*.

The training document contains 29 proceedings belonging to the local circumscriptions. These proceedings consist of a total amount of 1573 utterances. Furthermore, 2 speech records from the national diet consisting of a total of 363 speeches are also included in this file. This translates to a total of 1248 monetary expressions, which are our training samples. The class distribution of these samples is depicted in Table 4.1. The test document follows the same structure. A total of 760 utterances from local circumscriptions and 123 speeches from the national diet are included in this file. From all these transcriptions, 520 monetary expressions remain unlabelled in this document, which is the one used in the model evaluation of the BAM shared task.

4.4 Model Architecture

We propose a BERT-based [115] cascade model to undertake the complete BAM process (see Figure 4.1). All the BERT-based classifiers integrated in our cascade
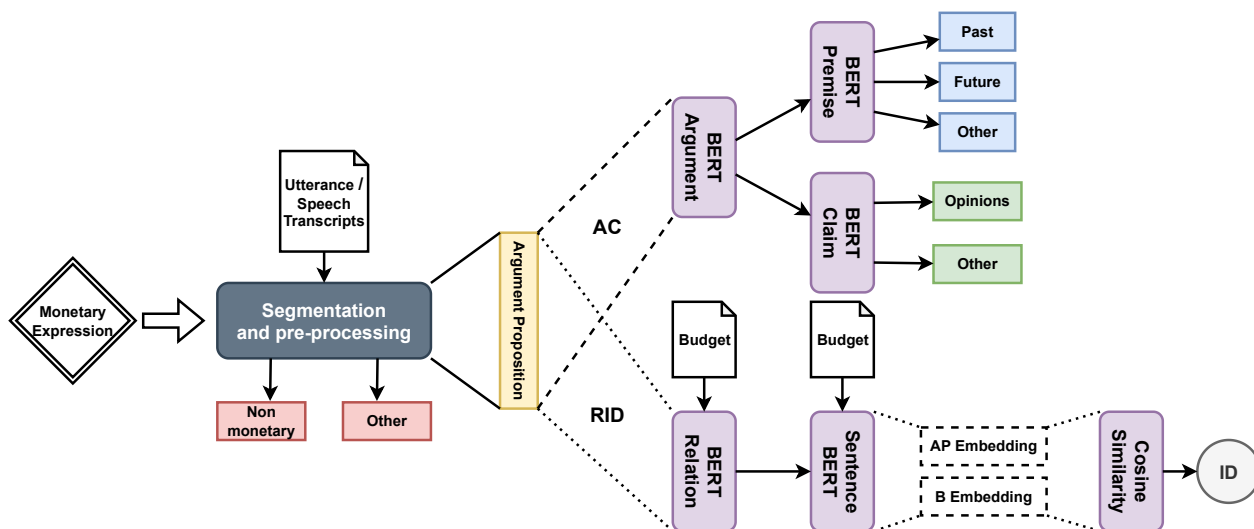
Figure 4.2: *rVRAIN* model architecture proposal.

model were fine-tuned from the *Inui Laboratory*[3] pre-trained BERT-large Japanese Language Model. In our approach, each monetary expression will be treated as the input and a class label and a related ID as the output. The proposed architecture aims to smooth the complexity of the classification task considering the size of the training corpus and the number of classes. Furthermore, it approaches independently the AC and the RID sub-tasks. Figure 4.2 synthesises the proposed architecture. The code implementation of the model architecture proposed in this paper is publicly available in GitHub[4].

Before tackling the AC and RID tasks, each monetary expression was analysed together with the discussion transcripts (i.e., local government utterances and national diet speeches) to produce segmented argument propositions. The segmentation was done by considering the set of full sentences belonging to the same utterance/speech where the monetary expression was contained. A set of handcrafted rules was applied during this pre-processing to determine if the proposition was either non monetary or other than a premise or a claim (e.g., detecting the existence of monetary Japanese kanji characters such as "円"). This way, the total number of remaining classes was reduced from seven to five.

Then, the AC is tackled by three different BERT-based classification models.

---

[3] https://github.com/cl-tohoku/bert-japanese/tree/v2.0
[4] https://github.com/raruidol/Budget-AM

A high-level BERT-based binary classifier was trained to detect if an argument proposition was either a premise or a claim. Once having assigned a high-level class to the sample, two low-level BERT-based classifiers were trained for 3-class premise classification and binary claim classification. This way, the high-level model focuses on the premise/claim discriminatory features, while the low-level focuses on more specific intra-class features. Furthermore, the class complexity of the problem is also decomposed from 5-class to 3-class and binary classifications.

Finally, the RID is tackled by a BERT-based binary classifier, and a cosine similarity calculation for pairs of Sentence-BERT [306] embeddings. In this second part of the task, the segmented argument propositions containing monetary expressions are paired together with the *item* and *description* features of the budget items. A binary classifier is used to determine if a given pair (i.e., argument proposition, budget item) could be related or not. Then, all the pairs classified as related are scored using the cosine similarity of the dense embeddings of argument propositions (AP) and budget (B) items (i.e., *item*+*description* features) generated using a Sentence-BERT model. The highest scored relation is used in our approach to produce the model's output.

Therefore, each monetary expression was completely analysed by our cascade model, classified into one of the seven argumentative classes, and related to one of the budget items in the list.

4.5   Results

The evaluation of the architecture proposed in this paper has been carried out at three different levels. First, we performed a local evaluation of the models aimed at having preliminary notions of how would our proposal behave with the test data. Second, we received feedback of our model's performance in an initial "*Dry-run*" phase of the BAM task. Finally, the "*Formal-run*" evaluation of the models corresponds to the last and definitive round of the shared task.

Experimental Setup

All the experiments and results reported in this paper have been implemented and run under the following setup. For the pre-processing of the corpus, we have used *pandas* [233] for handling data structuring, and *fugashi* [232] for the analysis of Japanese natural language text. For model training and transfer learning we have used the *PyTorch* and *Transformers* [405] libraries. This powerful deep learning tools have made possible to take advantage of existing large language models in Japanese, and adapt them to our specific task. For the semantic cosine similarity calculus in RID sub-task we used the *Sentence Transformers* [306] library. Finally, the local evaluation metrics (i.e., accuracy and macro f1) have been implemented using the *sklearn* library.

Local Evaluation

During the local evaluation, we have tested different model architectures. The most basic approach consisted of a 7-class BERT-based classification model (*7BERT*). We also experimented with a 5-class BERT-based classification model together with a set of handcrafted rules for the underrepresented classes (i.e., non monetary and other) (*5BERT*). Finally, we evaluated the BERT-based cascade model proposed in this work for tackling the BAM task (*rVRAIN*). Each of these models was also evaluated considering a balanced version of the corpus, where premise and claim training sample distributions were more balanced than the original corpus (*BD*). For the evaluation, we considered the accuracy and the Macro-F1 scores. This decision was made based on the strong unbalance between classes observed in the training corpus. Furthermore, we evaluated our models using a 10-fold cross validation. Table 4.2 summarises the obtained results during the local evaluation of our models.

We can observe how the best accuracy score was obtained by the *5BERT* model. However, *rVRAIN* achieved the best performance considering the Macro-F1 score. This means that our cascade model generalised better on this task, by doing a better classification of the samples belonging to underrepresented classes.

Table 4.2: Local evaluation of the different models for AC.

Model	Accuracy	Macro-F1
7BERT	0.71	0.19
7BERT(BD)	0.56	0.16
5BERT	**0.76**	0.25
5BERT(BD)	0.55	0.19
rVRAIN	0.47	**0.27**
rVRAIN(BD)	0.42	0.22

Dry-Run Evaluation

The Dry-Run evaluation phase of the BAM shared task used the test file to evaluate our submissions, and was divided into two different stages. During the early stage (see Table 4.3), the evaluation script assigned a unique score to the team submissions. This score combined the performance of the models in AC and RID sub-tasks. In the early stage, we evaluated the performance of the same models evaluated during the local evaluation, except for the balanced data versions. Our models achieved the 2nd and 3rd best scores for the BAM task. *RB* stands for the random baseline provided by the organisers of the task.

However, the evaluation script was updated the last month of the Dry-Run evaluation. The late stage (see Table 4.4) of the Dry-Run evaluation provided individual scores for the AC and RID tasks, together with a global evaluation of the performance of the model in the BAM task. Aimed at easing the readability of the results, we will only include the best performing approaches of each team in the tables. Our best performing model in the late stage was the one using the *5BERT* model for AC, but we could not evaluate the cascade architecture during this phase. Furthermore, the new evaluation script only considered those samples with both, the argument class and the relation ID correctly predicted, to increase the global score of the BAM task. This explains why our approach was the 3rd ranked with the best general score, even though it was the 2nd in the AC and the 1st in the RID sub-tasks.

Table 4.3: Dry-run (early) evaluation of the different models for BAM.

Team	Score AC+RID
fuys	0.51
rVRAIN (5BERT)	0.45
rVRAIN	0.40
OUC	0.33
rVRAIN (7BERT)	0.25
RB	0.09

Table 4.4: Dry-run (late) evaluation of the different models for BAM.

Team	Score AC+RID	AC	RID
fuys	**0.13**	**0.57**	0.17
OUC	**0.13**	0.37	**0.21**
rVRAIN (5BERT)	0.06	0.48	**0.21**
takelab	0.00	0.33	0.00
RB	0.00	0.13	0.00

Formal-Run Evaluation

During the Formal-Run evaluation, the same test file than with the Dry-Run was used. We achieved our best results using the proposed cascade model architecture for AC, together with the proposed semantic similarity calculation method for RID. As presented in Table 4.5, the *rVRAIN* achieved the 4th best performing position from a total of 6 participating teams. However, our approach was the best performing one from the teams that did not include task organisers.

Table 4.5: Formal-run evaluation of the different models for BAM. (*)The team contains task organisers.

Team	Score AC+RID	AC	RID
JRIRD*	**0.51**	**0.58**	0.61
OUC*	0.45	0.57	**0.66**
fuys*	0.23	0.57	0.34
rVRAIN	0.17	0.48	0.21
rVRAIN (5BERT)	0.06	0.48	0.21
takelab	0.04	0.39	0.06
SMLAB	0.00	0.38	0.00
RB	0.00	0.13	0.00

4.6 Discussion

We have described the participation of *rVRAIN*'s team at the *Budget Argument Mining* task organised in the *QALab PoliInfo 3* and the *NTCIR-16*. The organisers proposed a new instance of the argument mining task, a classic in the NLP area of research. In this new instance, the main goal was to correctly classify arguments containing monetary expressions and relate them to items in a list of political budgets. In this paper, we have proposed a new approach to this task relying in the latest advances in NLP (i.e., Transformer-based architectures). The proposed cascade model architecture achieved the fourth position in the performance ranking, and it was the best among teams without task organisers.

Several observations can be drawn from this paper's proposal and experimentation. First, we have seen how when dealing with highly unbalanced corpora, a system can benefit from defining a set of handcrafted rules and relaxing the class complexity of the task. Instead of approaching the complete problem with the use of a unique classifier. Second, we have also observed that no improvement could be achieved by forcing the balance of the corpus. When using the balanced version,

the score dropped significantly. This is most probably because the real distribution that the model has to predict is not balanced, but the corpus size limitation can also have a major role in this issue.

Finally, we foresee the implementation of some communication between the models for AC and RID during their training as a future work improvement of the model's performance on the BAM task. Furthermore, we also consider that using different test sets for each phase of the shared task (i.e., Dry-Run and Formal-Run) would be beneficial for the generalisation of the findings in this topic.

Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation

RAMON RUIZ-DOLZ, JOSE ALEMANY, STELLA HERAS AND ANA GARCÍA-FORNES

Abstract

Argument Mining is defined as the task of automatically identifying and extracting argumentative components (e.g., premises, claims, etc.) and detecting the existing relations among them (i.e., support, attack, rephrase, no relation). One of the main issues when approaching this problem is the lack of data, and the size of the publicly available corpora. In this work, we use the recently annotated US2016 debate corpus. US2016 is the largest existing argument annotated corpus, which allows exploring the benefits of the most recent advances in Natural Language Processing in a complex domain like Argument (relation) Mining. We present an exhaustive analysis of the behavior of transformer-based models (i.e., BERT, XLNET, RoBERTa, DistilBERT and ALBERT) when predicting argument relations. Finally, we evaluate the models in five different domains, with the objective of finding the less domain dependent model. We obtain a macro F1-score of 0.70 with the US2016 evaluation corpus, and a macro F1-score of 0.61 with the Moral Maze cross-domain corpus.

135

5.1 Introduction

Computational Argumentation has proved to be a very solid way to approach several problems such as fake news detection [196], recommendation systems [298] or debate analysis [185] among others. However, in almost every domain, it is of great importance to be able to automatically extract the arguments and their relations from the input source. Argument Mining (AM) is the Natural Language Processing (NLP) task by which this problem is addressed. The Transformer model architecture [380] and its subsequent pre-training approaches have been a turning point in the NLP research area. Thanks to its architecture, it has been possible to capture longer-range dependencies between input structures, and thus the performance of systems developed for the most general NLP tasks (i.e., translation, text generation or language understanding) improved significantly. Therefore, the Transformer architecture has laid the foundations on which newer models and pre-training approaches have been proposed, defining the state-of-the-art in NLP. In this work, we analyze the behavior of BERT [115], XLNET [410], RoBERTa [218], DistilBERT [335] and ALBERT [203] when facing the hardest AM task: identifying relational properties between arguments.

Argument Mining was formally defined in [265] as the task that aims to automatically detect arguments, relations and their internal structure. As pointed out in [207], due to the complexity of AM, the whole task can be decomposed into three main sub-tasks depending on their argumentative complexity. First, the identification of argument components consists in distinguishing argumentative propositions from non-argumentative propositions. This allows to segment the input text into arguments, making it possible to carry out the subsequent sub-tasks. Second, the identification of clausal properties is the part of AM that focuses on finding premises or conclusions among the argumentative propositions. Third, the last sub-task is the identification of relational properties. Two different argumentative propositions are considered at a time, and the main objective is to identify which type of relation links both propositions. Different relations can be observed in argument analysis, from the classical attack/support binary analysis [99], to the identification of complex patterns of human reasoning (i.e., argumentation schemes [400]). Therefore, the identification of argumentative relations is the most complex part of AM [207], but its complexity may vary depending on

how the problem is instantiated.

One of the main problems when addressing any AM task is the lack of high quality annotated data. In fact, as the argumentative complexity of the task increases, it gets harder to find large enough corpus to do experiments that match the latest NLP advances. An important feature that characterizes the transformer-based models is that large corpora are needed to achieve the performance improvement mentioned above. Recently, in [387], a new argument annotated corpus of the United States 2016 debate (*US2016*) was published. This corpus contains data from the transcripts of the televised political debates and from internet debates generated around the same context. This is the first publicly available corpus with enough data to begin exploring the benefits of the most recent contributions in NLP, when applied to the identification of argumentative relations. Additionally, the *US2016* corpus has been annotated using Inference Anchoring Theory (IAT[1]), a standard argument annotation guideline that provides more information than the classic attack/support binary annotation. Learning a model to automatically annotate with the use of IAT, makes it possible to evaluate its performance not only with the test samples of the corpus, but also with other different corpus already analyzed and tagged using this standard (e.g., *Moral Maze* corpus). This way, it is our objective to both: evaluate the performance of these new models in the identification of argument relations task; and to find out which one is more robust to variations in the application domain.

In this work, we explore the benefits of the most recent advances in NLP applied to relation prediction in the AM domain. For this purpose we use the recently published *US2016* corpus, since it is to the best of our knowledge, the largest annotated corpus containing information about argumentative relations, and the *Moral Maze* cross-domain corpus. Then we do: (i) a pre-processing of the corpus in order to clean and structure the data for the requirements of our experiments; (ii) an analysis of the performance of the most relevant transformer-based models (i.e., BERT, XLNET, RoBERTa, DistilBERT and ALBERT) when learning to predict the relations between argumentative propositions defined by the IAT standard; and (iii) an evaluation of the obtained models in five different domains (*Moral Maze* corpus) with the objective of analyzing the domain dependency of the transformer-

---

[1]`https://typo.uni-konstanz.de/add-up/wp-content/uploads/2018/04/IAT-CI-Guidelines.pdf`

based models when facing this AM task.

5.2   Related Work

Argument Mining is one of the main research areas in Computational Argumentation. AM has caught the attention of many researchers since it is considered to be the first step towards autonomous argumentative systems. We identified many different approaches to the Argument (relation) Mining problem, which depend on the proposed methods (i.e., Parsing algorithms, Textual Entailment Suites, Logistic Regression, Support Vector Machines and Neural Networks), and the available corpus at each moment. Initial research on automatic identification of argument relations was done in [265] where parsing algorithms were used to determine the type of relation existing between two argument propositions. Some years later, AM started to gain relevance in the NLP community. We can observe the popularization of machine learning techniques for NLP purposes in [252], [353] and [234]. Support Vector Machines (SVMs) seemed to be the best performing machine learning technique for the purpose of argument relation identification. With the advent of Neural Networks (NNs), a performance gap between previous works and this new approach could be observed. In [99] the empirical results obtained by Recurrent Neural Network (RNN) models for AM were significantly better. However, there is an interesting observation to make emphasis on, which makes it hard to compare AM works. As it can be pointed out after looking at the results depicted in works like [256] or [99], the corpus used in each work has a lot of influence in the results. This is due to many different factors such as class distributions, variable language complexity (e.g., use of irony, enthymemes, etc.) or the own size of the corpus. Therefore, misleading results may be observed if the generalization of the model is not properly evaluated. On the other hand, deep learning algorithms require much more data to significantly increase its performance compared to classic neural, machine learning or statistical methods. Therefore, from all these past years of argument relation identification works, the performance has been improved not only with the use of new models or techniques, but also with the creation of better corpora. In [387], a new argument annotated corpus (*US2016*) was published, with enough data to begin exploiting the benefits of the most recent
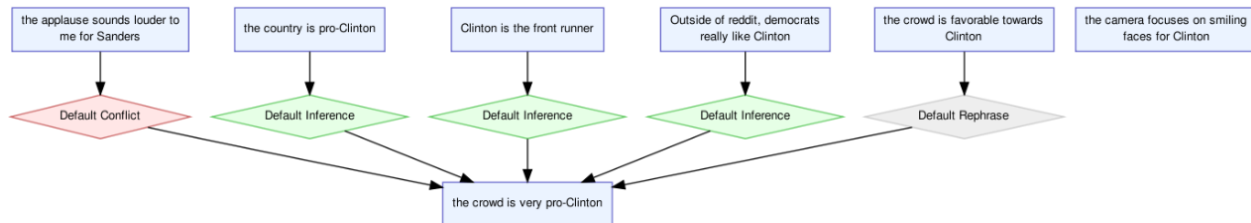
Figure 5.1: US2016 Argument Map Sample. ADUs are bounded by rectangles.
Relation types are contained in the rhombuses.

advances in NLP (i.e., transformer-based models) in the AM domain. To the best
of our knowledge, this is the first work addressing the Argument (relation) Mining
problem using Transformer-based models in more than a unique domain.

5.3   Data

Two different corpora have been used in this work: the *US2016* debate corpus and
the *Moral Maze* multi-domain corpus. Both corpora can be downloaded from the
Argument Interchange Format Database (AIFdb), an initiative of researchers from
the ARG-tech[2] with the objective of creating a standard formatted argument cor-
pus database [205]. This database contains 193 different argumentative corpora
structured using the AIF standard. Each corpus is divided into several argument
maps (**Figure 5.1**), and each argument map contains a set of Argument Discourse
Units (ADUs) with its argumentative relations annotated using the Inference An-
choring Theory (IAT). This annotation method considers the most important three
argumentative relations: inference (RA), conflict (CA) and rephrase (MA). An in-
ference relation between two propositions determines that one is used to support
or justify the other; a conflict relation indicates that two propositions have contra-
dictory information; and a rephrase between two propositions means that they are
equivalent from an argumentative point of view.

In order to adapt the AIFdb corpus to the needs of this task, we did some pre-
processing. Each argument map is stored in a JSON file, and represented as a
graph following the AIF standard. We generate a unique tab-separated values file

---

[2]https://www.arg-tech.org/

Table 5.1: Class distribution of the US2016 corpus, Train and Test partitions.

	US2016	Train	Test
RA	2744	2195	549
CA	888	710	178
MA	705	564	141
NO	8055	6444	1611
Total	12392	9913	2479

per corpus containing three different values: proposition1, proposition2 and label. In addition to the existing IAT relation labels, we decided to generate an additional relation: the no relation (NO) label. Since most of the pairs of propositions found in a debate are not related, we decided to generate a 65% of samples belonging to this new class. For this purpose, we mixed up the propositions that were not annotated with any of the IAT relation classes. This way, the resulting model will also be able to discriminate between related or not related propositions.

US2016 Debate Corpus

The *US2016*[3] corpus is an argument annotated corpus of the electoral debate carried out in 2016 in the United States. It contains both, transcriptions of the different rounds of TV debate, and discussions from the Reddit forums as detailed in [387]. The class distribution of the processed *US2016* corpus is depicted in **Table 5.1**. Since it is the largest publicly available argument annotated corpus in the literature, we used it to train the models. We decided to split the corpus with the 80% of the proposition pairs for training, and the remaining 20% for the evaluation.

---

[3]http://corpora.aifdb.org/US2016

Table 5.2: Multi-domain evaluation corpus (Moral Maze) class distribution.

	MM2012	B	E	M	P	W
RA	833	128	121	205	192	187
CA	200	26	36	30	45	63
MA	156	3	25	48	41	39
NO	2209	292	339	526	517	537
Total	3398	449	521	810	795	826

Moral Maze Multi-Domain Corpus

The *Moral Maze*[4] multi-domain corpus is an argument annotated corpus obtained from the transcriptions of the 2012 Moral Maze BBC discussion show. This corpus has been built from samples collected in five different broadcasts. The class distribution of the processed *Moral Maze* corpus is depicted in **Table 5.2**. This corpus is used to evaluate the domain robustness of the trained models across five different domain corpus: Bank (B), Empire (E), Money (M), Problem (P) and Welfare (W); each one focused on a specific debate topic and with a different distribution of classes.

5.4 Automatic Identification of Relational Properties

The problem addressed in this paper can be seen as an instance of the sentence pair classification problem. The sentence pair classification problem consists of assigning the most likely class to two text inputs at a time. In Argument Mining, after segmenting the text and defining the argument components, the argument graph must be built by identifying the relational properties between every two argument components. Therefore, given two argument components (i.e., sentences): $x_1^N = x_1, x_2, \ldots, x_N$ of length N where x_n is each word of the first component;

---

[4]http://corpora.aifdb.org/mm2012

and $y_1^M = y_1, y_2, \ldots, y_M$ of length M where y_m is each word of the second component, the classification problem can be modeled as defined in Equation 5.1,

$$(5.1) \qquad\qquad \hat{c} = \arg\max_{c \in C} p(c|x_1^N, y_1^M)$$

where $C = [RA, CA, MA, NO]$, so the four different relation types existing in the IAT labeling are considered: inference (RA), conflict (CA), rephrase (MA) and no relation (NO). To approach this problem, we decided to use transformer-based neural architectures. The most recent works in the literature tackle the Argument Mining problem using Recurrent Neural Networks (e.g., LSTMs, BiLSTMs, etc.). However, the Transformer architecture presents several interesting improvements with respect to the RNNs. The Transformer architecture uses multiple attention modules, which allow to capture longer range dependencies between words in a sentence. Given the nature of this work's task, we expect to have long input sentences since argumentative text is, generally, more complex than others. Therefore, we think attention mechanisms can be very useful for the identification of relational properties between argument components.

In this work, we apply Inductive Transfer Learning combined with different Transformer pre-training methods that allow us to learn our task not from scratch but using previously calculated weights. We decided to use the pre-training methods that performed the best in other NLP tasks such as Natural Language Understanding, Question Answering or Text Generation: BERT [115], XLNET [410], RoBERTa [218], DistilBERT [335] and ALBERT [203]. All these models have in common that they are based on the Transformer architecture, however different approaches have been considered in order to compute the initial weights. BERT, also known as Bidirectional Encoder Representations from Transformers, is pretrained on masked language model and next sentence prediction tasks. The model is designed to be able to fine-tune its weights on other different tasks by adding an additional output layer. XLNet is proposed after identifying a potential problem in BERT: the language modeling of the existing dependencies between the masked positions. XLNet combines both auto-regressive language modeling and auto-encoding techniques in order to overcome the detected potential issues. RoBERTa is an strong optimization of the BERT pre-training approach. After researchers did

Table 5.3: Transformer-based architectures configuration.

Model	TBlocks	HSize	AH	Params.
BERT-base [115]	12	768	12	110M
BERT-large [115]	24	1024	16	340M
XLNet-base [410]	12	768	12	110M
XLNet-large [410]	24	1024	16	340M
RoBERTa-base [218]	12	768	12	125M
RoBERTa-large [218]	24	1024	16	335M
DistilBERT-base [335]	6	768	12	66M
ALBERT-base [203]	12	768	12	11M
ALBERT-xxlarge [203]	12	4096	64	223M

a thorough analysis on the impact of the most important hyper-parameters, this new
model was able to obtain interesting results in most of the evaluated tasks. Finally,
both DistilBERT and ALBERT were proposed as smaller and faster versions of the
previous approaches. We find it interesting to also analyze and evaluate the behavior of these smaller versions, which have been designed to democratize the use of
transformer-based pre-training methods without significant loss of performance.

5.5   Evaluation

Experimental Setup

All the experiments carried out in this work have been run in a double NVIDIA
Titan V computer with an Intel Xeon W-2123 CPU and 62Gb of RAM. This way,
we can evaluate not only the performance of the models in the classification task,
but also their training computational cost in our specific task. The number of parameters of each model is directly related to the training computational cost. **Table
5.3** summarises the most relevant features that define each Transformer architecture considered in this research. The Transformer blocks (TBlocks) stand for the
number of layers; the hidden size (HSize) represents the number of hidden states
in each layer; the attention heads (AH) indicate the number of pointers used by the
attention layers; finally, the last feature is the total number of parameters (Params.)
of each architecture.

In our experiments, we explore the benefits of transfer learning applied to the argument relation mining task. For that purpose, during our training phase, we use the pre-trained encoder of each model with a linear layer on its top. The output size of the linear layer coincides with the number of classes considered in our instance of the problem (i.e., 4). With the *softmax* function, we are able to model the probability of belonging to one class or another for each pair of arguments (Equation 5.1).

We adapted the maximum sequence length and the batch size of our inputs in each experiment. These parameters were configured in order to use the whole available GPU memory. When training BERT-base models, we defined a maximum sequence length of 256 and a batch size of 64. When training BERT-large models, we halved those values to a maximum sequence length of 128 and a batch size of 32. We trained XLNet-base with a maximum sequence length of 256 and a batch size of 32, and XLNet-large with a maximum sequence length of 256 and a batch size of 8. RoBERTa-base was trained with a maximum sequence length of 256 and batch size of 32, and for training RoBERTa-large we used the same maximum sequence length but a batch size of 16. For DistilBERT we used a maximum sequence length of 256 and a batch size of 128. Finally, ALBERT-base was trained defining a maximum sequence length of 256 and batch size of 64, but in order to fit ALBERT-xxlarge in our available memory we had to define a maximum sequence length of 128 and a batch size of 4. We trained all these models for 50 epochs in our corpus. The best results (depicted in the following section) were obtained with a 1e-5 learning rate.

Results

In this section we present the empirical results obtained after running the experiments on all the previously defined models. In addition to the Transformer-based architectures, we have also trained a Recurrent Neural Network (RNN) as a baseline in our task. We used the best performing RNN architecture in argument relation mining proposed in [99], consisting of two Long Short-Term Memory (LSTM) networks working in parallel with each pair of arguments. We trained the baseline model for 50 epochs in our data, as the authors did in the original publication. In order to measure the performance of the different models, we have evaluated them

Table 5.4: Performance of the models in the automatic identification of argument relations, given in macro F1-scores.

Experiment	US2016-test	MM2012	Bank	Empire	Money	Problem	Welfare
LSTM (baseline)	.26	.24	.25	.22	.24	.25	.23
BERT-base-cased	.62	.53	.40	.45	.54	.47	.53
BERT-base-uncased	.65	.56	.42	.48	.54	.50	.54
BERT-large-cased	.61	.55	.45	.49	.53	.47	.51
BERT-large-uncased	.66	.57	.47	.49	.56	.49	.57
XLNet-base	.65	.56	.44	.49	.51	.54	.55
XLNet-large	.69	.57	.44	.51	.53	.53	.54
RoBERTa-base	.68	.58	.51	.52	.54	.52	.58
RoBERTa-large	**.70**	**.61**	**.53**	.53	**.59**	**.56**	**.59**
DistilBERT	.55	.42	.33	.39	.40	.43	.39
ALBERT-base-v2	.60	.54	.49	.45	.53	.47	.51
ALBERT-xxlarge-v2	.67	.59	.50	**.54**	.56	.48	**.59**

using the macro F1-score metric. Due to the huge class imbalance in our corpora, the use of the macro F1-score makes possible to avoid misleading results during the evaluation. Additionally, we also measured the training time required by each model when learning the task proposed in this work, in order to analyze if it can be worthwhile to sacrifice their performance in pursuit of faster training times or availability in lower resource environments.

The macro F1-scores obtained by each model are depicted in **Table 5.4**. In the first column, we can see every trained model. The second column represents the macro F1 obtained by each model when evaluated with the test partition of the same corpus used for training (i.e., *US2016*). The third column contains the scores obtained when the evaluation is performed on a different corpus (i.e., *MM2012*) containing a mixture of five domains. Finally, the last five columns are the macro F1-scores of the models when using each one of the five domain specific corpora (i.e., *Bank*, *Empire*, *Money*, *Problem* and *Welfare*) for evaluation.

With most of the models, we achieved state-of-the-art macro F1-scores for relation identification in Argument Mining [94]. Here, it is important to make emphasis that the way we considered to represent argumentative relations (i.e., IAT

labelling) make this task harder than most of the previous work (i.e., attack/support) in this area. We obtained a 0.70 macro F1-score with *RoBERTa-large*, outperforming the LSTM baseline used as a reference of previous research in argument relation identification. Furthermore, in order to have a more strong reference to compare with previous published results, we carried out an experiment using the same parameters but considering a binary instance of the problem (i.e., only attack and support relations). This way, *RoBERTa-large* achieved a macro F1-score of 0.81 highlighting the mentioned complexity gap between the two instances of the same problem. In general, we can observe that *RoBERTa* has performed very well in this task. When looking at the cross-domain evaluation, *RoBERTa-large* has also performed the best. We obtained a 0.61 macro F1-score when doing the evaluation with a different domain corpus. Moreover, the model has been able to keep a good performance with each one of the five domain specific corpora, even having different class distributions. With *ALBERT-xxlarge-v2* it has been possible to obtain a slightly better performance when evaluating with the *Empire* corpus. It is possible to observe how the scores obtained on the *Bank* and *Empire* corpus are slightly lower than the rest. This is mainly due to their smaller size, combined with the strong imbalance between classes. We also did experiments with cased and uncased models, in order to see the relevance of cased text in the relation identification task. As we can observe, the uncased models performed significantly better than the cased models, so we can point out that cased text did not help to improve the performance of the models in our task.

On the other hand, we obtained the worst results with *DistilBERT* and *ALBERT-base-v2*, as one might expect. We decided to use these models in order to see if the observed performance sacrifice was worthwhile in exchange for more feasible training times. **Table 5.5** contains the training times required by each model under our experimental setup. With *DistilBERT*, it was possible to achieve a significant reduction of training time in exchange for a huge drop in performance. However, with *ALBERT-base-v2* we could not observe a significant reduction of training time. From our experiments, we have not seen any significant advantage in using these *lite* models. We also observed that the computational cost of training *XLNet-large* and *ALBERT-xxlarge-v2* in our task was very expensive. *XLNet-large* was 5.1 times slower to train than *BERT-large*. As for *ALBERT-xxlarge-v2*, the training time was 7.1 times higher than *BERT-large*. This is due to its hidden size of 4096

Table 5.5: Training time of 50 epochs running in a double NVIDIA Titan V computer.

Experiment	Time
BERT-base	39m 11s
BERT-large	2h 19m 57s
XLNet-base	1h 52m 38s
XLNet-large	11h 51m 09s
RoBERTa-base	43m 17s
RoBERTa-large	4h 44m 33s
DistilBERT	16m 15s
ALBERT-base-v2	38m 04s
ALBERT-xxlarge-v2	16h 20m 22s

with respect to the 1024 sized large models. Thus, observing the performance of the models in means of their macro F1-score and the required training time, we still think that *RoBERTa* is the best approach to tackle both, domain-specific and cross-domain identification of relational properties between arguments. Even the *RoBERTa-base* version performed well in this task and it was 6.6 times faster than its large version on training.

Error Analysis

In an effort to conduct a thorough evaluation, we decided to analyze the errors made by *RoBERTA-large*, the best performing model. For this purpose we measured the volume of misclassifications found on each one of the four classes considered in this work. **Table 5.6** shows the error distribution detected when analyzing the results. Two important remarks can be pointed out when looking at the obtained error distributions. First of all, it is possible to observe how most of the misclassified argument pairs labeled with an inference relation were assigned the no relation class. Similarly, most of the misclassified argument pairs without relation were assigned the inference class by our model. We observed that many of

Table 5.6: Distribution of the misclassified samples per class using the *RoBERTa-large* model. Each column indicates the real class of the samples, each row indicates the assigned class by our model.

Pred. \ Real	RA	CA	MA	NO
RA	-	0.512	**0.603**	**0.730**
CA	0.200	-	0.138	0.226
MA	0.100	0.075	-	0.044
NO	**0.700**	0.412	0.259	-

these errors were due to a loss of contextual information. In an argumentative discourse, it is very common to refer to past concepts without explicitly mentioning them (i.e., enthymemes) or simplifying them with the use of pronouns. The lack of dialogical context can make the automatic identification of argument relations a harder task. For a better understanding of this problem we present the following example with two argument components labeled with inference relation:

P1: *I think **it**'s not going to help change the culture*

P2: *In banking **we**'ve a totally different situation*

Our system classified the pair as no related samples. In fact, by only reading the sentence pair, one may think there is not any argument relation between them. In these situations, it is evident that the key to avoid any possible error is to give additional information about the uttered propositions. In this case, depending on the background meaning of the "***it***" and "***we***" pronouns, the sentences may be related or not. The only way of considering this proposition pair related as an inference, is assuming that the ***it*** pronoun refers to the banking system. We also detected that in these situations the softmax outputs of our model gives very close probabilities for both, *RA* and *NO* classes. Another indicator of the existing model misunderstandings presented before, are the similar error distributions that conflicting arguments show with both inference and no relation classes. On the other

hand, we also pointed out that the rephrased argument pairs were mainly misclassified as inference related arguments. However, when analyzing them we observed that most of the relations could also be considered as inference related arguments depending on the interpretation. For example:

P1: **We** *do need curriculum reform*

P2: **RUBIO too** *believes in curriculum reform*

In this case, the sentence pair can be interpreted as a rephrase, assuming that "**We**" and "**RUBIO too**" are equivalent subjects. But it can also be interpreted as an argument from authority, with P2 supporting (inference relation) P1. In some situations the line that differences rephrase from inference may not be as clear as desired, and both types of relation can be considered correct. Additionally, with these second type of significant detected errors, it is also possible to observe the problem mentioned before. Therefore, the loss of information caused by the use of pronouns or enthymemes in the discourse can be determinant when approaching a task of this complexity.

5.6 Conclusion

The automatic identification of argument relations is an essential task in the whole computational argumentation process. It allows to automatically generate the argumentative structure from argument discourse units. In this work, we present how the automatic identification of argument relations, based on Inference Anchoring Theory labeling, can be approached using the latest advances in natural language processing. To the best of our knowledge, this is the first work using transformer-based pre-trained models to learn this task. For this purpose, we have used the largest publicly available argument annotated corpus to the date. Most of the trained models have been able to outperform the state-of-the-art baselines in argument relation mining ([94]), even with a more complex instance of the task. We observed a significant better performance with *RoBERTa* than other models, the best results were achieved with *RoBERTa-large*. We also made a cross-domain evaluation of the models, in order to find out their domain robustness. Even there

was a small drop in performance (most probably because of the significant variations of linguistic and class distributions between different domain corpora), the scores on different domains were still close to previous AM reports on a unique domain. This way, it is our objective to contribute on paving the way for finding models that do a better generalisation of this task. Finally, we analyzed the errors made by our best performing model. We have seen that two important groups of errors are caused by the loss of contextual information. We also pointed out that another important group of errors made by the model was due to possible multiple interpretations of the relations. We think that significant improvements in model performance can be achieved after analyzing the most common errors detected in this work. As future work, we propose the following modifications to the automatic identification of argument relations task: (i) pronoun replacement, to solve the loss of contextual information in some propositions; (ii) consider the possible classification ambiguity, in some cases, by accepting multiple correct relations if the interpretation leads to this conclusion; and (iii) incorporation of external information. In argumentation theory, an enthymeme is known as the omission of a claim or a support of an argument. In order to make the discourse more fluid, it is very common to use enthymemes in situations where the omitted information is considered to be known by all the participants. Therefore, without external information, the model may not be able to fully understand relations between enthymemes.

Automatic Debate Evaluation with Argumentation Semantics and Natural Language Argument Graph Networks

RAMON RUIZ-DOLZ, STELLA HERAS AND ANA GARCÍA-FORNES
Under Review in the Annual Meeting of the Association for Computational Linguistics 2023.
DOI:

Abstract

The lack of annotated data on professional argumentation and complete argumentative debates has led to the oversimplification and the inability of approaching more complex natural language processing tasks. Such is the case of the automatic evaluation of complete professional argumentative debates. In this paper, we propose an original hybrid method to automatically evaluate this kind of debates. For that purpose, we combine concepts from argumentation theory such as argumentation frameworks and semantics, with Transformer-based architectures and neural graph networks. Furthermore, we obtain promising results that lay the basis on an unexplored new instance of the automatic analysis of natural language arguments.

6.1 Introduction

The automatic evaluation of argumentative debates is a Natural Language Processing (NLP) task that can support judges in debate tournaments, analysts of

political debates, and even help to understand the human reasoning used in social media (e.g., Twitter debates) where argumentation may be difficult to follow. This task belongs to the computational argumentation area of research, a broad, multidisciplinary area of research that has been evolving rapidly in the last years [28]. Classically, computational argumentation research focused on formal abstract logic and computational (i.e., graph) representations of arguments and their relations. In this approach, the evaluation of arguments relied exclusively in logical and topological properties of the argument representations [16]. Furthermore, these techniques have been thoroughly studied and analysed, but from a theoretical and formal viewpoint considering specific cases and configurations instead of large, informal debates [382, 73].

The significant advances in NLP have enabled the study of new less formal approaches to undertake argumentative analysis tasks [207]. One of the most popular tasks that has gained a lot of popularity in the recent years is argument mining, a task aimed at finding argumentative elements in natural language inputs (i.e., argumentative discourse segmentation) [186], defining their argumentative purpose (i.e., argumentative component classification) [32], and detecting argumentative structures between these elements (i.e., argumentative relation identification) [325]. Even though most of the NLP research applied to computational argumentation has been focused in argument mining, other NLP-based tasks have also been researched such as the natural language argument generation [239], argument persuasive assessment [31], and argument summarisation [36], among others. However, it is possible to observe an important lack of research aimed at the evaluation of complete argumentative debates approached with NLP-based algorithms. Furthermore, most of the existing research in this topic has been contextualised in online debate forums, considering only short text arguments and messages, and without a professional human evaluation [172].

In this paper, we propose a hybrid method for evaluating complete argumentative debates considering the lines of reasoning presented by professional debaters and taking into account the evaluations provided by an impartial jury. Our method combines concepts from the classical computational argumentation theory (i.e., argumentation frameworks and semantics), with models and algorithms effectively used in other NLP tasks (i.e., Transformer-based sentence vector representations and graph networks). This way, we take a complete professional debate includ-

ing all the argumentation and rebuttal phases as an input, and predict the winning
stance (i.e., in favour or against) for a given argumentative topic. For that purpose,
we define the computational modelling of complete professional natural language
debates that makes possible to carry out the argumentative debate evaluation task
proposed in this paper. Finally, we present the automatic evaluation of a real pro-
fessional debate using the method proposed in this paper.

6.2   Related Work

Classically, the representation and evaluation of arguments is conducted through
argumentation frameworks and argumentation semantics [124]. However, this line
of research has been focused on abstract argumentation and formal logic-based ar-
gumentative structures, and has not been properly extended to the *informal* natural
language representation of human argumentation.

The automatic assessment of natural language arguments is a relatively new
topic of research that has been addressed from different NLP viewpoints. Most
of this research has focused on performing an individual evaluation of arguments
or argumentative lines of reasoning [390] instead of a global, *interactive* view-
point where complete debates consisting of multiple, conflicting lines of reasoning
are analysed. Typically, the automatic evaluation of natural language arguments
has been carried out comparing the convincingness of pairs of arguments [151];
analysing user features such as interests or personality to predict argument persua-
siveness [191]; and analysing natural language features of argumentative text to
estimate its persuasive effect [31]. Recently, a graph-based approach to evaluate
individual argument structures has been explored in [338].

The global (i.e., debate) approach on the evaluation of natural language argu-
ments was initially researched in [283] where Recurrent Neural Networks were
used to evaluate non-professional debates in a corpus of limited size and structure.
Following this trend, in [343], the authors propose a method based on the persua-
siveness to predict the outcome of online debates using a support vector machine.
Recently, in [172], the authors present an algorithm for predicting the outcome
of non-professional debates of limited length and depth in online forums. Further-
more, in the previous work the considered argumentative structures are simple, and

the proposed methods depend exclusively on natural language features. All these works have two main aspects in common: first, they are focused exclusively in online text-based debates, where information is easy to obtain, but very limited from an argumentative viewpoint; and second, the debates brought into consideration present short interactions and simpler arguments than the ones that can be found in a professional debate.

We have observed that concepts from computational argumentation theory are typically overseen in the literature, and the used corpora contain *debates* that are far from the concept of a professional debate. Thus, we propose a new method that combines the advantages of both areas of research (i.e., formal argumentation and NLP) aimed at approaching the automatic analysis of human argumentation. Furthermore, our proposal enables the analysis of more complex argumentative debates in both length and argumentative depth.

6.3 Data

In this paper, we approach the automatic evaluation of natural language professional debates in its full form. For that purpose, we use the *VivesDebate*[1] corpus [327] to conduct all the experiments and the evaluation of our proposed method. This corpus contains the annotations of the complete lines of reasoning presented by the debaters in a debate tournament based on the IAT [69] standard, and the professional jury evaluations of the quality of argumentation presented in each debate. It is important to emphasise this aspect, as the average length of the debates we analysed in this paper is 4819 words (30-40 minutes), and large language models have problems with long strings of natural language text [48]. Previously published corpora for the analysis of natural language argumentation always tended to simplify the annotated argumentative reasoning, by only considering individual arguments, pairs of arguments, or considering a small set of arguments, instead of deeper and complete lines of argumentative reasoning. For example, in argument mining (e.g., *US2016* [387]), argument assessment (e.g., *IBM-EviConv* [151]), or natural language argument generation/summarisation (e.g., *GPR-KB [260], DebateSum* , [318]). Furthermore, online debates with their crowd-sourced evalu-

---

[1]Available online in: `https://doi.org/10.5281/zenodo.6531487`

ations were compiled in [130], but argumentation was produced in short written paragraphs, and evaluations were based on anonymous votes from the community that did not require any justification. Therefore, the *VivesDebate* corpus is the only identified publicly available corpus that enables the study of the automatic natural language evaluation of professional argumentative debates in their complete form.

The *VivesDebate* corpus contains 29 complete argumentative debates (139,756 words) from a university debate tournament in Catalan annotated in their entire structure. Each debate is annotated entirely without partitions, and capturing the complete lines of reasoning presented by the debaters. The natural language text is segmented into Argumentative Discourse Units (7,810 ADUs) [274]. Each ADU contains its own text, its stance (i.e., in favour or against the topic of the debate), the phase of the debate where it has been uttered (i.e., introduction, argumentation, and conclusion), and a set of argumentative relations (i.e., inference, conflict, and rephrase) that make possible to capture argumentative structures, the sequentiality in the debate, and the existing major lines of reasoning. Additionally, each debate has the scores of the jury that indicate which team has proposed a more solid and stronger argumentative reasoning. An in-depth analysis of the corpus structure and statistics can be found in [327].

6.4   Method

The human evaluation of argumentative debates is a complex task that involves many different aspects such as the thesis solidity, the argumentation quality, and other linguistic aspects of the debate (e.g., oral fluency, grammatical correctness, etc.). It is possible to observe that both, the logic of argumentation and the linguistic properties play a major role in the evaluation of argumentative debates. Therefore, the method proposed in this paper is designed to capture both aspects of argumentation by combining concepts from argumentation theory and NLP. Our method is divided into two different phases: first, (i) determining the acceptability of arguments (i.e., their logical validity) in a debate based on their logical structures and relations; second, (ii) scoring the resulting acceptable arguments by analysing aspects of their underlying natural language features to determine the winner of a debate. Figure 6.1 presents an scheme with the most important
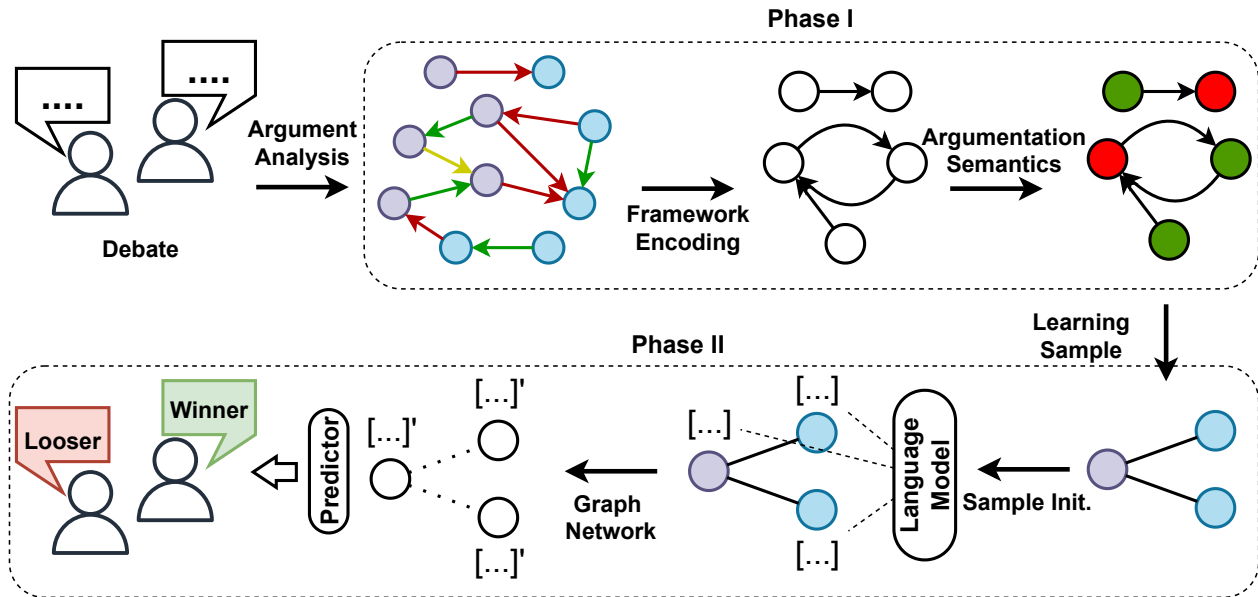
Figure 6.1: Structural scheme of the proposed automatic debate evaluation method.

phases and elements of the proposed method. The code implementation of the proposed method is publicly available in `https://github.com/raruidol/ArgumentEvaluation`.

Before describing both phases of our method, it is important to contextualise our proposal within the area of computational argumentation research. We assume that the whole argument analysis of natural language text has already been carried out: the argumentative discourse has been segmented, the argument components have been classified, and argument relations have been identified among the segmented argumentative text spans (see [237]). Thus, a graph structure can be defined from a given natural language argumentative input. As depicted in Figure 6.1, the *Argument Analysis* containing the text of the arguments (i.e., node content), their stance (i.e., node colour), inference relations (i.e., green edges), conflict relations (i.e., red edges), and rephrase relations (i.e., yellow edges) among arguments can be a valid starting point to the proposed method.

Phase I: Argument Acceptability

The first phase of the proposed method for the automatic evaluation of argumentative debates relies on concepts from computational argumentation theory. This phase can be understood as a pre-processing step from the NLP research viewpoint. Thus, the main goal of *Phase I* is to analyse the argumentative information contained in the argument graph, and to computationally encode this information focusing on the most relevant aspects for natural language argumentation (see Figure 6.1, *Framework Encoding* and *Argumentation Semantics*).

For that purpose, it is necessary to introduce the concept of an abstract argumentation framework and argumentation semantics. Originally proposed by Dung in [124], an argumentation framework is a graph-based representation of *abstract* (i.e., non-structured) arguments and their attack relations:

Definition 1 (Argumentation Framework) *An Argumentation Framework (AF) is a tuple $AF = < A, R >$ where: A is a finite set of arguments, and R is the attack relation on A such as $A \times A \to R$.*

Furthermore, argumentation semantics were proposed together with the AFs as a set of logical *rules* to determine the acceptability of an *abstract* argument or a set of arguments. In this paper, following one of the most popular notations in argumentation theory, we will refer to these sets as acceptable extensions. These semantics rely on two essential set properties: conflict-freeness and admissibility. Thus, we can consider that a set of arguments is conflict free if there are not any attacks between arguments belonging to the same set:

Definition 2 (Conflict-free) *Let $AF = < A, R >$ be an argumentation framework and $Args \subseteq A$. The set of arguments $Args$ is conflict-free iff $\neg\exists\alpha_i, \alpha_j \in Args : (\alpha_i, \alpha_j) \in R$.*

We can also consider that a set of arguments is admissible if, in addition of being conflict-free, it is able to defend itself from external attacks:

Definition 3 (Admissible) *Let $AF = < A, R >$ be an argumentation framework and $Args \subseteq A$. The set of arguments $Args$ is admissible iff $Args$ is conflict-free,*

and $\forall \alpha_i \in Args,\ \neg \exists \alpha_k \in A : (\alpha_k, \alpha_i) \in R$ *and* $(\alpha_i, \alpha_k) \notin R$*, or* $\neg \exists \alpha_j \in Args :$ $(\alpha_j, \alpha_k) \in R$

In this paper, we compare the behaviour of these two properties through the use of Naïve (conflict-free) and Preferred (admissible) semantics to compute all the acceptable extensions of arguments from the AF representations of debates. Naïve semantics are defined as maximal (w.r.t. set inclusion) conflict-free sets of arguments in a given AF. Similarly, Preferred semantics are defined as maximal (w.r.t. set inclusion) admissible sets of arguments in a given AF.

At this point, it is important to remark that the criteria of selecting both semantics for our method is oriented by the principle of maximality. Since acceptable extensions will be used as samples to train the natural language model in the subsequent phase of the proposed method, we selected these semantics that allow us to obtain the highest number of extensions, but keeping the most of the natural language information and maximising differences among the extensions (i.e., not accepting the subsets of a given maximal extension, which would result in data redundancy and hamper the distribution learning of the model).

Therefore, we encode the argument graphs, resulting from a natural language analysis of the debate, as abstract AFs using the proposed Algorithm 1. ADUs that follow the same line of reasoning (i.e., related with inference or rephrase) are grouped into *abstract* arguments, and the existing conflicts between ADUs are represented with the attack relation of the AF. Then, both Naïve and Preferred semantics are computed on the AF representation of the debate. This leads to a finite set of extensions, each one of them consisting of a set of acceptable arguments under the logic *rules* of computational argumentation theory. These extensions will be used as learning samples for training and evaluating the natural language model in the subsequent phase of the proposed method.

Phase II: Argument Scoring

The second phase of the method focuses on analysing the natural language arguments contained in the acceptable extensions, and determining the winner of a given debate. For that purpose, we use the Graph Network (GN) architecture combined with Transformer-based sentence embeddings generated from the

---

Algorithm 1 Argumentation Framework Encoding.

1: **function** GRAPHTOAF($ArgumentGraph$)
2:     $AG \leftarrow ArgumentGraph$
3:     $r \leftarrow AG.edges('conflict')$
4:     $AG.removeEdges(r)$
5:     $cc \leftarrow AG.connected\_components()$
6:     $AF \leftarrow NewGraph()$
7:     **for** $subgraph \in cc$ **do**
8:         $arg \leftarrow \{\}$
9:         **for** $node \in subgraph$ **do**
10:             $arg.append(node.Data())$
11:         **end for**
12:         $AF.addNode(arg)$
13:     **end for**
14:     $AF.addEdges(r)$
15:     **return** AF
16: **end function**

---

natural language arguments contained in the acceptable extensions. A GN is a machine learning algorithm aimed at learning computational representations for graph-based data structures [45]. Therefore, a GN receives a graph as an input containing initialised node features (i.e., $v_1, \ldots, v_i \in V$), edge data (i.e., $(e_1, r_1, s_1), \ldots, (e_k, r_k, s_k) \in E$, where e are the edge features, r is the receiver node, and s is the sender node), and global features (i.e., u); and updates them according to three learnt update ϕ and three static aggregation ρ functions:

$$(6.1) \quad \begin{aligned} e'_k &= \phi^e(e_k, v_{r_k}, v_{s_k}, u) & \bar{e}'_i &= \rho^{e \rightarrow v}(E'_i) \\ v'_i &= \phi^v(\bar{e}'_i, v_i, u) & \bar{v}' &= \rho^{e \rightarrow u}(E') \\ u' &= \phi^u(\bar{e}', \bar{v}', u) & \bar{v}' &= \rho^{v \rightarrow u}(V') \end{aligned}$$

This way, ϕ^e computes an edge-wise update of edge features, ϕ^v updates the features of the nodes, and ϕ^u is computed at the end, updating global graph features. Finally, ρ functions must be commutative, and calculate aggregated features, which are used in the subsequent update functions.

Thus, the first step in *Phase II* is to build the learning samples from the previously computed extensions of AFs (see Figure 6.1, *Learning Sample*). An extension is a set of logically acceptable arguments under the principles of conflict-freeness and/or admissibility. However, there are no explicit relations between the acceptable arguments, since AF representations only consider attacks between arguments, and the conflict-free principle states that there must be no attacks between arguments belonging to the same extension. Thus, in order to structure the data and make it useful for learning linguistic features for the debate evaluation task, we generate a complete bipartite graph from each extension. The two disjoint sets of arguments are determined by their stance (i.e., one set consisting of all the acceptable arguments in favour, and the other against), since argumentation semantics allow to define sets of logically acceptable arguments but do not guarantee that they will have the same claim or a similar stance.

The second step consists on initialising all the required features of the learning samples for the GN architecture (see Figure 6.1, *Sample Init.*). Thus, we define which features will encode edge, node, and global information of the previously processed bipartite graph samples. Edges do not contain any relevant natural language information, so we initialise edge features identically (similar to previous research [108]), so that node influence can be stronger when learning edge update functions. Nodes, however, are a pivotal aspect of this second phase since they contain all the natural language data. Node features are initialised from sentence embedding representations of the natural language ADUs contained in each node. Thus, we propose the use of Transformer-based Language Models (i.e., BERT, RoBERTa, XLNET, etc.) to generate dense vector representations of these ADUs, and initialise the vector features for learning the task. Finally, the global features of our learning samples encode the probability distribution of winning/losing a given debate (represented as acceptable extension-based bipartite graphs), and are a binary label that indicates the winning stance.

The final step in the *Phase II* of our proposed method is focused on learning the automatic evaluation of argumentative debates (see Figure 6.1, *Graph Network*). In a classical debate, there are always two teams/stances: in favour and against some specific claim. In this paper, we approach the debate evaluation as a binary classification task. Therefore, at the end of the proposed method, we model the classification problem as follows:

(6.2)
$$\hat{c} = \arg\max_{c \in C} P(c|G)$$

where $C = [\text{``}F\text{''}, \text{``}A\text{''}]$, depending on the winner of each debate (i.e., in "F"avour or "A"gainst). And G is a complete bipartite graph generated from the acceptable extensions of the AF pre-processing described in the *Phase I* of our method. Thus, we approach this probabilistic modelling with three Multi Layer Perceptrons (MLP) with two layers of 128 hidden units for each of the ϕ update functions. Since the debate evaluation is an instance of the graph prediction task, it is important to point out that the architecture of the two MLP approaching ϕ^e and ϕ^v are equivalent, and their parameters are learnt from the backpropagation of the MLP architecture for ϕ^u. Finally, the model has a 2-unit linear layer (for binary classification) and a *softmax* function (for modelling the probability distribution) on its top.

6.5   Experiments

Experimental Setup

All the experiments and results reported in this paper have been implemented using *Python 3* and run under the following setup. The initial corpus pre-processing and data structuring (i.e., *Phase I*) has been carried our using *Pandas* [233] together with *NetworkX* [167] libraries. Argumentation semantics have been implemented considering the *NetworkX*-based AF graph structures. Regarding *Phase II*, the language model and the dense sentence vector embeddings have been implemented through the *Sentence Transformers* library [306]. We used a pre-trained XLM-RoBERTa architecture [307] able to encode multilingual natural language inputs into a 768 dimensional dense vector space (i.e., word embedding size). Finally, the *Jraph*[2] library has been used for the implementation of the graph network architecture, for learning its update functions (i.e., Equation 6.1), and for the probabilistic modelling defined in Equation 6.2. We used an Intel Core i7-9700k computer with an NVIDIA RTX 3090 GPU and 32GB of RAM to run our experiments.

---

[2]`https://jraph.readthedocs.io/`

Furthermore, it is important to completely define the notion behind a learning sample, and how the data pipeline manages all these samples and structures them for training/evaluation. In our proposal, we defined a learning sample as an acceptable extension of a given debate. Thus, different debates may produce a different number of learning samples depending on the argumentation semantics and/or the argumentation framework topology. This way, learning samples can be managed from a debate-wise or an extension-wise viewpoint. Even though we used the learning samples individually (i.e., extension-wise) for the training of the proposed models, we will always consider debate-wise partitions of our data in our experimental setup. This decision has been made because it would be unfair to consider learning samples belonging to the same debate in both our train and test data partitions. The reported results could be misleading, and would not properly reflect the capability of generalisation of our method.

Therefore, we used 28 complete debates in our experiments (17 in favour, 11 against) that were divided following the 80% - 20% distribution for training and test partitions. Once having the partitions defined, we applied different semantics to generate the acceptable extensions (i.e., learning samples), and we trained the GN model for 2500 training steps.

Results

Regarding the *Phase I*, we have computed the Naïve and Preferred acceptable extension sets from the AF representations of the 28 debates. This led us to a total of 467 Naïve and 31 Preferred extensions. We can observe how the admissibility principle is much more strict than the conflict-free principle, and has a significant repercussion on the number of acceptable extensions (i.e., learning samples) produced. The 467 Naïve extensions are distributed as follows: 203 learning samples belonging to class 0 (i.e., in favour team wins in the 43.4% of the cases), and 264 samples belonging to class 1 (i.e., against team wins in the 56.6% of the cases). Similarly, the 31 Preferred extensions are distributed as follows: 19 learning samples belonging to class 0, and 12 samples belonging to class 1. Given the irregular generation of acceptable extensions per debate mentioned above, we carried out all of our experiments with a fixed train/test partition with the same 80% of the debates used for train and the 20% of the debates used for evaluation, such that the

produced acceptable extensions in both partitions did not significantly alter this distribution. We can observe an example of this irregularity with the AF encoded from *Debate3.csv* producing 92 Naïve extensions and the AF encoded from *Debate29.csv* with only 4 Naïve extensions. Given this restriction, and to provide solid results, we are not able to perform a K-Fold evaluation without considering data splits with significant differences in data distributions. Thus, in all of our experiments we used 23 debates producing 369 Naïve and 26 Preferred samples for training, and 5 debates for test.

With the data configurations defined above, we trained two different GN models in order to approach the task of automatic evaluating argumentative debates: the *Naïve-GN* and the *Preferred-GN*. Furthermore, we defined four baselines to compare the performance of our proposed method: a Random Baseline (*RB*) that assigned randomly a class to each extension; a Naïve Argumentation Theory Baseline (*Naïve-ATB*) that classified each Naïve extension by counting the majority number of acceptable arguments belonging to each stance; a Preferred Argumentation Theory Baseline (*Preferred-ATB*) that did a similar classification to the *Naïve-ATB* but considering Preferred extensions; and a Language Modelling Basline (*LMB*) which is implemented ignoring the *Phase I* of the proposed method, and the GN is trained directly over the whole argumentative analysis graphs. The results of our experiments are depicted in Table 6.1.

It is possible to observe how the best performing results are achieved by the *Naïve-GN* model, which is also the one with the most number of samples for train. We also observed that the *LMB* always learns to predict the majority class. Both *ATB*s performed worse, meaning that relying only in aspects from computational argumentation theory and logic might not be enough to conduct a proper automatic evaluation of natural language debates. Therefore, we can observe that the hybrid method proposed in this paper benefits from the logical aspects of argumentation theory to improve the available data from a given set of natural language debates, together with NLP techniques that enable a probabilistic modelling of the natural language and improve the automatic evaluation of human debates. The provided results can be used as a baseline for future research.

Experiment	Train		Eval. Metrics	
Model	D	S	Acc.	Macro-F1
Naïve-GN	23	369	0.72	0.48
Preferred-GN	23	26	0.40	0.40
Naïve-ATB	-	-	0.16	0.14
Preferred-ATB	-	-	0.20	0.16
LMB	23	23	0.60	0.37
RB	-	-	0.48	0.33

Table 6.1: Accuracy and Macro-F1 results of the automatic debate evaluation task. D and S indicate the number of debates and learning samples respectively used in the Train data partition in our experiments.

6.6 Automatic Evaluation of Argumentative Debates

In this section, we analyse the *Naïve-GN* model when used for the automatic evaluation of a complete real debate. For that purpose, we use the *Debate7.csv* file, which was randomly excluded from the previous experiments. Figure 6.2 presents an argumentative graph resulting from the preliminary analysis of the argumentative natural language in this debate file. The nodes represent ADUs and the edges argumentative relations between them. The size of the node in the argument graph indicates the phase of the debate where each ADU was uttered (i.e., introduction, argumentation, or conclusion), and its colour indicates the team stance. The colour of the edges indicates what kind of argumentative relation exists between different ADUs: a red arrow represents a conflict, a green represents an inference, and a yellow represents a rephrase.

PHASE I: The first step is to encode the natural language argumentation graph into a compact computational representation that simplifies the data structures and condenses all the argumentative information correctly. For that purpose, Algorithm 1 is applied taking the argumentation graph (Figure 6.2) as the input and producing
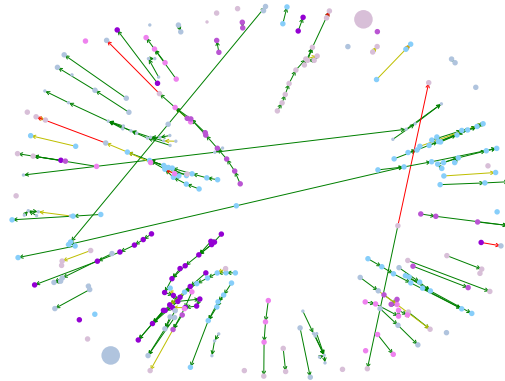
Figure 6.2: Argumentation graph resulting from a preliminary analysis of a natural
language debate.
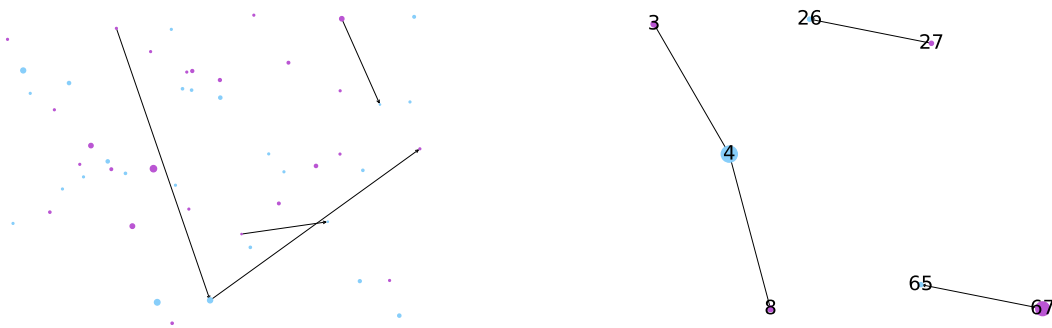
(a) Complete Argumentation Framework gener-
ated from the argumentation graph.

(b) Conflicting arguments of the Argumentation
Framework.

Figure 6.3: Argumentation Framework visualisation.

an Argumentation Framework (Figure 6.3a) as the output. Here, the size of the
nodes represent the natural language size of each *abstract* argument. Next, we
compute the acceptable extensions of the AF under Naïve semantics. Four different
extensions are obtained from this AF. All the nodes (i.e., arguments) that do not
belong to any attack relation are included in all of the acceptable extensions. Then,
the variations between extensions happen within the conflicting nodes. Under the
conflict-free principle, arguments [3, 8, 26, 65] are accepted within extension $\mathcal{E}_1$;
[3, 8, 26, 67] within extension $\mathcal{E}_2$; [3, 8, 27, 65] within $\mathcal{E}_3$; and [3, 8, 27, 67] within
$\mathcal{E}_4$ (see Figure 6.3b for node numeration).

PHASE II: The next step is to generate the learning samples from the four acceptable extensions. Thus, a complete bipartite graph is produced per each extension where the two disjoint sets contain all the acceptable arguments belonging to the same stance (i.e., $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_4$). Before predicting the outcome of the debate, each of these samples is initialised with the XLM-RoBERTa sentence embedding vector representations of the natural language arguments encoded within their nodes. Then, each of the four samples is fed into the *Naïve-GN* model trained in our experiments and the following predictions are produced: $\mathcal{S}_1 \rightarrow F$, $\mathcal{S}_2 \rightarrow F$, $\mathcal{S}_3 \rightarrow F$, and $\mathcal{S}_4 \rightarrow F$ (i.e., team in favour wins). In this specific case, the model has correctly predicted all the debate samples, but this might not be always the case. In these situations where there are conflicting predictions (e.g., $\mathcal{S}_1 \rightarrow F$, $\mathcal{S}_2 \rightarrow A$), a normalised aggregation of the *softmax* probability distributions for each different extension is done to determine the most probable class for a given debate.

6.7 Conclusions

In this paper, we propose an original hybrid method to approach the automatic evaluation of natural language professional argumentative debates. For that purpose, we present a new instance of the argument assessment task, where argumentative debates and their underlying lines of reasoning are considered in a comprehensive, undivided manner. The proposed method combines aspects from formal logic and computational argumentation theory, with NLP and Deep Learning. From the observed results, several conclusions can be drawn. First, it has been possible to determine that our method performed better than approaching independently the debate evaluation task from either the argumentation theory or the NLP viewpoints. Furthermore, we have observed in our experiments that conflict-free semantics produce a higher number of acceptable extensions from each AF compared to the admissibility-based semantics. This helped to improve the learning of the argument evaluation task in a similar way to that achieved by data augmentation techniques for Deep Learning. Thus, better probabilistic distributions of natural language features and dependencies that are not too constrained to formal logic and graph topology (as they are when using admissibility-based semantics)

can be learnt by our model. This paper represents a solid starting point of research in the evaluation of long natural language professional debates. As future work, we foresee to consider finer-grained features for the evaluation of argumentation such as thesis solidity, argumentation quality, and adaptability. We also plan to extend our method with acoustic features, considering aspects such as the intonation or the fluency. Finally, there is still a problem with the amount of annotated data. Subsequent research in this topic would significantly benefit from extending the number of completely annotated professional debates.

VivesDebate-Speech: A Corpus of Spoken Argumentation to Leverage Audio Features for Argument Mining

RAMON RUIZ-DOLZ AND JAVIER IRANZO-SÁNCHEZ

Abstract

In this paper, we describe VivesDebate-Speech, a corpus of spoken argumentation created to leverage audio features for argument mining tasks. The creation of this corpus represents an important contribution to the intersection of speech processing and argument mining communities, and one of the most complete publicly available resources in this topic. Moreover, we have performed a set of first-of-their-kind experiments which show an improvement when integrating audio features into the argument mining pipeline. The provided results can be used as a baseline for future research.

7.1   Introduction

The automatic analysis of argumentation in human debates is a complex problem that encompasses different challenges such as mining, computationally representing, or automatically assessing natural language arguments. Furthermore, human argumentation is present in different mediums and domains such as argumentative monologues (e.g., essays) and dialogues (e.g., debates), argumentation in text

169

(e.g., opinion pieces or social network discussions) and speech (e.g., debate tournaments or political speeches), and domains such as the political [162, 321], legal [285], financial [85], or scientific [6, 35] among others. Thus, human argumentation presents a linguistically heterogeneous nature that requires us to carefully investigate and analyse all these variables in order to propose and develop argumentation systems which are robust to these variations in language. In addition to this heterogeneity, it is worth mentioning that a vast majority of the publicly available resources for argumentation-based Natural Language Processing (NLP) have been created considering text features only, even if their original source comes from speech [387, 152]. This is a substantial limitation, not only for our knowledge on the impact that speech may directly have when approaching argument-based NLP tasks, but because of the significant loss of information that happens when we only take into account the text transcript of spoken argumentation.

In this work, we will focus on the initial steps of argument analysis considering acoustic features, namely, the automatic identification of natural language arguments. Argument mining is the area of research that studies this first step in the analysis of natural language argumentative discourses, and it is defined as the task of automatically identifying arguments and their structures from natural language inputs. As surveyed in [207], argument mining can be divided into three main sub-tasks: first, the segmentation of natural language spans relevant for argumentative reasoning (typically defined as Argumentative Discourse Units ADUs [274]); second, the classification of these units into finer-grained argumentative classes (e.g., major claims, claims, or premises [352]); and third, the identification of argumentative structures and relations existing between these units (e.g., inference, conflict, or rephrase [325]). Therefore, our contribution is twofold. First, we create a new publicly available resource for argument mining research that enables the use of audio features for argumentative purposes. Second, we present first-of-their-kind experiments showing that the use of acoustic information improves the performance of segmenting ADUs from natural language inputs (both audio and text).

The rest of the paper is structured as follows. Section 7.2 describes the *VivesDebate-Speech* corpus. Section 7.3 provides a thorough description of the problem approached in this work. Section 7.4 explains the proposed methodology used to integrate audio features into the argument mining process. Section 7.5 re-

ports and analyses the results observed during our experimentation. Lastly, Section
7.6 summarises the most important findings from our experiments.

7.2 The VivesDebate-Speech Corpus

The first step in our research was the creation of a new natural language argumentative corpus. In this work, we present *VivesDebate-Speech*, an argumentative corpus created to leverage audio features for argument mining tasks. The *VivesDebate-Speech* has been created taking the previously annotated *VivesDebate* corpus [327] as a starting point.

The *VivesDebate* corpus contains 29 professional debates in Catalan, where each debate has been comprehensively annotated. This way, it is possible to capture longer-range dependencies between natural language ADUs, and to keep the chronological order of the complete debate. Although the nature of the debates was speech-based argumentation, the *VivesDebate* corpus was published considering only the textual features included in the transcriptions of the debates that were used during the annotation process. In this paper, we have extended the *VivesDebate* corpus with its corresponding argumentative speeches in audio format. In addition to the speech features, we also created and released the BIO (i.e., *Beginning*, *Inside*, *Outside*) files for approaching the task of automatically identifying ADUs from natural language inputs (i.e., both textual and speech). The BIO files allow us to determine whether a word is the ***Beginning***, it belongs ***Inside***, or it is ***Outside*** an ADU.

The *VivesDebate-Speech* corpus is, to the best of our knowledge, the largest publicly available resource for spoken argument mining. Furthermore, combined with the original *VivesDebate* corpus, a wider range of NLP tasks can be approached taking the new audio features into consideration (e.g., argument evaluation or argument summarisation). Compared to the size of the few previously available speech-based argumentative corpora [217, 235] (i.e., 2 and 7 hours respectively), the *VivesDebate-Speech* represents a significant leap forward (i.e., more than 12 hours) for the research community (see Table 7.1). Therefore, the *VivesDebate-Speech* corpus consists of two parts. First, the text BIO files allow us to determine which span of the natural language debate is an ADU and which is

not. Second, using these BIO files and the text transcriptions of the debates, we have been able to align them with the audio of the debate and produce a set of *.wav* files containing the audio features of each ADU. The *VivesDebate-Speech* is released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0) and can be publicly accessed from Zenodo[1].

Text

The text-based part of the *VivesDebate-Speech* corpus consists of 29 BIO files where each word in the debate is labelled with a BIO tag. This way, the BIO files created in this work enable the task of automatically identifying argumentative natural language sequences existing in the complete debates annotated in the *VivesDebate* corpus. Furthermore, these files represent the basis on which it has been possible to achieve the main purpose of the *VivesDebate-Speech* corpus, i.e., to extend the textual features of the debates with their corresponding audio features.

We created the BIO files combining the transcriptions and the ADU annotation files of the *VivesDebate*[2] corpus. For that purpose, we performed a sequential search of each annotated ADU in the transcript file of each corresponding debate, bringing into consideration the chronological order of the annotated ADUs.

Speech

Once the revised transcription has been augmented with the ADU information, the transcription was force-aligned with the audio in order to obtain word level timestamps. This process was carried out using the hybrid DNN-HMM system that was previously used to obtain the VivesDebate transcription, implemented using TLK [113]. As a result of this process, we have obtained *(start,end)* timestamps for every *(word,label)* pair. We split the corpus into train, dev, and test considering the numerical order of the files (i.e., Debate1-23 for train, Debate24-26 for dev,

---

[1]https://doi.org/10.5281/zenodo.7102601
[2]https://doi.org/10.5281/zenodo.5145655

Set	# debates	Duration	# tags		
			B	**I**	**O**
Train	23	9.8h	4605	63305	21432
Dev	3	1.3h	692	8058	3752
Test	3	1.3h	640	8413	3102

Table 7.1: Set-level statistics of the VivesDebate-Speech corpus. Each debate is carried out between two teams, and two to four members of each team participate as speakers in the debate.

and Debate27-29 for test). The statistics for the final *VivesDebate-Speech* corpus are shown in Table 7.1.

7.3 Problem Description

For understanding argumentative discourses, the first phases of the human argumentative reasoning process are the identification and the analysis of natural language arguments [395]. From a computational viewpoint, these phases are typically framed into the argument mining area of research, which investigates how to identify and analyse arguments in natural language inputs. In this paper, we approach the segmentation and identification of ADUs in natural language debates. Furthermore, we approach this problem considering both, text-based and audio-based natural language features. The inclusion of audio-based natural language features into the argument mining pipeline extends the scarce previous existing research in this topic [217, 235], and allows to explore this new dimension, which has not been typically addressed from the computational viewpoint, but that represents an important source of information for the human argumentative reasoning process.

Therefore, we approached the identification of natural language ADUs in two different ways: (i) as a token classification problem, and (ii) as a sequence classification problem. For the first approach, we analyse our information at the token level. Each token is assigned a BIO label and we model the probabilities of a token belonging to each of these specific label considering the n-length closest

natural language contextual tokens. For the second approach, the information is analysed at the sentence level. In order to address the ADU identification task as a sequence classification problem we need to have a set of previously segmented natural language sequences. Then, the problem is approached as a 2-class classification task, discriminating argumentative relevant from non-argumentative natural language sequences. In the following section, we present our proposal to tackle this problem considering both approaches, token and sequence level analysis of natural language inputs.

7.4 Proposed Method

The use of audio information for argument mining presents significant advantages across 3 axes: efficiency, information and error propagation. Firstly, the segmentation of the raw audio into independent units is a pre-requisite for most Automatic Speech Recognition (ASR) system. If the segmentation produced in the ASR step is incorporated into the argument mining pipeline, we remove the need for a specific text-segmentation step, which brings significant computational and complexity savings. Secondly, the use of audio features allows us to take into account relevant prosodic features such as intonation and pauses which are critical for discourse segmentation, but that are missing from a text-only representation. Lastly, the use of ASR transcriptions introduces noise into the pipeline as a result of recognition errors, which can hamper downstream performance. Working directly with the speech signal allows us to avoid this source of error propagation.

Two different methods to leverage audio features for argument mining are explored in this paper. First, a standard end-to-end (E2E) approach that takes the text-based transcription of the spoken debate produced by the ASR as an input, and directly outputs the segmentation of this text into argumentative units. Second, we propose a cascaded model composed of two sub-tasks: argument segmentation and argument classification. In the first sub-task, the discourse is segmented into independent units, and then for each unit it is determined if it contains argumentative information or not. Both approaches produce an equivalent output, a sequence of BIO tags which is then compared against the reference. This work investigates how audio features can be best incorporated into the previously described process.
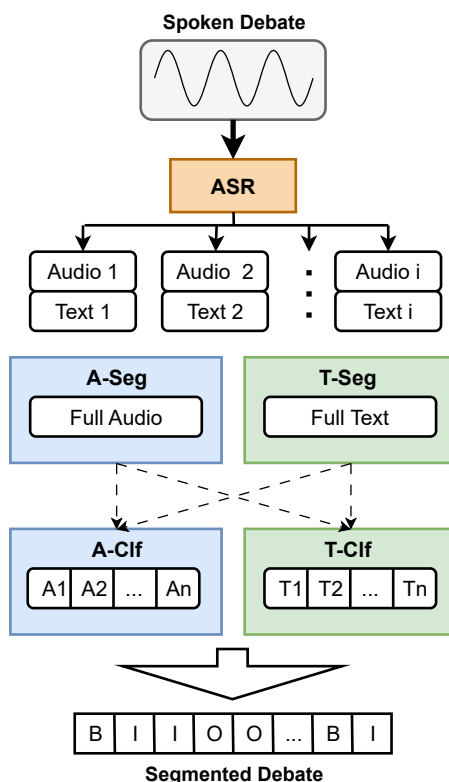
Figure 7.1: Overview of the proposed cascaded approach.

An overview of the proposed cascaded method is shown in Figure 7.1. As we can observe, a segmentation step follows the ASR step, which segments either the whole audio (A-Seg) or the whole text (T-Seg) into (potential) argumentative segments. A classification step then detects if the segment contains an argumentative unit, by using either the audio (A-Clf) or the text (T-Clf) contained in each segment. If efficiency is a major concern, the segmentation produced by the ASR step can be re-used instead of an specific segmentation step tailored for argumentation, but this could decrease the quality of the results. This process significantly differs from the E2E approach where the BIO tags are directly generated from the output of the ASR step. This way, our cascaded model is interesting because it makes possible to analyse different combinations of audio and text features.

The cascaded method has one significant advantage, which is that audio segmentation is a widely studied problem in ASR and Speech Translation (ST) for which significant breakthroughs have been achieved in the last few years. Cur-

rently, one of the best performing audio segmentation methods is SHAS [374], which uses a probabilistic Divide and Conquer (DAC) algorithm to obtain optimal segments. Furthermore, we have compared SHAS with a Voice Activity Detection (VAD) baseline, as well as with the non-probabilistic VAD method [282] using a Wav2Vec2 pause predictor [29], which performs ASR inference and then splits based on detected word boundaries. To complete our proposal, we have also explored text-only segmentation methods in which a Transformer-based model is trained to detect boundaries between natural language segments. This way, each word can belong to two different classes, boundary or not boundary.

The second stage of our cascaded method is an argument detection classifier, that decides, for each segment, if it includes argumentative content and should be kept, or be discarded otherwise. In the case that our classifier detects argumentative content within a segment, its first word is assigned the *B* label (i.e., *Begin*) and the rest of its words are assigned the *I* label (i.e., *Inside*). Differently, if the classifier does not detect argumentative content within a segment, all the words belonging to this segment are assigned the *O* label (i.e., *Outside*).

7.5 Experiments

Experimental Setup

Regarding the implementation of the text-based sub-tasks (see Figure 7.1, green modules) of our cascaded method for argument mining, we have used a RoBERTa [218] architecture pre-trained in Catalan language data. For the segmentation (T-Seg), we experimented with a RoBERTa model for token classification, and we used the segments produced by this model to measure the impact of the audio features compared to text in the segmentation part of our cascaded method. For the classification (T-Clf), we finetuned a RoBERTa-base model[3] for sequence classification with a two-class classification *softmax* function able to discriminate between the argument and the non-argument classes.

The implementation of the audio-based sub-tasks (see Figure 7.1, blue modules) is quite different between segmentation to classification. For audio-only seg-

---

[3]projecte-aina/roberta-base-ca-v2 was used for all RoBERTa-based models reported in this work

mentation (A-Seg), we performed a comparison between the selected algorithms:
VAD, DAC, and SHAS. For the hybrid DAC segmentation, two Catalan W2V ASR
models are tested, xlsr-53[4] and xls-r-1b[5]. For SHAS, we used the original SHAS
Spanish and multilingual checkpoints. Additionally, we trained a Catalan SHAS
model with the *VivesDebate-Speech* train audios, as there exists no other public
dataset that contains the unsegmented audios needed for training a SHAS model.
Regarding our audio-only classifier (A-Clf), Wav2Vec2[6] models have been fine-
tuned on the sequence classification task (i.e., argument/non-argument).

As for the training parameters used in our experiments with the RoBERTa mod-
els, we trained them for 50 epochs, considering a learning rate of 1e-5, and a batch
size of 128 samples. The best model among the 50 epochs was selected based on
the performance in the dev set. All the experiments have been carried out using an
Intel Core i7-9700k CPU with an NVIDIA RTX 3090 GPU and 32GB of RAM.

Evaluation

In order to evaluate the performance of the proposed methods, we use the reference
transcriptions and timestamps of the VivesDebate-Speech as input to our models.
This is necessary in order to be able to compare the system hypothesis with the
reference labels. If a real ASR system had been used, a different transcription
would have been obtained, and we would need a way of mapping the BIO tags of
the real, noisy transcription with those of the reference. There is no reliable way to
obtain such a mapping, as the BIO-annotated transcription and the reference will
have different lengths and consist of different words.

Table 7.2 shows the best performance on the dev set of the different audio seg-
mentation methods tested. Results are reported without using an argument classi-
fier, which is equivalent to a majority class classifier baseline, as well as an oracle
classifier which assigns the most frequent class (based on the reference labels) to a
segment. This allows us to analyze the upper-bound of performance that could be
achieve with a perfect classifier.

---

[4]softcatala/wav2vec2-large-xlsr-catala

[5]PereLluis13/wav2vec2-xls-r-1b-ca

[6]facebook/wav2vec2-xls-r-300m and facebook/wav2vec2-xls-r-1b

Method	Classifier			
	Majority		Oracle	
	Acc.	Macro-F1	Acc.	Macro-F1
Baseline (VAD)	0.53	0.35	0.70	0.53
DAC xlsr-53	0.61	0.34	0.81	0.65
DAC xls_r-1b	0.61	0.35	0.81	0.64
SHAS-es	0.64	0.37	0.83	0.66
SHAS-ca	0.64	0.36	0.82	0.64
SHAS-multi	0.65	0.37	0.83	0.66

Table 7.2: Audio segmentation methods performance on the dev set, as measured by accuracy (Acc.) and Macro-F1.

The results highlight the strength of the SHAS method, with the SHAS-es and SHAS-multi models which are working on a zero-shot scenario, outperforms the Catalan W2V models. The SHAS-ca model had insufficient training data to achieve parity with the zero-shot models trained on larger audio collections. As a result of this, the SHAS-multi model was selected for the rest of the experiments.

One key factor to study is the relationship between the maximum segment length (in seconds) produced by the SHAS segmenter, and the performance of the downstream classifier. Longer segments provide more context that can be helpful to the classification task, but a longer segment might contain a significant portion of both argumentative and non-argumentative content. Figure 7.2 shows the performance of the text classifier as a function of segment size, measured on the dev set. 5 seconds was selected as the maximum sentence length, as shorter segments did not improve results.

Once the hyperparameters of each individual model have been optimised, the best results for each system combination are reported on Table 7.3. The results are consistent across the dev and test sets. The end-to-end model outputs the BIO tags directly, either from fixed length input (of which 5 was also the best performing value), denoted as *E2E BIO-5*. Alternatively, the *E2E BIO-A* was trained considering the natural language segments produced by the SHAS-multi model instead of
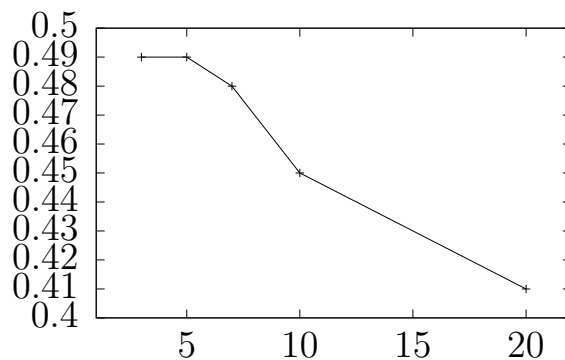
Figure 7.2: Dev set F1 score as a function of maximum segment length (s), SHAS-multi segmenter followed by text classifier.

Model	Dev		Test	
	Acc.	**Macro-F1**	**Acc.**	**Macro-F1**
E2E BIO-5	0.71	0.45	0.72	0.47
E2E BIO-A	0.72	0.48	0.75	0.49
T-Seg+A-Clf	0.59	0.41	0.58	0.41
T-Seg+T-Clf	0.64	0.49	0.69	0.49
A-Seg+A-Clf	0.60	0.42	0.58	0.43
A-Seg+T-Clf	0.67	**0.51**	0.70	**0.51**

Table 7.3: Accuracy and Macro-F1 results of the argumentative discourse segmentation task on both *dev* and *test* sets.

relying on a specific maximum length defined without any linguistic criteria. This way, it was our objective to improve the training of our end-to-end model through the use of linguistically informed audio-based natural language segments. It can be observed how this second approach leverages the audio information to improve test macro-F1 from 0.47 to 0.49.

For the cascade model, we test both audio and text segmenters and classifiers. Similarly to the E2E case, the use of audio segmentation consistently improves the results. For the text classifier, moving from text segmentation (*T-Seg + T-Clf*) to audio segmentation (*A-Seg + T-Clf*) increases test macro-F1 from 0.49 to 0.51. Likewise, when using an audio classifier *(A-Clf)*, audio segmentation improves the test results from 0.41 to 0.43 macro-F1. However, the relatively mediocre performance of the audio classification models with respect to its text counterparts stands out. Although the use of an audio classifier is significantly better than the majority baseline (see Table 7.2), there is still quite a gap with the performance achieved by the text models. We believe this could be caused due to the fact that speech classification is a harder task than text classification in our setup, because the audio classifier deals with the raw audio, whereas the text classifier uses the reference transcriptions as input. Additionally, the pre-trained audio models might not be capable enough for some transfer learning tasks, as they have been trained with a significantly lower number of tokens, which surely hampers language modelling capabilities.

7.6 Conclusions

We have presented the *VivesDebate-Speech* corpus, which is currently the largest speech-based argument corpus. This opens up exciting research opportunities for more realistic argument mining experiments taking into account audio information.

The experiments have shown how having access to audio information can be a source of significant improvement. Specifically, using audio segmentation instead of text-based segmentation consistently improves performance, both for the text and audio classifiers used in the cascade approach, as well as in the end-to-end scenario, where audio segmentation is used as a decoding constraint for the model.

In terms of future research, an exciting research direction is to better under-

stand the underperformance of the audio-based classifiers, as well as devising new techniques for bridging the gap with the results obtained by text classifiers.

Part IV

Argument-based Computational Persuasion

A Qualitative Analysis of the Persuasive Properties of Argumentation Schemes

RAMON RUIZ-DOLZ, JOAQUIN TAVERNER, STELLA HERAS, ANA
GARCÍA-FORNES AND VICENT BOTTI

Abstract

Argumentation schemes are generalised patterns that provide a way to (partially) dissociate the content from the reasoning structure of the argument. On the other hand, Cialdini's principles of persuasion provide a generic model to analyse the persuasive properties of human interaction (e.g., natural language). Establishing the relationship between principles of persuasion and argumentation schemes can contribute to the improvement of the argument-based human-computer interaction paradigm. In this work, we perform a qualitative analysis of the persuasive properties of argumentation schemes. For that purpose, we present a new study conducted on a population of over one hundred participants, where twelve different argumentation schemes are instanced into four different topics of discussion considering both stances (i.e., in favour and against). Participants are asked to relate these argumentation schemes with the perceived Cialdini's principles of persuasion. From the results of our study, it is possible to conclude that some of the most commonly used patterns of reasoning in human communication have an underlying persuasive focus, regardless of how they are instanced in natural language argumentation (i.e., their stance, the domain, or their content).

8.1 Introduction

A fundamental component of human social interaction is the cognitive ability to reason on the basis of different arguments. Human argumentative reasoning facilitates the exchange of ideas, beliefs, or opinions among others, with the purpose of defending a position and/or convincing or persuading other people. Argumentative reasoning has, therefore, an influence on the capacity for judgement and decision making in human beings.

Over the years, from the field of argumentation theory [395], which encompasses different disciplines including philosophy, linguistics, and psychology, several efforts have been made to study, define, and structure human argumentative reasoning. One of the most prominent reasoning-based structural classifications of arguments is the one proposed under the argumentation scheme concept. Argumentation schemes are common rules or inference patterns that underlie argumentative reasoning and can be articulated and classified providing structure to arguments [400]. Each scheme consists of a set of premises, a conclusion, and a set of connections between the premises and the conclusion mostly considering the underlying logic, which makes them independent of the context and the argumentative domain. More than sixty different argumentation schemes have been proposed and identified in the literature, compiled by D. Walton together with an elaborated meta-classification of such schemes [400, 398]. The thorough research carried out in this direction has enabled computational and Artificial Intelligence (AI) researchers to have a much more structured reference of arguments and argumentation when designing computational argumentation systems: for argument mining [206, 396]; for decision support [293]; and for automated reasoning [347].

A fundamental aspect when defending a position and/or trying to convince someone, is the power or capacity of persuasion. In the field of psychology, several approaches have been adopted to identify the generally used human persuasion strategies. In this sense, R. B. Cialdini [90] provided a theory in which he defined six principles of persuasion: Reciprocity, Authority, Commitment, Liking, Social Proof, and Scarcity. These principles are specific to persuasion, but not to any context nor domain. Furthermore, this solid definition and classification of persuasive strategies provided a richer dimension for computational persuasion researchers in the proposal of new paradigms of human-computer interaction [261, 379, 263, 92].

However, not much attention has been focused on the relationship between the different persuasion strategies and the different patterns of human reasoning from a domain-agnostic viewpoint. Most of the identified argument-based persuasive research is applied to a very specific domain or within a well defined context, which hampers the interpretation and the generalisation of the presented results. Thus, establishing a relationship between argumentation schemes and Cialdini's six principles of persuasion can: (i) improve the actual understanding of the persuasive properties of argumentative reasoning, (ii) provide a domain independent approach for argument-based persuasion, and (iii) support the evolution of new computational argumentation systems by improving the interaction with human users (e.g., through better persuasion). For this purpose, we conduct a qualitative analysis of the persuasive properties of argumentation schemes. Using the formal structures for argumentation schemes proposed by D. Walton, we explore the relationship of twelve of the most commonly used argumentation schemes in human communication to each of the six principles of persuasion of Cialdini through an experiment conducted with one hundred participants. In the experiment, we combined four different topics of discussion along with arguments in favour and against, with the aim of generalising the results to extrapolate them to computational argumentation models.

8.2 Related Work

Persuasion is a key element in computational argumentation research. The study of the persuasive properties of arguments and their computational representations make possible to have a better understanding of how human users perceive them when used in competitive (e.g., debate tournament), or cooperative (e.g., decision-making assistance) environments. Furthermore, the use of argumentative structures instead of simple messages make possible to deepen the direct interactions between humans and computer systems. Arguments and argumentative reasoning seen from the computational viewpoint provide a solid framework for approaching the process of human reasoning [320], having a more natural human-computer interaction [179], and generating more reliable explanations of decisions made by computer systems [297]. However, the concept of persuasion and its interpretation

may differ substantially from one work to another in the area of computational argumentation. For instance, in [244], persuasion is understood as the effectiveness of arguments in reaching a satisfactory agreement during a negotiation. The authors propose a Reinforcement Learning [359] argumentative agent that is trained to find the best argument for each negotiation step based on an internal set of preferences. This way, the most persuasive expected argument is chosen to improve the effectiveness of argumentation.

A different perspective on argumentative persuasion is presented in [385], in which authors analyze the mental engagement, emotions, and persuasive power of the arguments uttered during a debate. In that approach, the persuasive power is represented with the three major persuasive strategies proposed in [303]: Ethos, Pathos and Logos. Each one of this three strategies tries to be persuasive by considering different aspects. Ethos appeals to vocabulary and social positioning, Pathos appeals to emotions, and Logos appeals to logical reason.

In recent years, research conducted in the field of computational argumentation refers to the establishment of metrics to estimate the persuasive power of arguments based on empirical evidence. For example, in [368], the development of a new scale to measure human perception on message/argument persuasion is presented. This scale relies on three major factors which serve as indicators of the perceived persuasiveness: the effectiveness, the quality, and the capability of the message. Authors conducted an experiment with humans in which these three factors were measured on five different messages belonging to the healthy eating domain. In the same way, the work performed in [323], presents a new metric to measure the persuasive power of arguments and reasoning patterns when they are used in a social network domain for privacy concerns. That metric was derived from the results of an experiment in which the impact of intrinsic human characteristics (i.e., personality traits and social interaction) when rating persuasive power was evaluated. Another interesting approach for evaluating the persuasiveness of arguments based on domain concerns under the framework of abstract argumentation is presented in [164]. From the results of that experiment, it is possible to observe that using human user preferences (i.e., ranking) over the domain concerns make possible to improve the persuasiveness of an argumentative system. These findings are integrated into an argumentative chatbot aimed at persuading human users [179]. In that proposal, persuasion is achieved through a combination of user

modelling (i.e., beliefs and preferences over concerns), and a dialogue engine in charge of creating the most persuasive dialogue strategy depending on each user model.

A different approach to empirical metrics comes from the Natural Language Processing (NLP) area of research. Based on the textual properties of natural language inputs, researchers model text persuasion as a downstream task to establish the level of persuasiveness of arguments. For example, in [151], the authors present a new corpus for determining the most persuasive argument from a given pair of arguments. A neural network architecture is used to undertake the task of predicting and modelling persuasion from a natural language input. Another interesting approach is the one proposed in [31] that focuses on the analysis of the impact of the style of news editorial arguments on their persuasive power. For that purpose, five different NLP features are used to model style: Linguistic Inquiry and Word Count [277], a lexicon of emotions (i.e, anger, disgust, and fear) and sentiments (i.e,. positive and negative) [243], argumentative discourse units features (i.e., anecdotal, statistical, and testimonial evidence) [7], arguing elements (i.e., assessments, doubt, authority, and emphasis) [346], and text subjectivity (i.e., subjective or objective) [404]. These features are used to train a Support Vector Machine (SVM) [378] over a task aimed at predicting if a message will be persuasive or not. Finally, the research conducted in [191] presents a combination of user and text modelling. The authors use users' beliefs, interests, and personality traits, along with NLP feature engineering on natural language inputs to predict persuasiveness of arguments and users' resistance to persuasion.

As it can be observed, most of the existing research on computational persuasion and computational argumentation present a quantitative approach to the task. Previous research tries to measure and quantify the persuasiveness of specific arguments when used in specific domains. Thus, these metrics are dependent on the domain or the type of arguments used. Another important thread running through all of the reviewed works is that the definition of an argument does not match between them. In each research work, the argument definition depends on the needs, the available data, and the objectives of the researchers. Thus, there are important differences between the proposed argument instances (e.g., short messages, complete sentences, opinion and evidence, etc.), which makes it difficult to generalise their findings from an argumentative viewpoint. Therefore, it is interesting to have

general evaluations and analyses of the persuasive properties of arguments, in order
to establish new computational models of argumentation aimed at interacting with
human users. This can be achieved through the evaluation of generic reasoning pat-
terns or structures that are used to build arguments regardless of their domain, their
content, or their stance. The problem was partially addressed in previous research,
where a relation between domain-specific messages and persuasive properties was
established [367]. However, the results can not be easily generalised, due to the do-
main in which the study was conducted, and the specificity of the used messages.
In this work, we have focused on deepening in this line of research by providing
domain independent and stronger results. A complete analysis of these lines of
related work is provided at the end of the following section, after defining the two
concepts on which our research relies: the Cialdini's principles of persuasion and
the argumentation schemes.

8.3 Background: Principles of Persuasion and Argumentation Schemes

In this section, we define two major concepts which are the pillars of our inves-
tigation, and that allow to overcome the identified limitations in computational
persuasion and argumentation research. First, Cialdini's principles of persuasion
are the six fundamental concepts on which persuasive strategies can be developed.
Second, Walton's argumentation schemes provide a structured framework for ar-
gumentation. Both concepts were proposed regardless of the domain, generalised
for the areas of persuasion and argumentation research respectively.

Cialdini's Principles of Persuasion

Over the years, different approaches have been introduced in the field of psychol-
ogy to provide a generalisation of human persuasion strategies. One of the most
significant contributions adopted in this field is the one provided by R.B. Cial-
dini [90]. Cialdini defines six principles of persuasion: *Reciprocity*, *Authority*,
Commitment, *Liking*, *Social Proof*, and *Scarcity*. These principles are specific to
persuasion, but not to any domain or argument. Therefore, a qualitative analysis

of argumentative persuasion considering these six principles can be relevant to a
more generalised understanding of the persuasive properties of arguments.

Argumentation Schemes

An argument is the expression of an idea or reasoning that attempts to prove, jus-
tify or refute a thesis. The general structure of an argument is composed of a
premise (or a set of premises) and a conclusion [222]. This structure must allow
the conclusion to be derived from the premises. Argument structures are gen-
erally constructed using commonly accepted rules or inference patterns. These
patterns can be articulated and classified through different general argumentation
schemes. Argumentation schemes were proposed as structured representations of
arguments depending on their underlying reasoning pattern [104]. Each argumen-
tation scheme consists of a set of premises, a conclusion, a definition of the re-
lationships between the premises and the conclusion, and a set of critical ques-
tions. Argumentation schemes allow a general classification of arguments regard-
ing their underlying logic [398]. Moreover, while the definition of an argument
can be fuzzy, argumentation schemes have a well-defined structure and allow for
good integration into computational argumentation systems. Furthermore, argu-
mentation schemes were defined considering only the underlying logic. Thus, they
can be used regardless of the argumentative domain. One of the theorists who
conducted a significant contribution on the identification and definition of argu-
mentation schemes was D. Walton [400]. Over the years, Walton identified over
sixty different argumentation schemes commonly used in human argumentation
which have been widely used by the research community to generate computa-
tional argumentation models encompassing different fields such as the automatic
identification of arguments in natural language text [206, 396] (i.e., Argument Min-
ing), the computational representation of argumentative structures [251, 293], and
the automatic evaluation of natural language argumentative sources [347]. Thus,
the main tasks belonging to computational argumentation research have benefited
from the definition of argumentation scheme structures since they provide a struc-
tured framework which makes easier to classify, represent, and evaluate arguments
depending on their underlying reasoning pattern.

Cialdini's Principles and Argumentation Schemes for Computational Persuasion

The two concepts introduced in this section provide: (i) a solid psychology-based theory on the principles governing human persuasion that do not depend on any hand-crafted metric, or a hard to interpret (e.g., black-box-based) estimation; and (ii) well-defined formal structures commonly found in human argumentative reasoning that do not rely on the topic, the content, or the stance of the own argument. Both of them provide the necessary tools to deepen on the understanding of argument-based human persuasion strategies. Not much research aimed at generalising findings on the persuasive properties (i.e., qualitative) of arguments has been identified. But we have been able to find a few preliminary works focused in this direction.

Regarding the Cialdini's principles of persuasion, in [263, 92], the authors conducted a study with human participants in order to understand how personality, gender, and age could be affecting the perception of these six principles when trying to be persuaded in a survey and a text-based game respectively. Similarly, in [261], a complete study is made regarding the effectiveness of the different principles of persuasion when used with humans. However, argumentation was not formally taken into account in none of these studies.

On the other hand, argument persuasion is highly influenced by the underlying reasoning used for the elaboration of arguments. Which means that argumentation schemes may represent an important aspect when studying the persuasive properties of arguments. However, the persuasive aspect of argumentation was not explicitly taken into account when defining the stereotyped patterns of human reasoning on which argumentation schemes rely. Previous research analysed the existing relations between argumentation scheme-based adapted messages and Cialdini's persuasive principles in the healthy eating domain [366, 367]. The authors conducted an experiment with 29 participants to analyse the correlation between argumentation schemes and Cialdini's principles. In that study, modified versions of Walton's argumentation schemes were used to simplify the definition of persuasive messages, and to ease the integration of the persuasive principles (e.g., the argumentation scheme of practical reasoning was adapted to a new version of practical reasoning with liking, integrating the liking principle of persuasion

into its formal premises and claim). These findings were integrated into a persuasive computational argumentation system to assist human users eat healthy [365]. However, the study and the experiments were carried out considering a unique domain and adapted versions of the original argumentation schemes, which makes harder to generalise their findings. In the present work, we have based in the related work's proposed methodology (i.e., [367] study) and extended the variety of discussion topics to make our results less domain-dependent. Furthermore, our approach relies on a more formal viewpoint of argumentation schemes, instead of using messages created specifically for our experiments or domains. Considering our approach, the implementation of argument-based persuasive systems and the definition of user-tailored persuasive strategies, would benefit from the qualitative analysis of the persuasive properties of (domain independent) argumentation schemes. This way, it will be possible to generalise the findings, and to create domain-specific arguments from general patterns of human reasoning (i.e., argumentation schemes) with knowledge of their specific persuasive properties.

8.4 Study Design

In this paper, we have created a study to bridge the gap between the areas of persuasion and argumentation theory research. For that purpose, we bring together the two concepts that allow us to do a qualitative and domain-independent analysis: Cialdini's principles of persuasion and argumentation schemes. Thus, we present a new study with the objective of overcoming the previously identified limitations, and to consolidate strong relationships between the argument-based persuasion and the underlying logic of argumentation. The study design was motivated by the search for the answer to four different research questions related to the persuasive properties of argumentation schemes: (**RQ1**) How do argumentation scheme reasoning structures relate to the Cialdini's principles of persuasion?; (**RQ2**) How does the topic in which argumentation schemes are instanced into natural language arguments influence on the human perception of persuasive principles?; (**RQ3**) How does the stance of argumentation schemes instanced into natural language arguments influence on the human perception of persuasive principles?; (**RQ4**) Do gender and/or age have an effect on human perception of persuasive principles in

arguments?.

Measures and Instruments

In our study, we considered twelve different argumentation schemes. The selection of this specific set of schemes was motivated both by previous related work [366] and a thorough analysis conducted by the five authors on the well-known compendium of argumentation schemes proposed by Walton in [400], focusing on those which are most frequently found in human communication. Taking this into consideration, the final set of schemes proposed for the present study are: *Argument from Popular Opinion* (AFPO), *Argument from Popular Practice* (AFPP), *Argument from Position to Know* (AFPK), *Argument from Expert Opinion* (AFEO), *Argument from Commitment* (AFCM), *Argumentation from Values* (AFVL), *Argument from Practical Reasoning* (AFPR), *Argument from Waste* (AFWS), *Argument from Sunk Costs* (AFSC), *Argument from Threat* (AFTH), *Argument from Cause to Effect* (AFCE), and *Argument from Rules* (AFRL) (see [400] for the formal definition of the argumentation schemes included in our study). From our selection of argumentation schemes, we can observe significant differences between the underlying logic used in their definitions. For example, AFPP and AFPO are both built taking socially popular and acceptable aspects as premises. But, if we look at the premises of AFPK and AFEO schemes, we can observe that they rely on someone's (i.e., informed or expert source) viewpoint or opinion. Diversely, the premises of an AFCM depend on a previous commitment of the arguer; AFVL's premises are built upon positive or negative judgement values; AFPR's and AFCE's premises rely on conditional logic that justifies the claim; the premises of AFWS and AFSK schemes are defined from a previously done effort or commitment which can be wasted or inconsistent if the claim is not accepted; AFTH's premises combine conditional logic with a threatening position that influence the claim of the scheme; and finally, the premises of the AFRL scheme mainly depend on a previously established rule which leads to the conclusion of the argument. Furthermore, in order to analyse the persuasive properties of argumentation schemes, the six Cialdini's principles of persuasion (i.e., reciprocity, authority, commitment, liking, social proof, and scarcity) were also included in the design of our study. This way, it was our goal to find any existing relation between the underlying logic of arguments

and its persuasive approach.

An important issue when analysing an argumentation scheme instanced into a natural language argument is the independence of its message. It is necessary to establish mechanisms to control the bias produced by the message of the argument. We have identified two types of biases that can be produced when instantiating an argumentation scheme: the bias produced by the content of the message and the bias produced by the topic of the message. To mitigate such biases, we have decided to use multiple natural language arguments instantiating each argumentation scheme. Thus, to find out the persuasive principles associated with the selected argumentation schemes, we created these arguments considering four different topics of current relevance: (**T1**) *Should COVID-19 Coronavirus vaccination be mandatory?*; (**T2**) *Should euthanasia be legalized?*; (**T3**) *Should you take care of your physical appearance to achieve personal and professional success?*; and (**T4**) *Should you do intermittent fasting to lose weight?*

We selected two more controversial discussion topics (i.e., **T1** and **T2**) where people tend to be more polarised, and two more neutral discussion topics (i.e., **T3** and **T4**). Furthermore, we have created two stances (in favour and against) for each of the twelve natural language arguments, representing the argumentation schemes in our four topics. Therefore, we have generated a total of ninety-six arguments instantiating twelve argumentation schemes to perform our experiment. Note that we have designed the premises and conclusions of each natural language argument according to the original argumentation schemes' structures.

We put together all these concepts in a unique questionnaire aimed at measuring the relation between argumentation schemes and Cialdini's principles of persuasion. For that purpose, our questionnaire was structured into six stages (Figure 8.1). *Stage 0* was designed for registration, the participants must indicate their identification number in order to be able to keep an individual tracking of all of them and to retrieve their personal features (i.e., age and gender). In *Stage 1*, a description of the subject of the study along with the instructions that the participants had to follow to complete the experiment was provided. In the task description, we provided a brief introduction to argumentation schemes and Cialdini's principles as well as a definition of each principle of persuasion. The remaining four stages were the core part of our questionnaire, where a total amount of a hundred and two questions were distributed along our four different topics (one topic per
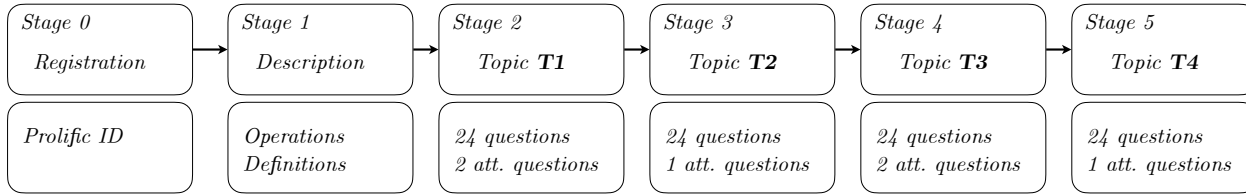
Figure 8.1: Stages of the experiment. Note that *att.* refers to attention check questions

stage). From the total number of items, our questionnaire was composed of ninety-six questions regarding our twelve argumentation schemes in favour and against each topic (i.e., twenty-four per topic), and six attention check questions (created following the guidelines of the online recruitment platform) randomly distributed along the stages of the study. The questions were distributed as follows: *Stage 2* consisted of twenty-four questions related to the topic **T1** along with two attention check questions; *Stage 3* included twenty-four questions related to the topic **T2** together with one attention check question; *Stage 4* contained twenty-four questions related to the topic **T3** along with two attention check questions; and *Stage 5* consisted of twenty-four questions related to the topic **T4** together with one attention check question.

To validate our findings, we used the *Free-Marginal Multirater Kappa* (K_{free}) inter-annotator agreement score [302]. With this metric, it is possible to understand how strongly are perceived the relations between the argumentation schemes and the principles of persuasion, by means of the observed agreement between the participants of our study. Furthermore, the K_{free} was originally proposed as an improved agreement score for statistical studies similar to the one conducted in our research [367], with measurements done in the nominal scale (i.e., principles of persuasion) and more than two different "annotators" (i.e., participants).

Finally, to study the influence of the variables sex and age on the selection of principles of persuasion for each argumentation scheme, we used Pearson's chi-squared statistic (χ^2) [342]. Pearson's chi-squared statistic is a non-parametric test designed to measure differences between groups when using categorical variables. That test can be used both to measure the "goodness of fit" (i.e., the level of disagreement between an observed and a theoretical distribution) and to measure the independence of two variables by the use of contingency tables.

Context of the Study and Participants

Data was collected through the online platform Prolific [264]. The Prolific plat-
form manages recruitment, payment, and personal information of participants such
as age or gender. In addition, it offers several tools to customize experiments, in-
cluding a set of filters to select the suitable participants for an experiment. If par-
ticipants successfully complete the experiment they are paid, otherwise they are
not paid and they get a penalty that will negatively affect their eligibility for future
experiments.

In our experiment we established a filter to recruit participants whose first lan-
guage was Spanish. Furthermore, we also required our participants to have at least
completed fifty Prolific questionnaires, and to have a minimum of a 90% accep-
tance ratio from past questionnaires in Prolific. The average reward per hour for
the participants in the experiment was £8.34.

One hundred and seventeen participants completed the experiment. Seventeen
of these participants were excluded for failing one or more attention check ques-
tions. The remaining one hundred participants were distributed as 44 women and
56 men ranging in age between 19 and 62 years old ($\mu = 30.9$, $\sigma = 12.0$). The
age distribution in the different quartiles were: quartile Q1 $= 22.75$, quartile Q2
$= 26.00$, and quartile Q3 $= 38.25$ years old.

Procedure

At the beginning of the experiment the instructions were presented to the partici-
pants along with an explanation of the purpose of the experiment, a brief introduc-
tion to the principles of persuasion, and a description of each of the six principles
of persuasion of Cialdini. The description of each persuasion principle of Cialdini
showed a short definition with the general idea of the principle and a longer defi-
nition that included some examples. Once the instructions of the experiment were
shown, the questionnaire was displayed.

In each block the questions were displayed randomly. Each question showed
an argument resulting from instantiating one of the twelve argumentation schemes
along with seven possible options (see Figure 8.2).

Figure 8.2: Experiment layout

In each question, participants were asked to classify an argument into one of the six Cialdini principles by selecting one of the six possible options or to select "other" in case the participant considers that the argument does not correspond to any of the six principles of persuasion.

Attention check questions followed the same pattern as the rest of the questions: an argument was shown along with the six options to select the persuasion principle. However, in these questions, the answer that participants had to select was indicated as part of the text of the argument.

8.5   Results

We divided the analysis of the study results into four sections: first, we measured the relation between the twelve argumentation schemes and the six Cialdini's principles of persuasion by means of the K_{free} agreement on the perceived principles by the participants; second, we measure the influence of the four topics (i.e., **T1**, **T2**, **T3** and **T4**) and the two stances (i.e., in favour or against) on the Cialdini's principles selection; third, we analysed the dependency of the gender and the age of our participants on the observed results; and, finally, we analysed the non-related argumentation schemes to better understand the main reasons why no strong agreement is found on the relation between these arguments and the persua-

sive principles.

The Principles of Persuasion of Argumentation Schemes

From the results of our study, we have been able to estimate the existing relationships between argumentation schemes and Cialdini's principles of persuasion, as perceived by an heterogeneous set of participants in an heterogeneous set of argumentation domains. Thus, we present a qualitative analysis of the persuasive properties of argumentative reasoning that can be easily generalized to other domains. Our first step for identifying and validating such relations has been to aggregate all the observations and calculate the K_{free} agreement score considering the complete population of our study. The K_{free} metric is interpreted as a reasonable agreement if $0.4 \leq K_{free} \leq 0.7$ and, a strong agreement, if $0.7 < K_{free}$. Therefore, we have only considered a valid relation between an argumentation scheme and a persuasive principle if the agreement is, at least, above the minimum of a reasonable agreement (i.e., $0.4 \leq K_{free}$).

As depicted in Table 8.1, it has been possible to identify six relationships between argumentation schemes and persuasive principles. In the relations found with a moderate and strong agreement (K_{free} scores highlighted in bold in Table 8.1), the number of observations captured in our study are mainly concentrated in a unique principle (Cialdini's principles percentages highlighted in bold in Table 8.1). This analysis indicates us that a specific argumentation scheme is approaching human persuasion considering a specific persuasive principle by its own underlying logic, rather than its domain-related argument instances (i.e., natural language content). Thus, regarding the research question **RQ1**, we observed that AFPO and AFPP are related to the Social Proof Cialdini's principle; AFPK and AFEO are related to the Authority principle of persuasion; and AFCM and AFSC are related to the Commitment principle of persuasion. According to experiment results, the remaining arguments considered in our study do not appear to be related by definition to a specific principle of persuasion.

Table 8.1: Relationship between arguments and principles of persuasion. The first columns show the percentage of participants that chose each option for each argument. The last column shows the resulting *Free-Marginal Multirater Kappa* (K_{free}).

Argumentation	Cialdini's principles							
Scheme	Reciprocity	Authority	Commitment	Liking	Social Proof	Scarcity	Other	K_{free}
AFPO	5.375	3.25	4.5	7.5	**75.0**	1.75	2.625	**0.513**
AFPP	4.0	9.0	5.125	11.75	**67.375**	1.125	1.625	**0.444**
AFPK	2.5	**77.75**	2.625	5.875	9.5	0.875	0.875	**0.573**
AFEO	0.5	**88.25**	1.75	5.75	2.75	0.375	0.625	**0.769**
AFCM	9.625	3.0	**70.625**	7.625	4.375	2.0	2.75	**0.477**
AFVL	15.5	4.875	43.5	15.125	6.5	6.75	7.75	0.150
AFPR	21.125	8.0	19.625	20.875	10.75	6.125	13.5	0.087
AFWS	23.875	5.625	21.75	14.0	7.75	22.75	4.25	0.112
AFSC	5.75	2.25	**75.5**	7.5	4.25	3.5	1.25	**0.535**
AFTH	32.625	7.875	8.5	13.75	20.25	10.375	6.625	0.131
AFCE	15.25	5.875	20.75	15.625	14.75	14.875	12.875	0.116
AFRL	6.75	44.125	20.625	4.75	13.75	4.125	5.875	0.192

The Impact of the Content on the Perceived Principles of Persuasion

To measure the impact of the content of each argumentation scheme on the human perception of the persuasive principles, we performed an analysis similar to the one described in the previous section but considering each topic (i.e., **T1**, **T2**, **T3** and **T4**) independently (**RQ2**). Furthermore, we also analysed how did the stance (i.e., in favour or against) influence our participants' perception in each specific topic (**RQ3**).

While no significant variations have been observed on the agreement scores of the argumentation schemes that were not related to a specific persuasive principle (i.e., AFVL, AFPR, AFWS, AFTH, AFCE, and AFRL), we observed interesting agreement variations on the previously identified relations depending on the topic in which each argumentation scheme was instanced. Table 8.2 condenses our findings regarding the impact of the content on the perceived persuasive principles in our study. The percentages of the non-related argumentation schemes have not been included in the table since there is not a unique principle monopolising the

participants' perceived relation between argumentation schemes and persuasive principles. Topics **T2** and **T4** present the weaker agreement scores on the participants' perception of relations compared to topics **T1**, and **T3** which has the most solid K_{free} agreement (K_{free} moderate and strong agreements are highlighted in bold). Furthermore, if we look at the stance, it is possible to observe two different situations. First, there is the case where the percentages of the participants' selections are balanced regardless of the stance. For example, in our first topic **T1**, AFEO were related to the Cialdini's principle of Authority a 96% of the times for the arguments in favour, and a 90% of the times for the arguments against. Similarly, in our second topic **T2**, the agreement for AFPP related with the Social Proof principle experienced a significant decrease, but the percentages between both stances were still balanced with a 66% for the in favour argument version and a 55% for the against version of the argument. In these situations, it is not possible to conclude that the stance is having any relevant impact on the perception of persuasive principles, but we can infer that the content of the argument regarding each topic is the main cause of these variations. Second, there is the case where the variations of the K_{free} agreement score within a specific topic is associated with a huge drop in a unique stance of the argument instance. For instance, in our fourth topic **T4**, it is possible to observe a decrease of the agreement on the relation between AFPP and the Social Proof principle, perceived only by a 31% of the participants for the argument in favour, but by the 73% of the participants for its version against. Likewise, in the **T2** topic, the perceived relation between AFCM and the persuasive principle of Commitment shows an important decrease of the K_{free} agreement, where only a 34% of the participants selected this principle for the argument in favour. These situations represent a strong evidence of the influence of the stance on the human perception of persuasive properties of arguments. Therefore, from the analysis of the results of our study, it can be concluded that both the topic and the stance of argumentation schemes, instanced into a natural language argument may have a significant influence on the human perception of the persuasive principles related to them.

Table 8.2: Relationship between arguments and principles of persuasion disaggregated by topics and stances. For each stance, the selection percentages of participants for each related principle are depicted. For each topic, the *Free-Marginal Multirater Kappa* (K_{free}) is indicated independently.

Argumentation	T1			T2			T3			T4			Cialdini's
Scheme	T1-F	T1-A	K_{free}	T2-F	T2-A	K_{free}	T3-F	T3-A	K_{free}	T4-F	T4-A	K_{free}	Principle
AFPO	79.0	59.0	**0.419**	87.0	81.0	**0.664**	78.0	68.0	**0.478**	80.0	68.0	**0.491**	*Social Proof*
AFPP	76.0	70.0	**0.474**	66.0	55.0	0.300	86.0	82.0	**0.662**	31.0	73.0	0.340	*Social Proof*
AFPK	68.0	51.0	0.294	86.0	83.0	**0.672**	78.0	81.0	**0.596**	91.0	84.0	**0.731**	*Authority*
AFEO	96.0	90.0	**0.844**	95.0	64.0	**0.621**	95.0	95.0	**0.886**	75.0	96.0	**0.725**	*Authority*
AFCM	53.0	81.0	**0.417**	34.0	71.0	0.256	78.0	84.0	**0.611**	78.0	86.0	**0.625**	*Commitment*
AFVL	-	-	0.231	-	-	0.057	-	-	0.252	-	-	0.062	
AFPR	-	-	0.047	-	-	0.041	-	-	0.193	-	-	0.067	
AFWS	-	-	0.120	-	-	0.104	-	-	0.115	-	-	0.109	
AFSC	83.0	83.0	**0.642**	71.0	78.0	**0.495**	85.0	95.0	**0.721**	47.0	68.0	0.281	*Commitment*
AFTH	-	-	0.071	-	-	0.140	-	-	0.159	-	-	0.153	
AFCE	-	-	0.216	-	-	0.070	-	-	0.130	-	-	0.049	
AFRL	-	-	0.363	-	-	0.085	-	-	0.196	-	-	0.124	

The Impact of the Gender and Age on the Perceived Principles of Persuasion

Human intrinsic characteristics, such as gender or age, can also have an influence on the understanding, perception, and the ability to relate argumentation schemes and principles of persuasion. To study the effect of these intrinsic factors on the selection of the principles of persuasion, we performed a comparative statistical analysis using the chi-squared test.

The Impact of the Gender

For gender, we proceed from the experiment sample with a gender distribution of 44 women and 56 men and we analyzed the dependence of the gender variable, with two options (i.e., female and male), and the principle of persuasion variable with seven possible options (i.e., the six principles of persuasion and the option "other"). Considering these two variables, we defined the null hypothesis (h_0) as "gender and principle of persuasion variables are independent" and the alternative hypothesis (h_1) as "gender and principle of persuasion have some degree of dependence".

Table 8.3: Results for the Chi-Squared test for the variables principle of persuasion, gender, and sex. And the K_{free} values for the age and gender clusters. The theoretical value for the gender χ^2 test with a level of risk of 5% and six degrees of freedom was $\chi^2_{0.05,6} = 12.592$. For the age χ^2 test, the theoretical value with a level of risk of 5% and eighteen degrees of freedom was $\chi^2_{0.05,18} = 28.869$.

Argumentation Scheme	Chi-Squared Test				K_{free} Test						Cialdini's Principle
	Gender		Age		Gender		Age				
	χ^2 value	p-value	χ^2 value	p-value	Female	Male	C1	C2	C3	C4	
AFPO	**15.532**	**0.016**	23.098	0.187	**0.571**	**0.468**	**0.482**	**0.545**	**0.506**	**0.510**	*Social Proof*
AFPP	**17.734**	**0.007**	**45.774**	**0.000**	**0.513**	0.395	**0.452**	**0.474**	**0.468**	**0.405**	*Social Proof*
AFPK	11.205	0.082	22.936	0.193	**0.585**	**0.562**	**0.544**	**0.621**	**0.576**	**0.552**	*Authority*
AFEO	10.711	0.098	**30.066**	**0.037**	**0.816**	**0.732**	**0.756**	**0.847**	**0.741**	**0.752**	*Authority*
AFCM	8.724	0.190	**58.289**	**0.000**	**0.550**	**0.422**	**0.480**	**0.503**	**0.491**	**0.431**	*Commitement*
AFVL	4.563	0.601	**33.042**	**0.016**	0.146	0.155	0.190	0.126	0.098	0.186	-
AFPR	4.12	0.660	**35.674**	**0.008**	0.107	0.067	0.099	0.090	0.086	0.076	-
AFWS	**19.546**	**0.003**	**32.832**	**0.017**	0.135	0.105	0.134	0.105	0.083	0.135	-
AFSC	**22.243**	**0.001**	**36.556**	**0.006**	**0.625**	**0.467**	**0.450**	**0.603**	**0.536**	**0.547**	*Commitement*
AFTH	**14.162**	**0.028**	**42.115**	**0.001**	0.114	0.143	0.156	0.096	0.137	0.139	-
AFCE	5.05	0.537	25.773	0.105	0.110	0.121	0.149	0.081	0.136	0.101	-
AFRL	**20.609**	**0.002**	24.484	0.140	0.182	0.201	0.193	0.191	0.187	0.188	-

Table 8.3 shows the results of the performed chi-squared test (significant results highlighted in bold i.e., $p \leq 0.05$). The theoretical value with a level of risk of 5% and the 6 degrees of freedom (i.e., seven possible answers and two possible genders) for gender and principle of persuasion was $\chi^2_{0.05,6} = 12.592$. Therefore, concerning the research question **RQ4**, with a confidence level of 95% we rejected the null hypothesis h_0, i.e. there seems to be a certain degree of dependence on gender in the selection of the persuasion principle, in argumentation schemes: AFPO, AFPP, AFWS, AFSC, AFTH, and AFRL. For the rest of the argumentation schemes, the null hypothesis h_0 cannot be rejected. Therefore, for those arguments we considered that there is no dependence on gender when selecting the principle of persuasion.

The differences between both genders could also be appreciated in the K_{free} agreement (moderate and strong agreements highlighted in bold in Table 8.3). For example, for AFPO, the K_{free} agreement for the female gender group was 0.513 (greater than 0.4) while for the male gender group it was 0.395 (less than 0.4). In addition, it appears that K_{free} agreement was generally higher in the female group than in the male group. Despite these existing differences in the selection of persuasion principles according to gender for some argumentation schemes, in

both groups the predominant persuasion principles selected for each argumentation scheme was the same for those argumentation schemes in which the K_{free} agreement was greater than 0.4 (see Table 8.1).

The Impact of the Age

As described in Section 8.4, the distribution of quartiles for the age variable was: quartile Q1 $= 22.75$, quartile Q2 $= 26.00$, and quartile Q3 $= 38.25$. According to the distribution of these quartiles, we classified the participants into four balanced clusters. We performed a chi-squared test considering the variables age (separated into the four clusters) and persuasion principle. Similarly to the previous case, we set the null hypothesis (h_0) as "age and principle of persuasion variables are independent" and the alternative hypothesis (h_1) as "age and principle of persuasion have some degree of dependence".

For this chi-squared test, the theoretical value with a level of risk of 5% and the 18 degrees of freedom (i.e., seven possible answers and four age clusters) was $\chi^2_{0.05,18} = 28.869$. Thus, regarding the research question **RQ4**, with a confidence level of 95% we can reject the null hypothesis h_0, i.e. there seems to be a certain degree of dependence on age in the selection of the persuasion principle, in argumentation schemes (see Table 8.3): AFPP, AFEO, AFCM, AFVL, AFPR, AFWS, AFSC, and AFTH. For the remaining four, we cannot reject the null hypothesis h_0. Therefore, those argumentation schemes did not appear to be age-dependence in selecting the principle of persuasion.

As in the case of gender, in all four age groups, the predominant persuasion principle chosen for each argumentation scheme was the same for those argumentation schemes in which the K_{free} agreement was greater than 0.4 in Table8.1. In this case, the differences in K_{free} agreement value were not as evident as in the case of gender. Although, we found that the cluster with the highest level of agreement was the *C2* cluster and the lowest agreement was in cluster *C4*, probably because of the dispersion of age in cluster *C4* (i.e., 6.05) was higher than for the other clusters.

An Analysis of the Non-Related Argumentation Schemes

Finally, we analyse the six argumentation schemes that have not reached a reasonable agreement (i.e., $0.4 \leq K_{free}$) among our participants on the perceived persuasive principle. For that purpose, we will look at the central columns of Table 8.1, which depict the percentage of the selection of the persuasive principles related to each argumentation scheme in our study. From the set of six argumentation schemes that are not related to a specific principle of persuasion by their underlying logic, we have identified two ruling patterns. Three of these argumentation schemes have a dominant Cialdini's principle of persuasion associated with them, even though no agreement has been observed: in a 43.5% of the selections of our participants the Commitment principle was associated to the AFVL; AFTH have been related to the Reciprocity principle in a 32.6% of the cases; and a 44.1% of the study responses imply an association of the AFRL with the Authority principle. On the other hand, the remaining three argumentation schemes are not dominated by any specific principle of persuasion: AFPR and AFCE have been related to the principles of Reciprocity, Commitment and Liking in the 15-21% of the cases; and AFWS were associated with the Reciprocity, Commitment, and Scarcity principles in the 21-23% of the cases.

Two major conclusions can be drawn from this last section of the analysis of the results. Some argumentation schemes (e.g., AFVL, AFTH, and AFRL) present weak relations with a Cialdini's principle, but the weight of their underlying logic on these relations is not enough compared to other influences such as the content of the natural language argument instances. Other argumentation schemes (e.g., AFPR, AFCE, and AFWS) might not be related to any principle of persuasion by their underlying logic. In these cases, the weight of the persuasiveness of an argument relies almost completely on *external* elements such as their natural language content or stance.

8.6 Discussion

At the beginning of this work, we emphasised the importance of having a solid knowledge of the persuasive properties of arguments for designing and implementing new approaches on argument-based human-computer interaction. This

knowledge makes possible to improve the interactions made by computer systems by being more natural, effective, and user-friendly. However, one of the main limitations identified along the reviewed research on argument persuasiveness is the lack of easy-to-generalise results, and a strong focus on quantitative approaches that are usually constrained by the application domain. Aimed at making a contribution to these identified limitations, we carried out a qualitative analysis of the persuasive properties of argumentation schemes. This way, we research into *how* are arguments trying to persuade human users rather than quantifying their abstract *persuasive power*. For that purpose, we raised four different research questions that have been answered throughout this paper. First, we identified the existing relations between twelve different argumentation schemes and the six principles of human persuasion (**RQ1**), Table 8.4 summarises this findings. We also carried out an analysis of how did the content of the argumentation scheme (i.e., how the reasoning pattern is instanced with natural language text) influence the perceived persuasive strategy (**RQ2** and **RQ3**). Thus, we found out that the perceived persuasive properties of arguments are quite sensitive to these aspects, and that the process of instancing reasoning patterns with natural language text is a delicate process that must be done correctly in order to keep its properties. Finally, in this work we also analysed how did intrinsic human features (i.e., the gender and the age) influence the perceived persuasive principles of arguments (**RQ4**). Even though we were able to discover that a certain degree of dependence could exist between these features and the human perception of some of the argumentation schemes, there were no significant variations on the principles of persuasion related to these argumentation schemes. However, persuasion is a hard to understand concept, that is not universal, and that may suffer variations from one human user to another. Several features can influence the perceived persuasion of arguments such as the age, the personality, the emotions, or the social context among others. Previous research has explored the impact of features such as the personality, the gender, or the age in message susceptibility [263, 92]. From the results of these studies it is possible to understand which principle of persuasion is more or less effective for different user models composed of the previously mentioned features. These user models can be easily represented by computational argumentation systems (e.g., [324]) that can bring into consideration these findings together with the results of this work to define better strategies and improve their persuasive capabilities.

Table 8.4: Argumentation schemes' principles of persuasion. Cialdini's principles
with an asterisk (*) indicate weak findings that might be highly influenced by the
natural language instance (i.e., topic and/or stance) of the argumentation scheme.

Argumentation Scheme	Cialdini's Principle of Persuasion
AFPK, AFEO	Authority
AFCM, AFSC	Commitment
AFPO, AFPP	Social Proof
AFVL	*Commitment**
AFTH	*Reciprocity**
AFRL	*Authority-Commitment**
AFPR, AFWS, AFCE	None

Previous research in computational persuasion with Cialdini's principles-based
messages enable the investigation of a new dimension to the results presented in
this work. For instance, in work [367], the authors point out that humans found
the Authority principle the most persuasive, and the Liking principle the least in
the healthy eating domain. A different study states that the Commitment princi-
ple is also strong in this domain [261]. Furthermore, research on argumentation
schemes' persuasive power for teenagers in the privacy domain [323] pointed out
that AFCQ and AFEO were the most persuasive, while AFPP and AFPO the least
persuasive. As we can observe, some of these findings coincide, and make possible
to draw stronger conclusions. The aggregation of the findings in previous research
with the new results presented in this work make possible to tailor computational
argumentation systems to have a better engage in human interaction. For example,
with the knowledge of that Authority and Commitment principles performed well
in the health domain, it would be interesting to design a system able to interact
with human users through AFPK or AFSC. Thus, all these research and empirical
results may lead to new formalisations of argument-based computational models
for human persuasion.

8.7 Conclusion

In this paper, we present a novel qualitative analysis of the persuasive properties of twelve different argumentation schemes commonly used in human reasoning for its use in computational argumentation models. Throughout the observed results of our study, it has been possible to identify six different relations between arguments' underlying logic and Cialdini's principles of persuasion. Furthermore, instead of presenting specific domain-based results, the way our study and analysis is defined allows us to generalise our findings to any domain, topic or context. This is possible thanks to the qualitative nature of our research, which does not tell us which argument is *more or less* persuasive, but *how* does each reasoning pattern try to persuade in interactions with human users. We also explored how did parameters such as the age, the gender of our participants, the topic, and the stance of the considered arguments can influence the way humans perceive persuasive arguments. All these observations can be of utmost importance for the definition of new argument-based human-computer interactive systems, where humans need to be persuaded (e.g., decision support systems, intelligent assistants, etc.). However, it remains future work to formally integrate the findings of this study into a computational model of argumentation. It will also be an interesting future line of research to deeply explore further human dimensions that may influence the perceived persuasion of arguments such as the mental state, the emotions, or the personality.

Toward the Prevention of Privacy Threats: How Can We Persuade Our Social Network Platform Users?

RAMON RUIZ-DOLZ, JOSE ALEMANY, STELLA HERAS AND ANA GARCÍA-FORNES

Abstract

Complex decision-making problems such as the privacy policy selection when sharing content in online social networks can significantly benefit from artificial intelligence systems. With the use of Computational Argumentation, it is possible to persuade human users to modify their initial decisions to avoid potential privacy threats and violations. In this paper, we present a study performed over 186 teenage users aimed at analysing their behaviour when we try to persuade them to modify the publication of sensitive content in Online Social Networks (OSN) with different arguments. The results of the study revealed that the personality traits and the social interaction data (e.g., number of comments, friends, and likes) of our participants were significantly correlated with the persuasive power of the arguments. Therefore, these sets of features can be used to model OSN users, and to estimate the persuasive power of different arguments when used in human-computer interactions. The findings presented in this paper are helpful for personalising decision support systems aimed at educating and preventing privacy violations in OSNs using arguments.

9.1 Introduction

Deciding which privacy policy is the best when making a publication in an Online Social Network (OSN) is not an easy task for human users because it requires to take multiple factors into account (i.e., the potential receivers, the information to be shared, the users' preferences, etc.). In many situations, the information regarding those factors can be incomplete or unknown, as the reachability of the publication or other users' preferences. Another relevant feature that characterises online communication is that, once the content is published online, it can be downloaded and stored by anyone with access to it. Therefore, it is important to make sure that the content published does not cause any future privacy issues. Additionally, if more than one user appears in the publication, it is even easier to violate any privacy preference of the rest of the users involved, leading to privacy conflicts between users. The multi-party privacy conflicts [358] are a common type of privacy threats happening in OSNs. This problem combined with the great increase of users in OSNs, mostly teenagers who are initiating in their usage and have limited abilities for self-regulation and complex decision-making [294], has raised the interest of privacy management assistance research.

A natural way to approach the existing privacy management problem in OSNs is with the use of Computational Argumentation [320]. Computational Argumentation research investigates how the human argumentative reasoning process can be approached from a computational viewpoint. Using Computational Argumentation techniques, it is possible to establish an argument-based human-computer interaction. This approach can be seen as a direct improvement of recommendation technologies [87] since added to the recommendation, a justification (i.e., an argument) is also provided to the user. An effective way to avoid and reduce the number of potential privacy threats (i.e., disclosure of sensitive information) is to persuade the author to adapt the initial privacy configuration since it may be harmful to him/her or any of the other users involved. The best way to persuade the author is by making him/her understand the reasons why the privacy threat is happening with the use of arguments. Using different messages and warnings make possible to persuade OSN users to modify their initial decisions [10]. However, the perceived persuasive power of these messages may vary from one message to another [367] or even from different representations or structures of

these messages [369]. In the OSN privacy domain, these persuasive messages may approach different privacy aspects. Based on the previous definition given in [324], up to four different types of argument might be considered depending on the source from which they can be supported: *Privacy*, *Content*, *Risk*, and *Trust*. Furthermore, arguments can be represented and structured according to different reasoning patterns. Argumentation schemes group the most common patterns of human reasoning [398].

In this paper, we study the persuasive power of different argumentation schemes and argument types when used to educate teenagers on privacy management in an OSN environment. In our study, we consider two different user models that help us to encode human behavioural data into a computational system: the personality and the social interaction data. Previous work such as [406] show how users' personality can be a key factor when directly interacting with them. Therefore, we have investigated any potential existing correlations between personality traits and arguments. Additionally, since obtaining those traits may not be possible in some social network environments and behaviour is usually influenced by personality [153, 47], a study of the correlation between the most common social interaction features with the persuasive power of arguments has also been performed. Therefore, the main contributions presented in this work are the following:

- Quantify and measure the persuasive power of arguments used as a privacy threat prevention mechanism.

- Analyse the existing significant correlations between the persuasive power of arguments and the Big Five personality traits.

- Analyse the existing significant correlations between the persuasive power of arguments and thirteen different social interaction features.

All these key findings are contextualised in the OSN educational domain with teenager participants. This is one of the most important target populations when working on this domain since they are very active and easy to convince to share their personal information.

The rest of the paper is structured as follows. Section 9.2 reviews the most relevant work regarding privacy management and argumentation in the OSNs do-

main. Section 9.3 introduces the background of our research, proposes the research questions, and presents the design of the study carried out in this work. Section 9.4 describes the observed results and analyses their implications and interpretation. Finally, Section 9.5 summarises the most important conclusions reached at the end of this paper and the future research directions.

9.2 Related Work

Multiple approaches have been considered in the literature regarding the problem of finding optimal privacy policies in OSNs aimed at avoiding potential privacy violations [13]. A collaborative privacy preserving tool proposed in [331]. This system allows to provide recommendations to users that do not endanger their privacy. In [19] the authors propose an algorithm for predicting and preventing privacy violations in OSNs. This system detects a potential privacy violation and warns the involved users to prevent further damage to the parties involved. A different approach to the privacy problem was recently introduced in [389], where the authors propose an algorithm that combines user features such as the age or gender with the trust between users to determine the risk of sharing a publication in an OSN. Another collaborative approach to provide privacy recommendations to users is proposed in [349]. The authors propose *CoPE*, a collaborative privacy management system where each user can decide a specific privacy configuration for each publication. The system decides the best policy considering the most voted configuration. Finally, some of the existing automated privacy management systems that rely on an internal negotiation process. For example, *PriArg* [195] is a multi-agent algorithm which has an underlying negotiation protocol to compute the best privacy configuration for a specific situation. In *PriArg*, the negotiation is approached with argumentation. The agents represent real social network users that have an ontology with information from the network, the relationships between users and the content being published. Considering all these data, each agent can generate arguments to achieve a deal trying to satisfy the user privacy preferences. Images were also brought into consideration in [200], where an autonomous agent uses the tags and image features to prevent privacy violations in OSNs. There are some common weak points in all these privacy management systems. All of them

are focused on privacy conflicts where multiple users are involved in the same
publication. But the case of a user choosing a dangerous configuration for itself
is not considered. There is also an important limitation if we seek to provide the
user with an explanation of why the configuration should be changed. None of the
analysed privacy management tools gives the user a reasoned explanation nor tries
to persuade him/her. A recent explainable approach was proposed in [249, 250],
but it was focused exclusively on collective privacy violations.

When trying to reach an agreement, explaining our viewpoint, or trying to con-
vince another person, it is very common to make use of arguments. An argument is
defined as a set of propositions that can support the veracity of the main statement
(the conclusion). Thus, with arguments, it is possible to provide a set of coherent
reasons supporting some specific idea. Therefore, the use of Computational Argu-
mentation can be seen as the natural way to approach a decision-making problem
in which a human user must be persuaded. In [164], it is possible to observe the
relevance of analysing the persuasive power of arguments and user preferences,
when developing decision making assistance AI systems. Several works using ar-
gumentation in the OSNs domain can be identified in the literature. As described
in [360], argumentation in OSNs can be very useful, such as enhancing dialogues
or helping to structure user opinions. It is also possible to use Computational Argu-
mentation techniques to model the dialogue between different users sharing their
preferences in an OSN, and to persuade students to use specific learning objects
in an educational environment [168]. Therefore, as [142] proposes, argumenta-
tion seems the most coherent way to approach a persuasion problem framed in an
educational context in OSNs. In [195], an argumentation protocol to define the
best privacy policy when a multi-party privacy dispute is triggered is proposed.
However, not many works in which all the topics of our research converge (i.e.,
privacy management, Computational Argumentation, and human user persuasion)
have been identified. In addition to the main flaws identified before, the existing
related work in argumentation in social networks is mainly focused on studying
the multi-party privacy conflicts too [195, 250]. As described in [402], it is very
common to find users regretting their own publications in OSNs. Since we are
focused on an educational domain, we need a system that considers not only pri-
vacy disputes between different users involved in the same publication but also
potential self-privacy violations. When defining an argument, several parameters

should be considered (e.g., the content, the reasoning pattern, the language, etc.) to maximise its persuasive power. The reasoning pattern of an argument is defined by the underlying logic of its elements. Argumentation schemes were conceived as common patterns on human reasoning. In [398], up to sixty generally accepted argumentation schemes that can be found in common dialogues have been identified. Therefore, the use of argumentation schemes is a convenient way to define the reasoning patterns of the arguments of our study.

Finally, persuasion plays a major role on the effectiveness of arguments when used in a dialogue. Different users may perceive arguments differently, so it is very important to be able to understand, and to adapt our arguments to each user if we want an effective human-computer interaction. In [365], an argumentative system to make users change their behaviour in the healthy eating domain is proposed. The persuasiveness evaluation of the semi-automatic generated arguments is described in [369]. Furthermore, a study of the impact of personality, age, and gender on message type susceptibility [92] has also been done. Considering these works altogether, it is possible to infer relations between elements like personality and effectiveness of argumentation schemes. Finally, in [329], the authors explore the persuasive principles underlying some of the most common patterns of human argumentative reasoning (i.e., argumentation schemes). However, to the best of our knowledge, no one has directly analysed the persuasive power of arguments on teenagers, but behaviour may differ substantially between a teenager and an adult in the OSNs domain [88].

Therefore, with this paper, we put together these three research topics and present new results which will are helpful to push forward all the identified limitations on these topics: (i) with our arguments, we consider both self-disclosure and multi-party privacy conflicts; (ii) we approach the privacy management assistance problem from a more explainable and educational perspective; and (iii) we study teenager persuasion with arguments in OSNs, which has not been analysed in the literature yet. Our study results provide a new perspective on human (i.e., teenager) persuasion in the privacy management domain. We propose different, but related, user models based on two human aspects which we use to analyse the persuasive power of arguments: the personality and the social interactions. This way, it is possible to optimise the chosen argument by the privacy management assistance system for each specific user.

9.3   Study Design

Background

Aimed at preventing privacy conflicts and minimising the number of privacy violations, an Argumentation Framework for Online Social Networks was proposed in [324]. It is defined as a tuple $\langle A, R, P, \tau_p \rangle$ where:

- A is a set of n arguments $[\alpha_1, \ldots, \alpha_n]$

- R is the attack relation on A such as $A \times A \to R$

- P is the list of e profiles involved in an argumentation process (i.e., a privacy dispute) $[p_1, \ldots, p_e]$

- τ_p is a function $A \times P \to [0, \ldots, 1]$ that determines the score of an argument α for a given profile p

A complete definition of all the parameters that define the Argumentation Framework for Online Social Networks is presented in [324]. As a solid motivation for the research conducted in this paper, it is important to mention that this framework models each user by their personality and their OSN usage statistics, which are the features that we analyse in this work. The personality of each user is represented with a 5-dimension vector modelled with the Big Five personality traits [317]. These traits are *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*, that represent the five most significant aspects of human personality. The process of generating arguments is thoroughly explained in [322] and starts when a potential privacy violation is detected when publishing content in the social network. Then, the set of relevant information is gathered and retrieved from the OSN. Once all the arguments are generated, the system determines the set of *acceptable* arguments based on the score function τ_p. Finally, the system *translates* the arguments in their computational shape into human readable text with the use of templates. The final step is the human-computer interaction. To interact with a human user, the argumentation system has the available the set of *acceptable* arguments. However, the system needs to know which argument is more effective during the interaction process. The present work attempts to shed

light on the persuasive power of argument types and argumentation schemes, to be able to define better dialogue strategies prioritising the most persuasive arguments.

Research Questions

The previously defined theoretical framework was proposed to be integrated into PESEDIA, an educational social network [12]. However, deciding the dialogue strategies when interacting with human users is still a challenge. Therefore, we carry out this study to answer the following research questions that arise when designing this interaction:

RQ1. Which reasoning pattern (i.e., *argumentation scheme*) is more persuasive for teenage OSN users?

RQ2. Which topic (i.e., *argument type*) is more persuasive for teenage OSN users?

RQ3. How does the personality traits of teenage users influence the persuasive power of arguments?

RQ4. How does the online social interaction behaviour of teenage users influence the persuasive power of arguments?

If it is possible to find any behavioural pattern regarding these questions, the arguments could be generated by the argumentation system following different strategies for each user depending on their personality traits or their social interaction behaviour.

Measures and Instruments

To answer the proposed research questions, we designed the following study based on three questionnaires and the social network usage. Questionnaires were used to retrieve the personality traits of the participants (Big Five personality test), the persuasive power of argumentation schemes (Questionnaire A), and the persuasive power of types of arguments (Questionnaire B). Participants also used the social

network PESEDIA [12] during one month from where we collected the online so-
cial interaction data. PESEDIA is an educational OSN aimed at teaching its users
the basic privacy competences in social networks. This social network provides a
free environment like other OSNs (e.g., Facebook, Instagram, etc.). The chosen
way to teach users is by gamification, with scores and a global ranking to reward
the most active and participatory users. It is possible then, to nudge the users to do
activities and participate in debates without forcing them [10]. To find answers to
the research questions proposed, this study has been carried out in PESEDIA with
teenage participants ranged from 12-15 years old. The study lasted one month,
with the social network active and accessible 24/7 for participants. An ethics and
law committee from the Universitat Politècnica de València reviewed and approved
the study performed. Specifically, they reviewed that the social network PESEDIA
met the GDPR laws about users' privacy protection and management of their data.

Therefore, we measured their personality and online social interactions to
model the participants of our study. A Big Five personality traits test aimed
at measuring the personality traits (Openness, Conscientiousness, Extraversion,
Agreeableness, and Neuroticism) of children and teenagers [223] has been used.
Furthermore, we have also divided the participants into clusters based on their
personality traits. Four major clusters have been recently identified in the litera-
ture: *Average*, *Self-centered*, *Reserved*, and *Role model* [149]. This clustering is
proposed as a way to group samples with similar social perceptions and with sim-
ilar expected behaviour. Our hypothesis to use these clusters in our study is that
among similar characterised participants, it can be possible to observe stronger
behavioural patterns, reducing the noise and leading to more solid findings. Thus,
we have split our samples into four different personality-based groups to observe if
those same clusters could be found in our population and if any behavioural pattern
towards argument persuasion could be detected in each specific cluster. We ran the
K-Means algorithm until its convergence to generate the mentioned clusters.

In some situations, it may not be possible to retrieve users' personality traits.
Therefore, in our study, we have also considered the data from their social interac-
tion behaviour in PESEDIA. Thirteen different features representing participants'
social interaction in the OSN have been used in our study to model PESEDIA users
as an alternative to their personality: number of friends (#friends), number of sta-
tus updates (#status_upd), number of likes (#likes), number of shares (#shares),

number of comments (#comments), number of private posts (#ppprivate), number of public posts (#pppublic), number of posts shared with friends (#ppfriends), number of posts shared with collections of friends (#ppcollections), number of uploaded photos (#photos), number of posts deleted (#deletes), the average length of text posts (avg_textsize), and the time spent on the network (time_spent). Previous work identified in the literature pointed out that these features could be closely related to user personality [153, 2, 175, 47]. Therefore, these features represent an alternative dimension to personality from which it is possible to model OSN users.

Finally, the persuasive power of arguments (for schemes and types) has been computed as the number of times an argument beats others. Our metric is based on [117] work. Therefore, we define the persuasive power for an argument α_i as follows,

$$(9.1) \qquad s(\alpha_i) = \frac{\sum_{j \in C} b_{ij}}{|P| \cdot (|C| - 1)}$$

where b_{ij} refers to the number of times the argument α_i beats another argument α_j ($i, j \in C$, $i \neq j$). An argument α_i beats another argument α_j if it is considered more persuasive by our participants in the questionnaires. In our study, the classes C are represented as argumentation schemes and types of arguments. Regarding the parameters $|P|$ and $|C|$, they represent the number of participants and the number of options inside a class and they are used to compute the maximum number of times an argument class can beat each other. The result is a 0-1 normalised value. We have used questionnaires to measure the persuasive power of arguments, different ones for schemes and types. There, participants were faced with the same situation (Figure 9.1): they are going to make a publication in the network and they are told not to do it. The way of persuading the participant not to make the publication is with the use of arguments, so they had to rank these arguments, from the most persuasive argument (1) to the least persuasive one ($|C|$). Next, we describe how these questionnaires have been designed.
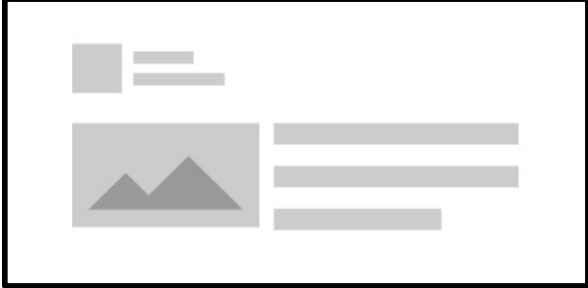
Questionnaire A (Schemes)

This questionnaire has been designed to capture the persuasive power of different argumentation schemes on a user (RQ1). We decided to consider the following five

Figure 9.1: Template of the persuasive power questionnaires.

schemes in our study: *Argument from Consequences* (AFCQ), *Argument from Popular Practice* (AFPP), *Argument from Popular Opinion* (AFPO), *Argument from Expert Opinion* (AFEO), and *Argument from Witness Testimony* (AFWT). With these schemes, it is possible to capture users behaviour when facing some of the most common reasoning patterns [400] used in social network privacy-related persuasive dialogues. Furthermore, we are able to analyse how practical reasoning and different source-based arguments are able to persuade teenager OSN users. By using these schemes, our goal was to see if teenagers were more concerned about recommendations based on the consequences of their actions, an expert opinion, similar user experiences, popular behaviours, or previously affected users.

Arguments from Consequences show the participant the consequences of doing some specific action, sharing some content in our case. With this scheme, we can measure the importance each participant gives to the effect of their actions in the social network. *Arguments from Popular Practice* try to persuade evidencing that there is a common popular practice among other similar people regarding

some specific topic. In this case, with AFPP we can observe the importance that participants give to an argument based on their friends' activity. Similarly, *Arguments from Popular Opinion* try to persuade with the use of a generally accepted opinion. Therefore, AFPO allows us to observe participants preferences towards the generally accepted opinion regarding their privacy. *Arguments from Expert Opinion* base their reasoning pattern on some expert opinion regarding a specific topic. These argumentation schemes make it possible to observe users reliance in a privacy domain expert. Finally, *Arguments from Witness Testimony* make the reasoning taking into account the experience of a person in the same knowledge position. With this scheme, it is possible to measure the trust that our participants give to someone with their similar expertise level in privacy management.

In this first questionnaire, the arguments that represent these five argumentation schemes in the OSN domain and that participants ranked by their perceived persuasive power are the following (You should not make this publication because...):

- Making the publication could have bad consequences for your privacy (AFCQ)

- Most of your friends would not publish this content (AFPP)

- Everyone knows that publishing this is a mistake (AFPO)

- The monitors are experts in social networks and they believe that making publications of this type could be dangerous (AFEO)

- A user of the PESEDIA network who has made similar publications considers that it can be dangerous (AFWT)

Questionnaire B (Types)

This questionnaire has been created to observe the persuasive power on our participants of the four different types of arguments considered by the argumentation framework (RQ2). These types are: *Privacy*, *Trust*, *Risk*, and *Content* arguments. *Privacy* arguments are generated regarding each user privacy preferences towards

the audience of his/her publications (i.e., private, friends, public, or friends col-
lection). Therefore, *Privacy* arguments will try to persuade the participants con-
sidering their privacy preferences and configuration. *Trust* arguments are the ones
generated taking friendships between users into account. This type of arguments
will try to persuade the participant making him/her understand that other persons
may be harmed if the content gets published. *Risk* arguments consider the publica-
tion reachability in the network, computed as explained in [11, 14]. Then, a *Risk*
argument is generated if the scope of the publication exceeds the user expected au-
dience. Finally, *Content* arguments are generated regarding the own content of the
publication. Six different types of content (i.e., location, medical, alcohol/drugs,
personal information, family/association, offensive) [72] are considered by our ar-
gumentation system. In this case, the degree of participants' persuasion may vary
with the type of content included in the publication due to its sensitivity [340].
The arguments that participants ranked by their perceived persuasive power in this
questionnaire and represent the four argument types are the following: (You should
not make this publication because...)

- you have chosen public privacy settings. (Privacy)

- some of the people who appear might get upset. (Trust)

- it could be read by strangers. (Risk)

- you are revealing your location. (Content: Location)

- you are giving out personal medical information. (Content: Medical)

- others may think you consume alcohol/drugs. (Content: Alcohol/Drugs)

- you are revealing personal data about yourself. (Content: Pers. Information)

- you are revealing a friend's personal information. (Content: Fam./Assoc.)

- you might offend some other user. (Content: Offensive)

where items represented as (●) refers to Privacy, Trust, and Risk types of ar-
guments and items represented as (○) refers to the different contents (Location,

Medical, Alcohol/Drugs, Personal Information, Family/Association, and Offensive) of Content-type arguments. This questionnaire was done by participants as many times as different contents of Content-type of arguments are in order to avoid biases on users' perception of information sensitivity [340].

Procedure

The study was carried out on the PESEDIA social network where teenage users used it during one month. To prevent interferences, we included a registry controller (using a secret token) to avoid undesired registrations that could affect the security of the participants and the study. The questionnaires described above to measure participants' features were integrated in the own social network and they were progressively enabled in the on-site sessions. They were not required to complete them at any specific moment, but participants were motivated through gamification techniques. During the whole period of the study, the participants had fully access to the PESEDIA social network to share their experiences and feelings.

We organised three on-site sessions of 90 min in equipped labs at the university to use as control points of the study. These three on-site sessions were distributed at three points in time: session 1, at the beginning of the one-month period; session 2, in the middle; and session 3, at the end. The aim of these sessions was to clarify any doubts that might arise among the participants about the functionality and features of the social network. Each session started with a brief explanation of the potential activities that they could do related to testing and understanding functionalities of the social network, and then participants had time to interact using the social network. In the first session, we introduced PESEDIA to the participants and they signed up on the social network. Then, they had to complete basic activities that focused on customising their user profiles, setting up their general setting options, and building their friendship relations. Before finishing the first session, the personality test was made available for the participants to complete it. In the second session, we requested participants to complete the questionnaires about persuasive power (Questionnaires A and B). In Questionnaire A, participants ranked the five argumentation schemes in a decreasing persuasive ordering. In Questionnaire B, participants faced six different instances of the questionnaire considering one specific content category at a time. They ranked the four argument types in a

decreasing persuasive ordering in each instance of the questionnaire. Arguments
were displayed in a different order in each round to avoid the order effects. Fi-
nally, in the third (and last) session, we presented the participants with a summary
regarding their behaviours and answers to the questionnaires to conclude the study.

Participants

A total of 218 teenagers participated in the study. From this total population, 215
participants completed the personality test and 212 completed both questionnaires
A and B. We excluded the participants who did not complete all of the control ses-
sions and the proposed questionnaires (29 participants) as well as the participants
who decided not to participate (3 participants did not log into Pesedia). Finally,
186 participants completed the study[1] (103 males, 83 females, M_{age}= 13.15, range:
12–15 years old). We included the participants in the experiment taking into ac-
count their age in order to have a sample of the teenage population (participants
older than 12 years old). All of the selected participants were attending high school
in different school centres of the Valencia area at the time of the experiment. In
our study, we modelled our participants considering two different dimensions: the
personality and the social interaction behaviour in the OSN. Furthermore, we in-
vestigated if stronger behavioural patterns could be identified when grouping our
population by gender (i.e., male/female) and by personality clusters (i.e., *Average*,
Self-centered, *Reserved*, and *Role model*).

   The first modelling dimension considered in this research is the personality. We
used the Big Five personality traits to represent the personality of our participants.
From these five personality trait values, we grouped our participants into four dif-
ferent personality clusters. Those clusters had the following composition: the *Av-
erage* cluster (C_1 = 44, 56.8% males); the *Self-centered* cluster (C_2 = 38, 68.4%
males); the *Reserved* cluster (C_3 = 52, 48.1% males); and the *Role model* cluster
(C_4 = 52, 63.5% males). Figure 9.2 shows the Big Five personality traits distri-
bution of the clusters found in our study. Each cluster is defined by the means of
averages of each personality trait z-score. Therefore, it is possible to observe how
depending on the cluster (i.e., *Average*, *Self-centered*, *Reserved* and *Role model*)
the personality trait average z-scores of its members follow different distributions.

---

[1]Contact the authors to get access to an anonymised version of the data gathered in this study.
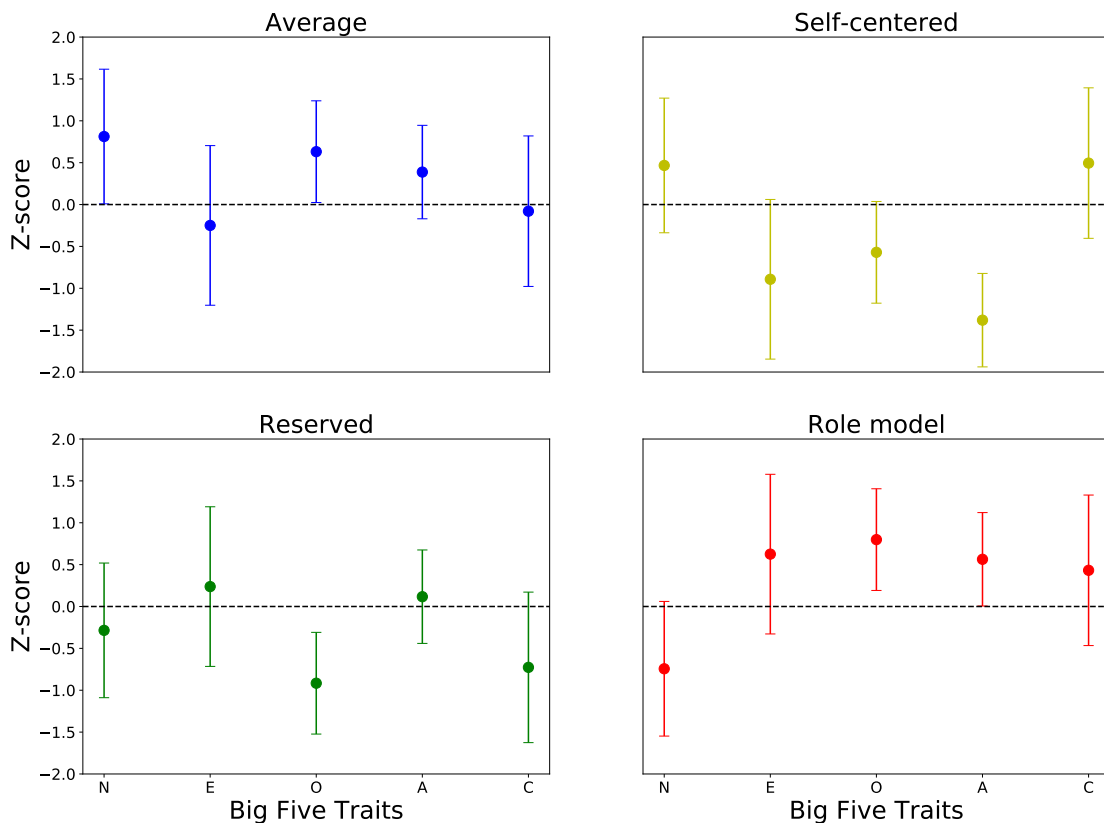
Figure 9.2: Personality clusters observed in our participants' data. (•) Is the position of cluster centres represented as the average z-score of each cluster personality traits. The error bars represent the standard deviation of each trait in each cluster. The dotted lines represent global average values (Z=0) for each personality trait.

Comparing the clusters found in this work with the clusters proposed in [149], it is possible to observe strong similarities between them. The Silhouette Coefficient (SC) [319] of the computed clusters is 0.173 meaning that some clusters could be overlapping (SC ≈ 0) but the samples are not being miss-classified (SC>0). The *Reserved* personality type is characterised by negative z-score values on Neuroticism and openness, while the rest of the traits (Extraversion, Agreeableness, and Conscientiousness) are slightly higher to 0. The *Role model* personality type is characterised by negative z-score values on Neuroticism and positive z-score values for the rest of the traits. For both clusters, the personality traits of our participants followed the same distributions as [149] clusters. The *Average* personality

type is characterised by z-score values close to 0 for all personality traits. In our
study, this cluster follows this trend with slightly higher z-score values on Neu-
roticism and Openness. Finally, the *Self-centered* personality type is characterised
by negative z-score values on all the personality traits except for the Extraversion
trait. By comparing it with our cluster, we found some differences between those.
However, we found a strong relationship with the original cluster in which *Self-
centered* was based, called *Undercontrolled*, which was introduced in [27] work.
In [149], the *Undercontrolled* personality group is said to strongly influence the
new proposed *Self-centered* cluster. From the clusters observed in our population,
we can support this statement. The *Self-centered* cluster observed in this work has
a positive Neuroticism z-score, similar to the original *Undercontrolled* group. Fur-
thermore, significant differences were observed regarding the Conscientiousness
trait in our *Reserved* and *Self-centered* clusters. Studies have shown how Con-
scientiousness is the most variable trait with the age [120]. Therefore, we think
the observed differences were mainly due to the important age gap between the
participants of both studies.

Finally, the second dimension used to model OSN users in our analysis is their
online social interaction behaviour. During our study, a total number of 2195 likes,
7650 comments, 1309 shares, 846 photos uploaded, and 7788 status updates (from
them 761 were private, 769 were public, 5774 were disclosed to friends, and 484
were disclosed to specific lists of friends) were registered in the PESEDIA database.
The participants had a mean of 12 friendships and regretted 2761 actions made
(which they undid/delete). The most common social interactions were comments
and status updates. We observed an average of 41 comments and 42 status updates
per user. It is also interesting to observe the high average number of deletes per user
(i.e., 15), which represents a high number of regrets of the content published in the
network. We also observed that in general, users preferred to share publications
with friends only rather than publicly, privately or considering specific collections
of friends.

At the end of the study, we collected 186 different combinations of the Big
Five personality traits and 18942 different OSN interactions. Furthermore, we also
collected: 930 persuasive pairwise comparisons of argumentation schemes, one
per participant (186) and argumentation schemes (5); and 4464 persuasive pairwise
comparisons of argument types, one per participant (186), argument types (4), and

contents variation of Content-type of arguments (6).

9.4 Results

Aimed at finding an answer to our research questions, we calculated the persuasive power of the five argumentation schemes and four argument types using the persuasive power equation (Equation 9.1). Furthermore, we did a correlation analysis between our user modelling features (i.e., personality and online social interaction features) and the previously calculated persuasive power values. In this section, we present the observed results after completing this process using the data gathered at the end of our study.

Persuasive Power of Arguments

From the results of the study, we have calculated the persuasive power of all the argumentation schemes and types considered in this work. Therefore, it is possible sort the five argumentation schemes taking into account their persuasive power as follows (RQ1): AFCQ $\succ$ AFEO $\succ$ AFWT $\succ$ AFPP $\succ$ AFPO. *Argument from Consequences* seemed to be the most effective scheme for persuading our participants with a score of 0.61. Following, we have the *Argument from Expert Opinion* scored with 0.53, *Argument from Witness Testimony* with a score of 0.47, *Argument from Popular Practice* with a score of 0.46 and finally, *Argument from Popular Opinion* was the less persuasive scheme with a score of 0.43. These results mean that teenagers, in general, can be persuaded easier by showing the consequences of their actions or with recommendations made by experts rather than nudging them with recommendations made by someone similar to them or with popular trends or opinions. Table 9.1 represents the direct comparison of the persuasive ranking between pairs of argumentation schemes. For that purpose, we measure the number of times an argument is ranked in a higher position than another. We can notice how this value is higher when arguments with stronger persuasive power are compared with lower persuasive power arguments and vice-versa.

On the other hand, we sorted the argument types taking their persuasive power into account as follows (RQ2): *Content* $\succ$ *Privacy* $\succ$ *Risk* $\succ$ *Trust*. *Content* argu-

Table 9.1: Pairwise rank comparative between argumentation schemes. This table represents the number of times an argumentation scheme (rows) beats another argumentation scheme (columns).

	AFCQ	AFEO	AFWT	AFPP	AFPO	TOTAL
AFCQ	-	106	114	112	123	**455**
AFEO	80	-	104	102	109	**395**
AFWT	72	82	-	96	99	**349**
AFPP	74	84	90	-	92	**340**
AFPO	63	77	87	94	-	**321**

ments were the most persuasive with a score of 0.59. Following, we have *Privacy* arguments with a score of 0.52, *Risk* arguments with score of 0.47 and *Trust* arguments with a score of 0.42. Meaning that teenagers are more concerned about sharing sensitive content rather than being read by unknown users or endangering other parties privacy. Similar to the previous analysis with argumentation schemes, Table 9.2 represents a direct comparison between the ranking position of every pair of argument types. Here, we also observe how arguments with a higher persuasive power score are ranked, in general, in a higher position than the rest. If we consider each round of the questionnaire B independently to analyse the effect of each content type on the persuasion of the argument, the following persuasive ordering is observed: Offensive ≻ Personal ≻ Family ≻ Medical ≻ Alcohol/Drugs ≻ Location. Therefore, although Content arguments were found as the most persuasive type of arguments, depending on which type of content was considered in each round, users' susceptibility was different. Our study revealed that teenagers are more concerned about sharing offensive content with a score of 0.64, closely followed by sharing personal information with a score of 0.62. The concern with these specific types of content matches the new trends in social networks of self-presentation [89]. The next most concerning types of content were family/association and medical content with scores of 0.59 and 0.58 respectively. Finally, revealing alcohol/drug consumption or location information, seemed to be the less relevant types of content for our participants with scores of 0.56 and 0.53 respectively.

Table 9.2: Pairwise rank comparative between argument types. This table represents the number of times an argument type (rows) beats another argument type (columns).

	Content	Privacy	Risk	Trust	TOTAL
Content	-	623	659	681	**1963**
Privacy	493	-	605	634	**1732**
Risk	457	511	-	616	**1584**
Trust	435	482	500	-	**1417**

Personality Impact on Argument Persuasion

To be able to adjust our argumentation system to increase the persuasive power of the arguments for our target population, we analysed the personality impact on the persuasive power (RQ3) of argumentation schemes and argument types. For this purpose, we have calculated the Spearman ρ rank correlation between the persuasive power of arguments and the Big Five personality traits. In order to ease the interpretation and visualisation of the results, we have grouped correlations into three correlation-strength categories based in the ones proposed in [103]. Weak correlations stand for correlation values between 0 and 0.2; we consider a Moderate correlation if its correlation value is between 0.2 and 0.6; finally, a Strong correlation stands for correlation values higher than 0.6.

The significant correlations found between argumentation schemes and argument types, with personality traits are represented in Table 9.3 and Table 9.4 respectively. As we can observe, different correlations have been found for different groups of users and associated with different personality traits. It is possible to observe how in some cases the personality correlates with the perceived persuasive power, which means that personality features could serve as predictors of persuasiveness when defining persuasive policies. We can also observe a smaller number of significant correlations between the argument types that the argumentation schemes. A possible cause for this pattern in our findings is that variations among different argumentation schemes have a greater impact on the perceived persuasiveness of an argument than the variation between argument types. Furthermore, studying specific groups of users categorised by descriptive features such as the

Table 9.3: Significant correlations of argumentation schemes persuasive power and personality traits. The significance is represented as: * $= p < 0.05$, ** $= p < 0.01$. The correlation strength is represented as: Weak = $+/-$; Moderate = $++/--$; Strong = $+++/---$

	Participants	O	C	E	A	N
	All	-	-	−AFEO**	−AFPP**	-
Gender	Male	-	-	−AFEO* +AFWT*	−AFPP* +AFWT*	-
	Female	-	-	-	-	+AFPP*
Personality Cluster	Average	-	-	-	-	-
	Reserved	-	-	−−AFEO*	-	++AFPP** −−AFEO*
	Self-centered	-	−−AFCQ*	-	-	-
	Role model	-	−AFPO* ++AFEO**	-	-	-

Table 9.4: Significant correlations of argument types persuasive power and personality traits. The significance is represented as: * $= p < 0.05$, ** $= p < 0.01$. The correlation strength is represented as: Weak = $+/-$; Moderate = $++/--$; Strong = $+++/---$

	Participants	O	C	E	A	N
	All	-	-	−Privacy*	-	-
Gender	Male	-	-	-	-	-
	Female	-	-	-	-	-
Personality Cluster	Average	-	-	-	-	-
	Reserved	−Risk*	-	-	-	-
	Self-centered	-	-	++Content*	-	-
	Role model	-	-	-	-	-

gender or the personality allow to draw more informed conclusions than considering the whole heterogeneous group of users. Thus, personalisation is a key aspect to improve the effectiveness of the human-computer interactions of an argumentation system.

Social Interaction Impact on Argument Persuasion

In some environments, obtaining the Big Five personality traits may not be possible. Therefore, in order to model our OSN users before analysing the persuasive power of arguments, we proposed an alternative to the personality based on the social interaction behaviour of our participants [153, 2, 175, 47]. This way, we analysed if there existed any correlation between the persuasion of arguments towards each participant depending on their social interaction behaviour (RQ4). To measure the impact of these thirteen features on the persuasive power of arguments, we have calculated the Spearman ρ rank correlation between them and the persuasive power of arguments. The interpretation of the correlation values is done the same way as the previous section. Furthermore, if personality traits are available, we have also considered making a complete analysis, taking into account personality clusters. This way, it is possible to combine the results of both analysis, thus observing even more useful correlations to define dialogue strategies.

The significant correlations found between argumentation schemes and argument types, with OSN interaction data have been represented in Table 9.5 and Table 9.6 respectively. As in the previous analysis, social interaction data has proven to be a good predictor of variations in perceived persuasion for different user models. Again, it was harder to find significant correlations when considering the argument types than the argumentation schemes. This pattern reinforces the hypothesis that argumentation schemes contain more persuasive information that argument types, and are aligned with the recent findings described in [329].

Interpretation of the Results

This work sets the starting point to develop the human interaction part of argumentative educational systems to help with privacy management in OSNs. The findings observed in this paper reveal that personality traits and social interaction data are relevant user modelling features useful for estimating the perceived persuasive power of arguments by different user models. Therefore, these features represent a powerful way to model human users when approaching a problem of these specifications. These findings are consistent with recent research in similar topics [170]. In Table 9.7, we present four different OSN user models consider-

Table 9.5: Significant correlations of argumentation schemes persuasive power and social interaction data. The significance is represented as: * = $p < 0.05$, ** = $p < 0.01$) The correlation strength is represented as: Weak = $+/-$; Moderate = $++/--$; Strong = $+++/---$

	Participants	#friends	#status_upd	#likes	#comments	#ppprivate	#pppublic	#ppfriends	#ppcollections	avg_textsize
	All	-	-	-	-	+AFPO*	-	-	−AFPP**	−AFPO*
Gender	**Male**	-	-	−−AFPP*	-	-	-	-	−−AFPP**	-
	Female	-	++AFEO*	-	++AFPO**	++AFPO**	-	−−AFCQ**	-	-
Personality Cluster	**Average**	++AFEO*	-	-	-	−−AFPP*	-	++AFEO*	-	-
	Reserved	++AFCQ*	++AFCQ*	-	-	-	−−AFWT*	++AFCQ*	-	-
	Self-centered	-	-	-	-	−−AFEO*	-	-	-	−−AFCQ** ++AFWT**
	Role model	-	-	-	++AFEO*	-	-	-	−−AFCQ** ++AFPO**	++AFCQ**

Table 9.6: Significant correlations of argument types persuasive power and social interaction data. The significance is represented as: * = $p < 0.05$, ** = $p < 0.01$) The correlation strength is represented as: Weak = $+/-$; Moderate = $++/--$; Strong = $+++/---$

	Participants	**#status_upd**	**#comments**	**#ppprivate**	**#ppcollections**	**#deletes**
	All	-	-	-	-	-
Gender	**Male**	++Trust*	-	-	++Trust*	−−Risk*
	Female	-	-	-	-	-
Personality Cluster	**Average**	-	-	-	-	-
	Reserved	-	-	-	-	-
	Self-centered	-	-	−−Risk*	-	-
	Role model	++Trust*	++Trust*	-	-	-

ing the features proposed in this work. From the found correlations presented in the previous section, we estimate potential trends in the persuasive power of arguments for these users as depicted in Table 9.8. We can observe how different user models may perceive arguments with a modified persuasive power. Thus, with the observed results, we can adapt the available argumentation schemes and argument types following different user tailored persuasive policies which are more effective than the one based on the general persuasive power of arguments.

Argumentation schemes have been previously investigated and classified by experts of many different disciplines such as spanning philosophy, communica-

Table 9.7: Four different user models. (-) represents an average value, (↑) represents a value above the average and (↓) represents a value below the average.

Model Features	User 1	User 2	User 3	User 4
Gender	Male	Female	Female	Male
Cluster	Self-centered	Role model	Reserved	Average
Openness	-	-	↓	-
Conscientiousness	↓	↑	-	-
Extraversion	↓	-	-	↑
Agreeableness	-	-	-	↑
Neuroticism	↑	↓	↑	-
#friends	-	-	↑	↓
#status_upd	↑	-	↓	-
#likes	↓	-	-	-
#shares	-	-	-	-
#comments	-	↑	-	-
#ppprivate	↓	-	↑	-
#pppublic	-	-	-	-
#ppfriends	↑	-	↑	-
#ppcollections	-	-	-	-
#deletes	-	-	-	↑
#photos	-	-	-	-
avg_textsize	-	↑	-	↓
time_spent	-	-	-	-

Table 9.8: Persuasive power of argumentation schemes and argument types for four different users. (-) represents an unmodified value, (↑) represents an increased persuasive power and (↓) represents a decreased persuasive power.

Persuasive Power	User 1	User 2	User 3	User 4
AFCQ	↑	↑	↓	-
AFEO	↑↑	↑↑	↓↓	↓
AFWT	↓	-	-	↑
AFPP	-	↓	↑↑	↓↓
AFPO	↓	-	-	↑
Content	↓	-	-	-
Privacy	↑	-	-	↓
Risk	↑	-	↑	↓
Trust	↑	↑	-	-

tion studies, linguistics, computer science and psychology [398]. Thus, several clusters of schemes have been defined grouped according to their general category. The schemes we work with belong to the general categories of "practical reasoning arguments" (AFCQ); and "source-dependent arguments", concretely, to its subcategories of "arguments from position to know" (AFEO and AFWT) and "arguments from popular acceptance" (AFPO and AFPP). Recently, a relation between this classification and Cialdini's principles of persuasion has been established [366, 329]. Thus, the "Consistency" principle of persuasion, by which people like to be consistent with the things they have previously said or done, relates to one's practical behaviour (AFCQ); the principle of "Authority", by which people follow the lead of credible, knowledgeable experts, relates to source-based arguments (AFEO and AFWT); and the principle of "Consensus", by which individuals are conformed to what the majority regards as acceptable, relates to arguments from popular acceptance (AFPO and AFPP). First, we can see how our findings detect a preference order "Consistency" > "Authority" > "Consensus" for the persuasion principles in our social media domain. Second, although there is still no specific research that orders these persuasion principles by their persuasion power in the context of social media, similar research in the healthy eating domain concluded an order "Authority" > "Consensus" and "Consistency" (no significant difference between these two) and stated that persuasive power is highly influenced by the domain [366]. Based on these mappings between argumentation schemes and persuasive principles, we can contextualise our findings within essential concepts of persuasive psychology research. Thus, any significant correlation detected between user descriptive features (i.e., personality and social interaction) and argumentation schemes can be also interpreted as a correlation between the user features and the three persuasive principles related to the five argumentation schemes analysed in our study. As an example, from our findings we can interpret that the "Authority" (AFEO) principle has shown an increased persuasive power for *Reserved* users with a low value on Extraversion. Although we have not been able to identify much prior research focused on this same purpose, the work presented in [92] sheds light on existing correlations between Big Five personality traits and Cialdini's persuasive principles. Albeit the populations of both studies differ substantially in age, some similarities can be identified among the significant correlations detected in both works. In Section 9.4, we have identified the fol-

lowing correlations: a negative correlation between AFEO and Extraversion trait; a positive correlation between AFEO and Conscientiousness for our *Role model* participants; and a positive correlation between AFPP and Neuroticism for our *Female* and *Reserved* participants. All these detected correlations have also been found in [92] work. However, the age gap and the significant differences between populations make it harder to compare the findings on both works and some of the found correlations remain unexplained.

9.5 Conclusion and Future Work

At the beginning of this work, we raised four different research questions aimed at having a better understanding of human persuasion in OSNs using arguments. With our findings, we have been able to answer the four research questions, and to have a better understanding of the persuasiveness of arguments when used with different user models. Personalisation plays a major role in effective human-computer interactions. In this paper, we have been able to observe that using representative user modelling features (i.e., personality and social interaction data) it is possible observe variations in the effectiveness of arguments. Therefore, the user models analysed in this work provide a solid basis for developing personalised argumentation systems aimed at educating and preventing privacy violations in an OSN environment. Given the nature of arguments and argumentation, the findings observed in our study lay the basis for developing powerful tools for education and decision-making assistance.

As future work, we foresee to deepen the analysis presented in this paper by extending our study to an adult population, and to implement an argument-based persuasive algorithm able to generate personalised persuasive policies aimed at maximising the efficiency of human-computer interactions.

Persuasion-enhanced Computational Argumentative Reasoning through Argumentation-based Persuasive Frameworks

RAMON RUIZ-DOLZ, JOAQUIN TAVERNER, STELLA HERAS AND ANA GARCÍA-FORNES

Abstract

One of the greatest challenges of computational argumentation research consists of creating persuasive strategies that can effectively influence the behaviour of a human user. From the human perspective, argumentation represents one of the most effective ways to reason and to persuade other parties. Furthermore, it is very common that humans adapt their discourse depending on the audience in order to be more persuasive. Thus, it is of utmost importance to take into account user modelling features for personalising the interactions with human users. Through computational argumentation, we can not only devise the optimal solution, but also provide the rationale for it. However, synergies between computational argumentative reasoning and computational persuasion have not been researched in depth. In this paper, we propose a new formal framework aimed at improving the persuasiveness of arguments resulting from the computational argumentative reasoning process. For that purpose, our approach relies on an underlying abstract argumentation framework to implement this reasoning and extends it with persuasive features. Thus, we combine a set of user modelling and linguistic features

through the use of a persuasive function in order to instantiate abstract arguments following a user-specific persuasive policy. From the results observed in our experiments, we can conclude that the framework proposed in this work improves the persuasiveness of argument-based computational systems. Furthermore, we have also been able to determine that human users place a high level of trust in decision support systems when they are persuaded using arguments and when the reasons behind the suggestion to modify their behaviour are provided.

10.1 Introduction

Computational Argumentation is a multidisciplinary area of research that investigates every phase of human argumentation from the computational viewpoint [28, 320]. Research in this area is done from different perspectives, such as Natural Language Processing (NLP), formal logic, Knowledge Representation and automated Reasoning (KRR), and Human-Computer Interaction (HCI), to approach specific tasks and solve concrete problems. On the one hand, argumentation-based NLP research has mostly been focused on identifying, classifying, and structuring arguments in natural language input sources. This line of research is aimed at solving Argument Mining tasks, and it is usually constrained to a specific domain [207]. It is also possible to identify NLP research by investigating how to estimate argument persuasiveness based on the natural language of the argument [151] and to automatically generate natural language arguments [190]. On the other hand, formal logic and KRR research in computational argumentation has been targeting the computational representation of arguments and the simulation of the human argumentative reasoning process from the computational viewpoint. Argumentation has typically been encoded using argumentation frameworks [124] (i.e., graphs), and the human argumentative reasoning has been approached using argumentation semantics [38] (i.e., graph algorithms with underlying logic rules). Finally, argumentation-based HCI research is mostly focused on understanding human behaviour when using arguments on different platforms such as decision support systems [322] or chatbots [82]. For that purpose, human users are modelled through the use of features (i.e., personality, concerns, emotions, beliefs, or online behaviour) that can have an impact on the perception of arguments and their persuasive power [164, 323]. However, most of the research carried out on this topic

focuses on a very specific perspective and does not explore the potential existing synergies among advances in different areas. Taking the human argumentative reasoning process as a reference [395], we consider that transversal computational argumentation research is of utmost importance in leveraging the advances and proposals made in each specific area of research. For example, by integrating the algorithms proposed for computationally approaching the human argumentative reasoning, with user modelling and predictive techniques.

Recently, Computational Argumentation research is investigating how the computational approaches to the different aspects of human argumentation (e.g., identification, analysis, evaluation, or invention [395]) can benefit from combining the advances contributed independently in each specific domain (e.g., NLP, formal logic, HCI, persuasion, etc.). Approaches that extend the specific tasks of argument mining have been investigated in search of a convergence between natural language argument structures and argumentation frameworks [100]. Furthermore, recent research reports the benefits of combining argumentation semantics with NLP algorithms for improving the automatic evaluation of argumentative debates [326]. However, argument-based computational persuasion research has still not explored such synergies in-depth. Most of the research aimed at persuading human users through argumentation independently explore the use of machine learning for estimating the most persuasive argument [119], analyses human behaviour with empirical studies [368], or explores the use of interactive chatbots for behaviour change [83]. A common feature in all of these independent approaches is the modelling of human users, which plays a major role in the personalisation of computational persuasion systems [177].

In this paper, we propose an extension for formal argumentation frameworks and their semantics that enables argument-based computational persuasion. With this extension, it is possible to bridge the gap between formal computational argumentation and HCI research. For that purpose, we introduce the Argumentation-based Persuasive Frameworks (APF), which rely on the argumentative reasoning provided by any underlying abstract argumentation framework and generates user-tailored natural language arguments. This goal is achieved through user modelling, which plays a fundamental role in our proposal and enables a personalised interaction between the human user and the argumentative system. We model our users using their personality and their online behaviour (e.g., number of friends, com-

ments, or likes). Then, natural language arguments are created taking into account the logical principles of admissibility and conflict-freeness [38] of abstract arguments encoded in the argumentation framework. The abstract arguments are instantiated into natural language arguments using a set of linguistic features that allow the perceived persuasiveness of the produced arguments to be increased for each different user profile. In addition to the formalisation, we do a complete integration of the APF in the Online Social Network (OSN) domain for the prevention of privacy violations. Furthermore, we evaluate the performance of an argumentation system with an underlying APF when trying to persuade human users not to disclose specific potential privacy threatening publications. We observed a significant improvement in the persuasiveness of arguments when using the proposed APF to engage Human-Computer Interaction instead of relying exclusively on an argumentation framework without any type of explicit personalisation. Furthermore, we have also observed a high level of trust from human users towards the argumentation system when modifying their initial decisions after reading the arguments.

The rest of the paper is structured as follows: Section 10.2 reviews the previous work done in the intersection of computational argumentation and computational persuasion; Section 10.3 introduces the formal background, and provides a formal definition of the Argumentation-based Persuasive Framework; Section 10.4 presents a use case of our framework in the online privacy domain and proposes a complete implementation of the proposed framework in a real argumentation system; Section 10.5 evaluates our proposal in terms of behaviour change and human persuasion; Section 10.6 discusses the obtained results; and Section 10.7 summarises the most important conclusions of this paper.

10.2 Related Work

Persuasion represents one of the most important goals of human argumentation. When engaging in an argumentative dialogue, a common goal is to persuade other participants [231]. From a computational perspective, persuasion is typically studied as a cornerstone of HCI systems. In computational argumentation research specifically, persuasion has been investigated from different viewpoints [177, 191].

The automatic estimation of the persuasiveness of a natural language argument
has been widely studied in the NLP area of research. In [151], the authors present
a corpus that is specifically designed for determining the most persuasive argu-
ment from a given pair of arguments. A neural network architecture is trained to
learn linguistic features and solve the task of predicting and modelling persuasion
from natural language input. Another approach is proposed in [31], where the
authors focus on the analysis of the impact of style on the persuasive power of
news editorial arguments. For that purpose, five different NLP features are used to
model style: Linguistic Inquiry and Word Count, a lexicon of emotions (i.e, anger,
disgust, and fear) and sentiments (i.e,. positive and negative), argumentative dis-
course units features (i.e., anecdotal, statistical, and testimonial evidence) [193],
arguing elements (i.e., assessments, doubt, authority, and emphasis) [346], and text
subjectivity (i.e., subjective or objective) [404]. These features are used to train a
Support Vector Machine (SVM) [378] on a task aimed at predicting whether or not
a message will be persuasive. Finally, we can observe a combination of NLP and
user modelling in [191]. The authors propose an approach that uses users' beliefs,
interests, and personality traits, along with NLP feature engineering on natural
language inputs to predict the persuasiveness of arguments and users' resistance to
persuasion. However, the analysed research only takes into consideration natural
language and user models and does not take argumentative reasoning into account.

A different approach aimed at understanding specific aspects of the persuasive
properties of computationally generated arguments comes by the hand of empirical
studies. In [368], the authors propose a scale to measure the persuasive power of
different argumentative messages in the health and security domains. The scale is
developed after conducting a study where users were asked to provide information
related to three different factors of the perceived persuasiveness of different mes-
sages: their effectiveness, their quality, and their capability. A study of the impact
of the personality, the age, and the gender of human users on their susceptibility to
persuasive messages is done in [92]. Combined with the results presented in [329],
we can learn more about the persuasion of arguments when used in an argumen-
tative interaction with a human user based on personal characteristics. Another
interesting approach is presented in [323], where the authors propose a metric
for measuring the persuasive power of different reasoning patterns and arguments
based on a study with human participants. The study makes an analysis of how

human features (i.e., personality and social interaction) are related to perceived persuasive power. Finally, in [164], the authors present a series of empirical studies that are designed to measure how different preferences and concerns of human users can be a factor of influence in perceived persuasion when reading specific arguments.

Persuasion has also been studied as the utility function of argumentation dialogues and negotiation. In [165], the authors present a framework for argumentation-based decision-making assistance. This framework relies on decision trees for modelling the dialogue and improves its persuasiveness when the user model is combined with emotional features. In a dialogue, choosing which argument are more persuasive can be modelled as an optimisation of a strategy learning problem. With regard to this, Reinforcement Learning (RL) [359], is a promising technique for learning persuasive dialogue strategies. In [244], persuasion is defined as the effectiveness of arguments when used in a negotiation for reaching a satisfactory agreement. In that work, an argumentative agent learns to use the most persuasive argument in a given step of the dialogue through RL. Similarly, RL is used for learning dialogue strategies in [8]. Furthermore, in [166], the authors retake the belief-concern user model of [164] and propose a Monte Carlo tree search for finding the optimal persuasive policies for specific user models. The belief-concern user model was also considered in [179], where a general framework for computational persuasion is presented. This framework is instantiated into an argumentative chatbot for the purpose of behaviour change in the domains of cycling and university fees. In a recent work, a machine learning approach to argument-based persuasion was proposed in [119], where bi-party decision trees are used for predicting an argument's utility (i.e., persuasiveness) in a dialogue. The proposed model is evaluated in a simulated environment. Finally, in a recent work, a visual interactive system for making persuasive analyses of online discussions has been proposed [407]. This system makes it possible to improve the persuasive strategies of users through a complete visualisation of different persuasive features of arguments when used in a dialogue.

From the previous literature review, two major limitations are identified. First, there is only limited research on how computational argumentative reasoning can be extended to a persuasive argumentative system. Research on this topic is relevant for deepening computational persuasion research, where a system could per-

form argumentative reasoning before interacting with a human user. Second, there are not many evaluations of behaviour change with real humans. Even though argument-based computational persuasion has been explored from many different viewpoints, only a few works have conducted a complete evaluation of their proposal when trying to persuade human users. Furthermore, it has not been possible to identify many works where concepts from computational argumentation theory are combined with HCI and argument-based persuasion such as [164] and [316]. In [164], argumentation frameworks are used for computationally representing arguments as a graph. However, this work only considers this concept as a data structure, and the automatic argumentative reasoning is not carried out using argumentation semantics. In contrast, in [316], the authors propose an argumentative agent that uses a formal argumentation framework and its semantics for approaching argumentative reasoning, together with a Partially Observable Markov Decision Process for learning persuasive strategies. This agent is evaluated when interacting with real human users, but only a very small population is used. Our research extends this line of work by providing a formal framework for generalising the integration of argumentation frameworks with persuasive systems, combined with user modelling for personalising the interactions. We present an implementation of our proposal with a complete evaluation of its persuasiveness when interacting with human users, which has been evaluated in a sample population of 50 participants.

10.3   Formalisation

In this section, we present all of the formal definitions that support the research conducted in this paper. First, we introduce all of the required background concepts in order to have a complete understanding of the scope of our proposal and our experimentation. Second, we formalise our Argument-based Persuasive Framework.

Background

Before defining our proposal for an Argument-based Persuasive Framework, it is of the utmost importance to introduce some fundamental formal aspects of the com-

putational abstract argumentation theory. The concept of *argumentation frameworks* can be considered as a cornerstone in this topic, from which most of the research in computational argumentation and logic has been based. As proposed in [124], an *argumentation framework* makes it possible to computationally represent the logical aspects behind human argumentation from an abstract perspective:

Definition 4 (Abstract Argumentation Framework) *An Abstract Argumentation Framework (AAF) is a tuple* $AAF = \langle A, R \rangle$ *where:* A *is a set of arguments, and* R *is the attack relation on* A *such that* $A \times A \rightarrow R$.

Thus, an *argumentation framework* can be instantiated as a directed graph, where nodes are arguments and edges are attack relations between arguments. This representation eases the computational encoding of an argument-based reasoning. However, *argumentation frameworks* are just data structures and representations and do not enable an analysis of their underlying reasoning *per se*. The set of (topo)logical rules or conditions that make it possible to carry out the analysis of an argument that is instantiated into an *argumentation framework* are the *argumentation semantics*. Through the semantics, it is possible to determine the set of *acceptable* (and *defeated*) arguments. In this paper, we emphasise the fundamental properties behind *argumentation semantics*, but a thorough review of the most important semantics is conducted in [38]. This way, the *argumentation semantics* defines the conditions required to determine the set of *acceptable* (and *defeated*) arguments belonging to an *argumentation framework*. These conditions rely on two basic properties that are related to sets of (abstract) arguments: the conflict-free principle, and the principle of admissibility.

Definition 5 (Conflict-free) *Let* $AF = \langle A, R \rangle$ *be an argumentation framework and* $Args \subseteq A$. *The set of arguments* $Args$ *is conflict-free iff* $\neg\exists\alpha_i, \alpha_j \in Args : (\alpha_i, \alpha_j) \in R$.

Definition 6 (Admissible) *Let* $AF = \langle A, R \rangle$ *be an argumentation framework and* $Args \subseteq A$. *The set of arguments* $Args$ *is admissible iff* $Args$ *is conflict-free, and* $\forall\alpha_i \in Args, \neg\exists\alpha_k \in A : (\alpha_k, \alpha_i) \in R$, *or* $\exists\alpha_k \in A : (\alpha_k, \alpha_i) \in R$ *and* $\exists\alpha_j \in Args : (\alpha_j, \alpha_k) \in R$ *(i.e,. defends* $Args$*).*

This way, it is possible to define a conflict-free set of arguments whenever no attack relations can be observed among the arguments included in the set, and an admissible set of arguments whenever the arguments belonging to a conflict-free set also defend themselves from external attacks. It is important to point out that admissible sets of an AF are always among the conflict-free sets of the same AF. Let us illustrate these formal definitions with an example. Assuming a situation where four different arguments are encoded in an AF, i.e., $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in A$; and the relations $(\alpha_1, \alpha_2), (\alpha_2, \alpha_3), (\alpha_3, \alpha_4), (\alpha_4, \alpha_3) \in R$; it could be possible to define two groups of acceptable arguments depending on which principle is brought into consideration. The conflict-free sets of arguments are $\{\alpha_1, \alpha_3\}, \{\alpha_1, \alpha_4\}$, and $\{\alpha_2, \alpha_4\}$ since there are no attack relation among the arguments included in these sets. In contrast, the admissible set of arguments would only be $\{\alpha_1, \alpha_4\}$ because only these two arguments are conflict-free and are able to defend themselves from external attacks.

From these properties, two major families of semantics for abstract *argumentation frameworks* arise, conflict-free and admissibility-based semantics. Some significant examples of these semantics are Complete, Preferred, Grounded, and Ideal for admissibility-based semantics, and Naïve, Stage, and CF2 for conflict-free based semantics (see [38] for more detail in their formalisation and properties). Depending on each domain and/or the nature of the encoded argument, the suitability of *argumentaton semantics* can differ. However, in general, the admissibility principle is of the utmost importance when defining consistent sets of arguments from a framework since they can defend themselves.

Finally, in order to completely understand the experimentation carried out in this work, it is important to introduce the Argumentation Framework for Online Social Networks (AFOSN). This framework was originally proposed in [324] as the basis of an argumentation system aimed at the prevention of privacy threats in online environments. Its underlying mechanism is based on the theory behind the QBAFs [43] and allows the acceptabilty of an abstract argument to be determined depending on a quantitative feature. For this purpose, in addition to abstract arguments and attacks, the AFOSN relies on information that is extracted from the social network (i.e., publication features and user profiles) and on an argument scoring function for determining the acceptability of the arguments. The AFOSN is formally defined as follows:

Definition 7 (Argumentation Framework for Online Social Networks) *We de-fine an argumentation framework for online social networks as a tuple $AFOSN =$ $\langle A, R, P, \tau \rangle$ where: A is a set of n arguments $[\alpha_1, \ldots, \alpha_n]$; R is the attack relation on A such that $A \times A \to R$; P is the list of e profiles involved in an argumentation process $[p_1, \ldots, p_e]$; and τ is a function $A \times P \to [0, \ldots, 1]$ that determines the score of an argument α for a given profile p.*

An argument $\alpha \in A$ is instantiated by the framework as a 3-tuple $\alpha = (\beta, T, D)$: β represents the claim (i.e., +1 if the argument is in favour and -1 if the argument is against sharing); T indicates the type of the argument (i.e., Privacy, Risk, Trust and Content); and D encodes the support of the argument (i.e., a numerical value distilled from the Online Social Network environment). Each user profile $p \in P$ is also instantiated as a 3-tuple $p = (\nu, \rho, M)$ where the preference values ν, the personality of a user profile ρ and a set of general information M (e.g., age, likes, statistics) are used to model human users. Finally, the argument scoring function τ is defined as follows:

$$(10.1) \qquad \tau(\alpha, p) = \alpha_\beta \cdot \alpha_D \cdot p_{\nu_i}$$

The resulting product of the claim, the support of the argument, and the prefer-ence value of a specific human user towards each topic will determine the strength of an argument in the AFOSN. Then, it is possible to define defeat for an argument as follows:

Definition 8 (Defeat (AFOSN)) *An argument $\alpha_i \in A$ defeats another argument $\alpha_j \in A$ in a context determined by a user profile p iff $(\alpha_i, \alpha_j) \in R \land |\tau(\alpha_i, p)| > |\tau(\alpha_j, p)|$.*

The collective defeat for a set of arguments w.r.t. another set of arguments is defined as follows:

Definition 9 (Collective Defeat (AFOSN)) *The set of arguments $Args_i \subset A$ de-feats the set of arguments $Args_j \subset A$ in a context determined by a user pro-file p iff $\forall \alpha_i \in Args_i, \forall \alpha_j \in Args_j, (\alpha_i, \alpha_j) \in R \land \sum_{\forall \alpha_i \in Args_i} |\tau(\alpha_i, p)| > \sum_{\forall \alpha_j \in Args_j} |\tau(\alpha_j, p)|$.*

Thus, from these defeat definitions, it is possible to define acceptance (considering defeat) and collective acceptance (considering collective defeat) in an AFOSN:

Definition 10 (Acceptance (AFOSN)) *An argument $\alpha_i \in A$ is acceptable iff $\forall\, \alpha_j \in A \wedge defeat(\alpha_j, \alpha_i) \rightarrow \exists \alpha_k \in A \wedge defeat(\alpha_k, \alpha_j)$ or $\forall\, \alpha_j \in A \wedge \nexists\, defeat(\alpha_j, \alpha_i)$.*

Definition 11 (Collective Acceptance (AFOSN)) *The set of arguments $Args_i \subset A$ is acceptable iff $\neg\exists\, Args_j \subset A$; $Args_i \cap Args_j = \emptyset \wedge defeat(Args_j, Args_i)$.*

It is important to emphasise that collective defeat and collective acceptance are the core of an AFOSN since there will always be two sets of arguments, one in favour of sharing and one against doing it. This way, the AFOSN will result in a bipartite graph, granting the properties of conflict-freeness and admissibility to the acceptable arguments defined under collective acceptance.

Argument-based Persuasive Framework

Abstract argumentation frameworks and semantics provide the formal tools to encode human argumentative reasoning from a computational viewpoint. However, most of the research in formal argumentation focuses on proposing models for approaching argumentative reasoning instead of deepening the focus on how the output of the underlying reasoning could be used in a direct human-computer interaction. In this paper, we formalise the Argument-based Persuasive Framework as a higher-level framework that enables Human-Computer Interaction and that can be instantiated on top of any abstract argumentation framework. Our proposal brings into consideration any underlying formal argumentation framework that is in charge of approaching the argumentative reasoning, a human user model for personalising and adapting the interaction, and a set of linguistic features to concretise the abstract arguments. All of these features are combined by a persuasive function as described below:

Definition 12 (Argument-based Persuasive Framework) *We define an Argument-based Persuasive Framework as a tuple $APF = \langle AF,\ U,\ L,\ \gamma \rangle$ where: AF is the underlying argumentative framework; U is the human user*

*model; L is a set of linguistic features; and γ is a persuasive function that pro-
duces a persuasive Natural Language Argument (NLA) such that $U \times Args \times L$
$\rightarrow NLA$.*

Each user model U contains a set of user descriptive features (e.g., personality,
behavioural patterns, or emotions) that may vary depending on the application
environment and domain, and the availability of such features. The set of linguistic
definitions L (e.g., argumentation schemes, argument templates or databases, or
logical structures) contains different non-abstract representations of the arguments
that are included in the argumentation process. Finally, the γ function is aimed at
estimating the most persuasive natural language argument given a set of arguments
and natural language features for a specific user profile:

$$(10.2) \qquad\qquad \gamma(U, Args, L) = \hat{Ar}$$

which takes as input the user descriptive features associated with a human pro-
file U, the set of acceptable arguments $Args \in A$ (where A is the argument set of
the underlying AF), and the set of linguistic features L, to produce a persuasive
argument $\hat{Ar} \in |NLA|$ belonging to the domain of Natural Language Arguments.
Using the APF, a new dimension to formal computational argumentation research
can be unlocked. This framework makes it possible to leverage the computational
argumentative reasoning provided by any argumentation framework (which may
vary depending on our needs, the application domain, or the available information)
for defining better informed persuasive strategies through the use of arguments and,
thus enable an effective argument-based HCI.

10.4 Implementation of the Argument-based Persuasive Framework

To validate our formal proposal and to depict how the Argument-based Persua-
sive Framework can be instantiated and implemented in a real situation, we have
chosen the domain of privacy management in Online Social Networks (OSNs).
Privacy violations in OSNs are a threat of major concern that has been thoroughly

researched in the literature. Different viewpoints and approaches can be identified
when dealing with this problem, e.g., automatic agent-based negotiations [195],
privacy nudges [1], persuasive argumentation systems [324], and the multi-party
privacy conflict [250], among others. As discussed in Section 10.3, an AFOSN
provides the underlying reasoning mechanism of an argumentation system that
is aimed at identifying and preventing privacy violations in OSNs [324]. In this
paper, we retake this domain to instantiate the Argument-based Persuasive Frame-
work (APF) on top of the AFOSN and to evaluate its power of behaviour change
when preventing privacy violations.

For that purpose, we instantiate the APF (i.e., $\langle AF, U, L, \gamma \rangle$) as follows:

- The computational argumentative reasoning engine (AF) is managed by an
  AFOSN. Whenever a new post is being shared in the network, it generates a
  set of abstract arguments from the data retrieved from the OSN as described
  in [322]. Then, the set of acceptable arguments is defined (see *Collective
  Acceptance*, Definition 11) to determine if a potential privacy violation is
  happening.

- The user model (U) is instantiated tacking into account two different help-
  ful aspects for user behaviour modelling: the Big Five personality traits
  model [317] (i.e., Openness, Conscientiousness, Extraversion, Agreeable-
  ness, Neuroticism), and their OSN interaction data. As proven in previous
  research [323], both aspects are helpful in identifying variances in the per-
  ceived persuasive power of arguments and reasoning patterns.

- The set of linguistic definitions (L) enables the natural language represen-
  tation of the abstract arguments provided by the argumentation framework.
  In our experiments, we consider the four argument types supported by the
  AFOSN (i.e., Privacy, Risk, Trust, and Content) and five different argumen-
  tation schemes [400] (i.e., patterns of human reasoning) in order to define
  a database of 45 domain-specific natural language arguments. We selected
  five commonly used argumentation schemes that suited our application do-
  main and that were researched in previous studies [323]: the *Argument from
  Consequences* (AFCQ), the *Argument from Expert Opinion* (AFEO), the *Ar-*

gument from Popular Practice (AFPP), the *Argument from Popular Opinion*
(AFPO), and the *Argument from Witness Testimony* (AFWT).

- The persuasive function (γ) is approached in two steps: persuasive policy
learning and natural language argument generation. This way, in our approach, we first estimate a persuasive policy for each specific user, and then
we generate natural language arguments by combining the predicted policies
and the argumentative linguistic definitions. Both steps are described in the
following sections.

Persuasive Policy Learning

The Persuasive Policy Learning Task.

Our first step for approaching the γ function is to learn user-specific persuasive
policies. For that purpose, we need to consider both the user model U and the
linguistic definitions L. Furthermore, depending on the content and nature of any
privacy threatening publication, the set of coherent arguments may vary (e.g., if
a publication does not involve more than one person, it would not be acceptable
to argue against sharing the publication by reasoning that some other user that
appears in the publication could be offended). Our objective is to be able to always
use the most persuasive coherent argument for each given author of any conflicting
publication. For this purpose, we need to estimate the persuasive policies π^s and π^t
for the whole set of argumentation schemes (s) and argument types (t) considered
in this work, respectively. We define a persuasive policy $\pi \in \mathbb{R}^l$, where l are
argumentative features in L, as follows: $\pi = [\alpha_1, \alpha_2, \ldots, \alpha_{|l|}]$, where $pp(\alpha_1) \geq$
$pp(\alpha_2) \geq \ldots \geq pp(\alpha_{|l|})$ being $pp(\alpha)$ the perceived persuasive power of an argument
α by a human user U. We consider two different sets of linguistic features L: five
argumentation schemes ($l_s = 5$) and four argument types ($l_t = 4$). Furthermore, we
use the persuasive power definition presented in [323], where the persuasiveness
of an argument is represented as a quantitative score based on the position of each
argument in a persuasive ranking indicated by human users. Thus, our persuasive
policies are represented as lists with orderings of arguments based on their assigned
persuasive power.

In this work, we model the persuasive policy learning as a maximisation of the conditional probability described in Equation 10.3. For each user model U, we need to estimate the probabilistic distributions of the persuasive power of both the argumentation schemes π^s and argument types π^t.

$$(10.3) \qquad \hat{\pi}_U^{s,t} = \arg\max_{j \in J} P(\pi_j | U)$$

where J is the total number of possible combinations for a given set of linguistic features (i.e., $5!$ for the argumentation schemes, and $4!$ for the argument types). Then, each user U is modelled by combining the two features described above (i.e., Big Five and OSN interaction data), which will be the input for the probabilistic models in our experiments. To sum up, we approach the persuasive policy learning as a pattern recognition task. The goal is to identify any existing pattern in the different user models that allow us to determine the optimal privacy policy for each specific user model.

The OSNAP-400 Dataset.

To learn user-specific persuasive policies and to approach this task as the probabilistic modelling proposed in Equation 10.3, we have developed a new dataset for argument persuasion. A total of 400 adults (194 males, 206 females) from 18 to 76 years old completed a study designed for the creation of the Online Social Network Argument Persuasion (OSNAP-400) dataset[1]. This study was aimed at adult OSN users. The study from which we created the OSNAP-400 dataset consisted of the 50-item personality inventory [154], two persuasive questionnaires for argumentation schemes (*Questionnaire A*), and argument types (*Questionnaire B*), and an OSN interaction questionnaire (*Questionnaire C*). In the persuasive questionnaires, the participants had to order the arguments (i.e., schemes and types) displayed in a randomised way based on their perceived persuasiveness. Furthermore, we included attention check questions in all of the questionnaires in order to validate their submissions.

---

[1]Contact the authors for data availability inquiries.

For the elaboration of the OSNAP-400, we first calculated the Big Five personality traits of all of the participants from the results of the 50-item personality test. Then, with the answers provided in *Questionnaires A* and *B*, we also calculated the persuasive power of the five argumentation schemes and the four argument types following the definition presented in [323], from which we generated the ground truth persuasive policies for each specific user. Finally, we encoded the OSN interaction answers of *Questionnaire C* to discrete normalised values in the range from 0 to 1. Thus, the OSNAP-400 consists of 400 samples. Each sample of the dataset represents a different OSN user modelled with the Big Five and the OSN interaction data and is associated with two persuasive policies (one for argumentation schemes π^s, and the other for argument types π^t).

Before approaching the persuasive policy learning task, we conducted a descriptive analysis of the OSNAP-400 data. First of all, we analysed the user descriptive features (see Figure 10.1). For the OCEAN Big Five personality traits (i.e., OCEAN stands for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) of our samples, we observed almost all of the possible values in the allowed range for every trait (see Figure 10.1a). However, we also were able to observe that Extraversion and Neuroticism traits tend to have lower values than the rest in our dataset. For the social network interaction data, we included twelve different user modelling features that represent the online behaviour of each human user: the number of friends, the number of status updates, the number of likes, the number of comments, the number of publications shared in private, the number of publications shared in public, the number of publications shared with friends only, the number of publications shared with a specific collection of friends, the number of publications deleted, the number of photos uploaded, the average length of the text in the publications, and the average time spent using OSN. Some interesting insights can be observed: how users prefer to share content with friends rather than the whole network; that it is easier for users to give likes than to comment on other users' publications; and that there is an important number of publication regrets that lead to deleting the previously shared content (see Figure 10.1b). Furthermore, it is important to emphasise that the age distribution of the samples used in our experiments is not uniform (see Figure 10.1c); most of the samples are within in the 22-34 age interval. Finally, for the gender distribution, we have a balanced population of 194 male samples and 206 female samples

(see Figure 10.1d).

We also analysed the distribution of the observed persuasive policies π^s and π^t in the OSNAP-400, in order to describe how balanced the dataset is. Figure 10.2 depicts the frequency at which each persuasive policy appears in the dataset. We observed that regardless of being argumentation schemes or argument types, there is a very strong imbalance in the data. We found that the most frequent persuasive policy of argumentation schemes (with a total of 22 occurrences) was the following one: AFCQ > AFPO > AFEO > AFWT > AFPP. It was closely followed by the second most frequent persuasive policy for argumentation schemes with (21 occurrences): AFCQ > AFEO > AFPO > AFWT > AFPP. We observed how the arguments from consequences are in general perceived to be the most persuasive pattern of human reasoning in our domain. On the other hand, regarding argument types, we observed that the most frequent persuasive policy (with a total of 60 occurrences) is dominated by the arguments containing content references: *Content > Trust > Privacy > Risk*. The strong imbalance observed between the existing persuasive policies of argumentation schemes and argument types makes the persuasive policy learning a hard task to perform a probabilistic modelling on, as the following section shows.

Experimental Results.

Finally, we present the results obtained in the proposed persuasive policy learning task. For that purpose, we trained five different models to predict how a given user should perceive the persuasive power of both argumentation schemes and argument types and generate the subsequent user-specific persuasive policies π^s and π^t. Considering the probabilistic modelling defined in Equation 10.3, the user modelling features were used as the input for our models, and an optimised persuasive policy was generated as the output. Based on the findings of a previous study on the persuasive power of arguments in the OSN domain [323], we modeled our users by combining their Big Five personality traits together with twelve different features that represent their social behaviour in online environments.

Thus, four classical machine learning algorithms have been used in our persuasive policy learning experiments: Support Vector Regression, Stochastic Gradient Descent Linear Regression, K-Neighbours Regression, and Random Forests.
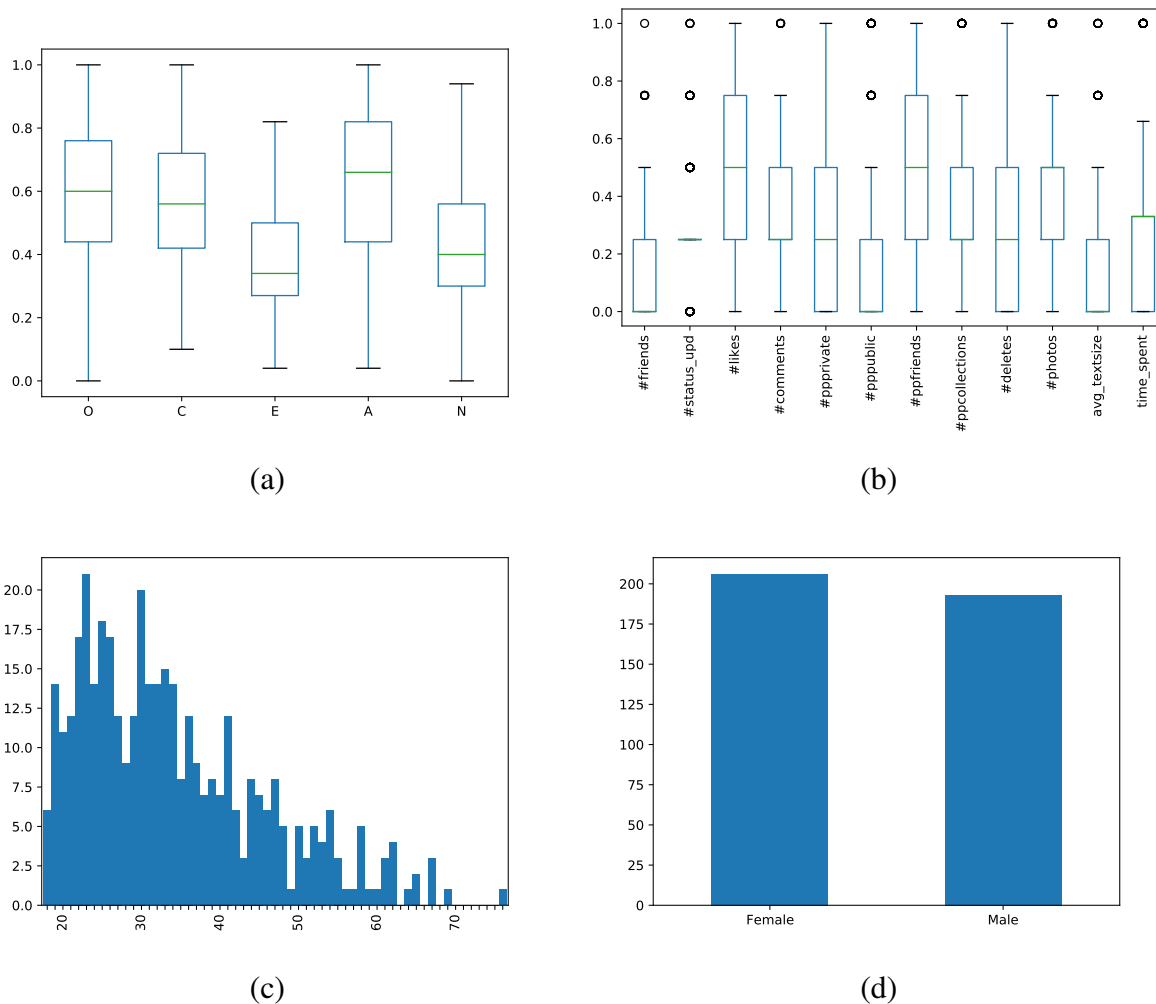
Figure 10.1: (a) Box and whiskers diagram of the OCEAN Big Five personality traits observed among the samples of the OSNAP-400 dataset. (b) Box and whiskers diagram of the OSN interaction data observed among the samples of the OSNAP-400 dataset. (c) Age distribution of the OSNAP-400 dataset samples. (d) Gender distribution of the OSNAP-400 dataset samples.

(a)

(b)
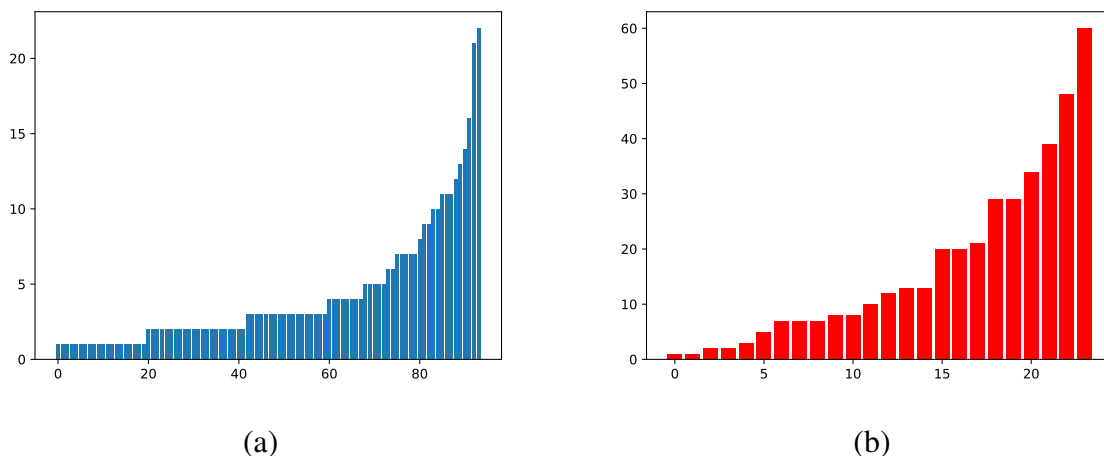
Figure 10.2: Distribution of the number of occurrences of the observed persuasive policies. Figure (a) stands for argumentation schemes and Figure (b) for argument types. The Y axis represents the number of occurrences of each different persuasive policy. The X axis represents each different observed persuasive policy. Each policy is represented by a unique *id* from 0 (the least frequent) to N-1 (the most frequent), being N the number of different persuasive policies observed in our data.

Support Vector Regression (SVR) [122] is a Maximum Margin Regression model which has shown good performance in a wide variety of tasks. After optimising its hyperparameters, we used the linear kernel, C = 100 and 1e-9 tolerance values. Stochastic Gradient Descent Linear Regression (SGDLR) [64] is a technique by which a linear model is optimised with stochastic gradient descent on minimising a regularised empirical loss. In our experiments, we obtained the best results minimising the huber loss function with a 1e-3 tolerance value and a 1e-5 alpha. *K*-Nearest Neighbours Regression (*k*-NNR) is a regression method that is based on the *k*-Nearest Neighbours algorithm [106]. The estimated value for an unobserved sample is based on the *k* samples that are the closest to it. In our experiments, we considered the 32 nearest neighbours weighted by their distance to the new observation. The last classical approach considered in this work are Random Forests [67]. A random forest is a meta-learning technique which fits a specific number of decision trees on different subsets of the original dataset. In our experiments, we used 10000 decision trees to estimate the value that minimises the Mean Absolute

Error loss for each tree split. We used the *sklearn*[2] implementations of all of the described classical machine learning algorithms.

In addition to these four classical machine learning models, we also experimented with a neural network model. We implemented a feed-forward Multi-Layer Perceptron (MLP) to approach the persuasive policy learning task. The chosen architecture for our model consists of 3 hidden layers (32, 32 and 64 units per layer) with *ReLU* activation functions and a total amount of 4196 parameters. The input layer has as many units as the size of our input (i.e., 17 user modelling features). The output layer has 4 or 5 units depending on the persuasive policy being learnt (π^t or π^s, respectively) and a *sigmoid* activation function.

The performance results of the described models on the persuasive policy learning task are depicted in Table 10.1. In addition to the five models, we also considered two baselines: a random baseline and a majority baseline. First, the random baseline assigns a random persuasive power (i.e., a value in range [0,1]) to each one of the arguments and generates a persuasive policy by ordering them by their randomly assigned persuasive power. Second, the majority baseline uses the most common persuasive policy of both argumentation schemes and argument types for all users regardless of their descriptive features. Three different metrics were used to evaluate different aspects of the persuasive policy learning task: the Mean Absolute Error (MAE, lower is better), the Hit Rate (HR, higher is better), and the Spearman ρ correlation (higher is better). These are common metrics that are used to evaluate recommendation systems with similar requirements [160]. The MAE indicates the quality of the model predictions tacking exclusively into account the persuasive power estimations of each individual argument. However, it is not possible to draw significant conclusions about the performance on the persuasive policy learning task considering the MAE alone. The Hit Rate (HR) measures the number of *hits* observed in the predicted persuasive policies. We consider a *hit* to be whenever an argument (scheme or type) is correctly placed in the predicted persuasive policy compared to the ground truth persuasive policy for a given human user. This metric is most revealing when it comes exclusively to the performance of our models in the persuasive policy learning task. Finally, to complement the previously described metrics, we also considered the Spearman ρ correlation mea-

---

[2]https://scikit-learn.org/stable/index.html

sure between predicted and ground truth persuasive policies. With the Spearman ρ metric, it is possible to evaluate how good the models are at learning partial orderings in the predicted persuasive policies. For example, assuming the ground truth persuasive policy $\pi_u = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ and the estimated persuasive policy $\pi'_u = [\alpha_2, \alpha_1, \alpha_4, \alpha_3]$, then $\text{HR}(\pi'_u) = 0$ but $\rho(\pi'_u) = 0.6$, since the estimated persuasive policy does not have any argument placed in its correct position, but the persuasive partial orderings of arguments are decently estimated. This way, it is possible to understand how well the models are performing, not only when predicting persuasive policies, but also when predicting the individual persuasive power of arguments, and retaining partial ordering dependencies between different arguments.

Table 10.1: Results obtained on the persuasive policy learning task (Schemes π^s/ Types π^t). The depicted results represent the average of a 10-Fold evaluation.

Model	MAE (π^s / π^t)	Hit Rate (π^s / π^t)	Spearman ρ (π^s / π^t)
Random Baseline	0.32/0.31	0.22/0.23	-0.02/0.04
Majority Baseline	-	0.20/0.19	**0.22**/-0.01
SVR	**0.16/0.17**	**0.34**/0.38	**0.10/0.09**
SGDLR	0.17/0.17	0.32/0.38	0.06/0.07
k-NNR	0.17/0.17	0.32/**0.40**	0.04/0.07
RandomForest	0.17/0.17	0.33/0.38	0.08/0.06
MLP	0.18/0.18	0.33/0.38	0.09/0.05

It can be observed in Table 10.1, in general, the models perform better than the proposed baselines. Furthermore, it can also be observed that all of the models perform similarly after a 10-Fold evaluation using the OSNAP-400 dataset. We attribute this behaviour to model convergence and a limited size of training sam-

ples. However, the proposed models achieved an improvement with respect to the baselines of 42%-50% regarding the prediction of the individual persuasive power of arguments (i.e., MAE), an improvement of 54%-110% regarding the accuracy when estimating persuasive policies (i.e., HR), and an improvement of 125% when learning partial orderings in the estimated persuasive policies (i.e., Spearman ρ). These results are reported when learning persuasive policies for both argumentation schemes and argument types (π^s and π^t, respectively). An exception in the Spearman ρ performance of the majority baseline for argumentation scheme persuasive policy estimation can also be observed. It presents outstanding results compared to the rest of approaches. This may be because of the data distribution of ground truth persuasive policies of argumentation schemes (see Figure 10.2a), where the most common occurrences are slight variations preserving similar partial orderings. However, it performs significantly worse than the rest of the models regarding the Hit Rate, even worse than the random baseline. Thus, even though it outperforms our models when learning partial orderings of the persuasive policies, it is not a solid alternative to bring into consideration when approaching the persuasive policy learning task.

Natural Language Argument Generation

Our second step in this work is the generation of natural language arguments. Once we have computed the user-specific persuasive policies ($\pi_U^{s,t}$), we need to be able to automatically generate a natural language argument for each abstract argument produced by the AFOSN in order to persuade the human user. For that purpose, we defined a database of 45 natural language arguments by combining the four types of arguments supported by the AFOSN with the five argumentation schemes selected for the OSN domain. This way, the persuasive function γ takes into account the user model U, the set of acceptable arguments provided by the AFOSN $Args$, and the set of linguistic features L.

Our approach is then able to generate a different natural language argument for each user model depending on the predicted privacy policies (both π^s and π^t for argumentation schemes and argument types, respectively). As depicted in Figure 10.3, when engaging a persuasive interaction with a human user, our system selects from the argument database the most (potentially) persuasive argument consider-
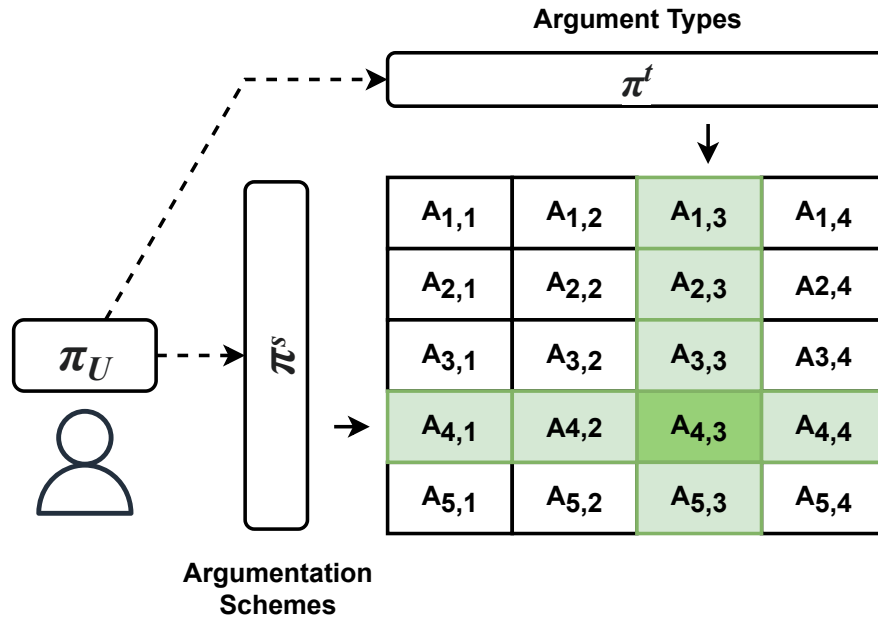
Figure 10.3: Scheme of the proposed natural language argument generation method.

ing the persuasive policy estimations. Thus, our argumentation system retrieves the natural language argument tacking into account the most persuasive argumentation scheme (rows) and the most persuasive argument type (columns) from the set of acceptable arguments. Our proposed method for generating natural language arguments only considers arguments that are coherent with each privacy threatening situation. Therefore, the argumentation system will select the most persuasive argument type provided by π^t, from only the ones that are included in the set of acceptable arguments $Args$ produced by the AFOSN (see Definition 11). Thus, we avoid the problem of using arguments that are not coherent with a situation where a potential privacy violation is happening and whose persuasiveness would be nil. The persuasive aspect related to coherence is therefore granted by the underlying computational argumentative reasoning.

10.5 Persuasive & Behaviour Change Evaluation

To evaluate the persuasive power of the arguments generated by our Argument-based Persuasive Framework w.r.t. behaviour change, we have designed a study that is divided into two stages. The APF is used to persuade OSN users in order to prevent potential privacy violations. In the first stage, we collect user modelling inputs (i.e., personality traits and OSN behaviour); in the second stage, we measure the persuasive power of the arguments generated by our APF by considering the user modelling inputs and comparing them with a random selection method. For this purpose, a set of abstract arguments is generated for each potential privacy-threatening publication using an AFOSN, and its semantics are used to determine the set of acceptable arguments. Then, the persuasive γ function is used to improve the persuasiveness of the argumentative reasoning provided by the argumentation framework. In view of the results of the persuasive policy learning task, we decided to use the SVR model to estimate the optimal persuasive policies for the users who were participating in our evaluation.

To analyse the influence the content of the post on the persuasive power of the arguments, six different types of content were included in the experiment: location, medical, alcohol/drugs, personal, family/association, and offensive.

Participants

For this experiment, 50 participants (25 male and 25 female) ranging in age between 18 and 44 years old ($\mu = 25.72$, $\sigma = 5.18$) were recruited. We required the participants to have experience using at least one social network.

In order to keep parity between age and gender, we divided the participants into two groups: experimental and control. The experimental group consisted of 30 participants (15 males and 15 females) ranging in age between 20 and 33 years old ($\mu = 25.87$, $\sigma = 4.22$). The control group was composed of 20 participants (10 males and 10 females) ranging in age between 18 and 44 years old ($\mu = 25.5$, $\sigma = 6.48$).

Figure 10.4: Experiment layout.

Materials

For the first stage concerning the acquisition of user modelling inputs, we designed
an online questionnaire that was composed of two sections. In the first section, we
asked the participants to answer a set of questions based on the 50-item person-
ality inventory [154] along with three attention check questions using the same
questionnaire as in Section 10.4; in the second section, we asked the participants
to complete the OSN interaction questionnaire (*Questionnaire C* described in Sec-
tion 10.4) along with one attention check question.

In the second stage, in which the persuasive power of the arguments generated
by our argument-based persuasive framework was evaluated, we designed an on-
line questionnaire composed of fourteen sections. In each section, a scenario in
which a post (consisting of a message and an image) containing sensitive material
that could violate the user's privacy was presented (see Figure 10.4). The post was
followed by an argument that attempted to convince the user to modify the original

post in order to preserve his or her privacy. To evaluate the persuasive power of the argument, the participants were asked whether or not they would publish the post after reading the argument and also their degree of trust regarding this decision. To capture the degree of trust, we used a 5-item Likert scale ranked from "not very convinced" to "very convinced". To measure the impact of the arguments on the participants' decisions, at the end of the section, the participants were asked whether or not the argument had influenced their decision.

There were fourteen sections in total: two sections were for attention monitoring, and twelve sections represented the six types of arguments (two sections per type of argument content) that were randomly distributed. The sections dedicated to attention monitoring followed a similar pattern to the twelve sections in order to determine if the participants were actually reading the questions carefully and not answering randomly.

With regard to the selection of the arguments to be presented to the participants during the second stage of the experiment, the experimental group received arguments that were generated by the Argument-based Persuasive Framework. The control group received arguments whose reasoning pattern was randomly chosen and instantiated to natural language. Likewise, the type of argument was also randomly selected, but only those types that made sense with the context of the question were considered.

Procedure

The two stages of the experiment were performed on different days to avoid biases. At the beginning of each stage of the experiment, the participants were provided with the instructions describing the task to be accomplished. Then the participants were asked to complete the questionnaires without a time limit.

Results

The results of the experiment show differences between the control group and the experimental group when making the decision of whether or not to publish a post on a social network. Thus, we observed that, in the control group, the participants who chose to modify the post after reading the argument reported that the argument

had influenced their decision (30.41% of the group). This result contrasts with the 37.7% obtained in the experimental group. Therefore, by personalising the arguments to the users' characteristics, we obtained better effectiveness in modifying their behaviour. To analyse the statistical difference in the participants' behaviour according to the arguments in the two groups, we performed a chi-square test. The results of the analysis show significant statistical evidence between the control group and the experimental group with a chi-square value of 10.57 and a p-value of 0.014 (for a critical value of 7.82 and 3 degrees of freedom). These results also confirm that arguments that are generated according to user-specific persuasive policies improve the persuasiveness of an argumentation system.

With regard to the type of content of the arguments, we found that, in general, there was a greater change in user behaviour in the experimental group compared to the control group in five of the six types analysed (all except personal content). In the sections related to medical content, 28.33% of the participants in the experimental group modified their behaviour after being influenced by the argument compared to 15% of the control group. The same can be observed for the offensive content, where 66.67% of the participants of the experimental group modified there behaviour compared to 55% of the control group. For family/association and alcohol/drugs, the experimental group was influenced by the argument (26.67% and 35%, respectively), while the control group was only influenced by 17.5% and 22.5%, respectively. However, in the case of personal content, we found that 48.88% of participants in the experimental group modified their behaviour after being influenced by the argument versus 50% in the control group. This may be due to the sensitivity of the content of the post. We observed that the in the experimental group the posts related to personal content and to offensive content were more sensitive since, in general, the participants modified their behaviour (49% and 62%, respectively). In contrast, the medical content, the family content, and the location content showed less sensitivity and less probability of behavioural change influenced by an argument (23%, 23%, 17%, respectively).

With regard to the level of trust, we found that the mean of the degree of trust that users showed when modifying their behaviour based on an argument was $\mu = 4.23$ (with $\sigma = 0.85$) out of a maximum of 5. In contrast, the mean of the degree of trust of the participants who decided not to modify their behaviour was only $\mu = 2.58$ (with $\sigma = 0.81$). This is an interesting result, which indicates that

the use of arguments to persuade users' behaviour reinforces their degree of trust in their decision when modifying their behaviour on a social network. These results highlight the importance of research into the use of persuasive argumentation systems in applications that seek to study, interpret, or modify human behaviour.

10.6 Discussion

Abstract argumentation frameworks have been extensively used in the field of computational argumentation to encode argumentative data and to approximate argumentative reasoning through the use of argumentation semantics. Research on this topic has been focused on proving and refuting logical properties and formulae, rather than extending their functions to other areas such as natural language processing or computational persuasion.

The ideas of extending formal computational argumentation concepts to the area of computational persuasion have been explored in recent research [177]. The authors propose a general framework for computational persuasion for behaviour change applications where computational argumentation is introduced as a promising approach to solve this problem. A complete analysis of the existing research and proposed techniques is done, but no specific proposal or implementation is presented. Some of these ideas are further developed in [164]. However, argumentation frameworks are considered to be mere graph data structures, and argumentation semantics are removed from the computational argumentative reasoning process. Thus, it is not possible to explore the benefits of combining the coherence and rationality provided by argumentative reasoning together with personalised persuasive interactions that are aimed at behaviour change. A more ambitious effort at combining aspects from formal computational argumentation theory and computational persuasion is done in [316]. The authors propose a persuasive agent that approaches argumentative reasoning through a weighted argumentation framework and its quantitative semantics. Arguments are then used in a dialogue with human users following strategies learnt by a partially observable Markov decision process. The results achieved by the agent show 20% of cases where human users decided to change their behaviour. However, a small population was used to evaluate the argumentative agent (i.e., 15 participants).

In order to overcome the identified limitations, we have proposed a gener-
alised framework for extending formal computational argumentation techniques
to the area of computational persuasion. The main contributions of our proposal
are twofold. First, we have formalised a general framework for argument-based
computational persuasion that is designed to work with any underlying argumen-
tation framework considering different user models. Our APF is not constrained
to any specific argumentation framework, semantics, or user model, and it can be
instantiated on top of any computational argumentative algorithm that provides a
set of acceptable arguments, regardless of the domain or how the algorithm is ap-
proached (i.e., quantitative or qualitative). Furthermore, the APF also includes a
persuasive function that is not constrained to any specific implementation. It is
important to emphasise that our approach to the persuasive function γ is not the
only valid one. Throughout Sections 10.4 and 10.4, we presented an implemen-
tation proposal of the γ of the APF's that is formally defined at the beginning of
this paper. However, other approaches for generating a natural language argument
from the set of acceptable arguments of an argumentation framework can also be
proposed. The only requirement is that the γ function approach must take into
account a user model and a set of linguistic features in addition to the acceptable
abstract arguments. Second, we provide a complete implementation of the APF in
a real case study and a persuasive evaluation with real human users. In our pro-
posal, we model our human users considering two different sets of user modelling
features: personality and online behaviour (e.g., number of friends, comments, or
likes). Through our implementation, it is possible to observe how the different
parameters of the APF need to be instantiated. Furthermore, at the end of our ex-
periments, we validated the proposed persuasive framework since it significantly
improves the persuasiveness of an argumentation system that is aimed at prevent-
ing privacy violations in OSNs.

Compared to previous research, our approach enables the use of computational
argumentative reasoning techniques for approaching and improving the computa-
tional persuasion task. Our proposal and results present a significant contribution
to the user modelling and personalised computational interaction of argumentative
systems. However, there are some limitations in our work. First, the proposed im-
plementation and results of the evaluation are constrained to our domain. We have
implemented the APF for the domain of privacy management in OSNs, and our

implementation cannot be extrapolated to any other different domain. The same goes for the results. The reported improvement in persuasive performance caused by the use of the APF might differ between different domains and implementations. For example, using different user models or taking a different approach to the implementation of the persuasive function γ may result in significant variations of the perceived persuasiveness of our system by human users. Second, our implementation of the APF has been evaluated using a series of one-shot interactions with the users. Our experiments have not been designed to investigate the definition of persuasive strategies in a dialogue but to estimate persuasive policies in order to persuade user with individual arguments.

10.7 Conclusion

In this paper, we have proposed Argument-based Persuasive Frameworks. APFs extend the computational argumentative reasoning provided by argumentation frameworks and enable a persuasive interaction with human users. Thus, an argumentation system can computationally approach human argumentative reasoning through an argumentation framework and its semantics and broaden its purposes to persuasive and personalised interaction with human users. In addition to the definition, we have proposed a use case of the APF that is framed within the domain of privacy management in OSNs, and we have provided a complete implementation of the framework in a real situation. We implemented the APF on top of an argumentation framework that is specifically defined for its use in OSNs (i.e., AFOSN), and we modelled our users taking into account their personality and their online behaviour (e.g., number of friends, comments, or likes). Furthermore, we conducted a persuasive evaluation of our proposal, where we observed that the use of an APF on top of an argumentation framework improves the persuasiveness of the arguments used by the argumentation system during the interaction with human users. We have also observed that the trust placed by human users in an interactive system that provides arguments for behaviour change is really high, meaning that argumentation is a powerful technique for designing trusted and reliable decision support systems. Therefore, the extension of argumentation frameworks for their use in persuasive systems represents a step forward that helps in the convergence

between formal computational argumentation and Human-Computing Interaction research.

With all of these findings, we foresee further research at the intersection of the two research areas of computational argumentation and computational persuasion. Specifically, these include analysing different user models, linguistic features, and persuasive functions, in addition to research on the relation between these variables and the application domain. We also find it important to investigate how the APF could be implemented or extended to interact directly with human users in argumentative dialogues.

Part V

Discussion

Discussion

At the beginning of this thesis, we proposed three main objectives that have been addressed throughout the research work described in the central chapters of this document. Our first objective was to review, analyse, and classify the existing research in computational argumentation in a way that it could be easily followed and understood from the human argumentative point of view. This analysis is presented in Part II, Chapter 2, where we identify three major clusters where the research in computational argumentation can be consistently grouped: argument mining, Argument-based KRR, and Argument-based HCI. This way, argument mining involves all the research aimed at segmenting natural language argumentative inputs, classifying natural language argumentative propositions, and detecting relations and structures between these propositions; Argument-based KRR encompasses all the research that proposes data structures and algorithms for computationally encoding arguments and approaching the logical aspects of human argumentative reasoning (e.g., identifying logical properties such as admissibility, or even estimating the winner of a debate); finally, Argument-based HCI includes all the research focused on improving the interaction with human users through the use of arguments, from the automatic generation of natural language arguments to the study and analysis of the persuasive power of different arguments when used in any argumentative dialogue.

Our second objective was to propose new techniques for the automatic analysis of human argumentative discourses. This objective has been addressed throughout the chapters included in Part III. In Chapter 3, we describe *VivesDebate*, an argumentative corpus that was annotated in the frame of this thesis, and publicly

released to the computational argumentation research community. Compared to the previously existing natural language argumentative corpora, the *VivesDebate* corpus enables the analysis of complete, undivided argumentative debates belonging to a debate tournament. In addition to the typical labels used in argumentation-based NLP, we also released the objective evaluations provided by an impartial jury for each of the debates. The publication of this corpus allows to approach new problems from the natural language viewpoint such as the automatic evaluation of natural language debates, and to provide a new perspective to formal argumentation algorithms in an informal setup. Furthermore, in Chapter 4 and Chapter 5 we propose a new architecture for segmenting and classifying arguments in Japanese language political discussions, and evaluate different Transformer-based architectures on the identification of argumentative relations in English debates respectively. The experiments conducted with the Japanese corpus were framed into the political budget argument mining research project carried out during the visiting research stay at the National Institute of Informatics in Tokyo. Conversely, the experiments described in Chapter 5 were conducted before the creation of the *VivesDebate* corpus, and were essential for the identification of the main limitations of existing natural language argumentative corpora, and the design and annotation of the *VivesDebate* corpus. This way, we performed experiments considering different architectures and algorithms aimed at approaching a complete analysis of natural language argumentative texts. In Chapter 6, we propose an original algorithm for automatically estimating the winner of a complete natural language debate. For that purpose, we combine concepts from NLP and formal argumentation theory and present promising results in this under-researched task. Finally, in Chapter 7, we create and release the largest corpora of argumentative speeches in audio format. The *VivesDebate-Speech* complements the *VivesDebate* corpus with the acoustic information of the debates, which is of utmost importance for the analysis of argumentation. After our successful experiments in argument mining and argumentative analysis in transcribed professional debates, we decided to extend the text features with audio in order to explore an additional dimension that remained unexplored in natural language computational argumentation research (i.e., speech). We approached argument segmentation and classification comparing both text-based and audio-based approaches, and combining them. These experiments leave the door open to a richer approach to argument mining and to the

automatic evaluation of spoken argumentative debates.

Our third and last objective was to study and improve the persuasiveness of interactions between humans and computer systems through the use of arguments and computational argumentation reasoning techniques. The research work in which this objective has been addressed is grouped together in Part IV, where three different research studies and the observed results have been described. First, in Chapter 8, we perform a qualitative analysis of the persuasive properties of argumentation schemes in which we tried to discover if logical structures underlying human argumentative reasoning endorsed any persuasive principle by their own definition rather than their natural language context. This study helped us to understand that some reasoning patterns commonly found in human argumentation are persuading other human users through their logical structure, while in other cases natural language carries most of the weight with respect to persuasion. This knowledge can be useful to define new formal models of persuasive computational argumentation, and to improve the persuasiveness of argumentation systems. Second, in Chapter 9, we conduct a study over teenager participants to evaluate and understand the persuasive power of arguments when used as a means of interaction between the computer and the human. This study was framed into the research projects TIN2017-89156-R, PROMETEO/2018/002, and PID2020-113416RB-I00 aimed at educating teenagers in privacy management and developing explainable persuasive technologies, and integrated into a real online social network with educational purposes. From the results of the study, we could detect significant correlations between the variations of the persuasive power of arguments and user modelling features such as their personality and their online social behaviour. These findings motivated the proposal and development of a persuasion-enhanced computational argumentation system. In Chapter 10, we describe the theoretical framework that serves as a bridge between computational argumentative reasoning and persuasive human-computer interactions. Our proposal was instantiated in the online privacy domain and evaluated with adult human participants that allowed us complement our previous study with teenagers. We were able to observe an important improvement of its persuasiveness compared to other approaches using arguments but without the proposed underlying persuasive framework.

Therefore, with this thesis, we present significant contributions to the research in the whole computational argumentation process. We describe solid advances

and promising results in topics related to argument mining, Argument-based KRR, and Argument-based HCI. With our proposals, we have been able to establish connections between the argument-based NLP and the formal argumentation theory areas of research, and between the formal argumentation theory and computational persuasion research. With these transversal approaches, we have not only been able to establish the mentioned connections, but also to approach new problems that were previously not explored in the literature (e.g., the automatic analysis and evaluation of complete natural language argumentative debates) as well as to improve existing algorithms and techniques for the requirements of the tasks approached in this thesis. Furthermore, we have released a completely new corpus that improved the existing available resources in size and in the quality and detail of annotation, enabling a deeper analysis of natural language argumentation in both text and speech.

Nevertheless, some limitations can also be identified. This thesis presents a broad collection of research experiments, studies, analyses, and results that represent a starting point to the development of a complete software engine for the automatic analysis of natural language argumentative discourses, but further investigations should be carried out before releasing such a complete software. Moreover, some of our contributions have suffered from data limitations such as the size of the available resources, their language, or their own strongly unbalanced distributions. Finally, some of the scientific studies carried out in Part IV are very useful to improve our knowledge and understanding of the human behaviour when interacting with them using different arguments, but conducting complementary experiments in real use cases with a more direct engagement, instead of relying exclusively on laboratory experiments, could improve the significance of our findings. Some of these identified limitations are proposed as future work in the final chapter to continue on the line of research started with this thesis.

Part VI

Conclusion and Future Work

Conclusion and Future Work

"This party's over." – Mace Windu, Jedi Master.

Argumentative reasoning plays a fundamental role in human intelligence and communication. Studied by classic philosophers such as Aristotle [25] thousands of years ago (∼350 BCE), we humans have never stopped researching and studying this subject. Argumentation has been investigated from different viewpoints belonging to different fields of research and study, such as philosophy, linguistics, or logic. With the emergence of Artificial Intelligence, argumentation also began to be studied from a computational point of view, giving rise to the research area of computational argumentation. Having its origins in concepts from the data structures of graph theory and formal logic among others, computational argumentation has been integrating and exploring new emerging concepts in AI, such as intelligent agents and multi agent systems, or machine learning and deep learning algorithms. This way, computational argumentation resulted in a very heterogeneous, multidisciplinary area of research, where different techniques approached various sub-tasks underlying argumentation from the computational viewpoint. In some cases, computational argumentation could be understood as the technique used to address a problem rather than the problem to be solved itself. Thus, in this thesis, we have given a structure to the research in computational argumentation that makes it possible to understand and contextualise the role of individual research papers in the whole argumentative process from a human reasoning viewpoint. In addition, we proposed new techniques and algorithms aimed at finding common ground between different approaches in the literature, usually investigated independently. From the findings observed in this thesis, it is now possible to achieve a more informed and complete analysis of natural language argumentative discourses, and to improve the persuasiveness of argument-based HCI systems. Furthermore, this thesis sets an starting point in computational argumentation transversal research. However, there still remain important open challenges that we plan to address in future work belonging to both, the automatic analysis of argumentative discourses, and the persuasiveness of argumentation systems.

First of all, with the published *VivesDebate* corpus, it is now possible to perform an automatic analysis and evaluation of complete natural language debates. The corpus contains the revised transcriptions of the speeches performed during 29 different debates. Furthermore, with the *VivesDebate-Speech*, we extend the corpus with the acoustic information of the debates, producing the largest available audio-based argument mining corpus, and running initial experiments on argument

mining from speech in addition to text. Spoken argumentation relies in linguistic features such as the intonation, confidence in speech, or pronunciation to make a stronger and more forceful discourse than the opponent's. In fact, all these features are taken into account by the jury to determine the winner of a debate in a tournament [327]. However, it remains future work to study the implementation of audio features in our proposed automatic debate evaluation models.

In second place, in this thesis we presented a series of analyses and experiments on the whole process of argument mining, but approached as independent offline tasks. This is how NLP tasks are typically addressed. However, with the annotation of the *VivesDebate*, one of the new possibilities that remained unapproachable with the previous corpora is the real-time argument mining. Following the footsteps of other NLP areas of research, we foresee to integrate all the findings observed in this thesis into a complete real-time system aimed at identifying and analysing arguments in a real-time environment (e.g., a debate). This approach will allow the development of live argumentative assistants to help judge, analyse, and better understand the arguments used in debates or political campaigns, assist in the evaluation of competitive debates, or visualise the argumentation in a trial.

Related to this analysis of argumentation, it also remains future work to explore the automatic generation of conclusions/summaries of a given debate. The conclusions are usually presented by each of the participating sides as a biased summary of the whole debate that favours their stance. To the best of our knowledge, some works address the generation of natural language argument summaries, but automatically summarising entire natural language debates remains an unexplored challenge. With the *VivesDebate*, it is possible to investigate this aspect, as a last step of our extension to the analysis of human argumentation presented in this thesis.

Finally, related to argument-based computational persuasion, we foresee that one of the next steps to be taken in this topic will be to study and analyse the role of emotions in argument-based human-computer interactions. Human beings are not purely rational, emotions play an important role in decision making and reasoning, and their influence might lead to an unexpected output from a purely rational viewpoint [66]. Therefore, a complete study and analysis of how emotions are triggered when interacting with human users through arguments will enable a better modelling of the argument-based interactions, and improve the persuasiveness

of computational argumentation systems. In the case of human debates, we have started to explore if observable emotional patterns in the audience are related to the decision of the jury and the outcome of the debate. This emotional approach to argument-based interactions can be relevant for the proposal and development of new empathetic argument-based personal assistants, and for improving the human trust on these new human-computer interactive technologies.

Bibliography

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman M. Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv.*, 50(3):44:1–44:41, 2017.

[2] Sibel Adali and Jennifer Golbeck. Predicting personality with social behavior. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 302–309. IEEE, 2012.

[3] Aseel Addawood and Masooda N. Bashir. "what is your evidence?" A study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining, ArgMining@ACL*. The Association for Computer Linguistics, 2016.

[4] Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP*, pages 118–128. Association for Computational Linguistics, 2017.

[5] Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP*, pages 91–96. Association for Computational Linguistics, 2017.

[6] Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. Argument mining for scholarly document processing: Tak-

ing stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, 2021.

[7] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, 2017.

[8] Sultan Alahmari, Tommy Yuan, and Daniel Kudenko. Reinforcement learning for dialogue game based argumentation. In *Proceedings of the 19th Workshop on Computational Models of Natural Argument, CMNA@PERSUASIVE*, volume 2346, pages 29–37. CEUR-WS.org, 2019.

[9] Sultan Alahmari, Tommy Yuan, and Daniel Kudenko. Reinforcement learning of dialogue coherence and relevance. In *Proceedings of the 19th Workshop on Computational Models of Natural Argument, CMNA@PERSUASIVE*, volume 2346, pages 38–48. CEUR-WS.org, 2019.

[10] José Alemany, Elena del Val, Juan Alberola, and A García-Fornes. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human-Computer Studies*, 129:27–40, 2019.

[11] José Alemany, Elena del Val, Juan Alberola, and Ana García-Fornes. Estimation of privacy risk through centrality metrics. *Future Generation Computer Systems*, 82:63–76, 2018.

[12] José Alemany, Elena del Val, and Ana García-Fornes. Assisting users on the privacy decision-making process in an osn for educational purposes. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 379–383. Springer, 2020.

[13] José Alemany, Elena del Val, and Ana García-Fornes. A review of privacy decision-making mechanisms in online social networks. *ACM Comput. Surv.*, 55(2):37:1–37:32, 2023.

[14] José Alemany, Elena Del Val, Juan M Alberola, and Ana Garćia-Fornes. Metrics for privacy assessment when sharing information in online social networks. *IEEE Access*, 7:143631–143645, 2019.

[15] Gianvincenzo Alfano, Sergio Greco, and Francesco Parisi. Efficient computation of extensions for dynamic abstract argumentation frameworks: An incremental approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, pages 49–55. ijcai.org, 2017.

[16] Gianvincenzo Alfano, Sergio Greco, Francesco Parisi, and Irina Trubitsyna. Argumentation frameworks with strong and weak constraints: Semantics and complexity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6175–6184, 2021.

[17] Alaa Alhamzeh, Mohamed Bouhaouel, Elöd Egyed-Zsigmond, Jelena Mitrovic, Lionel Brunie, and Harald Kosch. A stacking approach for cross-domain argument identification. In *Database and Expert Systems Applications - 32nd International Conference, DEXA 2021, Virtual Event, September 27-30, 2021, Proceedings, Part I*, volume 12923 of *Lecture Notes in Computer Science*, pages 361–373. Springer, 2021.

[18] Alaa Alhamzeh, Elöd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrovic, Lionel Brunie, and Harald Kosch. Empirical study of the model generalization for argument mining in cross-domain and cross-topic settings. *Trans. Large Scale Data Knowl. Centered Syst.*, 52:103–126, 2022.

[19] Ali Altalbe and Faris Kateb. Assuring enhanced privacy violation detection model for social networks. *Int. J. Intell. Comput. Cybern.*, 15(1):75–91, 2022.

[20] Leila Amgoud and Jonathan Ben-Naim. Ranking-based semantics for argumentation frameworks. In *Scalable Uncertainty Management - 7th International Conference, SUM*, volume 8078, pages 134–147. Springer, 2013.

[21] Leila Amgoud and Jonathan Ben-Naim. Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.*, 99:39–55, 2018.

[22]   Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23(10):1062–1093, 2008.

[23]   Leila Amgoud and Srdjan Vesic. A new approach for preference-based argumentation frameworks. *Ann. Math. Artif. Intell.*, 63(2):149–183, 2011.

[24]   Leila Amgoud and Srdjan Vesic. Rich preference-based argumentation frameworks. *Int. J. Approx. Reason.*, 55(2):585–606, 2014.

[25]   Aristotle. *Aristotle's Politics*. Oxford: Clarendon Press, 1905.

[26]   Aristotle. *Prior analytics*. Hackett Publishing, 1989.

[27]   Jens B Asendorpf, Peter Borkenau, Fritz Ostendorf, and Marcel AG Van Aken. Carving personality description at its joints: Confirmation of three replicable personality prototypes for both children and adults. *European Journal of Personality*, 15(3):169–198, 2001.

[28]   Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Ricardo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Mag.*, 38(3):25–36, 2017.

[29]   Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020.

[30]   Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG*, pages 54–64. Association for Computational Linguistics, 2019.

[31]   Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Analyzing the persuasive effect of style in news editorial argumentation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, pages 3154–3160. Association for Computational Linguistics, 2020.

[32] Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6354–6364. Association for Computational Linguistics, 2021.

[33] Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10437–10449. Association for Computational Linguistics, 2022.

[34] Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. Argument pair extraction with mutual guidance and inter-sentence relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3923–3934, 2021.

[35] Jianzhu Bao, Jingyi Sun, Qinglin Zhu, and Ruifeng Xu. Have my arguments been replied to? argument pair extraction as machine reading comprehension. In *ACL 2022*, pages 29–35, 2022.

[36] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 4029–4039. Association for Computational Linguistics, 2020.

[37] Roy Bar-Haim, Dalia Krieger, Orith Toledo-Ronen, Lilach Edelstein, Yonatan Bilu, Alon Halfon, Yoav Katz, Amir Menczel, Ranit Aharonov, and Noam Slonim. From surrogacy to adoption; from bitcoin to cryptocurrency: Debate topic expansion. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 977–990. Association for Computational Linguistics, 2019.

[38] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4):365–410, 2011.

[39] Pietro Baroni and Massimiliano Giacomin. Semantics of abstract argument systems. In *Argumentation in Artificial Intelligence*, pages 25–44. Springer, 2009.

[40] Pietro Baroni, Massimiliano Giacomin, and Giovanni Guida. Scc-recursiveness: a general schema for argumentation semantics. *Artif. Intell.*, 168(1-2):162–210, 2005.

[41] Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. On topology-related properties of abstract argumentation semantics. A correction and extension to dynamics of argumentation systems: A division-based method. *Artif. Intell.*, 212:104–115, 2014.

[42] Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1736–1743. AAAI Press, 2018.

[43] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. Automatic evaluation of design alternatives with quantitative argumentation. *Argument Comput.*, 6(1):24–49, 2015.

[44] Paul Bator. Aristotelian and rogerian rhetoric. *College Composition and Communication*, 31(4):427–432, 1980.

[45] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti,

David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[46] Dorothea Baumeister, Daniel Neugebauer, Jörg Rothe, and Hilmar Schadrack. Verification in incomplete argumentation frameworks. *Artif. Intell.*, 264:1–26, 2018.

[47] Seren Başaran and Obinna H. Ejimogu. A neural network approach for predicting personality from facebook data. *SAGE Open*, 11(3):21582440211032156, 2021.

[48] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[49] Trevor J. M. Bench-Capon. Value-based argumentation frameworks. In *9th International Workshop on Non-Monotonic Reasoning (NMR)*, pages 443–454, 2002.

[50] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artif. Intell.*, 171(10-15):619–641, 2007.

[51] M. Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. Emotions in argumentation: an empirical evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI*, pages 156–163. AAAI Press, 2015.

[52] Mohamed S. Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. Persuasive argumentation and emotions: An empirical evaluation with users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality - 19th International Conference, HCI*, volume 10271, pages 659–671. Springer, 2017.

[53] Philippe Besnard and Anthony Hunter. A logic-based theory of deductive arguments. *Artif. Intell.*, 128(1-2):203–235, 2001.

[54] Philippe Besnard and Anthony Hunter. Argumentation based on classical logic. In *Argumentation in Artificial Intelligence*, pages 133–152. Springer, 2009.

[55] Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. Argument invention from first principles. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 1013–1026. Association for Computational Linguistics, 2019.

[56] Yonatan Bilu and Noam Slonim. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*. The Association for Computer Linguistics, 2016.

[57] Pierre Bisquert, Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Change in argumentation systems: Exploring the interest of removing an argument. In *Scalable Uncertainty Management - 5th International Conference, SUM*, volume 6929, pages 275–288. Springer, 2011.

[58] Stefano Bistarelli and Francesco Santini. Well-foundedness in weighted argumentation frameworks. In *Logics in Artificial Intelligence - 16th European Conference, JELIA*, volume 11468, pages 69–84. Springer, 2019.

[59] Guido Boella, Souhila Kaci, and Leendert W. N. van der Torre. Dynamics in argumentation with single extensions: Abstraction principles and the grounded extension. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 10th European Conference, ECSQARU*, volume 5590, pages 107–118. Springer, 2009.

[60] Guido Boella, Souhila Kaci, and Leendert W. N. van der Torre. Dynamics in argumentation with single extensions: Attack refinement and the grounded extension (extended version). In *Argumentation in Multi-Agent Systems, 6th International Workshop, ArgMAS*, volume 6057, pages 150–159. Springer, 2009.

[61] Filip Boltuzic and Jan Snajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argument Mining ArgMining@ACL*, pages 49–58. The Association for Computer Linguistics, 2014.

[62] Filip Boltuzic and Jan Snajder. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL*, pages 110–115. The Association for Computational Linguistics, 2015.

[63] Andrei Bondarenko, Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.

[64] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012.

[65] Rihab Bouslama, Raouia Ayachi, and Nahla Ben Amor. Using convolutional neural network in cross-domain argumentation mining framework. In *Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compiègne, France, December 16-18, 2019, Proceedings*, volume 11940 of *Lecture Notes in Computer Science*, pages 355–367. Springer, 2019.

[66] Andreas Brännström, Timotheus Kampik, Ramon Ruiz-Dolz, and Joaquín Taverner. A formal framework for designing boundedly rational agents. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, ICAART 2022, Volume 3, Online Streaming, February 3-5, 2022*, pages 705–714. SCITEPRESS, 2022.

[67] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[68] Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014,*

Reykjavik, Iceland, May 26-31, 2014, pages 917–924. European Language Resources Association (ELRA), 2014.

[69] Katarzyna Budzynska and Chris Reed. Whence inference. *University of Dundee Technical Report*, 2011.

[70] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *The 50th Annual Meeting of the Association for Computational Linguistics*, pages 208–212. The Association for Computer Linguistics, 2012.

[71] Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 5427–5433. ijcai.org, 2018.

[72] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46. ACM, 2014.

[73] Martin Caminada. On the issue of reinstatement in argumentation. In *European Workshop on Logics in Artificial Intelligence*, pages 111–123. Springer, 2006.

[74] Martin Caminada. Semi-stable semantics. In *Computational Models of Argument: Proceedings of COMMA*, volume 144, pages 121–130. IOS Press, 2006.

[75] Giuseppe Carenini and Johanna D. Moore. Generating and evaluating evaluative arguments. *Artif. Intell.*, 170(11):925–952, 2006.

[76] Valeria Carofiglio and F d de Rosis. Combining logical with emotional reasoning in natural argumentation. In *3rd Workshop on Affective and Attitude User Modeling*. Citeseer, 2003.

[77] Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL*, pages 29–34. The Association for Computational Linguistics, 2015.

[78] Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *CoRR*, abs/1401.3838, 2014.

[79] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 8th European Conference, ECSQARU*, volume 3571, pages 378–389. Springer, 2005.

[80] Federico Cerutti, Sarah Alice Gaggl, Matthias Thimm, and Johannes Peter Wallner. Foundations of implementations for formal argumentation. *FLAP*, 4(8), 2017.

[81] Federico Cerutti, Nava Tintarev, and Nir Oren. Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *ECAI 2014 - 21st European Conference on Artificial Intelligence*, volume 263, pages 207–212. IOS Press, 2014.

[82] Lisa A. Chalaguine and Anthony Hunter. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *Computational Models of Argument - Proceedings of COMMA*, volume 326, pages 9–20. IOS Press, 2020.

[83] Lisa A. Chalaguine, Anthony Hunter, Henry W. W. Potts, and Fiona Hamilton. Impact of argument type and concerns in argumentation with a chatbot. In *31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI*, pages 1557–1562. IEEE, 2019.

[84] Günther Charwat, Wolfgang Dvorák, Sarah A Gaggl, Johannes P Wallner, and Stefan Woltran. Implementing abstract argumentation-a survey. *Institut Fur Information Systeme, Tech. Rep.*, 2013.

[85] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-16 finnum-3 task: investor's and manager's fine-grained claim detection. 2022.

[86] Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8914–8922. Association for Computational Linguistics, 2022.

[87] Carlos Iván Chesñevar, Ana Gabriela Maguitman, and María Paula González. Empowering recommendation technologies through argumentation. In *Argumentation in Artificial Intelligence*, pages 403–422. Springer, 2009.

[88] Emily Christofides, Amy Muise, and Serge Desmarais. Hey mom, what's on your facebook? comparing facebook disclosure and privacy in adolescents and adults. *Social Psychological and Personality Science*, 3(1):48–54, 2012.

[89] Trudy Hui Hui Chua and Leanne Chang. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior*, 55:190–197, 2016.

[90] Robert B Cialdini. The science of persuasion. *Scientific American*, 284(2):76–81, 2001.

[91] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*. Morrow New York, 1993.

[92] Ana Ciocarlan, Judith Masthoff, and Nir Oren. Actual persuasiveness: Impact of personality, age and gender on message type susceptibility. In *Persuasive Technology 14th International Conference PERSUASIVE, Proceedings*, volume 11433, pages 283–294. Springer, 2019.

[93]   Jonathan Clayton and Rob Gaizauskas. Predicting the presence of reasoning markers in argumentative text. In *Proceedings of the 9th Workshop on Argument Mining, ArgMining@COLING 2022, Online and in Gyeongju, Republic of Korea, October 12 - 17, 2022*, pages 137–142. International Conference on Computational Linguistics, 2022.

[94]   Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. A dataset independent set of baselines for relation prediction in argument mining. *CoRR*, abs/2003.04970, 2020.

[95]   Oana Cocarascu, Kristijonas Cyras, and Francesca Toni. Explanatory predictions with artificial neural networks and argumentation. 2018.

[96]   Oana Cocarascu, Antonio Rago, and Francesca Toni. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, pages 1261–1269. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[97]   Oana Cocarascu, Andria Stylianou, Kristijonas Cyras, and Francesca Toni. Data-empowered argumentation for dialectically explainable predictions. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325, pages 2449–2456. IOS Press, 2020.

[98]   Oana Cocarascu and Francesca Toni. Argumentation for machine learning: A survey. In *Computational Models of Argument - Proceedings of COMMA*, volume 287, pages 219–230. IOS Press, 2016.

[99]   Oana Cocarascu and Francesca Toni. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1374–1379. Association for Computational Linguistics, 2017.

[100] Oana Cocarascu and Francesca Toni. Mining bipolar argumentation frameworks from natural language text. In *Proceedings of the 17th Workshop on Computational Models of Natural Argument co-located with ICAIL*, volume 2048 of *CEUR Workshop Proceedings*, pages 65–70. CEUR-WS.org, 2017.

[101] Andrea Cohen, Sebastian Gottifredi, Alejandro Javier García, and Guillermo Ricardo Simari. A survey of different approaches to support in argumentation systems. *Knowl. Eng. Rev.*, 29(5):513–550, 2014.

[102] Andrea Cohen, Simon Parsons, Elizabeth I. Sklar, and Peter McBurney. A characterization of types of support between structured arguments and their relationship with support in abstract argumentation. *Int. J. Approx. Reason.*, 94:76–104, 2018.

[103] Gregory W Corder and Dale I Foreman. Nonparametric statistics for non-statisticians, 2011.

[104] Angelo Costa, Stella Heras, Javier Palanca, Jaume Jordán, Paulo Novais, and Vicente Julián. Argumentation schemes for events suggestion in an e-health platform. In *International Conference on Persuasive Technology*, pages 17–30. Springer, 2017.

[105] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. Prudent semantics for argumentation frameworks. In *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 568–572. IEEE Computer Society, 2005.

[106] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13(1):21–27, 1967.

[107] Dennis Craandijk and Floris Bex. AGNN: A deep learning architecture for abstract argumentation semantics. In *Computational Models of Argument - Proceedings of COMMA*, volume 326, pages 457–458. IOS Press, 2020.

[108] Dennis Craandijk and Floris Bex. Deep learning for abstract argumentation semantics. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1667–1673, 2021.

[109] Robert Craven and Francesca Toni. Argument graphs and assumption-based argumentation. *Artif. Intell.*, 233:1–59, 2016.

[110] Kristijonas Cyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Syst. Appl.*, 127:141–156, 2019.

[111] SE Fulladoza Dalibón, DC Martinez, and GR Simari. An approach to emotion-based abstract argumentative reasoning. *13th Argentine Symposium on Artificial Intelligence*, 2012.

[112] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, 2008.

[113] Miguel A. del Agua, Adrià Giménez, Nicolás Serrano, Jesús Andrés-Ferrer, Jorge Civera, Alberto Sanchís, and Alfons Juan. The translectures-upv toolkit. In *Advances in Speech and Language Technologies for Iberian Languages, IberSPEECH*, pages 269–278, 2014.

[114] Jérôme Delobelle. *Ranking-based Semantics for Abstract Argumentation. (Sémantique à base de Classement pour l'Argumentation Abstraite)*. PhD thesis, Artois University, Arras, France, 2017.

[115] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[116] Yannis Dimopoulos, Jean-Guy Mailly, and Pavlos Moraitis. Control argumentation frameworks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4678–4685. AAAI Press, 2018.

[117] Regina Dittrich, Walter Katzenbeisser, and Heribert Reisinger. The analysis of rank ordered preference data based on bradley-terry type models die

analyse von präferenzdaten mit hilfe von log-linearen bradley-terry modellen. *OR-Spektrum*, 22(1):117–134, 2000.

[118] Dragan Doder, Srdjan Vesic, and Madalina Croitoru. Ranking semantics for argumentation systems with necessities. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*, pages 1912–1918. ijcai.org, 2020.

[119] Ivan Donadello, Anthony Hunter, Stefano Teso, and Mauro Dragoni. Machine learning for utility prediction in argument-based computational persuasion. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5592–5599. AAAI Press, 2022.

[120] M Brent Donnellan and Richard E Lucas. Age differences in the big five across the life span: evidence from two national samples. *Psychology and aging*, 23(3):558, 2008.

[121] Sylvie Doutre and Jean-Guy Mailly. Constraints and changes: A survey of abstract argumentation dynamics. *Argument Comput.*, 9(3):223–248, 2018.

[122] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, pages 155–161. MIT Press, 1996.

[123] Lorik Dumani, Manuel Biertz, Alex Witry, Anna-Katharina Ludwig, Mirko Lenz, Stefan Ollinger, Ralph Bergmann, and Ralf Schenkel. The recap corpus: A corpus of complex argument graphs on german education politics. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 248–255. IEEE, 2021.

[124] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.

[125] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. In *Argumentation in artificial intelligence*, pages 199–218. Springer, 2009.

[126] Phan Minh Dung, Paolo Mancarella, and Francesca Toni. A dialectic procedure for sceptical, assumption-based argumentation. In *Computational Models of Argument: Proceedings of COMMA*, volume 144, pages 145–156. IOS Press, 2006.

[127] Phan Minh Dung and Phan Minh Thang. Towards (probabilistic) argumentation for jury-based dispute resolution. In *Computational Models of Argument: Proceedings of COMMA*, volume 216, pages 171–182. IOS Press, 2010.

[128] Paul E. Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael J. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artif. Intell.*, 175(2):457–486, 2011.

[129] Paul E. Dunne and Michael J. Wooldridge. Complexity of abstract argumentation. In *Argumentation in Artificial Intelligence*, pages 85–104. Springer, 2009.

[130] Esin Durmus and Claire Cardie. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy, July 2019. Association for Computational Linguistics.

[131] Esin Durmus and Claire Cardie. Exploring the role of prior beliefs for argument persuasion. *CoRR*, abs/1906.11301, 2019.

[132] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 2317–2322. Association for Computational Linguistics, 2017.

[133] Wolfgang Dvorák, Matthias König, Johannes Peter Wallner, and Stefan Woltran. Aspartix-v21. *CoRR*, abs/2109.03166, 2021.

[134] Ryo Egawa, Gaku Morio, and Katsuhide Fujita. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 422–428. Association for Computational Linguistics, 2019.

[135] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 11–22. Association for Computational Linguistics, 2017.

[136] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics, COLING*, pages 831–844. Association for Computational Linguistics, 2018.

[137] Valentinos Evripidou and Francesca Toni. Argumentation and voting for an intelligent user empowering business directory on the web. In *Web Reasoning and Rule Systems - 6th International Conference, RR*, volume 7497, pages 209–212. Springer, 2012.

[138] Jeanne Fahnestock and Marie Secor. The stases in scientific and literary argument. *Written communication*, 5(4):427–443, 1988.

[139] Bettina Fazzinga, Sergio Flesca, and Filippo Furfaro. Probabilistic bipolar abstract argumentation frameworks: complexity results. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 1803–1809. ijcai.org, 2018.

[140] Vanessa Wei Feng and Graeme Hirst. Two-pass discourse segmentation with pairing and global features. *CoRR*, abs/1407.8215, 2014.

[141] Seeger Fisher and Brian Roark. The utility of parse-derived features for automatic discourse segmentation. In *ACL Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. The Association for Computational Linguistics, 2007.

[142] Ricard L. Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Trans. Comput. Hum. Interact.*, 24(1):5:1–5:29, 2017.

[143] Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144, 2011.

[144] Sarah Alice Gaggl, Thomas Linsbichler, Marco Maratea, and Stefan Woltran. Design and results of the second international competition on computational models of argumentation. *Artif. Intell.*, 279, 2020.

[145] Fabrice Gaignier, Yannis Dimopoulos, Jean-Guy Mailly, and Pavlos Moraitis. Probabilistic control argumentation frameworks. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, pages 519–527. ACM, 2021.

[146] Alejandro Javier García and Guillermo Ricardo Simari. Defeasible logic programming: An argumentative approach. *Theory Pract. Log. Program.*, 4(1-2):95–138, 2004.

[147] Debela Gemechu and Chris Reed. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 516–526. Association for Computational Linguistics, 2019.

[148] Kallirroi Georgila and David R. Traum. Reinforcement learning of argumentation dialogue policies in negotiation. In *Interspeech, 12th Annual Conference of the International Speech Communication Association*, pages 2073–2076. ISCA, 2011.

[149] Martin Gerlach, Beatrice Farb, William Revelle, and Luís A Nunes Amaral. A robust data-driven approach identifies four personality types across four large data sets. *Nature human behaviour*, 2(10):735, 2018.

[150] Talmy Givón. *Topic continuity in discourse*. Amsterdam: John Benjamins, 1983.

[151] Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 967–976. Association for Computational Linguistics, 2019.

[152] Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4143–4149. ijcai.org, 2022.

[153] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262. 2011.

[154] Lewis R Goldberg et al. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28, 1999.

[155] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[156] Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications - 8th Hellenic Conference on AI, SETN*, volume 8445, pages 287–299. Springer, 2014.

[157] Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. The work-week is the best time to start a family – a study of gpt-2 based claim generation, 2020.

[158] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7805–7813. AAAI Press, 2020.

[159] Joseph E Grimes. *The Thread of Discourse.* ERIC, 1972.

[160] Asela Gunawardana and Guy Shani. Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer, 2015.

[161] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.

[162] Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 4684–4690. Association for Computational Linguistics, 2019.

[163] Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. Opponent modelling in persuasion dialogues. In *IJCAI, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 164–170. IJCAI/AAAI, 2013.

[164] Emmanuel Hadoux and Anthony Hunter. Comfort or safety? gathering and using the concerns of a participant for better persuasion. *Argument Comput.*, 10(2):113–147, 2019.

[165] Emmanuel Hadoux, Anthony Hunter, and Jean-Baptiste Corrégé. Strategic dialogical argumentation using multi-criteria decision making with application to epistemic and emotional aspects of arguments. In *Foundations of Information and Knowledge Systems - 10th International Symposium FoIKS, Proceedings*, volume 10833, pages 207–224. Springer, 2018.

[166] Emmanuel Hadoux, Anthony Hunter, and Sylwia Polberg. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *CoRR*, abs/2101.11870, 2021.

[167] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[168] Stella Heras, Paula Rodríguez, Javier Palanca, Néstor D. Duque, and Vicente Julián. Using argumentation to persuade students in an educational recommender system. In *Persuasive Technology - 12th International Conference, PERSUASIVE, Proceedings*, volume 10171, pages 227–239. Springer, 2017.

[169] Christopher Hidey and Kathy McKeown. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1756–1767. Association for Computational Linguistics, 2019.

[170] Danial Hooshyar, Yueh-Min Huang, and Yeongwook Yang. A three-layered student learning model for prediction of failure risk in online learning. *Human-centric Computing and Information Sciences*, 12, 2022.

[171] Yufang Hou and Charles Jochim. Argument relation classification using a joint inference model. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP*, pages 60–66. Association for Computational Linguistics, 2017.

[172] Fa-Hsuan Hsiao, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. Modeling inter round attack of online debaters for winner prediction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2860–2869. ACM, 2022.

[173] Xinyu Hua, Zhe Hu, and Lu Wang. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Conference of the*

Association for Computational Linguistics, ACL, pages 2661–2672. Association for Computational Linguistics, 2019.

[174] Xinyu Hua and Lu Wang. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 219–230. Association for Computational Linguistics, 2018.

[175] Chiungjung Huang. Social network site use and big five personality traits: A meta-analysis. *Computers in Human Behavior*, 97:280–290, 2019.

[176] Anthony Hunter. Some foundations for probabilistic abstract argumentation. In *Computational Models of Argument - Proceedings of COMMA*, volume 245, pages 117–128. IOS Press, 2012.

[177] Anthony Hunter. Towards a framework for computational persuasion with applications in behaviour change. *Argument Comput.*, 9(1):15–40, 2018.

[178] Anthony Hunter. Generating instantiated argument graphs from probabilistic information. In *ECAI - 24th European Conference on Artificial Intelligence*, volume 325, pages 769–776. IOS Press, 2020.

[179] Anthony Hunter, Lisa A. Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. Towards computational persuasion via natural language argumentation dialogues. In *KI: Advances in Artificial Intelligence - 42nd German Conference on AI, Proceedings*, volume 11793, pages 18–33. Springer, 2019.

[180] Anthony Hunter, Sylwia Polberg, and Matthias Thimm. Epistemic graphs for representing and reasoning with positive and negative influences of arguments. *Artificial Intelligence*, 281:103236, 2020.

[181] Anthony Hunter and Matthias Thimm. Probabilistic reasoning with abstract argumentation frameworks. *J. Artif. Intell. Res.*, 59:565–611, 2017.

[182] Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. The mllp-upv

german-english machine translation system for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 418–424, 2018.

[183] Mathilde Janier and Chris Reed. Corpus resources for dispute mediation discourse. In *Tenth International Conference on Language Resources and Evaluation*, pages 1014–1021. European Language Resources Association, 2016.

[184] Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. Claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–30. Association for Computational Linguistics, 2018.

[185] Ria Jha, Francesco Belardinelli, and Francesca Toni. Formal verification of debates in argumentation theory. In *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, Online*, pages 940–947. ACM, 2020.

[186] Yohan Jo, Jacky Visser, Chris Reed, and Eduard H. Hovy. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL*, pages 11–24. Association for Computational Linguistics, 2019.

[187] Javier Jorge, Adrià Giménez, Javier Iranzo-Sánchez, Jorge Civera, Albert Sanchís, and Alfons Juan. Real-time one-pass decoder for speech recognition using LSTM language models. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3820–3824. ISCA, 2019.

[188] Shafiq R. Joty, Giuseppe Carenini, and Raymond T. Ng. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 904–915. ACL, 2012.

[189] Timotheus Kampik, Dov M. Gabbay, and Giovanni Sartor. The burden of persuasion in abstract argumentation. In *Logic and Argumentation - 4th*

International Conference, CLAR 2021, Hangzhou, China, October 20-22, 2021, Proceedings, volume 13040 of *Lecture Notes in Computer Science*, pages 224–243. Springer, 2021.

[190] Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4744–4754. Association for Computational Linguistics, 2021.

[191] Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 7067–7072. Association for Computational Linguistics, 2020.

[192] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-domain mining of argumentative text through distant supervision. In *NAACL HLT, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. The Association for Computational Linguistics, 2016.

[193] Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1351–1357. Association for Computational Linguistics, 2017.

[194] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Kazuma Kadowaki, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine. Overview

of the ntcir-16 qa lab-poliinfo-3 task. *Proceedings of The 16th NTCIR Conference*, 6 2022.

[195] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. An argumentation approach for resolving privacy disputes in online social networks. *ACM Trans. Internet Techn.*, 17(3):27:1–27:22, 2017.

[196] Neema Kotonya and Francesca Toni. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL*, pages 156–166. Association for Computational Linguistics, 2019.

[197] Venelin Kovatchev, M Antònia Martí, and Maria Salamó. Etpc-a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[198] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[199] Isabelle Kuhlmann and Matthias Thimm. Using graph convolutional networks for approximate reasoning with abstract argumentation frameworks: A feasibility study. In *Scalable Uncertainty Management - 13th International Conference, SUM, Proceedings*, volume 11940, pages 24–37. Springer, 2019.

[200] Abdurrahman Can Kurtan and Pinar Yolum. Assisting humans in privacy management: an agent-based approach. *Auton. Agents Multi Agent Syst.*, 35(1):7, 2021.

[201] Jean-Marie Lagniez, Emmanuel Lonca, Jean-Guy Mailly, and Julien Rossit. Introducing the fourth international competition on computational models of argumentation. In *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation co-located with (COMMA)*, volume 2672, pages 80–85. CEUR-WS.org, 2020.

[202] Jean-Marie Lagniez, Emmanuel Lonca, Jean-Guy Mailly, and Julien Rossit. Design and results of ICCMA 2021. *CoRR*, abs/2109.08884, 2021.

[203] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[204] John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. Aifdb: Infrastructure for the argument web. In *Computational Models of Argument - Proceedings of COMMA*, volume 245, pages 515–516. IOS Press, 2012.

[205] John Lawrence and Chris Reed. Aifdb corpora. In *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 465–466. IOS Press, 2014.

[206] John Lawrence and Chris Reed. Argument mining using argumentation scheme structures. In *COMMA*, pages 379–390, 2016.

[207] John Lawrence and Chris Reed. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818, 2019.

[208] John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argument Mining, ArgMining@ACL*, pages 79–87. The Association for Computer Linguistics, 2014.

[209] Dieu-Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP*, pages 121–130. Association for Computational Linguistics, 2018.

[210] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nat.*, 521(7553):436–444, 2015.

[211] Mirko Lenz, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. Towards an argument mining pipeline transforming texts to argument graphs. In *Computational Models of Argument - Proceedings of COMMA*, volume 326, pages 263–270. IOS Press, 2020.

[212] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, 2014.

[213] Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. Unsupervised corpus-wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP*, pages 79–84. Association for Computational Linguistics, 2017.

[214] Hengfei Li, Nir Oren, and Timothy J. Norman. Probabilistic argumentation frameworks. In *Theorie and Applications of Formal Argumentation - First International Workshop, TAFA*, volume 7132, pages 1–16. Springer, 2011.

[215] Yinzi Li, Wei Chen, Zhongyu Wei, Yujun Huang, Chujun Wang, Siyuan Wang, Qi Zhang, Xuanjing Huang, and Libo Wu. A structure-aware argument encoder for literature discourse analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 7093–7098. International Committee on Computational Linguistics, 2022.

[216] Bei Shui Liao, Li Jin, and Robert C. Koons. Dynamics of argumentation systems: A division-based method. *Artif. Intell.*, 175(11):1790–1814, 2011.

[217] Marco Lippi and Paolo Torroni. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2979–2985, 2016.

[218] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoy-

anov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[219] Martyn Lloyd-Kelly and Adam Wyner. Arguing about emotion. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 355–367. Springer, 2011.

[220] Stephanie M. Lukin, Pranav Anand, Marilyn A. Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 742–753. Association for Computational Linguistics, 2017.

[221] Andrea A Lunsford. Aristotelian vs. rogerian argument: A reassessment. *College Composition and Communication*, 30(2):146–151, 1979.

[222] Fabrizio Macagno, Douglas Walton, and Chris Reed. Argumentation schemes. history, classifications, and computational applications. *History, Classifications, and Computational Applications (December 23, 2017). Macagno, F., Walton, D. & Reed, C*, pages 2493–2556, 2017.

[223] Marta Maćkiewicz and Jan Cieciuch. Pictorial personality traits questionnaire for children (pptq-c)—a new measure of children's personality traits. *Frontiers in psychology*, 7:498, 2016.

[224] Nitin Madnani, Michael Heilman, Joel R. Tetreault, and Martin Chodorow. Identifying high-level organizational elements in argumentative discourse. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 20–28. The Association for Computational Linguistics, 2012.

[225] Lars Malmqvist. Afgcn: An approximate abstract argumentation solver. In *Fourth International Competition on Computational Models of Argumentation (ICCMA'21)*, pages 1–3. ICCMA, 2021.

[226] Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. Argument mining on clinical trials. In *COMMA*, pages 137–148, 2018.

[227] Tobias Mayer, Elena Cabrio, and Serena Villata. Evidence type classification in randomized controlled trials. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP*, pages 29–34. Association for Computational Linguistics, 2018.

[228] Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325, pages 2108–2115. IOS Press, 2020.

[229] Irene Mazzotta, Fiorella De Rosis, and Valeria Carofiglio. Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent systems*, 22(6):42–51, 2007.

[230] Irene Mazzotta, Vincenzo Silvestri, and Fiorella De Rosis. Emotional and non emotional persuasion strength. In *Proceedings of AISB*, volume 8, pages 14–21. Citeseer, 2008.

[231] Peter McBurney and Simon Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *J. Log. Lang. Inf.*, 11(3):315–334, 2002.

[232] Paul McCann. fugashi, a tool for tokenizing Japanese in python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online, November 2020. Association for Computational Linguistics.

[233] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[234] Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 4889–4896. AAAI Press, 2018.

[235] Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J Norman. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, 2021.

[236] Maria Miceli, Fiorella de Rosis, and Isabella Poggi. Emotional and non-emotional persuasion. *Applied Artificial Intelligence*, 20(10):849–879, 2006.

[237] Lenz Mirko, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. Towards an argument mining pipeline transforming texts to argument graphs. *Computational Models of Argument: Proceedings of COMMA 2020*, 326:263, 2020.

[238] Amita Misra, Brian Ecker, and Marilyn A. Walker. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the SIGDIAL Conference*, pages 276–287. The Association for Computer Linguistics, 2016.

[239] Koh Mitsuda, Ryuichiro Higashinaka, and Kuniko Saito. Combining argumentation structure and language model for generating natural argumentative dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 65–71, 2022.

[240] Sanjay Modgil. Reasoning about preferences in argumentation frameworks. *Artif. Intell.*, 173(9-10):901–934, 2009.

[241] Sanjay Modgil and Henry Prakken. The aspic+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[242] Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. Automatic detection of arguments in legal texts. In *The Eleventh International Conference on Artificial Intelligence and Law*, pages 225–230. ACM, 2007.

[243] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465, 2013.

[244] Ariel Monteserin and Analía Amandi. A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. *Expert Syst. Appl.*, 40(6):2182–2188, 2013.

[245] Gaku Morio and Katsuhide Fujita. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP*, pages 11–21. Association for Computational Linguistics, 2018.

[246] Gaku Morio and Katsuhide Fujita. Syntactic graph convolution in multitask learning for identifying and classifying the argument component. In *13th IEEE International Conference on Semantic Computing, ICSC*, pages 271–278. IEEE, 2019.

[247] Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 3259–3266. Association for Computational Linguistics, 2020.

[248] Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. End-to-end argument mining with cross-corpora multi-task learning. *Trans. Assoc. Comput. Linguistics*, 10:639–658, 2022.

[249] Francesca Mosca and Jose M. Such. ELVIRA: an explainable agent for value and utility-driven multiuser privacy. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 916–924. ACM, 2021.

[250] Francesca Mosca and Jose M. Such. An explainable assistant for multiuser privacy. *Auton. Agents Multi Agent Syst.*, 36(1):10, 2022.

[251] Pradeep K Murukannaiah, Anup K Kalia, Pankaj R Telangy, and Munindar P Singh. Resolving goal conflicts via argumentation-based analysis of

competing hypotheses. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, pages 156–165. IEEE, 2015.

[252] Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems - International Workshops: CMNA XV*, volume 9935, pages 16–25. Springer, 2015.

[253] Nona Naderi and Graeme Hirst. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, 2018.

[254] Fahd Saud Nawwab, Paul E. Dunne, and Trevor J. M. Bench-Capon. Exploring the role of emotions in rational decision making. In *Computational Models of Argument: Proceedings of COMMA*, volume 216, pages 367–378. IOS Press, 2010.

[255] Huy V. Nguyen and Diane J. Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5892–5899. AAAI Press, 2018.

[256] Vlad Niculae, Joonsuk Park, and Claire Cardie. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 985–995. Association for Computational Linguistics, 2017.

[257] Andreas Niskanen and Matti Järvisalo. Algorithms for dynamic argumentation frameworks: An incremental sat-based approach. In *ECAI - 24th European Conference on Artificial Intelligence*, volume 325, pages 849–856. IOS Press, 2020.

[258] Andreas Niskanen, Johannes Peter Wallner, and Matti Järvisalo. Synthesizing argumentation frameworks from examples. *J. Artif. Intell. Res.*, 66:503–554, 2019.

[259] Donald Nute. *Defeasible Logic*, page 353–395. Oxford University Press, Inc., USA, 1994.

[260] Matan Orbach, Yonatan Bilu, Ariel Gera, Yoav Kantor, Lena Dankin, Tamar Lavee, Lili Kotlerman, Shachar Mirkin, Michal Jacovi, Ranit Aharonov, and Noam Slonim. A dataset of general-purpose rebuttal. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language ProcessingEMNLP*, pages 5590–5600. Association for Computational Linguistics, 2019.

[261] Rita Orji, Regan L Mandryk, and Julita Vassileva. Gender, age, and responsiveness to cialdini's persuasion strategies. In *International Conference on Persuasive Technology*, pages 147–159. Springer, 2015.

[262] Wassila Ouerdane, Nicolas Maudet, and Alexis Tsoukiàs. Argumentation theory and decision aiding. In *Trends in Multiple Criteria Decision Analysis*, pages 177–208. Springer, 2010.

[263] Kiemute Oyibo, Rita Orji, and Julita Vassileva. Investigation of the influence of personality traits on cialdini's persuasive strategies. *PPT@ PERSUASIVE*, 2017:8–20, 2017.

[264] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[265] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings*, pages 98–107. ACM, 2009.

[266] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22, 2011.

[267] Alexandros Papangelis and Kallirroi Georgila. Reinforcement learning of multi-issue negotiation dialogue policies. In *Proceedings of the SIGDIAL Conference*, pages 154–158. The Association for Computer Linguistics, 2015.

[268] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argument Mining, ArgMining@ACL*, pages 29–38. The Association for Computer Linguistics, 2014.

[269] Joonsuk Park and Claire Cardie. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*. European Language Resources Association (ELRA), 2018.

[270] Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. TATHYA: A multiclassifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM*, pages 2259–2262. ACM, 2017.

[271] Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. Argumentative relation classification with background knowledge. In *Computational Models of Argument - Proceedings of COMMA*, volume 326, pages 319–330. IOS Press, 2020.

[272] Florian Pecune and Stacy Marsella. A framework to co-optimize task and social dialogue policies using reinforcement learning. In *IVA '20: ACM International Conference on Intelligent Virtual Agents*, pages 45:1–45:8. ACM, 2020.

[273] Andreas Peldszus. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argument Mining, ArgMining@ACL*, pages 88–97. The Association for Computer Linguistics, 2014.

[274] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7(1):1–31, 2013.

[275] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of*

the 1st European Conference on Argumentation, Lisbon, volume 2, pages 801–815, 2015.

[276] Andreas Peldszus and Manfred Stede. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 938–948. The Association for Computational Linguistics, 2015.

[277] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, University of Texas, 2015.

[278] Isaac Persing and Vincent Ng. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics ACL*, pages 543–552. The Association for Computer Linguistics, 2015.

[279] Isaac Persing and Vincent Ng. End-to-end argumentation mining in student essays. In *NAACL HLT, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394. The Association for Computational Linguistics, 2016.

[280] Livia Polanyi. A formal model of the structure of discourse. *Journal of pragmatics*, 12(5-6):601–638, 1988.

[281] Sylwia Polberg and Anthony Hunter. Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *Int. J. Approx. Reason.*, 93:487–543, 2018.

[282] Tomasz Potapczyk, Pawel Przybysz, Marcin Chochowski, and Artur Szumaczuk. Samsung's system for the IWSLT 2019 end-to-end speech translation task. In *Proceedings of the 16th International Conference on Spoken Language Translation, IWSLT*, 2019.

[283] Peter Potash and Anna Rumshisky. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2475, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[284] Nico Potyka. Abstract argumentation with markov networks. In *ECAI - 24th European Conference on Artificial Intelligence*, volume 325, pages 865–872. IOS Press, 2020.

[285] Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Gonçalves, and Paulo Quaresma. Echr: legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, 2020.

[286] Henry Prakken. Ai & law, logic and argument schemes. *Argumentation*, 19(3):303–320, 2005.

[287] Henry Prakken. Formal systems for persuasion dialogue. *Knowl. Eng. Rev.*, 21(2):163–188, 2006.

[288] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument Comput.*, 1(2):93–124, 2010.

[289] Henry Prakken. Historical overview of formal argumentation. *FLAP*, 4(8), 2017.

[290] Henry Prakken and Giovanni Sartor. The role of logic in computational models of legal argument: a critical survey. *Computational logic: Logic programming and beyond*, pages 342–381, 2002.

[291] Henry Prakken and Giovanni Sartor. Law and logic: A review from an argumentation perspective. *Artif. Intell.*, 227:214–245, 2015.

[292] Henry Prakken and Gerard Vreeswijk. Logics for defeasible argumentation. In *Handbook of philosophical logic*, pages 219–318. Springer, 2001.

[293] Henry Prakken, Adam Wyner, Trevor Bench-Capon, and Katie Atkinson. A formalization of argumentation schemes for legal case-based reasoning in aspic+. *Journal of Logic and Computation*, 25(5):1141–1166, 2015.

[294] Farzana Quayyum, Daniela S. Cruzes, and Letizia Jaccheri. Cybersecurity awareness for children: A systematic literature review. *Int. J. Child Comput. Interact.*, 30:100343, 2021.

[295] Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto, and Wolfgang Minker. Evaluation of argument search approaches in the context of argumentative dialogue systems. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC*, pages 513–522. European Language Resources Association, 2020.

[296] Niklas Rach, Klaus Weber, Louisa Pragst, Elisabeth André, Wolfgang Minker, and Stefan Ultes. EVA: A multimodal argumentative dialogue system. In *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI*, pages 551–552. ACM, 2018.

[297] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni. Argumentative explanations for interactive recommendations. *Artificial Intelligence*, 296:103506, 2021.

[298] Antonio Rago, Oana Cocarascu, and Francesca Toni. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 1949–1955. ijcai.org, 2018.

[299] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. Discontinuity-free decision support with quantitative argumentation debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR*, pages 63–73. AAAI Press, 2016.

[300] Iyad Rahwan, Mohammed Iqbal Madakkatel, Jean-François Bonnefon, Ruqiyabi Naz Awan, and Sherief Abdallah. Behavioral experiments for as-

sessing the abstract argumentation semantics of reinstatement. *Cogn. Sci.*, 34(8):1483–1502, 2010.

[301] Iyad Rahwan and Guillermo R Simari. *Argumentation in artificial intelligence*, volume 47. Springer, 2009.

[302] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*, 2005.

[303] Christof Rapp. Aristotle's rhetoric. *Stanford Encyclopedia of Philosophy*, 2011.

[304] Chris Reed and Derek Long. Content ordering in the generation of persuasive discourse. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI*, pages 1022–1029. Morgan Kaufmann, 1997.

[305] Chris Reed, Derek Long, and Maria Fox. An architecture for argumentative dialogue planning. In *International Conference on Formal and Applied Practical Reasoning*, pages 555–566. Springer, 1996.

[306] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 3980–3990. Association for Computational Linguistics, 2019.

[307] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[308] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th*

Conference of the Association for Computational Linguistics, ACL, pages 567–578. Association for Computational Linguistics, 2019.

[309] Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL*, pages 45–55. The Association for Computational Linguistics, 2015.

[310] Rodney A Reynolds and J Lynn Reynolds. What we know so far about evidence. *The Persuasion Handbook: Developments in Theory and Practice*, pages 427–432, 2002.

[311] Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 440–450. The Association for Computational Linguistics, 2015.

[312] Régis Riveret and Guido Governatori. On learning attacks in probabilistic abstract argumentation. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 653–661. ACM, 2016.

[313] Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. Cross-lingual argumentative relation identification: from english to portuguese. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 144–154. Association for Computational Linguistics, 2018.

[314] João António Rodrigues and António Branco. Transferring confluent knowledge to argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6859–6874, 2022.

[315] Ariel Rosenfeld and Sarit Kraus. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Trans. Interact. Intell. Syst.*, 6(4):30:1–30:33, 2016.

[316] Ariel Rosenfeld and Sarit Kraus. Strategical argumentative agent for human persuasion. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence*, volume 285, pages 320–328. IOS Press, 2016.

[317] Sebastiaan Rothmann and Elize P Coetzer. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*, 29(1):68–74, 2003.

[318] Allen Roush and Arvind Balaji. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online, December 2020. Association for Computational Linguistics.

[319] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[320] Ramon Ruiz-Dolz. Towards an artificial argumentation system. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 5206–5207, 7 2020. Doctoral Consortium.

[321] Ramon Ruiz-Dolz. A cascade model for argument mining in japanese political discussions: the qa lab-poliinfo-3 case study. *arXiv preprint arXiv:2207.01672*, 2022.

[322] Ramon Ruiz-Dolz, José Alemany, Stella Heras, and Ana García-Fornes. Automatic generation of explanations to prevent privacy violations. In *Proceedings of the 2nd EXplainable AI in Law Workshop (XAILA 2019)*, volume 2681. CEUR-WS.org, 2019.

[323] Ramon Ruiz-Dolz, José Alemany, Stella Heras, and Ana García-Fornes. On the prevention of privacy threats: How can we persuade our social network users? *CoRR*, abs/2104.10004, 2021.

[324] Ramon Ruiz-Dolz, Stella Heras, José Alemany, and Ana García-Fornes. Towards an argumentation system for assisting users with privacy management

in online social networks. In *Proceedings of the 19th Workshop on Computational Models of Natural Argument CMNA@PERSUASIVE 2019*, volume 2346, pages 17–28. CEUR-WS.org, 2019.

[325] Ramon Ruiz-Dolz, Stella Heras, José Alemany, and Ana García-Fornes. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70, 2021.

[326] Ramon Ruiz-Dolz, Stella Heras, and Ana García-Fornes. Automatic debate evaluation with argumentation semantics and natural language argument graph networks. *CoRR*, abs/2203.14647, 2022.

[327] Ramon Ruiz-Dolz, Montserrat Nofre, Mariona Taulé, Stella Heras, and Ana García-Fornes. Vivesdebate: A new annotated multilingual corpus of argumentation in a debate tournament. *Applied Sciences*, 11(15):7160, 2021.

[328] Ramon Ruiz-Dolz, Joaquín Taverner, Stella Heras, Ana García-Fornes, and Vicente J. Botti. A qualitative analysis of the persuasive properties of argumentation schemes. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022*, pages 1–11. ACM, 2022.

[329] Ramon Ruiz-Dolz, Joaquin Taverner, Stella Heras, Ana García-Fornes, and Vicente Botti. A qualitative analysis of the persuasive properties of argumentation schemes. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2022, Barcelona, Spain, July, 4-7, 2022, In press.* ACM, 2022.

[330] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.

[331] Pietro Russo, Lorenzo Bracciale, and Giuseppe Bianchi. Dare-to-share: Collaborative privacy-preserving recommendations with (almost) no crypto. *Secur. Priv.*, 4(3), 2021.

[332] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[333] Sougata Saha, Souvik Das, and Rohini K. Srihari. Dialo-ap: A dependency parsing based argument parser for dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 887–901. International Committee on Computational Linguistics, 2022.

[334] Patrick Saint-Dizier. A two-level approach to generate synthetic argumentation reports. *Argument Comput.*, 9(2):137–154, 2018.

[335] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019.

[336] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL*, pages 56–66. The Association for Computational Linguistics, 2015.

[337] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. End-to-end argument generation system in debating. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics ACL*, pages 109–114. The Association for Computer Linguistics, 2015.

[338] Ekaterina Saveleva, Volha Petukhova, Marius Mosbach, and Dietrich Klakow. Graph-based argument quality assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1268–1280, Held Online, September 2021. INCOMA Ltd.

[339] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics NAACL-HLT, pages 380–396. Association for Computational Linguistics, 2021.

[340] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. Internet users' perceptions of information sensitivity–insights from germany. *International Journal of Information Management*, 46:142–150, 2019.

[341] Lucius Annaeus Seneca. *Lucius Annaeus Seneca De beneficiis libri VII*. Berkeley: University of California Press, 1950.

[342] Donald Sharpe. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation*, 20(1):8, 2015.

[343] Daiki Shirafuji, Rafal Rzepka, and Kenji Araki. Debate outcome prediction using automatic persuasiveness evaluation and counterargument relations. In *LaCATODA/BtG@ IJCAI*, pages 24–29, 2019.

[344] Guillermo Ricardo Simari and Ronald Prescott Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artif. Intell.*, 53(2-3):125–157, 1992.

[345] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. An autonomous debating system. *Nature*, 591(7850):379–384, 2021.

[346] Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, SIGdial*, pages 26–34. Association for Computational Linguistics, 2007.

[347] Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, 2014.

[348] Afonso Sousa, Bernardo Leite, Gil Rocha, and Henrique Lopes Cardoso. Cross-lingual annotation projection for argument mining in portuguese. In *Progress in Artificial Intelligence - 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7-9, 2021, Proceedings*, volume 12981 of *Lecture Notes in Computer Science*, pages 752–765. Springer, 2021.

[349] Anna C Squicciarini, Heng Xu, and Xiaolong Zhang. Cope: Enabling collaborative privacy management in online social networks. *Journal of the American Society for Information Science and Technology*, 62(3):521–534, 2011.

[350] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT*, pages 21–25. Association for Computational Linguistics, 2018.

[351] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510, 2014.

[352] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 46–56. ACL, 2014.

[353] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Comput. Linguistics*, 43(3):619–659, 2017.

[354] Christian Stab, Tristan Miller, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *CoRR*, abs/1802.05758, 2018.

[355] Manfred Stede and Jodi Schneider. *Argumentation Mining*. Morgan & Claypool Publishers, 2018.

[356] Nikolaos Stylianou and Ioannis Vlahavas. Transformed: End-to-end transformers for evidence-based medicine and argument mining in medical literature. *Journal of Biomedical Informatics*, 117:103767, 2021.

[357] Rajen Subba and Barbara Di Eugenio. Automatic discourse segmentation using neural networks. In *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 189–190, 2007.

[358] Jose M Such and Natalia Criado. Resolving multi-party privacy conflicts in social media. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1851–1863, 2016.

[359] Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9(5):1054–1054, 1998.

[360] Suyono, H Nasrudin, B Yonata, and W B Sabtiawan. The claims statements from viral videos for instrument development to assess argumentation thinking skills. *Journal of Physics: Conference Series*, 1899(1):012174, 2021.

[361] Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING*, pages 5384–5396. International Committee on Computational Linguistics, 2020.

[362] Matthias Thimm. A probabilistic semantics for abstract argumentation. In *ECAI 2012 - 20th European Conference on Artificial Intelligence*, volume 242, pages 750–755. IOS Press, 2012.

[363] Matthias Thimm. Harper+: Using grounded semantics for approximate reasoning in abstract argumentation. In *Fourth International Competition on Computational Models of Argumentation (ICCMA'21)*, pages 1–2. ICCMA, 2021.

[364] Matthias Thimm and Serena Villata. The first international competition on computational models of argumentation: Results and analysis. *Artif. Intell.*, 252:267–294, 2017.

[365] Rosemary Thomas, Nir Oren, and Judith Masthoff. Argumessage: a system for automation of message generation using argumentation schemes. In *Proceedings of the 18th Workshop on Computational Models of Argument (CMNA 2018)*, pages 27–31, 2018.

[366] Rosemary Josekutty Thomas. *Personalised persuasive messages for behaviour change interventions: combining Cialdini's principles and argumentation schemes*. PhD thesis, University of Aberdeen, UK, 2019.

[367] Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. Adapting healthy eating messages to personality. In *Persuasive Technology - 12th International Conference, PERSUASIVE, Proceedings*, volume 10171, pages 119–132. Springer, 2017.

[368] Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. Can I influence you? development of a scale to measure perceived persuasiveness and two studies showing the use of the scale. *Frontiers Artif. Intell.*, 2:24, 2019.

[369] Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. Is argumessage effective? A critical evaluation of the persuasive message generation system. In *Persuasive Technology - 14th International Conference PERSUASIVE, Proceedings*, volume 11433, pages 87–99. Springer, 2019.

[370] Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 303–317. Association for Computational Linguistics, 2020.

[371] Francesca Toni. A tutorial on assumption-based argumentation. *Argument Comput.*, 5(1):89–117, 2014.

[372] Stephen E Toulmin. Reasoning in theory and practice. In *Arguing on the Toulmin model*, pages 25–29. Springer, 2006.

[373] Amine Trabelsi and Osmar R. Zaïane. Extraction and clustering of arguing expressions in contentious text. *Data Knowl. Eng.*, 100:226–239, 2015.

[374] Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110, 2022.

[375] Frans H Van Eemeren and Rob Grootendorst. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004.

[376] Frans H van Eemeren, Sally Jackson, and Scott Jacobs. Argumentation. In *Reasonableness and Effectiveness in Argumentative Discourse, Fifty Contributions to the Development of Pragma-Dialectics*, volume 27 of *Argumentation Library*, pages 3–25. Springer, 2015.

[377] Frans H Van Eemeren and T Kruiger. Identifying argumentation schemes. *PDA*, pages 70–81, 1987.

[378] Vladimir Vapnik. The support vector method of function estimation. In *Nonlinear modeling*, pages 55–85. Springer, 1998.

[379] John Paul Vargheese, Somayajulu Sripada, Judith Masthoff, and Nir Oren. Persuasive strategies for encouraging social interaction for older adults. *International Journal of Human-Computer Interaction*, 32(3):190–214, 2016.

[380] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[381] Bart Verheij. Two approaches to dialectical argumentation: admissible sets and argumentation stages. *Proc. NAIC*, 96:357–368, 1996.

[382] Bart Verheij. Evaluating arguments based on toulmin's scheme. *Argumentation*, 19(3):347–371, 2005.

[383] Maria Paz Garcia Villalba and Patrick Saint-Dizier. A framework to extract arguments in opinion texts. *Int. J. Cogn. Informatics Nat. Intell.*, 6(3):62–87, 2012.

[384] S Villata et al. Using argument mining for legal text summarization. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, page 184. IOS Press, 2020.

[385] Serena Villata, M. Sahbi Benlamine, Elena Cabrio, Claude Frasson, and Fabien Gandon. Assessing persuasion in argumentation through emotions and mental states. In *Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference, FLAIRS*, pages 134–139. AAAI Press, 2018.

[386] Serena Villata, Elena Cabrio, Imène Jraidi, M. Sahbi Benlamine, Maher Chaouachi, Claude Frasson, and Fabien Gandon. Emotions and personality traits in argumentation: An empirical evaluation[1]. *Argument Comput.*, 8(1):61–87, 2017.

[387] Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. Argumentation in the 2016 US presidential elections: annotated corpora of television debates and social media reaction. *Lang. Resour. Evaluation*, 54(1):123–154, 2020.

[388] Jacky Visser, John Lawrence, Jean H. M. Wagemans, and Chris Reed. Revisiting computational models of argument schemes: Classification, annotation, comparison. In *Computational Models of Argument - Proceedings of COMMA*, volume 305, pages 313–324. IOS Press, 2018.

[389] Nadav Voloch, Nurit Gal-Oz, and Ehud Gudes. A trust based privacy providing model for online social networks. *Online Soc. Networks Media*, 24:100138, 2021.

[390] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 176–187. Association for Computational Linguistics, 2017.

[391] Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP*, pages 49–59. Association for Computational Linguistics, 2017.

[392] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING*, pages 3753–3765. Association for Computational Linguistics, 2018.

[393] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 241–251. Association for Computational Linguistics, 2018.

[394] Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP*, pages 68–78. Association for Computational Linguistics, 2018.

[395] Douglas Walton. Argumentation theory: A very short introduction. In *Argumentation in Artificial Intelligence*, pages 1–22. Springer, 2009.

[396] Douglas Walton. Argumentation schemes and their application to argument mining. *Studies in Critical Thinking, ed. JA Blair, Windsor Studies in Argumentation*, 8:177–211, 2019.

[397] Douglas Walton and Thomas F Gordon. The carneades model of argument invention. *Pragmatics & Cognition*, 20(1):1–31, 2012.

[398] Douglas Walton and Fabrizio Macagno. A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245, 2015.

[399] Douglas Walton and Chris Reed. Argumentation schemes and enthymemes. *Synth.*, 145(3):339–370, 2005.

[400] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[401] Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *Trans. Assoc. Comput. Linguistics*, 5:219–232, 2017.

[402] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the seventh symposium on usable privacy and security*, page 10. ACM, 2011.

[403] Cedric Waterschoot, Ernst van den Hemel, and Antal van den Bosch. Detecting minority arguments for mutual understanding: A moderation tool for the online climate change debate. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6715–6725. International Committee on Computational Linguistics, 2022.

[404] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*, pages 486–497. Springer, 2005.

[405] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

[406] Wen Wu, Li Chen, and Yu Zhao. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction*, 28(3):237–276, 2018.

[407] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. *CoRR*, abs/2204.07741, 2022.

[408] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, volume 11839 of *Lecture Notes in Computer Science*, pages 563–574. Springer, 2019.

[409] Huihui Xu, Jaromír Šavelka, and Kevin D Ashley. Using argument mining for legal text summarization. *Legal Knowledge and Information Systems JURIX*, pages 184–193, 2020.

[410] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 5754–5764, 2019.

[411] Yuxiao Ye and Simone Teufel. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678, 2021.

[412] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. Building more explainable artificial intelligence with argumentation. In *Proceedings*

of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), pages 8044–8046. AAAI Press, 2018.

[413] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. In *NAACL HLT, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141. The Association for Computational Linguistics, 2016.

[414] Ingrid Zukerman, Richard McConachy, and Sarah George. Using argumentation strategies in automated argument generation. In *INLG 2000 - Proceedings of the First International Natural Language Generation Conference*, pages 55–62. The Association for Computer Linguistics, 2000.

[415] Ingrid Zukerman, Richard McConachy, and Kevin B. Korb. Bayesian reasoning in an abductive mechanism for argument generation and analysis. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence AAAI*, pages 833–838. AAAI Press / The MIT Press, 1998.