Universitat Politècnica de València

Departamento de Sistemas Informáticos y Computación

PhD Dissertation

Juan Javier Sánchez Junquera

Detecting Deception, Partisan and Social Biases

Advisors
Paolo Rosso
Universitat Politècnica de València, Spain
Manuel Montes y Gómez
Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
Simone Paolo Ponzetto
University of Mannheim, Germany

May 2022

This PhD thesis was funded by the MISMIS-FAKEnHATE research project (PGC2018-096212-B-C31) of the Spanish Ministry of Science and Innovation.

Abstract

Today, the political world has as much or more impact on society than society has on the political world. Political leaders, or representatives of political parties, use their power in the media to modify ideological positions and reach the people in order to gain popularity in government elections. Through *deceptive* language, political texts may contain *partisan and social biases* that undermine the perception of reality. As a result, harmful political polarization increases because the followers of an ideology, or members of a social category, see other groups as a threat or competition, ending in verbal and physical aggression with unfortunate outcomes.

The Natural Language Processing (NLP) community has new contributions every day with approaches that help detect hate speech, insults, offensive messages, and false information, among other computational tasks related to social sciences. However, many obstacles prevent eradicating these problems, such as the difficulty of having annotated texts, the limitations of non-interdisciplinary approaches, and the challenge added by the necessity of interpretable solutions.

This thesis focuses on the detection of partisan and social biases, taking *hyperpartisanship and stereotypes about immigrants* as case studies. We propose a model based on a masking technique that can detect *deceptive* language in controversial and non-controversial topics, capturing patterns related to style and content. Moreover, we address the problem by evaluating BERT-based models, known to be effective at capturing semantic and syntactic patterns in the same representation. We compare these two approaches (the masking technique and the BERT-based models) in terms of their performance and the explainability of their decisions in the detection of hyperpartisanship in political news and immigrant stereotypes. In order to identify immigrant stereotypes, we propose a new taxonomy supported by social psychology theory and annotate a dataset from partisan interventions in the Spanish parliament. Results show that our models can help study hyperpartisanship and identify different frames in which citizens and politicians perceive immigrants as victims, economic resources, or threat. Finally, this interdisciplinary research proves that immigrant stereotypes are used as a rhetorical strategy in political contexts.

Resumen

En la actualidad, el mundo político tiene tanto o más impacto en la sociedad que ésta en el mundo político. Los líderes o representantes de partidos políticos hacen uso de su poder en los medios de comunicación, para modificar posiciones ideológicas y llegar al pueblo con el objetivo de ganar popularidad en las elecciones gubernamentales. A través de un lenguaje *engañoso*, los textos políticos pueden contener *sesgos partidistas* y *sociales* que minan la percepción de la realidad. Como resultado, los seguidores de una ideología, o miembros de una categoría social, se sienten amenazados por otros grupos sociales o ideológicos, o los perciben como competencia, derivándose así una polarización política con agresiones físicas y verbales.

La comunidad científica del Procesamiento del Lenguaje Natural (NLP, según sus siglas en inglés) contribuye cada día a detectar discursos de odio, insultos, mensajes ofensivos, e información falsa entre otras tareas computacionales que colindan con ciencias sociales. Sin embargo, para abordar tales tareas, es necesario hacer frente a diversos problemas entre los que se encuentran la dificultad de tener textos etiquetados, las limitaciones de no trabajar con un equipo interdisciplinario, y los desafíos que entraña la necesidad de soluciones interpretables por el ser humano.

Esta tesis se enfoca en la detección de sesgos partidistas y sesgos sociales, tomando como casos de estudio el *hiperpartidismo y los estereotipos sobre inmigrantes*. Para ello, se propone un modelo basado en una técnica de enmascaramiento de textos capaz de detectar lenguaje *engañoso* incluso en temas controversiales, siendo capaz de capturar patrones del contenido y el estilo de escritura. Además, abordamos el problema usando modelos basados en BERT, conocidos por su efectividad al capturar patrones sintácticos y semánticos sobre las mismas representaciones de textos. Ambos enfoques, la técnica de enmascaramiento y los modelos basados en BERT, se comparan en términos de desempeño y explicabilidad en la detección de hiperpartidismo en noticias políticas y estereotipos sobre inmigrantes. Para la identificación de estos últimos, se propone una nueva taxonomía con fundamentos teóricos en sicología social, y con la que se etiquetan textos extraídos de intervenciones partidistas llevadas a cabo en el Parlamento español. Los resultados muestran que los enfoques propuestos contribuyen al estudio del hiperpartidismo, así como a identificar cuándo los ciudadanos y políticos enmarcan a los inmigrantes en una imagen de víctima, recurso económico, o amenaza. Finalmente, en esta investigación interdisciplinaria se demuestra que los estereotipos sobre inmigrantes son usados como estrategia retórica en contextos políticos.

Resum

Avui, el món polític té tant o més impacte en la societat que la societat en el món polític. Els líders polítics, o representants dels partits polítics, fan servir el seu poder als mitjans de comunicació per modificar posicions ideològiques i arribar al poble per tal de guanyar popularitat a les eleccions governamentals. Mitjançant un llenguatge *enganyós*, els textos polítics poden contenir biaixos *partidistes i socials* que soscaven la percepció de la realitat. Com a resultat, augmenta la polarització política nociva perquè els seguidors d'una ideologia, o els membres d'una categoria social, veuen els altres grups com una amenaça o competència, que acaba en agressions verbals i físiques amb resultats desafortunats.

La comunitat de Processament del llenguatge natural (PNL) té cada dia noves aportacions amb enfocaments que ajuden a detectar discursos d'odi, insults, missatges ofensius i informació falsa, entre altres tasques computacionals relacionades amb les ciències socials. No obstant això, molts obstacles impedeixen eradicar aquests problemes, com ara la dificultat de tenir textos anotats, les limitacions dels enfocaments no interdisciplinaris i el repte afegit per la necessitat de solucions interpretables.

Aquesta tesi se centra en la detecció de biaixos partidistes i socials, prenent com a cas pràctic *l'hiperpartidisme i els estereotips sobre els immigrants*. Proposem un model basat en una tècnica d'emmascarament que permet detectar llenguatge *enganyós* en temes polèmics i no polèmics, capturant patrons relacionats amb l'estil i el contingut. A més, abordem el problema avaluant models basats en BERT, coneguts per ser efectius per capturar patrons semàntics i sintàctics en la mateixa representació. Comparem aquests dos enfocaments (la tècnica d'emmascarament i els models basats en BERT) en termes de rendiment i les seves solucions explicables en la detecció de l'hiperpartidisme en les notícies polítiques i els estereotips d'immigrants. Per tal d'identificar els estereotips dels immigrants, proposem una nova taxonomia recolzada per la teoria de la psicologia social i anotem un conjunt de dades de les intervencions partidistes al Parlament espanyol. Els resultats mostren que els nostres models poden ajudar a estudiar l'hiperpartidisme i identificar diferents marcs en què els ciutadans i els polítics perceben els immigrants com a víctimes, recursos econòmics o amenaces. Finalment, aquesta investigació interdisciplinària demostra que els estereotips dels immigrants s'utilitzen com a estratègia retòrica en contextos polítics.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The spread of information technology raises a range of changes in our society. For good or ill, information is used today to dominate most aspects of our daily life, for example, the political situation, our social relationships, healthcare, our emotional states, among others. One of the cons of this digital development, is how easily biased information is propagated and consumed across different social networks. A study from the last year [62] asked how much bias Americans believe is in the news source that they use. The 36% of respondents stated there was a fair amount of bias; a further 20% of those surveyed believed there was a great deal of bias in the news source that they use most.

The increasing amount of information, and the undeniable difficulty of discernment of the human being, make it impossible that people can manually check the veracity or distinguish all the time real from biased or intentionally uncertain information [210]. Since plenty of users fall into the misinformation and disinformation trap, one of the effects is that we live in times of increasing political and ideological polarization.

Natural Language Processing (NLP) tasks are developed to face these problems, but dealing with the detection of bias in social media represents a big challenge. Among others difficulties, it has to be taken into account different scenarios, domains, communicative strategies, and in some cases more than one task is involved. There are many areas in which bias and false information are present; in each one, the psychological, sociological, political, and linguistic aspects should be considered.

In this thesis, we aim to explore the detection of partisan and social biases from political narratives. It is known that partisanship has a powerful influence on attitudes and behaviours [24]. It not only shapes citizens' perceptions of the political world [14], but also of what is happening in society [115]. Journalists and politicians with a strong partisan loyalty represent the ideology and goals of their political party. For example, when politicians tell a story about a social group (e.g. immigrants), they focus very often on the

characteristics of the collective that are relevant for their partisan orientation. To do this, they make use, intentionally or not, of deceptive language that leads us to recreate a fragmented perception of reality.

Our research is focused, in particular, on the study of hyperpartisanship detection in news, and the identification of immigrant stereotypes from partisan interventions in public political debates. Since deceptive behaviour has been studied to be present also as part of political actions, we aim also to explore the detection of deceptive texts as a starting point of the approach that we propose.

The research is structured as follows. First, we address the deception detection task proposing a human-understandable method that we evaluate on datasets of different nature (i.e., different domains, and different psychological implications to the deceiver). The obtained results serve as solid evidence to extend the applicability of the method to factual and non factual information. Therefore, we extend the study to the detection of hyperpartisanship in news: we adapt the proposed method to compare both style and topic-based approaches. Next, we analyse the official communication strategy of the main politicians in the campaign of the Spanish election of 10th November 2019, and observe a prevalent potential bias regarding the immigration topic. We finally apply the proposed approach to the identification of immigrant stereotypes coming from partisan interventions in parliamentary speeches; to do this, we build a new annotated dataset. Furthermore, we propose a new approach to the study of immigrant stereotypes elaborating a taxonomy that we use to annotate the dataset. With this data we evaluate the effectiveness of our method to identify immigrant stereotypes, and analyse how the politicians make use of them as a rhetorical strategy in favour of their partisanship.

The rest of this chapter introduces a motivation of the relationship of the main tasks that this thesis involves. The next sections overview the reference methods for the detection of deception, hyperpartisan news, and immigrant stereotypes. We end the chapter presenting the research questions, the contributions, and the structure of the thesis.

1.1 Motivation

As of January 2021, Western and Northern Europe ranked first with a social media penetration rate of 79% each one, followed by Northern America with 74% [82]. Nowadays, people consume information with or without their intentions. The disseminated information has great repercussion on the perception of reality that we live, and therefore, in the decisions we make. Palpable examples of information that affects our day to day comes from news, rumors, personal opinions, reviews about products or services, political speeches, among others. In political contexts, the spread of information

generated to make a political position or candidate seem more attractive, could have a lasting impact with proven effects on voter behaviour and consequent political outcomes. This partisan bias is present in electoral campaigns by, for example, social media (e.g. Twitter and Facebook posts), fake news, hyperpartisan news, political debates, and parliamentary speeches.

By using specific linguistic means, politicians can fulfill their own political goals, which are intended to shape people's thinking and persuade them to act as they want [3]. The political action, in relation with the truth-telling and deception inevitably derives in conditions of partisan and hyperpartisan silencing [185]. The hyperpartisan news show an extreme manipulation of the reality based on an underlying and extreme ideology. They spreads much more successfully than mainstream news, and very often are inflammatory, emotional, and riddled with untruths [151]. Hyperpartisan and fake news detection have been framed into knowledge, context, and style-based paradigms which present, mostly the latter, common aspects with deceptive texts detection. To automatically prevent the misinformation and disinformation in the political context, and the ulterior consequences, computational efforts have examined linguistic attributes across multiple domains on political news by fact-checking statements of varying levels of graded deception [157]. Other works serve as examples of the applicability of deception detection features (e.g. readability scores) for distinguishing hyperpartisanship in real political news articles [151, 194].

Another information easily disseminated in social media and very often in the political context is regarding social phenomena: sometimes the spread of social bias helps politicians to gain popularity over them, or to gain more visibility because of their controversial point of view. For example, it is not casual the coincident rise of far-right wing political parties with the rapid rate of European immigration [41, 192]. These parties appeal to fears and anti-immigrant sentiments in the native population, and support the spread of offenses, incitements to hate, and violent speech [26]. In addition, other findings indicate that increased social and political trust are associated with low stereotyping and prejudice against immigrants [2]. *Stereotypes*, even if they reflect social realities, can *deceive* us and lead us to misperceptions and misjudgements [154]. At the end, stereotypes include general images about group of people, disregarding the great diversity of the group and highlighting a small set of their characteristics, which can be seen as fallacies or a kind of deception [139, 210, 211].

In this thesis, we investigate: (i) the possibility of detecting *deception* across different domains, including controversial and non-controversial topics; (ii) the detection of *hyperpartisanship* in news; (iii) and the identification of immigrant *stereotypes* expressed within *partisan* interventions in political debates.

1.2 Deception Detection

Deceiving is familiar to any human being regardless of variables such as the culture, social status or age. It can be expressed in multiple ways: omission of information, exaggeration, half-truths, literal truths designed to mislead one or more people, among others. It is a strategic act using the words or disregarding them. Examples of deception are found in daily life, in direct communication between people or through the Web in opinions about products or services, controversial opinions, statements in trials or interrogations, job interviews, financial reports, political campaigns, or user profiles [43, 210, 211, 221].

Among the vast scenarios in which we are deceived, we are willing a permanent psychological manipulation from politicians and political news as an influence technique created to change the behaviour or beliefs of its target audiences. Donald Trump, who has been a very studied politician in this regard, could be mentioned as an example, specially when he talked about immigrants [92]. Through distraction, deception, misrepresentation, exploiting rumors, conspiracy theories, fake news, and others, Donald Trump has sparked fanatical fears that have little basis in objective reality, but ring viscerally true to many people [79].

Psychologists have proven that when someone has the intention to deceive this has a cognitive load and psychological effects that are reflected in what and how it is said [43, 210]. Computational Linguistic approaches explode the use of textual features such as psycho-linguistic categories, sequences of words or characters (i.e., n-grams), part-of-speech tags (i.e., POS tags) and more complex representations that use deep learning models [97, 151]. Among the more referenced psycho-linguistic resources, the Linguistic Inquiry and Word Count (LIWC) dictionary [136] has been applied in deception detection in contexts such as fake reviews [164], and deceptive controversial opinions [109, 148, 149]. One of the deep learning methods that have been used in this task are the Transformers, whose text representations, in comparison to others, achieve very often the highest results [61]. In [13] for example, the authors used BERT, which beats the state of the art on the Deceptive Opinion Spam corpus [140], and obtained also interesting results about further part-of-speech analysis that indicates that deceptive texts are more formulaic and less varied than truthful texts.

Several cues of deception have been identified when it is taking place in specific domains or psychological contexts. For example, the use of pronouns in scenarios where the possibility of being caught implies serious risks tend to be associated to the truth-teller rather than the deceiver [78, 211]. However, as soon as the risk decreases the use of pronouns starts to reflect sort of tactics to better deceive and create more convincing stories [140]. All this makes the problem interesting and challenging because no identified cue can be taken as an universal indicator of deception [43, 211]. In addition, supervised

approaches of Machine Learning (ML) require annotated examples to learn the patterns to distinguish truth from deception. The required annotations are not only challenging because time consuming, but also because of the poor human skills as detectors and the need to design collection protocols for each domain of interest. Therefore, these problems to deal with deceptive texts from a computational perspective, demand to have methods with good versatility (e.g. less domain dependency) to be applied.

In this research we are aware of the fact that deception detection is relevant for other tasks concerning political actions as it was discussed in Section 1.1. In such scenarios, the deceptive behaviour can be then influenced by many variables such as the domain and different psychological contexts. For this reason, in this thesis we are interested in a deception detection approach having in mind those two aspects: domain and psychological implications of the deception. In particular, we propose a cross-domain deception detection model to evaluate the difficulty of detecting deceptive texts in domains with lack of annotated examples, but using a model trained with annotated data from another domain. To better analyse how the domain and psychological implications influence the usability of the learned patterns on different target domains, we use two sets of annotated data in the evaluation. First, three datasets of fake reviews about doctors, restaurants and hotels; and three datasets of deceptive controversial opinions about death penalty, abortion, and the best friend. In the three datasets on fake reviews, the deceivers do not have too many psychological implications of being caught; in the three datasets of deceptive controversial opinions, there could be high psychological implications for the deceiver after expressing the opposite of what he/she actually believes.

The proposed method is motivated by the good performance of *text-masking techniques* used for other tasks such as authorship attribution and thematic text clustering [76, 188]. With this technique, it is possible to configure to what one prefer to give more attention, if the writing style, or the content of what it is said, which could help in further analysis, specially in tasks that are emerging like hyperpartisanship and immigrants stereotypes detection.

1.3 Hyperpartisanship Detection in Political News

False information is easily spread and consumed today through several ways, each one with different particularities. Fake news, propaganda, rumors, clickbait, satire, and hoax are some examples of false information that have been studied from computational perspectives [134, 218]. Other types of false information sometimes are published in political news when they are extremely one-sided, i.e., hyperpartisan [151]. Hyperpartisanship in political news exhibits blind, *prejudiced*, and unreasoning devotion to one political party. The

way in which they are written stimulates emotional feelings among the readers in order to create a polarization among the content consumers [178].

Seminal work on hyperpartisanship finds out that the left-wing and right-wing documents have stylistically more similarity than documents from either orientation with mainstream documents [151]. Style-based approaches achieved better predictions than topic-based ones at distinguishing hyperpartisan from mainstream content, and topic-based approaches predicted better the partisan orientation in general. To shed light on the state of the art of the hyperpartisanship detection problem, SemEval-2019 Task 4 proposed new annotated datasets [94] and the participants provided different solutions to carry out the detection of hyperpartisanship. Among the different approaches, it was proposed an ELMO sentence representation convolutional network [87]; representations with lexical and semantic features [187]; linguistic features to measure the style [36]; while the authors of [106] proposed the de-noising of datasets weakly annotated. The report about SemEval-2019 Task 4 concludes by claiming that word embeddings were used to the best effect compared to other evaluated features, and that hyperpartisan news detection already demands approaches that include human-understandable explanations [94].

The authors of [50] proposed a model to detect hyperpartisan news using the same algorithms that they report to classify fake news. Some of the differences between hyperpartisan and non hyperpartisan news that were found, count with that the average number of words, sentences, adjectives in hyperpartisan news are higher than in the rest. With respect to the deceptive information contained in fake news, fake news contain on average fewer adjectives and superlatives than real news, but hyperpartisan articles contain more adjectives versus authentic news. Furthermore, in [105] it is used ELMO to develop a classification model that includes also bias features, in particular, bias word score generated from bias lexicon.

Recently, there have been more studies related to hyperpartisanship from other areas of science such as psychology, communication, and cognitive science. For example, the Facebook reactions were studied as emotional responses to hyperpartisan political news using Facebook pages before, during, and after the 2016 U.S. Presidential Election. Additionally to the emotional reactions, also were studied the political topics, rhetorical devices, stylistic devices, and emotionally charged content [216]. Moreover, the authors of [165] focused on the association between analytic thinking and the judgment of politically consistent hyperpartisan headlines.

In a recent work on detecting false information, in particular, propaganda in news articles [218], the authors propose to show the use of deception techniques as a way to offer interpretability in their solution. In a similar way, in this thesis we address the problem of hyperpartisanship detection in political news using the *text-masking technique* that we found effective at detecting deceptive texts. The flexibility of this technique makes it possible to com-

pare the style from the content-based approach. We address the problem also by evaluating BERT-based models that have been reported with good performances in other NLP tasks, and are effective at capturing semantic and syntactic patterns in the same representation. We compare these two different approaches (the masking technique and the BERT-based models) in terms of performance and also evaluating their explainable abilities.

1.4 Immigrant Stereotype Identification

Stereotyping is a complex social phenomenon involving over-generalized beliefs about a particular group of people, e.g. "Asians are good at math" or "women are bad drivers". We have known from the beginning of social psychology that stereotypes are at the base of *prejudice* and discrimination towards minorities, and that spreading prejudices is an efficient strategy for dogmatic groups and authoritarian ideologies [99, 184]. As a result, stereotyped minorities are victims of violence, hate speech, toxicity, among others. To mitigate and prevent all these consequences, many researches are improving and proposing automatic approaches to detect trolling, aggression, cyberbullying, violence incidents, offensiveness, and also implicit hate speech where stereotypes and figurative language are wrapped in irony or sarcasm [64, 124, 144]. Perhaps, the first step of all for the sake of the debiasing, is to be aware of the presence of social bias [9, 174, 214].

The automatic detection of social bias, in general, has been directed in many works to problems where two opposite social groups can be represented, e.g. gender and racial biases [20, 104]. Word embeddings are used very often to identify directions that primarily encode information used to create a semantic subspace to represent each opposite social group, e.g. women vs. men [18]. With these representations, it has been possible to measure and mitigate bias from a large number of texts that reflect the perception of the reality in society across time [70].

Motivated to study subtle ways of harmful social biases in which implied meanings are expressed, the authors of [175] introduce Social Bias Frames (SBF), a conceptual formalism that aims to model the pragmatic frames in which people project social biases and stereotypes onto others. SBF helps to distill potential language biases in a way that considers the offensiveness, intent of the speaker, as well as explanations of why the implication is biased, using knowledge about social dynamics and stereotypes. Some of the variables that the authors took into account in this formalism are: offensiveness, the intent to offend, lewd or sexual references, group implications, the mention of the targeted group, and examples representing the implied statements.

However, the presence of frame is not new in the academic literature. Research on framing is characterized by theoretical and empirical vagueness

[177]. Frame has been used as a concept with different nuances in social sciences and humanities [53, 68]. In Computational Linguistics there is also a variety in how frame is assumed. For example, the authors of [65, 175] use pragmatic frames as a set of implicit elements that give meaning to an event. In [29, 122, 134] framing refers to the definition of [53] in social science: *the selection of particular aspects of an issue that makes them salient in communicating a message*. In [100], the term is used as schemata of interpretation that is employed to structure experiences, interpret events, and make sense of ambiguous information. In this thesis, we also the concept of frame as in [53]. We share with [175] that to understand how stereotypes are used in natural language it is necessary to focus on non-explicit content, but we address it in a different way: the explicit is the frame, i.e. the situation or argument in which the social category is referred to (e.g. *economical resource* is many times a frame used to refer to immigrants); and the implicit is the process of constructing a meaning that this action of framing produces.

Nowadays, among the stereotypes about minority groups, those related to immigrants are one of the social biases more controversial in political speeches. Politicians create and recreate a frame [177], a kind of scenario, where they speak about immigrants, building a distorted image that in some cases is just a fallacy. However, most semantic narratives, to date, do not capture such pragmatic implications in which people express stereotypes [175]. Recently, the authors of [173] used a dataset related to immigration but mainly focused on the expressions of hate speech. Some other social biases have being taken into account in datasets used in [60, 133], but not being the immigrant stereotypes the main focus. Therefore, among the limitations that we have found, we see the need for datasets annotated considering the whole spectrum of immigrant stereotypes.

Besides some datasets that have been used related to immigration, they are mainly focused on the expressions of hate speech [173], and other social biases that include racism but not being the immigrant stereotypes the main focus [60, 133]. In this respect, we aim to build a dataset with parliamentary debates to analyse the immigrant stereotypes that take part in their *partisan* interventions.

As it has been point out in [176], members from the same party express similar viewpoints and support the motion under debate. In that work, the authors used BERT for political speech stance analysis combining semantic language representations and relations between debate transcripts, motions, and political party members. In this thesis, we think that same party members could express also similar viewpoints regarding immigration and immigrant lives.

We aim to address stereotypes about immigrants as a result of the activity of framing in *partisan* interventions of political speeches. Immigrants are very often seen as people whose more salient image is their presence in the country as a source of cultural, personal, or collective threat; victims of

suffering, people who lose their lives or live with serious problems; or also people whose main image is associated with the economy of the country that receives them. Taking these several images/frames about immigrants in mind, we think that immigrant stereotypes are integrated to *partisan bias* in political speeches. In particular, we think that some images about immigrants tend to be more frequently perceived by specific partisan ideologies. Therefore, it could be possible to identify immigrant stereotypes in partisan-biased texts, through an approach that has already been evaluated in the detection of hyperpartisanship.

We propose a taxonomy that focuses on different frames that politicians use to speak about immigrants. We use this taxonomy to annotate a new dataset (StereoImmigrant) that we create with political speeches from the principal parliament in Spain, the Congress of Deputies (*Congreso de los Diputados*). We evaluate the masking technique and transformers at identifying immigrant stereotypes in the StereoImmigrant dataset. Finally, we also provide examples of how our approach can help in supporting further analysis of human experts in social psychology. Similar to what the authors of [176] do to analyse political cohesion and partisan identities, we also use the attention mechanism to analyse the words that receive high scores in the decision making process for identifying stereotypes against immigrants. In order to address the problem of explainability [85, 215], we make a comparison between the results of both approaches, and compare if such different models look at the same words in their predictions.

1.5 Research Questions

The research questions we aim to answer in this thesis are:

- **RQ1:** *Can **deceptive language** be detected employing the masking technique taking into account both content and style in cross-domain scenarios?* Deceptive language is one of the strategic instruments used in political and social contexts. It is well known that how people deceive depends on many variables such as the domain, the culture, age, or gender [156]. In this thesis we propose an approach to detect deception in different domains and in cross-domain scenarios. In our proposal, we take into account the differences between truthful and deceptive texts regarding style and content from one domain to another.

- **RQ2:** *Can **hyperpartisanship in political news** be addressed from a deception detection perspective?* Besides several approaches that have been evaluated recently to detect hyperpartisanship, there are still effective state of the art methods applied in other tasks that have not been evaluated with hyperpartisan texts. Based on the studies that

confirm the presence of deceptive behaviour in political actions that derive in hyperpartisan silencing, we propose to detect hyperpartisanship considering a deception detection perspective

- **RQ3:** *How can be approached the detection of social biases like* ***stereotypes*** *against immigrants in political speeches considering the manipulative strategies of this kind of narratives?* Immigrant stereotypes have been approached without considering the whole spectrum of this complex phenomenon. We propose and compare two models that take into account both positive and negative beliefs about immigrants, and also the variability and subtle way in which stereotypes are reflected in political speeches.

- **RQ4:** *Can the masking and Transformer-based models* ***help human experts to further analyse*** *the above problems?* Similar to applications of healthcare and security, in the above tasks (deception detection, hyperpartisanship detection in political news, and immigrant stereotype identification in partisan interventions) it is not enough to achieve high results, but it is also necessary that results could be understood by human experts in the domain of study (e.g. social psychologists). We compare how two approaches diametrically opposite to each other can help to this goal.

1.6 Contributions

In this section we summarize the main contributions of the thesis.

We show that **deceptive language** from different domains can be represented as a combination of relevant content and style-related features. We proved that, in cross-domain scenarios, the proposed masking technique is effective at learning deceptive-related cues in the source domain, and can be employed to detect deception in the target domain. We used benchmark datasets to carry out our experiments in domains where the deceiver can feel different psychological implications at elaborating the lie. The results show that our approach can capture deceptive behaviour in narratives where the moral and ethic of the person can be judged or criticized (e.g. in controversial opinions).

Another contribution of this thesis is made in the context of **partisan bias**, in particular, we study the hyperpartisanship detection in political news. We proved that the partisan orientations can be predicted from a *deceptive detection* viewpoint achieving comparable results to the state of the art. We show that the masking technique offers versatility in the sense of being adaptable to address the problem with both a style or a topic-based approach. Moreover, we will see that the results indicate that Transformers can capture more complex patterns achieving the highest results. Addition-

ally, we prove that hyperpartisan news can be predicted with the beginning of the news.

We also studied the detection of **social bias** contributing specifically in the task of immigrant **stereotype identification**. From the theoretically viewpoint, we propose a new taxonomy that covers the whole spectrum of beliefs that make up the immigrant stereotype. We propose the first work that addresses the classification of this social bias taking into account the narrative contexts instead of the characteristics attributed to the group. The results show that our approach is effective at identifying the frames in which the immigrants are placed in political speeches. Furthermore, we propose the StereoImmigrant dataset, which is publicly available, and has been annotated in collaboration with an expert in social psychology.

Finally, we show how our approach can help to human expert analysis in the three tasks that we focus on, to understand what are the linguistic patterns that are often used in texts that contain *deceptive* or *partisan* information, as well as *stereotypes*.

1.7 Structure of the Thesis

The core of this PhD thesis is presented as a compendium of research articles (from Chapter 2 to Chapter 6) which were published during the study phase of the PhD candidate. The structure that follows contains a chapter dedicated to each of the articles, a chapter with a general discussion of the results together with some new experiments, and finally the conclusions. Next, we briefly overview the content of the remaining chapters.

- **Chapter 2: Masking domain-specific information for cross-domain deception detection.**

  In this chapter we present our work published in the Pattern Recognition Letters journal. In that paper we addressed the problem of cross-domain deception detection. We used a masking technique to obtain a text representation general to both domains, in that way, the particularities of the source domain are ignored by the model. Only the deceptive features that both domains have, are those that the model learns in the source to be used in the target. The results show that this technique is useful for detecting deceptive language from a domain without annotated examples, but employing a model trained with examples from another domain. With this paper we also proved that this technique can be used not only in domains such as reviews, but also in others that contain non factual controversial information.

- **Chapter 3: Masking and Transformer-based Models for Hyperpartisanship Detection in News.**

In this chapter we present our work published in the conference of the Recent Advances in Natural Language Processing (RANLP). This publication describes our work in detecting hyperpartisan political news with the masking technique and BERT-based models. We compared the development of these methods in terms of results and compared the parts of the text that they focused more in the classification. We also tested the masking technique considering style and topic-based approaches.

- **Chapter 4: A Twitter Political Corpus of the 2019 10N Spanish Election.**

  This chapter describes our publication in the International Conference on Text, Speech, and Dialogue (TSD). In that work we collected tweets of the main five parties (PSOE, PP, Cs, UP and VOX) covering the campaign of the Spanish election of 10th November 2019. We analysed the different topics discussed in the tweets, and the sentiments and emotions employed in them. The results indicated that each party was biased to pay more attention in some specific topics than in others, and that some topics (e.g. Immigration) received exclusive attention from only the far-right party.

- **Chapter 5: How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants.**

  This chapter is composed by our work published in the Applied Sciences journal. In that work we studied the *partisan bias* toward the immigration topic in the speeches of the Congress of Deputies. We created a corpus focused on the annotation of immigrant stereotypes. We proposed a taxonomy that involves the whole view of this social bias. The new taxonomy, described in detail, was designed to make annotations not only in terms of Stereotype and Non-Stereotype labels, but also regarding different frames about immigrants (e.g. victims of suffering, economical resources, personal threats, among others). Preliminary experiments showed that for stereotype identification, the traditional classifiers achieved competitive results compared to BERT-based models.

- **Chapter 6: Masking and BERT-based Models for Stereotype Identification.**

  This chapter presents our publication in the conference of the Sociedad Española de Procesamiento del Lenguage Natural (SEPLN). In this work, we used the masking technique and the BETO Transformer model for Spanish to detect stereotypes about immigrants with two different explainable approaches. We made a comparison between the

relevant features of one model, and the features with the highest attention of the other. The results showed a trade-off between performance and explainability.

- **Chapter 7: Discussion of the Results.**

  In this chapter we discuss the results achieved throughout the doctoral thesis from an integrative perspective. We complement our study developed in the previous chapters with some further experiments about local explanations of the deceptive language using the attention mechanisms; we show the robustness of the masking technique in the detection of hyperpartisan news; add new experiments and results that make evident the differentiated effort that politicians have in terms of advertising close to electoral campaigns; and we study how the use of social bias like the immigrant stereotypes are used as a rhetorical strategy according to the partisan bias.

- **Chapter 8: Conclusions and Future Work.**

  In this chapter we answer the research questions of Section 1.5, and draw some conclusions. Moreover, we comment the open research lines for possible future works.

Chapter 2

Masking Domain-specific Information for Cross-domain Deception Detection

This chapter presents the masking-based model that we propose in this thesis. In particular, here we focus on using the model for the detection of deceptive texts. We show the performance of the model in cross-domain scenarios with annotated data in texts concerned with facts, and also texts that give personal interpretations and beliefs on controversial topics. In addition, we show examples of how this technique can be used to visualize the relevant deceptive cues.

The work presented in this chapter was published in the following paper:

- **Sánchez-Junquera J.**, Villaseñor-Pineda L., Montes-y-Gómez M., Rosso P., Stamatatos E. (2020) Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, vol. 135, pp. 122-130 (**Impact Factor: 3.756 Q1**)

Abstract

The facilities provided by social media and computer-mediated communication make easy the dissemination of deceptive behaviour, after which different entities or people could be affected. The deception detection by supervised learning has been widely studied; however, the scenario in which there is one domain of interest and the labeled data is in another domain has received poor attention. This paper presents, to our knowledge, the first domain adaptation approach for cross-domain deception detection in texts. Our proposal consists in modifying original texts from the source and target domains in a form in which common content and style information is maintained, but domain-specific information is masked. In order to adequately select domain-specific terms to be masked, the proposed method uses unlabeled instances from both domains. Our experiments demonstrate that the masking technique is a good idea for detecting deception in cross-domain scenarios; and the performance could be further improved if unlabeled information from the target domain is considered.

Keywords: Deception detection, Domain adaptation, Masking information

2.1 Introduction

Over the years, human beings have found in deception a tool that provides either protection or another type of personal gain. Today, the presence of deception is becoming increasingly noticeable and harmful, e.g. due to the facilities provided by technology and the web. Deception refers to the attempt to create in another a belief which the communicator considers false [210]. For example, the fake service reviews that try to deliberately mislead customers; or lies that protect oneself from disapproval and manage others impressions outside boundaries of honesty [43]. In many cases, the importance of catching liars is due to the undesirable consequences of deception in online reviews, trial hearings, predatory communication, among others [140, 148, 166].

Text classification techniques have been extensively used to detect deception. For this approach it is necessary to acquire labeled data sets, which are traditionally constructed from manual labeling. Manual labeling is complex and expensive, especially in deception detection, due to the poor human skills as detectors and the need to design collection protocols for each domain of interest. Given this difficulty, it is essential to be able to use cross-domain solutions which employ labeled data from one domain for the classification of deception in other domain.

Previous work has shown that cross-domain approaches present a difficulty in detecting deception. The problem is that many cues to deception change from one domain to another due to the change in content, the consequences if the deceiver is getting caught lying, and the emotions experienced

by the deceiver due to the topic [43, 211]. For example, pronouns in one domain (e.g. essays on abortion) can be an indicator of deception while in another domain (e.g. reviews on hotels) they can describe truthful texts [140].

However, works such as [55, 149] have shown that both the style-related information and content words may be relevant for the detection of deception in cross-domain scenarios. These works have evaluated their proposals taking into account only characteristics of the source domain and ignoring those from the target domain. Hence, it might be possible to identify common characteristics (related to content or style) between the source and target domains, to obtain a more general representation of their texts.

We propose, to our knowledge, the first domain-adaptation method for deception detection that uses information from source and target domains. This method is a contribution to both the deception detection task and the cross-domain problem. Our method is inspired by the text distortion approach successfully used in thematic text clustering and authorship attribution [77, 189], but modified to be more suitable to deception detection and to be used as a domain adaptation approach. Its main idea is to transform original texts from the source and target domains by masking domain-specific terms. Source and target domains are observed to pick out the terms specific to only one of them. While the textual structure, the style-related information, and the common content words are maintained, the picked out domain-specific terms are masked obtaining a more general text representation. Our experiments show that the proposed method can improve the cross-domain classification between domains of online reviews or essays about controversial topics.

2.2 Related Work

Computational works have shown important results in deception detection. First of all, such works confirm that human judges make more mistakes in detecting deception in comparison to automated methods [140]. However, supervised learning studies are limited due to lack of appropriate corpora for deception detection. Furthermore, different kinds of features have been explored for the text representation in order to detect deception.

Earlier works mainly focused on single-domain scenarios for which many of them propose traditional techniques based on simple text representations. [149] and [140] demonstrated that truthful and deceptive texts are separable, through word n-grams, psycholinguistic features from LIWC, and part-of-speech features. More sophisticated features, such as deep syntactic patterns [55], argumentative features [34], and word embeddings [160], were also successfully evaluated. More recently, the character n-grams features have shown a good trade-off between simplicity and performance for detect-

ing deception in online reviews and essays on controversial topics [25, 171]. All these works found that both content and style are important factors to distinguish deception from truth.

There are few works that have reported results on cross-domain deception classification. They merely evaluated how the performance decreases when their models are trained on a source domain, and no information from the target domain was observed [109, 160]. These works showed interest in whether a relatively richer annotated domain could be used to train effective deception detection models for other domains, and how good the generalization ability of their models was. They suggested that the performance was affected because the target domain generally encoded some type of features different to the ones found in the source domain.

When the domain of labeled examples is different from that with the instances of interest (i.e., the cross-domain scenario), the results are affected by topic differences. This problem has been addressed in other classification tasks such as sentiment analysis and authorship attribution with domain adaptation approaches. On the one hand, a common idea in sentiment analysis is to search words from each domain that share a similar connotation [141]; or to separate the vocabulary into general words (i.e., domain-independent features) and specific words (i.e, domain-specific features) for a different usage of those specific words from source domain [200, 217]. On the other hand, in authorship attribution, [189] proposes a text distortion method which masks the occurrences of the least frequent words of the language; thus, the algorithm compresses topic information and maintains textual structure related to personal style.

2.3 Masking Domain-specific Terms for Deception Detection

Masking techniques have been applied to different tasks. On the one hand, [77] focused on masking frequent words to enhance performance in text clustering. On the other hand, based on the opposite perspective, [188] focused on masking the least frequent words to highlight style information that is used in authorship attribution. In our case, since both content and style information could be useful for detecting deception, we want to maintain both factors depending on the common information from source and target domains. For example, considering reviews about hotels and doctors as the source and target domain respectively, we could maintain function words (e.g. *the*, *my*) and common content words between the two domains (e.g. *staff*, *family*) and mask domain-specific words (e.g. *doctor*, *hotel*).

In this section, we present our domain adaptation approach. We first use the Frequently Co-occurring Entropy (FCE) to pick out domain-specific features and then we employ a distortion method to mask them.

2.3.1 Domain-specific Terms Filtering

In this work, we consider general terms as in [141]: they should occur frequently and act similarly in both the source and target domains. Subsequently, domain-specific terms are those that do not satisfy this condition. In order to achieve a trade-off between frequency and similarity of terms, we use FCE, proposed in [200]. The general formula is as follows:

$$FCE_w = \log\left(\frac{P_S(w) \times P_T(w)}{|P_S(w) - P_T(w)|}\right) \tag{2.1}$$

where $P_S(w)$ and $P_T(w)$ are the probabilities of the term w in the source and the target domain respectively[1]. In this work, as in [200, 217], we compute $P_S(w)$ and $P_T(w)$ as follows:

$$P_S(w) = \frac{N_w^S + \alpha}{N^S + 2 \times \alpha} \quad \text{and} \quad P_T(w) = \frac{N_w^T + \alpha}{N^T + 2 \times \alpha} \tag{2.2}$$

where N_w^S and N^S are the number of instances where w occurs at least once and the total number of instances, respectively, in the source domain; and N_w^T and N^T are the number of instances where w occurs at least once and the total number of instances, respectively, in the target domain. We set $\alpha = 0.0001$ in order to overcome overflow, which appears for infrequent terms in a large corpus. On the other hand, β is included[2] to deal with the extreme case when $P_S(w) = P_T(w)$:

$$FCE_w = \log\left(\frac{P_S(w) \times P_T(w)}{|P_S(w) - P_T(w)| + \beta}\right) \tag{2.3}$$

Table 2.1 shows a simple example taking reviews on hotels and doctors as the source and the target domains respectively (details of the used corpora are given in Table 2.5). We can see that *my*, *ever*, and *I* could be considered as more general terms; *needs*, *life*, and *helped* are less frequent terms and are more related to the content of both domains; however, *spa*, *consultation*, and *tests* are infrequent in at least one domain or have dissimilar occurring probability.

2.3.2 Text Distortion Methods

The main idea of the proposed method is to transform the original texts to a domain-abstract form where textual structure, related to a general style of deceivers or honest persons, is maintained while infrequent words, corresponding to domain-specific information, are masked. To this end, all the

[1] Defined as the probability of taking an instance from the corpus with the given term. No labeled data is necessary for this task.

[2] We take up on the values set in [200], i.e., $\alpha = \beta = 0.0001$.

Table 2.1: Examples of FCE results in Hotel and Doctor corpora.

w	N_w^{Hotel}	N_w^{Doctor}	FCE_w	Rank
my	987	351	2.50	1
ever	171	60	0.96	2
I	1340	402	−0.41	6
needs	35	35	−7.65	405
life	34	34	−7.69	409
helped	31	28	−7.74	425
spa	64	0	−25.15	2381
consultation	0	31	−26.67	10567
tests	0	9	−26.71	10605

occurrences (in both training and test corpora) of domain-specific terms are replaced by symbols.

Let W_k be a set of k general terms. A text is tokenized and all $w \notin W_k$ will be masked according to a specific text distortion technique. We describe Distorted View with Multiple Asterisks (DV-MA) and Distorted View with Single Asterisks (DV-SA); two text distortion methods introduced by [188]:

DV-MA: Every $w \notin W_k$ is masked by replacing each of its characters with an asterisk ($*$). Every digit in the text is replaced by the symbol #.

DV-SA: Every $w \notin W_k$ is masked by replacing each word occurrence with a single asterisk ($*$). Every sequence of digits in the text is replaced by a single symbol #.

We modify these methods by treating any token that includes punctuation marks in a special way. If the token is found to be domain-independent (e.g. commas and periods) then it is maintained. On the other hand, if it is found to be domain-specific (e.g. quotes, parentheses, or compound terms like *and/or*), it is replaced by the symbol @. Furthermore, to consider all the numeric details usually given by truthful communicators [209], we mask numerals (e.g. *one, two, three*, etc.) with a single symbol +.

An example of transforming a sentence, according to these text distortion variants, is provided in Table 2.2. In this case, W_k includes the 400 most general terms from reviews on hotels (source domain) and reviews on doctors (target domain). Table 2.3 shows an example of transforming the same input text according to DV-MA algorithm using different values of k.

We can note that $k = 0$ means that every term is considered domain-specific, therefore, even punctuation marks will be masked. However, when $k = 400$, mainly function words (e.g. *My*, *in*, *a*, *The*, *I*, *after*) and some punctuation marks are maintained, because they are not associated to a particular domain. Finally, by expanding the set of general terms to $k = 1000$, content-related terms associated with both domains are also maintained (e.g.

Table 2.2: An example of transforming a doctor review, according to two distortion techniques, observing reviews on doctors and hotels. In these transformations $k = 400$.

My Neck/S-Lift procedure performed in March 2009 was handled in a very professional manner, and I was able to attend a social event three weeks after surgery.	
DV-MA	*My @ ********* ********* in ***** ### was ******* in a very ********** ******, and I was able to ****** a ****** ***** +++++ ***** after *******.*
DV-SA	*My @ * * in * # was * in a very * *, and I was able to * a * * + * after *.*

Table 2.3: An example of transforming an input text according to DV-MA using different values of k.

My Neck/S-Lift procedure performed in March 2009 was handled in a very professional manner, and I was able to attend a social event three weeks after surgery.	
k=0	*** @ ********* ********* ** ***** ### *** ******* ** * **** ********** ******@ *** * *** **** ** ****** * ****** ***** +++++ ***** ***** *******@*
k=400	*My @ ********* ********* in ***** ### was ******* in a very ********** ******, and I was able to ****** a ****** ***** +++++ ***** after *******.*
k=1000	*My @ ********* ********* in ***** ### was handled in a very professional ******, and I was able to ****** a ****** ***** +++++ weeks after *******.*

handled, *professional*, *weeks*). Note that the terms *Neck/S-Lift*, *2009*, and *three*, are always masked.

Table 2.4 shows another example of a sentence, taken from a hotel review, transformed according to DV-MA when two different target domains are considered. Observe how the set of domain-independent terms changes when the target domain concerns reviews on either doctors or restaurants. For example, *help* is a general term in reviews on both doctors and hotels, but not on restaurants; and *location* is a general term in reviews on both restaurants and hotels, but not on doctors.

Table 2.4: A hotel review is transformed according to DV-MA with $k = 1000$ for two different target domains. Highlighted (in yellow) the general terms depending on the target domain.

Hotel review	**Target Domain**	
	Doctors	**Restaurants**
Superb location and proximity to local attractions. Staff is always friendly and eager to help.	***** ******** and ********* to ***** **********. Staff is always friendly and eager to help.	*Superb location and* ********* *to local* **********. *Staff is always friendly and eager to* ****.

2.4 Experiments

2.4.1 Datasets

We use benchmark datasets in English that include two genres: reviews (opinion spam) and essays (controversial opinions). The former comprises three domains, namely Hotel, Restaurant, and Doctor. The latter also comprises three domains, namely Abortion, Death Penalty, and Best Friend. Table 2.5 shows the statistics of the six datasets.

The datasets of reviews are parts of those collected by [109]. The truthful reviews were mined from a set of real customers and the deceptive ones were collected by crowd-sourcing. For each domain, *turkers* were asked to describe a fake experience as if it had been real.

All essays were also collected using crowd-sourcing. For Abortion and Death Penalty, participants were asked to express both their personal opinion and the opposite on that topic, imagining that they were taking part in a debate. In the Best Friend domain, participants were asked to write about their best friend and describe the detailed reasons for their friendship. Subsequently, they were asked to think about a person they could not tolerate and describe her/him as if s/he was their best friend [149].

Table 2.5: Statistics of the datasets. The number of deceptive (D) and truthful (T) instances, the average vocabulary size (per instance), as well as the average length of instances (either characters or words) are given.

		Instances		**Vocabulary**		**Length(ch)**		**Length(w)**	
Type	**Domain**	**T**	**D**	**T**	**D**	**T**	**D**	**T**	**D**
Spam	Hotel	800	800	101	95	821	791	172	164
	Doctor	200	356	66	75	465	593	97	119
	Restaurant	200	200	97	89	762	709	160	146
Controversial	Abortion	100	100	64	50	499	359	101	73
	Best Friend	100	100	51	40	337	266	72	57
	Death Penalty	100	100	60	54	463	395	93	78

2.4.2 Experimental Setup

This section presents the experimental setup that we use in our experiments.

Preprocessing: We convert all words to lowercase letters and do not remove any character (e.g. symbol, punctuation mark, number or delimiter).

Text Representation: The proposed method uses two parameters: k indicates the top general terms which will not be masked and n is the order (length) of character n-grams that represent the masked texts. We empirically select the values of k and n by performing grid search for each pair of source-target domains: $k \in \{0, 100, 200, ..., 1000\}$ and $n \in \{3, 4, 5, 6, 7\}$. Except for Figure 2.4, all reported results were using $n = 4$ and the best value for k for each case. After the masking stage, we represent the transformed texts without removing any character n-gram feature and use a binary[3] weighting scheme.

Classifier: We use the Naïve Bayes (NB) classifier. Similar performance has been obtained based on Support Vector Machines, thus we only report results for NB.

Evaluation: We use 80% of the unlabeled target domain instances and all the source domain instances for picking out domain-specific terms in an unsupervised manner (information about the class, D or T, of each instance is not used). Then, we apply masking in all the texts of both training and test sets. We train the classifier using only the source domain instances and we apply the learned model to the unobserved (20%) target domain instances. In all the experiments, to avoid over-fitting, we randomly select 80% (for the masking process) and 20% (as the test set) unlabeled instances from target domain creating two disjoint subsets; we repeat this procedure 10 times ensuring that each instance was classified two times. The results reported in all experiments are average results of these 10 individual results. We use F_1 as the evaluation measure.

Baseline: Our baseline method is based on the same text representation and classifier but without applying any distortion method. It does not use any information from the target domain.

2.4.3 Results and Discussion

Figure 2.1 shows the average and standard deviation of F_1 in cross-domain deception detection for all pairs of source and target domains in reviews and essays: the blue (DV-MA) and red (DV-SA) bars indicate the results of the domain adaptation by the proposed approach, and the green bars indicate results of our baseline. The Figure also shows a line chart with the F_1 results in the single-domain scenario for each target domain (using the same representation); e.g. above the bars of Dr->H and Rest->H (results of DV-MA, DV-SA, and baseline respectively), a line chart indicates that we

[3] *tf* and *tf-idf* were also tested, but obtained slightly lower results.

obtained $F_1 = 0.89$ when both training and test instances come from the Hotel domain.

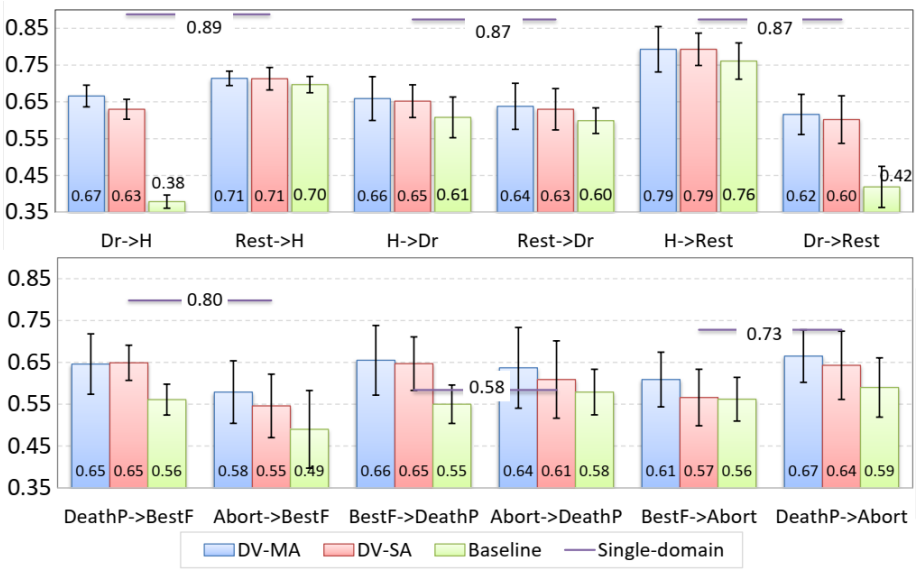


Figure 2.1: Evaluation results of the proposed approach and our baseline; the three bars show the average of F_1 and the standard deviation for the cross-domain problem (e.g. Dr->H means that Doctor is the source domain and Hotel is the target domain). For each target domain, a line chart indicates the single-domain performance.

The two variants of masking domain-specific terms, i.e. DV-MA and DV-SA, do not show significant differences in F_1. However, DV-MA tends to perform slightly better. From Figure 2.1 we can note that the proposed method always improves the performance of the baseline (in average by 14%). We suppose that the cases in which the proposed approach is only slightly better than the baseline are due to the similarity between the specific source and target domains and the little descriptive power of the source domain patterns over the target domain. Despite the fact that the proposed method demonstrates the usefulness of exploiting information from both domains and masking the domain-specific information, the differences of obtaind results with respect to those of single domain cases indicate that there is a lot of space for improvement.

Surprisingly, our method achieved higher F_1 than the single-domain evaluation in the Death Penalty corpus. We guess the reason is the difficulty of finding relevant patterns in this corpus, so the information obtained from other domains improve the performance. Similar behaviour with these controversial topics can be found in [149].

2.4.3.1 Sensitivity to the Distribution of Observed Instances

In order to filter out domain-specific terms, this work uses FCE, which does not require labeled instances. Therefore, this work can assume that there are no labeled instances from the target domain. On the other hand, deceptive and truthful instances have many terms with dissimilar distribution. In this section, we try to answer the question: is the classification accuracy affected if the majority of the observed instances in the target domain belong to a certain class? To answer this question, we compare (see Figure 2.2) the results of evaluating the proposed approach based on DV-MA by observing deceptive instances exclusively, truthful instances exclusively, or an equal number of instances of these two classes.

The results of Figure 2.2 do not consistently indicate whether or not it is better that the set of observed instances of the target domain is balanced with respect to the classes. In general, it can be noted that the results are comparable. This suggests that our method is robust to the distribution of target domain instances over the classes and the selection of domain-specific terms is not affected when more deceptive/truthful instances are included in the unlabeled data.

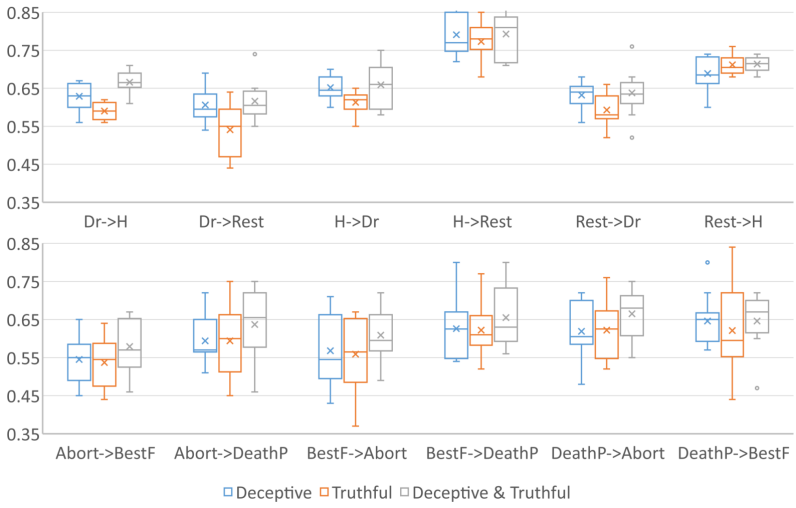


Figure 2.2: Results of DV-MA when the unlabeled instances of the target domain are exclusively deceptive, truthful, or belong to any of those classes.

2.4.3.2 The Contribution of Observing the Target Data

The proposed method differs from the baseline in two aspects. First, unlabeled data from the target domain are observed; second, a masking technique is applied. In this section we try to clarify if the improvement of our method over the baseline is due to the data observed from the target domain, the

masking technique, or both. To this end, we compare the performance of the masking method using DV-MA without access to target domain data and the proposed method using DV-MA with access to target domain data. The former does not depend on the target domain and masks the less frequent words of the English language[4]. The latter extracts domain-specific terms by applying FCE to the source and target domains.

Table 2.6 compares the results of these two methods with the baseline. The third and fourth columns are compared by printing in bold the highest value for each pair of domains and the fifth column shows, in bold and asterisk, those cases in which our method obtains the best results by observing unlabeled data from the target domain. We can see that by masking frequent words of the language our method improves the baseline (in 10 cases out of 12) by 9% in average. Although in this way domain adaptation is not actually performed since no information from the target domain is used, it can be concluded that the masking technique itself is useful to enhance performance in cross-domain deception detection. Furthermore, the results improve even more (in 9 cases out of 12) by 5% in average when information from the target domain is used and the terms to be masked are picked out accordingly. Therefore, if it is possible to observe unlabeled information from the target domain, the performance is enhanced. If, on the other hand, such information is not available, the masking technique is still useful.

Table 2.6: Average and standard deviation of F_1 results using different strategies for selecting the terms to be masked.

		Baseline	**Most frequent words in English**	**FCE**
Unlabeled data from both domains				✓
Masking technique			✓	✓
Source	**Target**			
Hotel	Restaurant	0.761 ± (0.050)	**0.779** ± (0.034)	**0.793*** ± (0.062)
	Doctor	0.608 ± (0.055)	**0.645** ± (0.044)	**0.659*** ± (0.060)
Restaurant	Hotel	0.697 ± (0.022)	**0.726** ± (0.023)	0.714 ± (0.020)
	Doctor	**0.599** ± (0.035)	0.596 ± (0.055)	**0.638*** ± (0.063)
Doctor	Restaurant	0.419 ± (0.056)	**0.554** ± (0.049)	**0.616*** ± (0.055)
	Hotel	0.379 ± (0.018)	**0.540** ± (0.026)	**0.666*** ± (0.030)
Abortion	Best Friend	0.490 ± (0.093)	**0.579** ± (0.080)	0.579 ± (0.075)
	Death Penalty	0.579 ± (0.055)	**0.647** ± (0.064)	0.637 ± (0.096)
Death Penalty	Abortion	0.590 ± (0.071)	**0.640** ± (0.058)	**0.665*** ± (0.063)
	Best Friend	0.561 ± (0.037)	**0.645** ± (0.084)	**0.646*** ± (0.072)
Best Friend	Abortion	**0.562** ± (0.053)	0.544 ± (0.075)	**0.609*** ± (0.065)
	Death Penalty	0.550 ± (0.046)	**0.594** ± (0.085)	**0.655*** ± (0.083)

[4]We extract the most frequent words of the BNC corpus (https://www.kilgarriff.co.uk/bnc-readme.html). For each pair of domains, we report the higher F_1 varying $k \in \{0, 100, 200, ..., 500, 1000, 2000, ..., 5000\}$ following the practice of [189].

2.4.3.3 Presence of Masks in Discriminatory Features

The authors of [77] concluded that in cases the textual structure was not maintained, the performance of clustering decreased. In a similar way, one may suspect that, even by transforming the original texts, the most discriminatory features are n-grams that do not include the masking characters (i.e., *, @, + , #). That way, it would not be necessary to maintain the domain-specific terms (either original or masked) in the representation (i.e., they might be removed). However, a deeper look in the most discriminatory character n-grams makes possible to note that there are many of them that include masking symbols.

Table 2.7 shows some character n-grams with high information gain for the hotel and restaurant domains as well as examples of sentences in which they occur. As it can be seen, truthful reviews on hotels or restaurants are characterized by providing numerical information, and explanatory phrases enclosed in parentheses. Thanks to the masking technique, the proposed method is not distracted by specific numbers or what was the particular clarification given. In general, it captures an abstract type of information commonly used by real customers.

Table 2.7: Features with high information gain in Hotel and Restaurant domains, with $k = 400$ and $n = 4$. The underscore symbol indicates a blank space in char n-grams. Examples of sentences where these char n-grams occur are highlighted (in yellow).

Class	**4-gram**	**Examples of information captured from original texts**	
		Hotel (source domain)	**Restaurant (target domain)**
True	*#_ **	*with 2 bathrooms*	*for lunch 2 days later*
	**...	*for the romantic couple...*	*or the salmon...*
	*_$##	*only $15 per day*	*to $10 and steaks closer to $30*
	*_(**	*in a (dark) corner*	*very (very) few places*
	*mall	*The room was very small*	*the small plates*
Deception	*_my_	*I made my reservation at*	*on my next visit*
	*I_wi	*and I will choose other*	*I will be back again*
	*I_wa	*I was able to relax*	*but I was pleasantly*
	*anyo	*if anyone carried my bags*	*to anyone looking for*
	*_rec	*I'd only recommend this*	*I would recommend this*

2.4.3.4 Effect of the Parameters' Values

In the previous experiments, we empirically select the values of k by performing grid search for each pair of domains. Figure 2.3 shows boxplots with the distribution of F_1 with all pairs of review (opinion spam) domains on one hand, and essay (controversial opinion) domains on the other. In these cases, we used DV-MA with character 4-gram features and varying $k \in \{0, 100, 200, ..., 1000\}$. For the baseline, we used character 4-gram features too and the Figure shows the average of F_1 since the baseline does not depend on k.

Figure 2.3: F_1 of DV-MA (varying k values) and baseline models.

As can be seen, the performance varies with different values of k. We conclude that it is always important to mask a set of terms ($k > 0$), possibly because two different domains have at least some terms with dissimilar distribution. At the same time, performance in general decreases with $k > 600$ for the examined corpora, indicating that terms with relatively low FCE score are actually distractful and it is better to mask them. Interestingly, with almost all pairs of domains evaluated, the proposed method improved the performance of the baseline for a wide range of values of k ($50 < k < 600$).

Similarly, Figure 2.4 shows boxplots with the distribution of F_1 of the proposed approach based on DV-MA with $k = 400$ and the baseline for various n-gram lengths. We can note that similar performance is obtained for all examined n values, which further proves the robustness of the proposed method.

2.4.3.5 Comparison to Other Works

In previous works, there are cross-domain deception detection results reported for the reviews corpora we used. Reported results on the essays corpora refer to a different evaluation setup using two source domains [149].

Table 2.8 shows the cross-domain deception detection results reported by [25] with the same versions of the review datasets we used in this work.

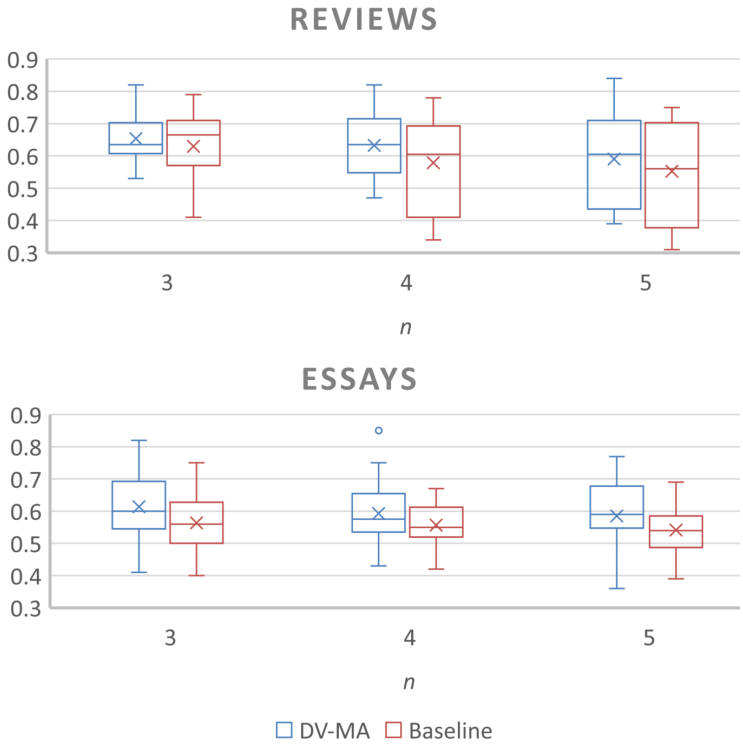


Figure 2.4: F_1 of DV-MA (varying n values) and baseline models.

Authors of [25] proposed an efficient representation for this task and trained their model using only the source domain. The proposed method is also trained using only the source domain, however, the terms to be masked are selected by observing unlabeled data from the target domain. To get a fair comparison with [25], we show in Table 2.8 the obtained results of our method in two cases: when no target domain information is used (the most frequent words of language are masked) and when unlabeled data from the target domain are used (based on FCE).

The third and fourth columns are compared by printing in bold the highest value for each pair of domains, and it is possible to note that in five out of six cases, the F_1 reported by [25] is improved by our method when no target domain information is used (the most frequent words of the language are masked). The fifth column shows, in bold and asterisk, those cases in which observing unlabeled data from the target domain, our method obtains a higher score than indicated in the two previous columns.

Finally, it is important to point out that other cross-domain results have been reported by [109] and [160]. However, these authors used the original versions of the three review (opinion spam) datasets, which contain more instances; therefore, their results cannot be directly compared with the ones

Table 2.8: Comparison of the performance of the proposed approach (either with or without access to target domain information) to the results reported by [25] on review datasets.

		Cagnina and Rosso (2017)	**Most frequent words in English**	**FCE**
Unlabeled data from both domains				✓
Masking technique			✓	✓
Source	**Target**			
Hotel	Restaurant	0.64	**0.779** ± (0.034)	**0.793*** ± (0.062)
	Doctor	0.50	**0.645** ± (0.044)	**0.659*** ± (0.060)
Restaurant	Hotel	0.66	**0.726** ± (0.023)	0.714 ± (0.020)
	Doctor	0.50	**0.596** ± (0.055)	**0.638*** ± (0.063)
Doctor	Restaurant	**0.57**	0.554 ± (0.049)	**0.616*** ± (0.055)
	Hotel	0.42	**0.540** ± (0.026)	**0.666*** ± (0.030)

obtained in this study[5].

2.5 Conclusions

This paper is a contribution to the cross-domain deception detection, a doubly challenging task due to the cross-domain problems and the difficulty at detecting deception. The proposed method improves the cross-domain classification performance in which labeled instances from the target domain are not given. The suitability of our method is due to we apply a text distortion technique that transforms original texts in a form in which distractful information is masked. We demonstrate that the masking technique is a good idea for detecting deception in cross-domain scenarios. Moreover, the performance is further improved if we consider *unlabeled* information from the target domain in order to pick out the terms to be masked. The method is robust to the distribution of the classes in the unlabeled data that is observed and to the parameter n (length of the n-grams used as features).

To our knowledge, this is the first domain adaptation approach that combines information from the source and target domain for a better text representation in the deception detection task. More data are needed to study more carefully how k depends on specific corpora characteristics.

[5]The original versions are currently unavailable. [109] reported 0.784 (H->Rest) and 0.679 (H->Dr), whereas [160] reported 0.826 (H->Rest) and 0.676 (H->Dr).

Chapter 3

Masking and Transformer-based Models for Hyperpartisanship Detection in News

This chapter of the thesis studies the detection of *hyperpartisanship* in political news using a technique effective at detecting *deceptive* language. We adapt the masking-based model to be able to compare the style vs. the content of what is discussed in the news. We use a dataset of political news with a reliable annotation with respect to the political orientation. Moreover, we compare the masking-based model with BERT-based models in terms of performances and explainability.

The work presented in this chapter was published in the following paper:

- **Sánchez-Junquera J.**, Rosso P., Montes M., Ponzetto S. (2021) Masking and Transformer-based Models for Hyperpartisanship Detection in News. Proc. Int. Conf. on Recent Advances in Natural Language Processing, RANLP-2021, Bulgaria, September 1-4, pp. 1244-1251.

  A preliminary version of this work (Unmasking Bias in News) was accepted at the 20th Int. Conf. on Computational Linguistics and Intelligent Text Processing, CICLing-2019, La Rochelle, France, April 7-13 (**CORE B**) [in press].

Abstract

Hyperpartisan news show an extreme manipulation of reality based on an underlying and extreme ideological orientation. Because of its harmful effects at reinforcing one's bias and the posterior behaviour of people, hyperpartisan news detection has become an important task for computational linguists. In this paper, we evaluate two different approaches to detect hyperpartisan news. First, a text masking technique that allows us to compare style vs. topic-related features in a different perspective from previous work. Second, the transformer-based models BERT, XLM-RoBERTa, and M-BERT, known for their ability to capture semantic and syntactic patterns in the same representation. Our results corroborate previous research on this task in that topic-related features yield better results than style-based ones, although they also highlight the relevance of using higher-length n-grams. Furthermore, they show that transformer-based models are more effective than traditional methods, but this at the cost of greater computational complexity and lack of transparency. Based on our experiments, we conclude that the beginning of the news show relevant information for the transformers at distinguishing effectively between left-wing, mainstream, and right-wing orientations.

3.1 Introduction

Media such as radio, TV channels, and newspapers control which information spreads and how it does it. The aim is often not only to inform readers but also to influence public opinion on specific topics from a hyperpartisan perspective.

Social media, in particular, have become the default channel for many people to access information and express ideas and opinions. The most relevant and positive effect is the democratization of information and knowledge but there are also undesired effects. One of them is that social media foster information bubbles: every user may end up receiving only the information that matches his/her personal biases, beliefs, tastes and points of view. Because of this, social media are a breeding ground for the propagation of fake news: when a piece of news outrages us or matches our beliefs, we tend to share it without checking its veracity; and, on the other hand, content selection algorithms in social media give credit to this type of popularity because of the click-based economy on which their business are based. Another harmful effect is that the relative anonymity of social networks facilitates the propagation of toxic, hate and exclusion messages. Therefore, social media contribute to the misinformation and polarization of society, as we have recently witnessed in the last presidential elections in USA or the Brexit referendum. Clearly, the polarization of society and its underlying discourses are not limited to social media, but rather reflected also in political dynamics (e.g. like those found in the US Congress [5]): even in this domain, however,

social media can provide a useful signal to estimate partisanship [81].

Closely related to the concept of controversy and the "filter bubble effect" is the concept of bias [9], which refers to the presentation of information according to the standpoints or interests of the journalists and the news agencies. Detecting bias is very important to help users to acquire balanced information. Moreover, how a piece of information is reported has the capacity to evoke different sentiments in the audience, which may have large social implications (especially in very controversial topics such as terror attacks and religion issues).

In this paper, we approach this very broad topic by focusing on the problem of detecting hyperpartisan news, namely news written with an extreme manipulation of the reality on the basis of an underlying, typically extreme, ideology. This problem has received little attention in the context of the automatic detection of fake news, despite the potential correlation between them. Seminal work from [152] presents a comparative style analysis of hyperpartisan news, evaluating features such as characters n-grams, stop words, part-of-speech, readability scores, and ratios of quoted words and external links. The results indicate that a topic-based model outperforms a style-based one to separate the left, right and mainstream orientations.

More recently, in [94], the features that participants used in SemEval-2019 task 4 on hyperpartisan news detection have been summarized: n-grams, word embeddings, stylometry (e.g. punctuation and article structure), sentiment and emotion features, named entities, quotations, hyperlinks, and publication date. Using the same dataset from SemEval-2019, [6] evaluated features like bag-of-words, bag-of-clusters, word embeddings and contextual character-based embeddings, POS n-grams, stylistic features and the sentiment; the authors found that dense document representations work better across domains and tasks than traditional sparse representations. Finally, [83] found effective to use personality information in hyperpartisan news detection after topic-based sub-sampling of the news training data. The datasets proposed in [94] were manually labeled and the largest one was labeled in a semi-automated manner via distant supervision.

Instead of employing the datasets from [94], we build upon previous work and use the dataset from [152]: this way we can investigate hyperpartisan-biased news (i.e., extremely one-sided) that have been manually fact-checked by journalists from BuzzFeed, and contrast our results with what they achieved. The articles originated from 9 well-known political publishers, three each from the mainstream, the hyperpartisan left-wing, and the hyperpartisan right-wing. To detect hyperpartisanship, we aim to explore the trade-off between the performance of the models and the transparency of their results. Taking this into account, we apply two approaches diametrically opposite to each other in the text classification state of the art. On the one hand, we use three transformer-based models, which have shown outstanding performance, but high complexity and lack of transparency. On the other hand, we

use a masking-based model that requires fewer computational-resources and showed a good performance in related tasks such as authorship attribution [188].

The masking technique transforms the original texts in a form where the textual structure is maintained, while letting the learning algorithm focus on the writing style or the topic-related information. This technique makes it possible for us to corroborate previous results that content matters more than style. Moreover, we aim to find explainable predictions of hyperpartisanship with the attention mechanism of the transformer-based models. With this purpose, we expect to derive the explanation by investigating the scores of different features used to output the final prediction. Based on this, we contrast the transparency of both approaches by comparing the relevant parts of the texts that they highlight.

The rest of the paper is structured as follows. In Section 3.2 we describe our method to hyperpartisan news detection based on masking. Section 3.3 presents details on the dataset and the experimental setup. In Section 3.4 we show the obtained results and discuss about them. Finally, Section 3.5 concludes with some directions for future work.

3.2 Masking and Transformer-based Models

3.2.1 Investigating Masking for Hyperpartisanship Detection

The masking technique that we propose here for the hyperpartisan news detection task has been applied to text clustering [77], authorship attribution [188], and deception detection [169] with encouraging results. The main idea of the proposed method is to transform the original texts to a form where the textual structure, related to a general style (or topic), is maintained while content-related (or style-related) words are masked. To this end, all the occurrences of non-desired terms are replaced by symbols. Let W_k be the set of the k most frequent words, we mask all the occurrences of a word $w \in W_k$ if we want to learn a *topic-related model*, or we mask all $w \notin W_k$ if we want to learn a *style-based model*. Whatever the case, the way in which we mask the terms in this work is called *Distorted View with Single Asterisks* and consists in replacing w with a single asterisk or a single # symbol if the term is a word or a number, respectively. For further masking methods, refer to [188].

Table 3.1 shows a fragment of an original text and the result of masking style-related information or topic-related information. With the former we obtain distorted texts that allow for learning a *topic-based model*; on the other hand, with the latter, it is possible to learn a *style-based model*. One of the options to choose the terms to be masked or maintained without masking is to take the most frequent words of the target language [188]. In the original

text from the table, we highlight some of the most frequent words in English.

Table 3.1: Examples of masking style-related information or topic-related information.

Original text	**Masking topic-related words**	**Masking style-related words**
Officers **went after** Christopher **Few after** watching **an** argument **between him and his** girlfriend outside **a** bar **just before the** 2015 shooting	* went after * Few after * an * between him and his * a * just before the # *	Officers * * Christopher * * watching * argument * * * * girlfriend outside * bar * * * 2015 shooting

3.2.2 Transformer-based Models

Transformer-based models have been trained with huge general language datasets. Such is the case of the Bidirectional Encoder Representations from Transformers (BERT). BERT is designed to pretrain deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers [47]. This text representation allows the model to capture complex patterns going beyond merely the use of words and capturing semantic and syntactic patterns in the same representation.

The framework of BERT consists of two steps: pre-training and fine-tuning. For the pre-training, the collected data included BooksCorpus (800M words) and English Wikipedia (2,500M words). The BERT$_{BASE}$ model has 12 layers with 12 self-attention heads, and uses 768 as hidden size, with a total of 110M parameters; and the BERT$_{LARGE}$ model has 24 layers with 16 self-attention heads, and uses 1024 as hidden size, with a total of 340M parameters. The vocabulary contains 30K tokens. For fine-tuning, the model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream task, which in our case are 1555 news annotated with the political orientation. The first token of every sequence is always a special classification token ([CLS]), which is used as the aggregate sequence representation for classification tasks. In our work, we add to the [CLS] representation two dense layers and a Softmax function to obtain the binary classification.

In this paper we evaluate three transformer-based models: BERT; the multilingual BERT (M-BERT) [47], which was pretrained on the concatenation of monolingual Wikipedia datasets from 104 languages [150, 212]; and XLM-RoBERTa, which was pretrained on 2.5TB of newly created clean CommonCrawl data in 100 languages [31].

3.3 Experiments

We used the BuzzedFeed-Webis Fake News Corpus 2016 collected by [152] whose articles were labeled with respect to three political orientations: mainstream, left-wing, and right-wing (see Table 3.2). Each article was taken from one of 9 publishers known as hyperpartisan left/right or mainstream in a period close to the US presidential elections of 2016. Therefore, the content of all the articles is related to the same topic. During initial data analysis and prototyping we identified a variety of issues with the original dataset: we cleaned the data excluding articles with empty or bogus texts, duplicates. As a result, we obtained a new dataset with 1555 articles out of 1627.[1] Following the settings of [152], we balanced the training set using random duplicate oversampling.

Table 3.2: Statistics of the original dataset and its subset used in this paper.

	Left-wing	**Mainstream**	**Right-wing**	Σ
Original data [152]	256	826	545	1627
Cleaned data	252	787	516	1555

3.3.1 Masking Content vs. Style in Hyperpartisan News

In this section, we reported the results of the masking technique from two different perspectives. In one setting, we masked *topic-related information* in order to maintain the predominant writing style used in each orientation. We call this approach a *style-based model*. With that intention we selected the k most frequent words from the target language, and then we transformed the texts by masking the occurrences of the rest of the words. In another setting, we masked *style-related information* to allow the system to focus only on the topic-related differences between the orientations. We call this a *topic-based model*. For this, we masked the k most frequent words and maintained intact the rest.

After the text transformation by the masking process in both the training and test sets, we represented the documents with character n-grams and compared the results obtained with the *style-based* and the *topic-related models*.

3.3.2 Experimental Setup

Text Transformation: We evaluated different values of k ($k \in \{100, 200, ..., 5000\}$) for extracting the k most frequent words from English[2]. For

[1] The dataset is available at https://github.com/jjsjunquera/UnmaskingBiasInNews/blob/master/articles1555.rar.

[2] We use the BNC corpus (https://www.kilgarriff.co.uk/bnc-readme.html) for the extraction of the most frequent words as in [188].

the comparison of the results obtained by each model with the ones of the state of the art, we only showed the results fixing $k = 500$.

Text Representation: We used a standard bag-of-words representation with *tf* weighting and extracted character 5-grams with a frequency lower than 50.

Classifiers: We compared the results obtained with Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF); for the three classifiers we used the versions implemented in *sklearn* with the parameters set by default.

Transformers: We approached the hyperparameter tuning by grid search. The best results were obtained with: *learning rate* $= 3e-5$; the *batch size* $= 16$; and the *adam* optimizer. Moreover, we applied a dropout value of 0.3 to the last dense layer. We have selected a value of 200 for the max_length hyperparameter.

Evaluation: We performed 3-fold cross-validation with the same configuration used in [152]. Therefore, each fold comprised one publisher from each orientation (the classifiers did not learn a publisher's style). We used macro F_1 as the evaluation measure since the test set is unbalanced with respect to the three classes. In order to compare our results with those reported in [152], we also used accuracy, precision, and recall.

Baseline: Our baseline method is based on the same text representation with the character n-grams features, but without masking any word.

3.4 Results and Discussion

Table 3.3 shows the results of the proposed method and the system from [152][3] in our cleaned dataset (Section 3.3), both considering topic and style-based methods. In order to compare our results with those reported in [152], we report the same measures the authors used. We also include the macro F_1 score because of the unbalance test set. For these experiments we extract the character 5-grams from the transformed texts, taking into account that as more narrow is the domain more sense has the use of longer n-grams. We follow the steps of [188] and set k = 500 for this comparison results.

Similar to [152], the topic-based model achieves better results than the style-related model. However, the differences between the results of the two evaluated approaches are much higher (0.66 vs. 0.57 according to Macro F_1) than those obtained from the system of [152] (0.63 vs. 0.61). The highest scores of the masking technique were consistently achieved using the SVM

[3] https://github.com/webis-de/ACL-18

Table 3.3: Results of the proposed masking technique ($k = 500$ and $n = 5$) applied to mask topic-related information or style-related information. NB: Naive Bayes; RF: Random Forest; SVM: Support Vector Machine. The last two rows show the results obtained by applying the system from [152] to our cleaned dataset (Section 3.3).

Masking Method	**Classifier**	**Macro F_1**	**Accuracy**	**Precision**			**Recall**			**F_1**		
				left	right	main	left	right	main	left	right	main
Baseline model	NB	0.52	0.56	0.28	0.57	0.81	0.49	0.58	0.56	0.35	0.57	0.66
	RF	0.56	0.62	0.28	0.61	0.80	0.36	0.72	0.63	0.32	0.66	0.70
	SVM	0.70	0.77	0.55	0.75	0.84	0.42	0.79	0.87	0.47	0.77	0.85
Style-based model	NB	0.47	0.52	0.20	0.51	0.73	0.28	0.65	0.49	0.23	0.57	0.59
	RF	0.46	0.53	0.24	0.58	0.64	0.36	0.34	0.73	0.29	0.43	0.68
	SVM	0.57	0.66	0.33	0.66	0.75	0.26	0.61	0.84	0.29	0.62	0.79
Topic-based model	NB	0.54	0.60	0.26	0.63	0.74	0.36	0.62	0.65	0.29	0.62	0.69
	RF	0.53	0.55	0.27	0.64	0.71	0.44	0.60	0.58	0.33	0.61	0.64
	SVM	0.66	0.74	0.48	0.73	0.81	0.38	0.78	0.82	0.42	0.75	0.82
System from [152] (applied to our cleaned dataset)												
Style	RF	0.61	0.63	0.29	0.62	0.71	0.16	0.62	0.80	0.20	0.61	0.74
Topic	RF	0.63	0.65	0.27	0.65	0.72	0.15	0.62	0.84	0.19	0.63	0.77
Transformer-based models												
M-BERT		0.76	0.83	0.65	0.75	0.93	0.49	0.86	0.92	0.56	**0.93**	0.80
XLM-RoBERTa		0.80	0.86	**0.80**	0.76	**0.95**	0.50	**0.91**	0.94	0.61	0.83	**0.95**
BERT		**0.86**	**0.89**	0.77	**0.87**	0.94	**0.75**	0.86	**0.96**	**0.76**	0.87	**0.95**

classifier and masking the style-related information (i.e., applying the topic-related model). This could be explained with the fact that all the articles are about the same political event in a very limited period of time. In line with what was already pointed out in [152], the left-wing orientation is harder to predict, possibly because this class is represented with fewer examples in the dataset.

Another reason why our masking approach achieves better results than the system from [152], could be that we use a higher length of character n-grams. In fact, comparing their results against our baseline model, it is possible to note that even without masking any word, the classifier obtains better results. This suggests that the good results are due to the length of the character n-grams rather than the use of the masking technique.

The last three rows of Table 3.3 show the results of the transformer-based models. As we can see, these models achieved the highest results, in particular the BERT model, with a Macro $F_1 = 0.86$. These models are known for their ability to capture complex syntactic and semantic patterns, therefore, these results are somehow justified to be the highest compared to the masking approach. However, what is interesting at this point is the effectiveness of the models at predicting the correct orientation using just the beginning of the news ($max\_length = 200$). This is aligned to the work of [71] that focused on analyzing the initial part of false news articles. The authors assumption is that false news tend to present a unique emotional pattern for each false information type in order to trigger specific emotions to the readers; in hyperpartisan news this probably happens to gain readers' attention and sympathy.

3.4.1 Relevant Features

Table 3.4 shows the features with the highest weights from the SVM (we used `scikit-learn`'s method to collect feature weights). It is possible to note that the mention of *cnn* was learned as a discriminative feature when the news from that publisher were used in the training (in the topic-based model). However, this feature is infrequent in the test set where no news from CNN publisher was included.

Table 3.4: Most relevant features to each class.

Baseline model		
left	**main**	**right**
_imag	_cnn_	e_are
_that	said_	lary_
e_tru	_said	_your
e_don	y_con	n_pla
_here	ry_co	e_thi
s_of_	_cnn_	s_to_
for_h	said_	illar
donal	_said	hilla
racis	ore_t	llary
_kill	story	_hill
_that	_said	_let_
_trum	tory_	_comm
trump	ed_be	lary_

Style-based model		
left	**main**	**right**
but_*	n_thi	y_**_
out_w	s_*_s	out_a
t_**_	_how_	as_to
you_h	at_he	o_you
t_and	m_*_t	ell_*
_is_a	*_*_u	and_n
h_*_a	e_#_*	hat_w
_of_#	and_*	*_#_#
or_hi	**_*	_it_t
for_h	t_the	e_of_
**_*_	and_*	o_you
_in_o	_*_tw	n_it_
hat_*	*_two	and_n

Topic-based model		
left	**main**	**right**
ant_*	_cnn_	hilla
lies_	ics_*	als_*
_*_ex	sday_	_*_le
etty_	ed_be	_dail
donal	_cnn_	_*_te
n_*_c	day_*	*_ame
onald	cs_*_	_*_am
ying_	ics_*	illar
thing	*_*_e	llary
â_*_*	ed_be	_*_le
eâ_*_	y_con	_*_ri
nâ_*_	tory_	_hill
tâ_*_	story	_bomb

The features related to Donald Trump (*donal* and *onald*), and Hillary Clinton (*llary* and *illar*) are more frequent in one of the hyperpartisan orientation, and none of them occurs frequently in the mainstream orientation. On the other hand, the relevant features from the style-based model involve function words that are frequent in the three classes (e.g. *out*, *you*, *and*, *of*) even if the combination between function words and other characters can lightly differ in different orientations.

3.4.2 Features with the Highest Attention Scores

Transformer-based models allow us to visualize different parts of the news according to the scores they received to obtain the final prediction. In Figure 3.1 we show examples of news predicted correctly by BERT (the model with the highest F_1 score). Due to space limitations, we provide fragments of six news, two per orientation. The more intense the color, the greater is the weight of attention given by the model.

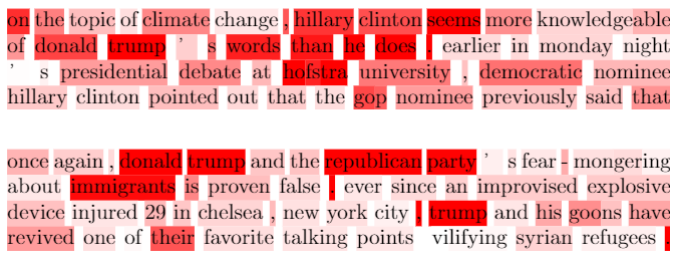
In the examples from 3.1a, the left-wing orientation remarks the names of the opposite politicians, and it is possible to see which of them is the favourite of the journalist. In particular, the leader of the right-wing (i.e., Trump) is referred in a negative way (he does not know his own words) while Hillary Clinton, the representative of the left-wing is favored by the news. Similar to this, examples 3.1c do the same but in the opposite direction; i.e., Hillary Clinton is put as a very negative “character" who *loves taxes*

Table 3.5: Fragments of original texts and their transformation by masking the k most frequent terms. Some of the features from Table 3.4 using the topic-related model are highlighted.

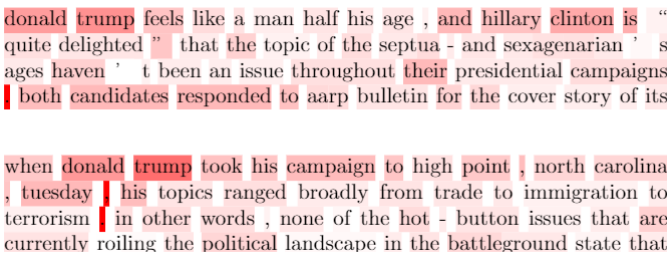
Topic-related model	
left	(...)which his son pretty much confirmed in a foolish statement. The content of those tax returns has been the subject of much speculation, but given Trump's long history of tax evasion and political bribery, it doesn't take much imagination to assume he's committing some kind of fraud
	* * son pretty * confirmed * foolish statement * content * * tax returns * * * subject * * speculation * * Trump * * * tax evasion * * bribery * doesn * * imagination * assume * committing * * * fraud
main	Obama proved beyond a shadow of a doubt in 2011 when he released his long-form birth certificate (...) CNN and Fox News cut away at points in the presentation. Networks spent the day talking about Trump's history as a birther (...) Before Friday, the campaign's most recent deception came Wednesday when campaign advisers told reporters that Trump would not be releasing results of his latest medical exam
	Obama proved beyond shadow * doubt * 2011 * * released * ** birth certificate (...) CNN * Fox News cut * *points * * presentation Networks spent * * talking * Trump * * birther (...) * Friday * campaign * recent deception * Wednesday * campaign advisers told reporters * Trump * * * releasing results * * latest medical exam
right	The email, which was dated March 17, 2008, and shared with POLITICO, reads: Jim, on Kenya your person in the field might look into the impact there of Obama's public comments about his father. I'm told by State Dept officials that Obama publicly derided his father on (...) Blumenthal, a longtime confidant of both Bill and Hillary Clinton, emerged as a frequent correspondent in the former secretary of (...)
	* email * * dated March 17 2008 * shared * POLITICO reads Jim * Kenya * * * * field * * * * impact * * Obama * comments * * * told * * Dept officials * Obama publicly derided * * (...) Blumenthal longtime confidant * * Bill * Hillary Clinton emerged * frequent correspondent * * former secretary * (...)

and is *the most despicable liar ever*. However, examples from 3.1c offer a comparison in which keep the reader in a neutral position. Moreover, in the second mainstream news, Trump's campaign is mentioned without describing the stance of the author whether Trump did well or not in his topic selection. This suggests that the style used to speak about the leaders can differ from the more biased (hyperpartisan) news to the less biased (mainstream).

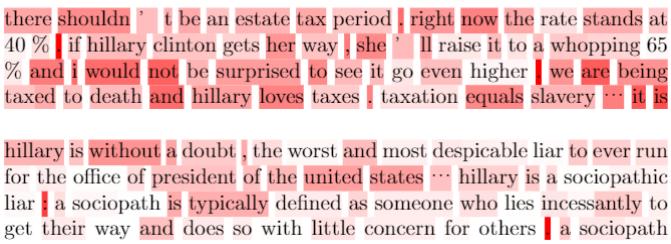
We can conclude that the attention mechanism of the transformers not only help in doing effective predictions, but offer some extra information that could be useful to understand some insights about hyperpartisanship. For example, the words with the highest scores can be used in other strategies to confirm the previous results that topic-based models outperform a style-based one at distinguishing left, right and mainstream orientations [152].



(a) Hyperpartisan (left-wing) news.



(b) Non-hyperpartisan (mainstream orientation) news.



(c) Hyperpartisan (right-wing) news.

Figure 3.1: Fragments of news (two for each political orientation) with the visualization of the attention learned by BERT. The more intense the color, the greater the weight of attention.

3.5 Conclusions

In this paper we presented initial experiments on the task of hyperpartisan news detection. In particular, we aimed to explore the trade-off between performance and transparency, and proposed a comparison of two different approaches. First, we explored the use of masking techniques to boost the performance of a lexicalized classifier. Our results corroborate previous research on the importance of content features to detect extreme content: masking, in addition, shows the benefits of reducing data sparsity for this task comparing our results with the state of the art. We evaluated different

values of the parameters and see that finally our baseline model, in which we extract character 5-grams without applying any masking process, achieves the better results. This seems to indicate a strong lexical overlap between different sources with the same orientation, which, in turn, calls for more challenging datasets and task formulations to encourage the development of models covering more subtle, i.e., implicit, forms of bias. Future datasets could consider more topics and different time spans to avoid the models learn from the topic, rather than the target classes.

Second, we used three transformer-based models (BERT, M-BERT, and XLM-RoBERTa) that are resource-hungrier than the masking technique, and achieved the highest results. We also presented some examples of how these models, through their attention scores, provide additional information about the relevant parts of the text for distinguishing their political orientation. Considering the high effectiveness of these models, and that they only observe the first part of the news, we will evaluate as future work how necessary is to use all the news (and not only the beginning), e.g. with the Transformer-XL model [39]. Moreover, we are motivated to take advantage of the attention scores to study in more detail the style used in hyperpartisan news in order to improve the predictions.

Chapter 4

A Twitter Political Corpus of the 2019 10N Spanish Election

In this chapter we present an analysis of the communication strategies of five Spanish political parties with different (*partisan*-biased) ideologies. We propose a dataset with tweets posted by the main political parties leaders covering the campaign of the Spanish election of 10th November 2019, and analyse the topics in which each party focused more. In addition, we investigate the use of emotions, observing that towards a topic, the same party expresses with opposite emotions indicating both, a viewpoint and a rhetorical strategy. The results show interest almost only from far-right leaders in the *immigration* topic.

The work presented in this chapter was published in the following paper:

- **Sánchez-Junquera J.**, Ponzetto S., Rosso P. (2020) A Twitter Political Corpus of the 2019 10N Spanish Election. Proc. 23rd Int. Conf. on Text, Speech and Dialogue, TSD-2020, Springer-Verlag, LNAI(12284), pp. 41-49.

Abstract

We present a corpus of Spanish tweets of 15 Twitter accounts of politicians of the main five parties (PSOE, PP, Cs, UP and VOX) covering the campaign of the Spanish election of 10th November 2019 (10N Spanish Election). We perform a semi-automatic annotation of domain-specific topics using a mixture of keyword-based and supervised techniques. In this preliminary study we extracted the tweets of few politicians of each party with the aim to analyse their official communication strategy. Moreover, we analyse sentiments and emotions employed in the tweets. Although the limited size of the Twitter corpus due to the very short time span, we hope to provide with some first insights on the communication dynamics of social network accounts of these five Spanish political parties.

Keywords: Twitter, Political text analysis, Topic detection, sentiment and emotion analysis

4.1 Introduction

In recent years, automated text analysis has become central for work in social and political science that relies on a data-driven perspective . Political scientists, for instance, have used text for a wide range of problems, including inferring policy positions of actors [113], and detecting topics [163], to name a few. At the same time, researchers in Natural Language Processing (NLP) have addressed related tasks such as election prediction [137], stance detection towards legislative proposals [203], predicting roll calls [98], measuring agreement in electoral manifestos [125], and policy preference labelling [1] from a different, yet complementary perspective. Recent attempts to bring these two communities closer have focused on shared evaluation exercises [135] as well as bringing together the body of the scholarly literature of the two communities [73]. The effects of these two strands of research coming together can be seen in political scientists making use and leveraging major advances in NLP from the past years [161].

The contributions of this paper are the following ones: (i) we introduce a corpus of tweets from all major Spanish political parties during the autumn 2019 election; (ii) we present details on the semi-automated topic and sentiment/emotion annotation process; and (iii) we provide a preliminary qualitative analysis of the dataset over different addressed topics of the election campaign. Building this preliminary resource of Spanish political tweets, we aim at providing a first reference corpus of Spanish tweets in order to foster further research in political text analysis and forecasting with Twitter in languages other than English.

In the rest of the paper we will describe how each tweet was annotated with topic information together with sentiments and emotions. Moreover, we will illustrate the preliminary experiments we carried out on topic detection.

Finally, we will present some insight about sentiment and emotion topic-related analyses.

4.2 Related Works

Twitter has been used as a source of texts for different NLP tasks like sentiment analysis [69, 195]. One work that is very related to our study is [118]. They collected a dataset in English for topic identification and sentiment analysis. The authors used distant supervision for training, in which topic-related keywords were used to first obtain a collection of positive examples for the topic identification. Their results show that the obtained examples could serve as a training set for classifying unlabelled instances more effectively than using only the keywords as the topic predictors. However, during our corpus development we noticed that keyword-based retrieval can produce noisy data, maybe because of the content and the topics of our tweets, and we then used a combination of both a keyword-based and a supervised approach.

4.3 Political Tweets in the 10N Spanish Election

In this paper, we focus on the Spanish election of November 10th, 2019 (10N Spanish Election, hereafter). For this, we analyse tweets between the short time span of October 10, 2019, and November 12, 2019. We focus on the tweets from 15 representative profiles of the five most important political parties (Table 4.1)[1]: i.e., Unidas Podemos (UP); Ciudadanos (Cs); Partido Socialista Obrero Español (PSOE); Partido Popular (PP); and VOX.

Table 4.1: Number of tweets of the five political parties. For each party, we use its official Twitter account, its leader, and the female politician that took part in the 7N TV debate.

Parties	**The main profiles**	**Tweets**
UP	@ahorapodemos, @Irene_Montero_, @Pablo_Iglesias_	671
Cs	@CiudadanosCs, @InesArrimadas, @Albert_Rivera	789
PSOE	@PSOE, @mjmonteroc, @sanchezcastejon	527
PP	@populares, @anapastorjulian, @pablocasado_	684
Vox	@vox_es, @monasterior, @santi_abascal	749
Total		3582

[1]The dataset is available at https://github.com/jjsjunquera/10N-Spanish-Election.

4.3.1 Topic Identification

Topic categories. We first describe how we detect the topic of the tweets on the basis of a keyword-based and supervised approach. In the context of the 10N Spanish Election, we focused on the following topics that were mentioned in the political manifestos of the five main Spanish parties: *Immigration, Catalonia, Economy (and Employment), Education (together with Culture and Research), Feminism, Historical Memory, and Healthcare.* We additionally include a category label *Other* for the tweets that talk about any other topic.

Manual topic annotation. We first manually annotate 1,000 randomly sampled tweets using our topic labels. Table 4.2 summarizes the label distribution across all parties. After removing the noisy tweets, we are left with only 765 posts. Many tweets in our corpus are not related to any of the topics of interest, and were assigned to the *Other* category. Moreover, during the annotation, we noticed in the manifestos of the five parties little information about topics such as research, corruption, renewable energy, and climate change.

Table 4.2: Total number of labelled tweets: the training set (i.e., manually annotated, and using keywords), and using automatic annotation. The last column has the total number of labelled tweets considering the training set and the classifier results.

Topic	Manual annotated	Keyword annotated	Automatically annotated	Total annotated
Catalonia	115	130	370	615
Economy	71	39	506	616
Education	2	19	23	44
Feminism	10	52	82	144
Healthcare	4	12	7	23
Historical Memory	12	16	30	58
Immigration	9	16	36	61
Other	541	153	1037	1731
Pensions	1	24	55	80
Total	765	461	2146	3372

Keyword-based topic detection. Due to the manual annotation is time consuming, we complement it by using topic-related keywords to collect tweets about each topic. We ranked the words appearing in the sections corresponding to the topics of interest with the highest Pointwise Mutual Information (PMI). PMI makes it possible to select the most relevant words for each topic, and is computed as: $PMI(T, w) = \log \frac{p(T,w)}{p(T)p(w)}$. Where $p(T, w)$ is the probability of a word to appear in a topic, $p(T)$ is the probability of a topic (we assume the topic distribution to be uniform), and $p(w)$ is the probability of w. For each topic, we collect the top-10 highest ranked keywords and manually filter incorrect ones (Table 4.3).

Table 4.3: Keywords used for collecting training data for topic identification.

Topic	Keywords
Catalonia	*autonómica; cataluña; civil*
Economy	*bienestar; discapacidad; energía; fiscalidad; impuesto; innovación; inversión; tecnológico*
Education	*cultura; cultural; educación; lenguas; mecenazo*
Feminism	*conciliación; familia; machismo; madres; discriminación; mujeres; sexual; violencia*
Healthcare	*infantil; sanitario; salud; sanidad; sanitaria; universal*
Historical Memory	*historia; memoria; reparación; víctimas*
Immigration	*ceuta; extranjeros; inmigrantes; ilegalmente*
Pension	*pensiones; toledo*

Supervised learning of topics. For each topic, we collect all tweets in our corpus in which at least one of its keywords appears. All retrieved tweets are then manually checked to ensure that the annotated tweets have a ground-truth.

Inspired by the work of [118], we use the topic-related keywords to obtain a collection of "positive" examples to be used as a training set for a supervised classifier. However, in our dataset, we noticed that keyword-based retrieval can produce much noisy data. Therefore, the keyword-based collected tweets are manually checked before training the classifier.

While our solution still requires the mentioned manual checking, the advantage of using keywords is that the labelling is more focused on tweets that are likely to be in one of the topics of interest, thus reducing the annotation effort associated with tweets from the *Other* category.

Table 4.2 summarizes in the second and third columns the number of tweets that we used as a training set. The second column represents the results after manually evaluating the tweets labelled by using the keywords. It is interesting that the annotated data reveal most attention towards some topics such as *Catalonia*, *Feminism* and *Economy*. Finally, the dataset used for training is composed of all the labelled tweets. To avoid bias towards the most populated categories we reduce their number of examples to 100 for training, for which we balance the presence of manually annotated and keyword-based annotated tweets.

We employ a SVM [2] to classifiy the still unlabeled tweets and leave-one-out cross-validation because of the small size of the corpus. We represent the tweets with unigrams, bigrams and trigrams, and use the *tf-idf* weighting scheme after removing the n-grams occurring only once.

Evaluation of topic detection. Table 4.4a shows the standard precision, recall, and F_1 scores. Table 4.2 shows in the fourth column the number of tweets annotated using our supervised model. The last column shows instead the total of labelled tweets for each of the topics – i.e., the overall number of labelled tweets obtained by combining manual, keyword-based annotations

[2]We used the implementation from *sklearn* using default parameter values for with a linear kernel.

Table 4.4: Results on topic classification and the total number of labelled tweets.

(a) Results on topic classification.

Topic	Precision	Recall	F1-score
Catalonia	0.72	0.86	0.78
Economy	0.56	0.7	0.62
Education	0.83	0.48	0.61
Feminism	0.8	0.73	0.77
Healthcare	1	0.38	0.55
Historical Memory	0.82	0.5	0.62
Immigration	0.92	0.44	0.59
Other	0.56	0.6	0.58
Pensions	0.85	0.68	0.76
macro avg.	0.78	0.6	0.65

(b) Total of labelled tweets.

Topic	UP	Cs	PP	PSOE	VOX
Catalonia	40	198	110	50	72
Economy	114	117	203	84	88
Education	12	12	11	5	4
Feminism	44	30	8	29	31
Healthcare	10	2	3	6	2
Historical Memory	25	7	2	8	16
Immigration	4	1	-	7	49
Other	258	262	200	174	243
Pensions	17	2	14	37	10

with the SVM classifier. We break down the numbers of these overall annotated tweets per party in Table 4.4b. The topic distributions seem to suggest that each party is biased towards specific topics. For instance, *Immigration* seems to be almost only mentioned by VOX, whereas parties like PP and Cs are mainly focused on *Catalonia* and *Economy*.

4.3.2 Sentiment Analysis

We next analyse the sentiment expressed by the parties about each topic. For this, we use SentiStrength to estimate the sentiment in tweets since it has been effectively used in short informal texts [202]. We compute a single scale with values from -4 (extremely negative) to 4 (extremely positive).

In order to compare for each topic the sentiment expressed by a party, we compute the average of the scores for the party on that topic. Only the topics with a precision greater than 0.6 (Table 4.4a), and the parties that wrote more than 10 tweets on the corresponding topic, were considered in this comparison. It means that we ignore, for instance, the sentiment showed towards *Economy* (precision lower than 0.6), and *Healthcare* (only UP wrote 10 tweets, see Table 4.4b, and the sentiment that Cs showed towards *Pensions* (only two tweets, see Table 4.4b).

Figure 4.1 shows the expressed sentiment for the parties for each topic. Sentiment scores seem to reveal some common dynamics of political communication from political parties in social networks in that generally, even when the party is known to be negative or have a critical stance with respect to a certain topic (e.g. a populist party on immigration), tweets receive a positive score. Specifically, we see that VOX was the only party addressing the *Immigration* topic, and we observe that in general, its sentiment is positive (i.e., solutions were commented). Also, just two parties show mainly negative sentiments, they are VOX and PP towards *Feminism* and *Pensions* respectively.

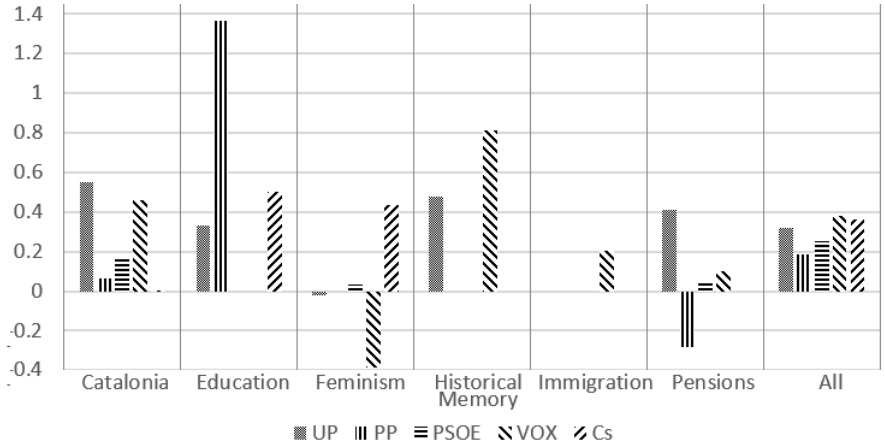


Figure 4.1: Expressed sentiment for each topic and party.

4.3.3 Emotion analysis

We finally analyse the emotions expressed by the parties for different topics using the Spanish Emotion Lexicon (SEL) [182]. SEL has 2,036 words associated with the measure of Probability Factor of Affective (PFA) concerning to at least one Ekman's emotions [52]: joy, anger, fear, sadness, surprise, and disgust. For each tweet, we compute the final measure for each of the six emotions by summing the PFA and dividing by the length of the tweet. We then compute the average PFA of all the emotions for each party and each topic.

Figure 4.2 (top image on the left) shows the emotions that the parties present in their tweets when talking about different topics. We analyse the emotions of the same pairs of parties and topics we analysed before in Section 4.3.2. Differently to the case of sentiment, there is a general trend shared in that joy and sadness are very much present across all parties. This could be due to several reasons. First, there is a bias in SEL towards joy (668 words related to joy vs. 391 for sadness, 382 for anger, 211 for fear, 209 for disgust, and 175 for surprise), and second, the terms that help to compute the SentiStrength score are not necessarily the same that are in SEL . Another interesting thing is the presence of joy and sadness in the same topic by the same parties. We attribute this behaviour to the fact that there are tweets describing the current problems and feelings present in the context of the election - e.g. using words like *sufrir* (to suffer), *muerte* (death), *triste* (sad), *grave* (grave), but also there are others with a propositive discourse about the problems - e.g. using words like *esperanza* (hope), *ánimo* (encouragement), *unión* (union), *fiesta* (party).

In Figure 4.2 we also highlight that PSOE shows contrasting emotions about Catalonia; and Cs shows high score of joy about topics related to feminism. The distribution of the emotions from VOX towards *Immigration*

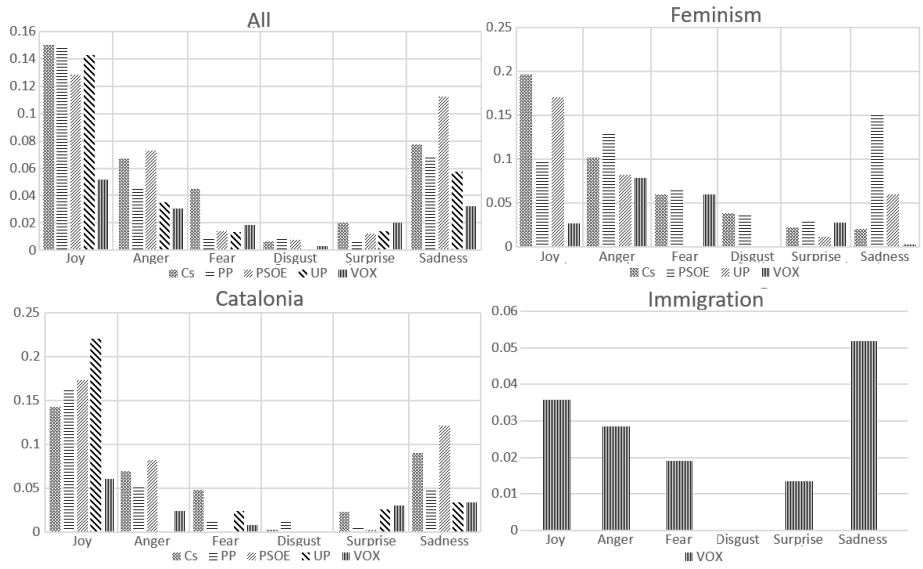


Figure 4.2: Emotions distribution across topics.

was omitted due to the space. However, despite the positive sentiment that VOX showed in this topic, the predominant expressed emotion was sadness.

4.4 Conclusions

In this paper we presented a first study about the most relevant topics that have been addressed in Twitter in the context of the 10N Spanish election for the five main political parties, together with their sentiments and emotions.

On the basis of the above analysis, we noticed that each party focused more on specific topics, expressing different sentiments and emotions. Our analysis, although preliminary, indicates potentially interesting dimensions of political communications on social networks such as the tendency towards positive tweets, as well the contrasted presence of problems vs. solutions. This work provides a first attempt towards analysing the political communication by the five main political parties in Spain on social networks using NLP techniques. Although we are aware of the limitations of this preliminary study due to the very short time span and the size of the corpus, we hope that this first analysis could contribute to understand how sentiments and emotions were expressed in Twitter by the politicians of the main five parties with respect to the topics mentioned in their manifestos during the political campaign of the 10N Election in Spain.

As future work we plan also to consider additional parties and languages (e.g. Catalan, Basque and Galician) to provide a more comprehensive resource as well as a comparative analysis.

Chapter 5

How do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants

This chapter describes our work at detecting *social bias* in political speeches where the *partisanship* of the speakers is already known. In particular, we are focused on *immigrant stereotype*, motivated by the results of the previous chapter where only one political party addressed the immigration topic in the collected tweets. In this chapter, we present the use of frames to identify stereotypes about immigrants. We collect speeches from parliamentary debates and create a taxonomy that encompasses exhaustively the frames of the immigrant stereotypes. Results indicate that the traditional classifiers achieve competitive predictions compared with transformers. Moreover, it makes us think that stereotypes, or *how* they are expressed, could be a rhetorical strategy used in *partisan* communication.

The work presented in this chapter was published in the following paper:

- **Sánchez-Junquera J.**, Chulvi B., Rosso P., Ponzetto S. (2021) How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Applied Science*, 11(8), 3610. (**Impact Factor: 2.679 Q2**)

Abstract

Stereotype is a type of social bias massively present in texts that computational models use. There are stereotypes that present special difficulties because they do not rely on personal attributes. This is the case of stereotypes about immigrants, a social category that is a preferred target of hate speech and discrimination. We propose a new approach to detect stereotypes about immigrants in texts focusing not on the personal attributes assigned to the minority but in the frames, that is, the narrative scenarios, in which the group is placed in public speeches. We have proposed a fine-grained social psychology grounded taxonomy with six categories to capture the different dimensions of the stereotype (positive vs. negative) and annotated a novel *StereoImmigrants* dataset with sentences that Spanish politicians have stated in the Congress of Deputies. We aggregate these categories in two supracategories: one is *Victims* that expresses the positive stereotypes about immigrants and the other is *Threat* that expresses the negative stereotype. We carried out two preliminary experiments: first, to evaluate the automatic detection of stereotypes; and second, to distinguish between the two supracategories of immigrants' stereotypes. In these experiments, we employed state-of-the-art transformer models (monolingual and multilingual) and four classical machine learning classifiers. We achieve above 0.83 of accuracy with the BETO model in both experiments, showing that transformers can capture stereotypes about immigrants with a high level of accuracy.

Keywords: social bias; stereotypes about immigrants; social psychology based taxonomy; stereoimmigrants dataset; transformer models; Spanish

5.1 Introduction

Social bias in information is receiving more and more attention in computational science. The information on the web has a strong impact on how people perceive reality and consequently on the decision they can make, the attitude they develop, and the prejudice they hold [22]. Some general examples where we can find bias include political news [170, 181], rumours [220], products reviews [172], among others. However, there is a kind of social bias which is massively present in everyday language, and of course on the web, which is the use of stereotypes. A recent work that measures stereotypical bias in pretrained language models has found that as the language model becomes stronger, so its stereotypical bias does too [133]. As the authors said, "this is unfortunate and perhaps unavoidable as long as we rely on real word distribution of corpora to train language models". The difficulty is clear but the need also: these stereotypes have a strong effect on the members of the stigmatised group, for instance, impacting the performance of individuals who face stereotype threats [44, 45, 190]. We have known from the beginning of social psychology that stereotypes are at the base of preju-

dice towards minorities and to spread prejudices is an efficient strategy for dogmatic groups and authoritarian ideologies [99, 184].

A long tradition of research in social psychology defines stereotype as a set of widespread beliefs that are associated with a group category [49, 199]. This set of beliefs facilitates the operation of prejudices and justifies them [22]. Research in prejudice has shown that this set of stereotyped beliefs may be both positive and negative [16]. The importance of positive beliefs in stereotyping social groups has been highlighted, especially in studies on gender stereotypes [74] but is less studied in relation to other stereotypes such as that of the social category of immigrants.

To understand how this social bias occurs in texts, we need to go beyond this common idea that a stereotype is a set of beliefs. A stereotype is a type of social bias that occurs when a message about a group disregards the great diversity existing within the members of this group and highlights a small set of features [199]. This process of homogenisation of a whole group of people is at the very heart of the stereotype concept [111]. We know from social science research that the main part of this definition process takes place in speeches from socially relevant actors [95]. Politicians, social movements, and mass media messages create and recreate a *frame* [177], a kind of scenario, where they speak about a group. Framing analysis [142] proposes that how citizens understand an issue—which features of it are central and which peripheral—is reflected in how the issue is framed.

Frame as a concept has a long tradition in psychology [15, 206] and in sociology [75]. Gamson defines *frame* as *"a central organizing idea or story line that provides meaning to an unfolding strip of events, weaving a connection among them. The frame suggests what the controversy is about, the essence of the issue"* [67]. As Kinder [95] resumes "frames are rhetorical weapons created and sharpened by political elites to advance their interest and ideas. *Frames* also lived inside the mind; they are cognitive structures that help individual citizens make sense of the issues that animate political life". From the cognitive-linguistic area, George Lakoff [103] had used intensively this concept of *frame* to explain the use of language in US politics. Lakoff argues that politicians invoke *frames* to dominate debates because they know that it is crucial: to attack the opponents' *frame* has the unwanted effect of reinforcing their message.

We aim to address stereotypes about immigrants as a result of this activity of *framing* in political and media speeches proposing a taxonomy that focuses on the different *frames* that politicians use to speak about immigrants. The concept of *frame* allows us to consider social cognition more a narrative process than a conceptual one. If as Jerome Bruner [23] states, the principle that organises the cognitive system of commonsense thinking is narrative rather than conceptual, we would consider narrative scenarios more than attributes assigned to a group in our detection of stereotypes about immigrants.

With the *framing* approach to stereotypes, we could detect how politicians built a *frame* that tells a story about the group focusing only on some features of the collective. In these speeches, they shape a stereotype without using explicit attributes for the group. These *frames* are subtle but powerful mechanisms to associate a group with some characteristics that are different dimensions of the stereotype.

The repeated use of this *frames* about a collective is present in the texts that computational systems process. This replication of a stereotyped vision of certain groups has an undesirable impact on people's life when they interact with technology. If stereotyping is a common bias difficult to fight in social communication, the data on the web is more likely to suffer from this lack of diversity because most of the information is created on the web by a minority of active users. For instance, 7% from a total of 40,000 users provide 50% of the total amount of posts in Facebook: it is not difficult to assume that this minority of users does not represent the knowledge and opinion of the majority [9].

In addition, recent studies have shown the difficulty of detecting ideological bias manifested in, for example, hyperpartisan news, that is news that tends to provide strongly biased information or exaggeration ending in fake news. If hyperpartisan news is easy to accept by the public that sees in them a confirmation of their own beliefs [152, 170] we can expect a great difficulty in mitigating the use of stereotypes.

In computational linguistics this problem has been addressed in some works where different techniques have been proposed to measure, represent, and reduce social bias, in particular, stereotypes and prejudice, concerning race, nationality, ethnic, and mostly gender and sex, among others [20, 104]. Most of them use a word embeddings representation and rely on the association of attributes to a social group to approach the stereotype (or other social bias) detection. We aim at approaching the problem of identifying stereotypes from a narrative perspective where computational linguistics could play a major role in analysing the complex process in which social actors create a stereotype placing a group in specific *frames*. Approaching the problem of stereotypes from this new perspective could also help to develop more sensitive tools to detect social bias in texts and new strategies to mitigate it.

We observed in the literature of computational linguistics a lack of datasets annotated with stereotypes and also works addressing the stereotypes about immigrants. We found that [173] created a dataset in Italian and included a binary stereotype annotation, but this work is mainly focused on hate speech and only annotates the existence or not of a stereotype belief about the hate speech target. In [133] it is proposed a dataset that includes the domain of racism (additionally to gender, religion, and profession), and report *immigrate* as one of the most relevant keywords that characterise such domain of bias. However, the authors do not focus on the study of stereotypes about

immigrants.

In order to detect how social bias about certain groups is present in everyday language, it is necessary to have a complex view of stereotypes taking into account both positive and negative beliefs and also how the different *frames* shape the stereotype. This more refined analysis of stereotypes would make it possible to detect social bias not only in clearly dogmatic or violent messages but also in other more formal and subtle texts such as news, institutional statements, or political representatives' speeches in a parliamentary debate.

In this paper we propose: (i) a social psychology grounded taxonomy (and an annotation guide) that considers the genesis of the stereotype taking into account the different *frames* in which the group is placed; (ii) *StereoImmigrants*, the first dataset annotated with dimensions of stereotypes about immigrants from political debates in a national parliamentary chamber; and (iii) a baseline for immigrant stereotype classification in the categories of the proposed taxonomy, using the state-of-the-art transformer models. For our experiments, we use some recent monolingual and multilingual transformer models (based on BERT) known for their effectiveness at the context-heavy understanding.

This paper aims to answer the following research questions:

RQ1: Is it possible to create a more fine grained taxonomy of stereotypes about immigrants from a social psychology perspective that focuses on *frames* and not on attributes defining the group?

RQ2: How feasible is to create a stereotype-annotated dataset relying on the new taxonomy?

RQ3: How effective classical machine learning and state-of-the-art transformers models are at distinguishing different categories of stereotypes about immigrants with this taxonomy?

The rest of the paper is structured as follows. Section 5.2 describes related work about stereotypes and social bias, both from social psychology and computational linguistics perspectives. Section 5.3 introduces the proposed social psychology grounded taxonomy and the annotation process that was employed to annotate the *StereoImmigrants* dataset. Sections 5.4 and 5.5 present the models that we use in the experiments and the experimental settings, respectively. In Section 5.6 we discuss the obtained results, and we conclude our work in Section 5.7 in which we also mention some directions for future work.

5.2 Related Work

From a computational perspective, there is a long list of works that address problems related to social bias like the detection of hate speech [168, 174],

aggressive language [102], abusive language [72], hostility [180], racism [117], and misogynistic language [7] among others. In this paper, we focus our attention on studying the genesis of stereotypes, specifically about immigrants.

In computational linguistics, stereotypes have been studied in images and text as well. For instance, [101] offers a study on the fairness of the algorithms that detect the descriptions of people appearing in images and their inferred gender; while [12] also focused on gender stereotypes, the authors study how the description is affected by the context in the image. Other works show the linguistic biases that are present in the way that one uses language as a function of the social group of the person(s) being described in the descriptions of images depicting people [17].

In [17], the authors describe some of the evidences of linguistic biases: (i) category labels and (ii) descriptions of behaviours. The former consists of labels used to refer to social categories, for example, explicitly marking unexpected gender roles or occupation when this one is inconsistent with the stereotypically expected role for the person's gender (e.g., *female* surgeon, *male* nurse); labels for individuals showing behaviours that violate the general stereotype (e.g., a *nice* Moroccan, a *tough* woman); and the use of nouns compared to adjectives to describe a person (e.g., being *a Jew* vs. Jewish, or Paul is *a homosexual* vs. is homosexual). The latter includes the description of the subject instead of an observable action (e.g., Jack *is flirtatious*, vs. Jack talks to Sue); the use of relatively more concrete language to describe behaviour that is inconsistent with the stereotype (e.g., *he has tears in his eyes* vs. the female consistent stereotype *she is emotional*); and the tendency to provide relatively more explanations in descriptions of inconsistent stereotype to make sense of the incongruity, among others.

To sum up, people reveal their stereotype expectancies in many subtle ways in the words they use. This fact can explain the effectiveness of several computational works at measuring social bias (e.g., gender, racial, religion, and ethnic stereotypes among others) by using word representations [20, 70, 110]. In [70], social bias are quantified by using embeddings of representative words such as women, men, Asians living in the United States, and white people (i.e., non-Hispanic subpopulation from the United States). The authors computed the average Euclidean distance between each representative group vector and each vector in a neutral word list of interest, which could be occupations or adjectives (this association of adjectives/occupations to the social group is consistent with [17] regarding stereotypes). The difference of the average distances is the metric they used for capturing personality trait stereotypes that were contrasted with historical surveys, gender stereotypes from 1977 and 1990, and ethnic stereotypes from 1933, 1951, and 1969. The authors found a correlation between the embedding gender bias and quantifiable demographic trends in the occupation participation in that period. Similar experiments were carried out with ethnic occupation. The results showed that several adjectives (e.g., *delicate*,

artificial, *emotional*, etc.) tend to be more associated to women than to men; also some occupations (e.g., *professor*, *scientist*, *engineer*, etc.) are more associated to Asians, and other occupations (e.g., *sheriff*, *clergy*, *photographer*, etc.) to white people.

In [20, 110], word embeddings are used to measure (and reduce) the bias. In [20], a methodology based directly on word embeddings is proposed to differentiate gender bias associations (e.g., biased association between *receptionist* and *female*) from associations of related concepts (e.g., between *queen* and *female*); a neutralisation and equalisation debias process. The same debias process is used in [110], but in this work a new representation to detect the bias is proposed: the authors include a contextualisation step and create two subspaces of the attribute (e.g., gender), one for male and other for female. The contextualisation relies on a large and diverse set of sentences in which the bias attribute words (e.g., *he/she, man/woman*) appear; for example, if in one subspace it is included a sentence containing *he*, the same sentence is included in the subspace for female but replacing *he* by *she*. In addition, in this line, [104] proposes some metrics over word embeddings representations to measure the bias. In this work, they distinguish two different biases: (i) *implicit bias*, in which we only have sets of target terms with respect to which a bias is expected to exist in the embedding space (e.g., $T1 = \{physics; chemistry; experiment\}$ and $T2 = \{poetry; dance; drama\}$ without any specification, one could expect in $T1$ and $T2$ a gender bias towards $\{man, father, woman, girl\}$); (ii) and *explicit bias*, in addition to sets $T1$ and $T2$, it is given one (e.g., $A = \{man, father, woman, girl\}$) or more (e.g., two opposite attribute sets $A1 = \{man, father\}$ and $A2 = \{woman, girl\}$) reference attribute sets.

Some other recent works on bias again face the gender stereotypes. A Gender Stereotype Reinforcement (GSR) measure was proposed in [54] to quantify the extent to which a search engine responds to stereotypically gendered queries with documents containing stereotypical language. Recently, in [37], the authors have compared the efficacy of lexicon-based approaches and end-to-end machine learning-based approaches (in particular BERT); the obtained results showed that the latter is significantly more robust and accurate, even when trained by moderately sized datasets. Differently, in [46] it is used Natural Language Inference (NLI) as the mechanism for measuring stereotypes [46]. The idea is that invalid inferences about sentences can expose underlying biases, and based on that, it is possible to see how gender biases affect inferences related to occupations. For example, a gender-biased representation of the word *accountant* may lead to a non-neutral prediction in which the sentence *The accountant ate a bagel* is an entailment or contradiction of the sentences *The man ate a bagel* and *The woman ate a bagel*; this could happen because of the gender bias concerning occupations. Therefore, the predictions of NLI on a set of entailment pairs that should be inherently neutral are used to compute the deviation from neutrality, which is assumed

as the gender bias.

A similar idea of [46] has been used in [133] but with a different perspective. In [133], the authors propose two different level tests for measuring bias: (i) the intrasentence test, in which there are a sentence describing the target group and a set of three attributes which correspond to a stereotype, an antistereotype, and an unrelated option; and (ii) the intersentence test, consisting of a first sentence containing the target group, a second sentence containing a stereotypical attribute of the target group, a third one containing an antistereotypical attribute, and lastly an unrelated sentence. These tests remain the aforementioned idea of [46] to use NLI to measure entailment, contradiction, or neutral inferences to quantify the bias. To evaluate their proposal, the authors of [133] collected a dataset (StereoSet) for measuring bias related to four domains: gender, profession, race, and religion. For this purpose, they use specific words to represent each social group.

However, stereotypes are not always merely the association of *words* (seen as attributes or characteristics) to two opposite social groups (e.g., women vs. men), and it is not always clear to define the opposite groups by using specific keywords, for instance in the case of immigrants vs. nonimmigrants, the set of words to represent nonimmigrants is not clear. There are few works related to the detection of stereotypes about immigrants. In [60], a system is proposed that allows one to see what was said about Muslims and Dutch people. The authors use the collection of descriptions that a single text provides on a given entity or event (it was called *microportraits*: labels, descriptions, properties, roles). For example, the expression *the pious Muslim smiled* contains the label *Muslim*, the property *pious*, and the role *smiling*. This is an interesting study that helps explain how prejudice works according to social psychologists. In [173] an Italian dataset was created that focused on hate speech against immigrants, that included the annotation *{yes, no}* about whether a tweet is a (mostly untrue) stereotype or not. In the HaSpeeDe shared task at EVALITA 2020 [116], six teams submitted their results for the stereotype detection task in addition to their hate speech models, using the above dataset. Most of those teams only adapted the same hate speech model to stereotype identification, representing (and reducing) stereotypes to characteristics of hate speech. The authors of [116] observed that stereotype appears as a more subtle phenomenon that needs to be approached also as nonhurtful text.

From a psychosocial perspective, the better-established model to analyze the language that shapes a group of stereotypes is the Stereotype Content Model (SCM) developed by Fiske and colleagues [56, 57, 58]. Fiske has developed his model arguing that in encounters with conspecifics, social animals—i.e., humans—must determine, immediately, whether the "other" is friend or foe (i.e., intends good or ill) and, then, whether the "other" can enact those intentions. Authors affirm that in answering these questions, humans use two universal dimensions of social cognition—warmth and competence—to judge

individuals and groups. People perceived as warm and competent elicit uniformly positive emotions and behaviour, whereas those perceived as lacking warmth and competence elicit uniform negativity. People classified as high on one dimension and low on the other elicit predictable, ambivalent affective and behavioural reactions. This theoretical framework has been completed with the ambivalent stereotypes hypothesis: many groups are stereotyped as high in one dimension and low in the other [58].

Cuddy, Fiske and Glik [38] also have investigated how stereotypes and emotions shape behavioural tendencies toward different groups and have proposed the BIAS Map (Behaviors from Intergroup Affect and Stereotypes).They did a correlation study with a representative US sample and conclude that warmth stereotypes determine active behavioural tendencies—attenuating active harm (harassing) and eliciting active facilitation (helping). Competence stereotypes determine passive behavioural tendencies—attenuating passive harm (neglecting) and eliciting passive facilitation (associating). Admired groups (warm, competent) elicit both facilitation tendencies; hated groups (cold, incompetent) elicit both harm tendencies. Envied groups (competent, cold) elicit passive facilitation but active harm and pitied groups (warm, incompetent) elicit active facilitation but passive harm. In this research, the authors also find that emotions predict behavioural tendencies more strongly than stereotypes do and usually mediate stereotype-behavioural-tendency links. In this research [38] immigrants are placed between the set of groups that are seen as "low warmth and low competence", with other social groups seen as poor, homeless, including Latinos, Muslims, and Africans, in the particular US context. It is predicted that groups placed in this position evoked disgust and contempt in terms of emotions [59]. However, how does one explain the appeal to fear that right-wing politicians use intensively when they speak about immigrants? Why be afraid of a group we see as incompetent? The authors do not include fear among the emotions linked to the low competence factor.

Stereotypes about immigrants present specific difficulties that Fiske addressed in an early work [107]: the internal variability existing between the members of the social category "immigrants" led the authors to study a more fine-grained taxonomy of the stereotype, based on nationalities and socioeconomic status (documented, undocumented, farm-workers, the tech industry, first-generation, and third-generation). This research concludes that people conceptualise immigrants at three levels at least: the "generic immigrant", who is equally low in *competence* and *warmth*; clusters of immigrant groups uniquely defined by one attribute, such as low or high competence, or high warmth; and immigrants by specific origin.

One interesting remark arises from this study: the group that received the least favourable stereotype across both dimensions was "undocumented immigrants". In contrast, "documented immigrants" were perceived similarly to an American. Legal status alone determines whether an immigrant is

perceived as a regular member of the mainstream society or as an outsider with the lowest status. The authors propose that one possible extension from this study could be the role of media framing of immigration status in perceived competition for finite amounts of societal resources. This idea of focusing on the framing activity is at the core of our research because we consider that it is this rhetorical activity of framing the one that enables the existence of what Lee and Fiske [107] define as "the generic immigrant".

Another interesting suggestion of these authors is that people's differing evaluations of documented and undocumented immigrants suggest that some dimensions (in this case, legal documentation) overwhelmingly bias judgement. The authors suggest for further research the question of which dimensions are most influential in perceiving immigrants when people receive information on multiple dimensions, for instance, if Asian immigrants are competent but undocumented immigrants are not, are undocumented Asian immigrants high or low in competence? They suspect that the more salient dimension would guide perception. In our research we will propose that the most salient dimension will be the result of a *frame* that presents the group in a given scenario. Recently, Kervyn, Fiske, and Yzerbyt [93] introduce—in their experiments on stereotypes about immigrants—symbolic and realistic threats and found that they improve the SCM´s prediction of warmth.

The realistic treat as origin of prejudice and stereotypes has a long tradition in the study of intergroup conflict [138, 219] and also in the Integrated Threat Theory [193] that proposes two types of perceived threat from outgroups. The first type comes from research on Realistic Group Conflict Theory [138], which posits that groups compete for scarce resources and, therefore, one group's success threatens other groups' well-being, resulting in negative outgroup attitudes. The second type of intergroup threat originates from research on Symbolic Racism, which considers racism as coming from conflicting beliefs and values rather than conflicting goals [96, 121]. Symbolic threat perceives the outgroup as threatening ingroup worldviews, assuming group differences in values, standards, beliefs, and attitudes.

Another line of research on prejudice towards minorities has been developed under the framework of Social Representations Theory by Moscovici [21, 128, 129]. For the study of prejudice, Moscovici [127] had the hypothesis that nature and culture constitute dimensions along which representations of human groups, that is to say, stereotypes, are organized in a sort of social ranking. Culture means "civilization" while Nature is "the primitive condition before human society". From this approach one of the key points to understand how people stereotype minorities is the role that these stigmatised groups play in the continuum between these two extremes of nature and culture. Perez, Moscovici and Chulvi [147, 155] have shown in their experiments that the majority group see itself nearest to the culture extreme of this vector and place the minority group nearest to nature. Other research in the dehumanisation process has shown that it is present not only in the

extreme manifestation of prejudice but it can also take subtle and everyday forms [80], for instance, differential attribution of uniquely human emotions to the ingroup versus an outgroup [108].

In the other extreme of the stereotyped continuum of minorities, we find the victimization bias [130, 131]. Most of the groups that have been considered deviant or marginal gained the status of victims from a historical process that according to Barkan [11] reaches its peak in 1990. One consequence of this shift is a change in the way that minorities are named; for instance, persons formerly labelled "handicapped" are now categorised as "differently abled" or "immigrants" are named "migrants" or "non-nationals" [183] in a social effort to give restitution to this minorities. The status of victims confers a feeling of moral superiority but according to Steele [191] "binds the victim to its victimization by linking the power to his status of the victim". When this status of victims becomes salient in the narratives about a collective, the counterstereotyped cases are made invisible.

All of this social bias that fills the content of a stereotype of immigrants and other minority groups has two main features: first, it serves to maintain minorities' discrimination, and second, it occurs in everyday language. As stereotypes are present in everyday language, they are also in the texts that systems use to classify and retrieve information. Those minorities as immigrants that suffer from this stereotyped vision of their collectives are now, with the extension of the web and massive use of social media [9], in the face of a loudspeaker that amplifies their stigma to infinity.

At the beginning of this section we mentioned some works that reduce the gender bias: similar attempts are needed to attenuate the social bias about immigrants. Nevertheless, we think that in this case more knowledge is needed for developing automatic systems that could be effective in mitigating and detecting social bias. Computational linguistics could play a major role in understanding how the immigrant's stereotype are employed—in its general formulation—as Lee and Fiske said [107]. To do this, it would be necessary to move the annotation process from the words that are used to qualify the group, to the narratives—stories about what is going on—that social psychology defines as *"frames"* [66, 67, 177] in which the minority is placed, insistently, again and again, by a social actor in the public discourse. In Computational linguistics, the concept of *frame* is also used by [175], where the authors propose Social Bias Frames, a novel conceptual formalism that aims to model pragmatic *frames* in which people project social biases and stereotypes on others.

To study more deeply this problem, the present work proposes an exhaustive taxonomy with different dimensions of the immigrants' stereotypes that have been used to annotate *StereoImmigrants*, a dataset of political speeches about immigration in Spain. Different from previous work, this work embraces the general picture about immigrants (in Spain) and not only the "negative" aspects of the stereotypes. Moreover, we do not rely on the at-

tributes, characteristics, and roles that are played by the immigrants, but we use the *frames* (labelled by humans following the taxonomy shown in Appendix 5.8) in which the immigrants are placed to detect the different dimensions of the stereotypes. More subtle examples of stereotypes are used to capture automatically more complex linguistic patterns in the way that stereotypes are present.

5.3 Social Psychology Grounded Taxonomy and StereoImmigrants Dataset

5.3.1 A Social Psychology Grounded Taxonomy

We have constructed a taxonomy trying to cover the whole attitudinal spectrum of stereotypes about immigrants, from the pro-immigrant attitude to the anti-immigrant attitude. Attitude is a theoretical concept that has been preeminent since the very beginnings of systematic research in social psychology, especially, in the study of prejudice [51]. If the stereotype is the cognitive aspect of the prejudice (a set of beliefs), the attitude expresses the effect (we could also say the feelings and emotions) that a group provokes. This taxonomy has six categories based on how the group is presented. We found that in public discourse immigrants could be presented as: (i) equals to the majority but the target of xenophobia (i.e., must have same rights and same duties but are discriminated), (ii) victims (e.g. people suffering from poverty or labour exploitation), (iii) an economic resource (i.e., workers that contribute to economic development), (iv) a threat for the group (i.e., cause of disorder because they are illegal and too many and introduce unbalances in our societies), or (v) a threat for the individual (i.e., a competitor for limited resources or a danger to personal welfare and safety). The sixth and last category presents immigrants as animals, excluding them—in whole or in part, explicitly or implicitly—from the supracategory "human beings".

The two first categories of the taxonomy (i.e., *Xenophobia's Victims* and *Suffering Victims*) hold a pro-immigrant attitude and we can aggregate them in a supracategory that we call *Victims*. Under this supracategory the goal of the speaker is to build a fair world. The speeches focus on xenophobic attitudes that are behind the problems of the minority and stress the causes of immigration. In the first category (*Xenophobia's Victims*) the speakers emphasise that the problem is not the minority but the racism and xenophobia from the majority. In this category, we include sentences such as *"We are ready to collaborate in all aspects that make life easier for our emigrants abroad, but at the same time we consider it important to work for the integration of immigrants in our country"* because they make a parallelism between the immigrant community in Spain and the Spaniards who emigrated focusing on the need to an integration strategy. In the second category (*Suffering*

Victims), we include sentences such as *"You can say what you like, but the migratory movements affecting the planet are almost exclusively linked to the phenomenon of poverty and misery"*.

The third category of the taxonomy (immigrants as an economical resource) holds an ambivalent attitude [91] that presents immigrants as an instrument for achieving societal goals. The goal of the speaker is to manage with efficiency a phenomenon difficult to avoid. In this category, we include sentences like *"how can you say that there should be no more immigrants regularisation's if, after all, reports such as that of La Caixa or BBVA indicate that the Spanish labor market needs foreigners?"*.

The three last categories of the taxonomy—immigrants as collective threat (iv), individual threat (v), or as less humans (vi)—hold the anti-immigrant attitude. We can aggregate these three categories in a supracategory that we call *Threat*. The goal of the speaker is the protection of the "national"group in front of immigrants that are presented as a danger or less human . Focused on the problems of the majority and critical about the immigrants, these speeches stress the negative effects of immigration. In the fourth category, we include sentences that consider immigration a source of problems such as *"it is clear that there is an increase in the number of people trying to enter Spain illegally"*. In the fifth category we include sentences that present immigrants as a threat not only for the collective but also for the health and security of the majority group in an explicit or implicit way: *"We need to tackle problems such as terrorism and immigration"*. In the last category of the taxonomy, which corresponds with the "dehumanization bias" we have not found examples in our dataset from the Spanish Parliament but some examples from statements made by Donald Trump could serve to illustrate the sense of this category. The former President of the United States said in a press conference at the White House: *"You wouldn't believe how bad these people are. These aren't people, these are animals, and we're taking them out of the country at a level and at a rate that's never happened before."* (NYT, 16 May 2018).

We have defined a finer level of granularity in each category to facilitate the annotation by humans (see Appendix 5.8). Each category contains a subset of *frames* that politicians used to speak about immigrants. These *frames* do not describe the group but convey a homogeneous picture of the group placing it in a particular scenario. For example, in the fifth category, defined as "a personal threat", we have identified three *frames*: (i) immigrants compete with the country's population for resources such as jobs, health services, etc.; (ii) immigrants bring diseases; and (iii) immigrants are associated with crime.

5.3.2 Annotation of the StereoImmigrants Dataset

We annotated political speeches presented in the ParlSpeech V2 dataset [158], a dataset that has been already used in other tasks like Sentiment Analysis [153, 167]. One of the peculiarities of ParlSpeech is that it is a transcription of a real debate between different and relevant social actors. Its dialogic nature makes it more difficult to approach from the perspective of computational linguistics, but it is also an opportunity to develop an interdisciplinary methodology that focuses on how social interaction takes place in language.

Specifically, we focused on the principal parliament in Spain, the Congress of Deputies (*Congreso de los Diputados*). This chamber is located in Madrid, has representatives from all regions, and elects the nation's prime minister. Using a list of 60 keywords (see Appendix 5.9) we selected all the speeches that contained at least one keyword. We obtain 5910 speeches from different years (see Figure 5.1).

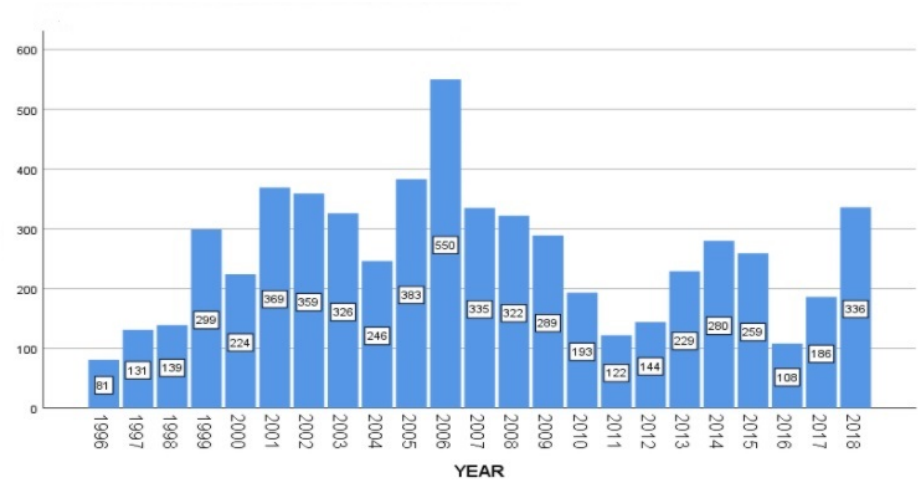


Figure 5.1: ParlSpeech V2 dataset: number of speeches with at least one immigrant-related keyword.

The year 2006 saw intense parliamentary activity on immigration. The Spanish media denominate this year the "Cayucos Crisis": the arrival of more boats than usual from different African countries to the coasts of the Canary Islands. The presence of these events in the media was very abundant, parliamentary activity on the issue was very intense, and Spanish public opinion was increasingly concerned about the issue of immigration (Based on the CIS Barometer: http://www.cis.es/cis/opencm/ES/11_barometros/index.jsp), accessed on April, 2021, (Figure 5.2).

Official data [159] on immigration for 2007 contrast with the climax that we see in the Spanish parliament and in the mass media that covered the "Cayucos Crisis". The total immigrant population at the beginning of 2007

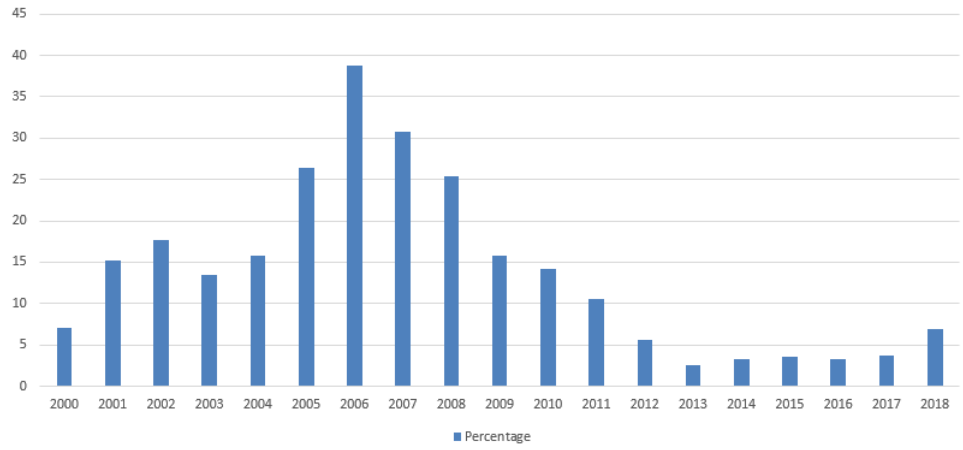


Figure 5.2: Percentage of people who consider immigration one of Spain's three main problems from a representative sample of the Spanish population. Source: CIS Barometer.

was 4.5 million. By origin, 40 percent came from Latin America and 33% from the European Union. Only 17% came from Africa despite the fact that the image of African people arriving in poor boats was the *frame* that was more used when the immigration phenomenon appeared in mass media and in political debates. The 2007 National Immigrant Survey indicates that only 10% of the immigrant's population was "undocumented" [159] but nearly 39% of the country's population was worried about illegal immigration how the CIS shows (see Figure 5.2). The 2007 National Immigrant Survey also shows that the immigrant population was quite similar to the Spaniards in some parameters: most parts of the immigrant population (59%) between 20 and 34 years old have completed lower and upper secondary education, 17% have higher education, and only 23% belong to the primary education or no education group. The different level of studies between the immigrant population and the Spanish population at this moment was not enormous: the group of Spaniards that has primary education or does not have any education in this age group was 13%. In fact, other studies, using other sources of data, highlight a situation in which the immigrant population has a very similar profile in terms of higher education to the Spanish population [32, 119].

This gap between the variability of the real situation and the image that was conveyed through the mass media and the parliamentary interventions led us to build the dataset focusing on the speeches from 1996 to 1998, from 2006 to 2008, and from 2016 to 2018. In that way, we could contemplate speeches of consecutive years and also how these vary from one decade to another. We selected 582 speeches with more than one keyword and manually discarded the ones in which immigrants were mentioned alongside other

groups but only tangentially. From these selected speeches, we manually extracted 3635 sentences for studying stereotypes about immigrants.

An expert in prejudice from the social psychology area annotated manually these sentences at the finest granularity of the taxonomy (i.e., identifying the different frames that fall into the same category, see Appendix 5.8), and selected also negatives examples (i.e., *Nonstereotype* label) where politicians speak about immigration but do not use explicitly or implicitly any category of the stereotypes about immigrants. After this expert annotation we use a procedure with some similarities with [133]: five nonexperts annotators read the label assigned by the expert to each sentence and decided if they agreed with it or considered that another label from the taxonomy was better suited for this sentence. The annotators where 77 undergraduate students from psychology, fine arts, and business. We only retained sentences where at least three validators agreed on the same category. In our dataset, each sentence is accompanied by the following information: politician's name, political party to which the speaker belongs, and date of the parliamentary session. This metainformation was hidden also for the expert annotator and for the nonexperts.

Table 5.1 depicts the distribution of instances per label. We include the mean of the length of the instance based on tokens to help us to define the hyperparameters of the models in Section 5.5. We can observe an imbalance across the dimensions of stereotype, where dimension 5 (i.e., *Personal threat*) is the smallest set of instances with only 81, and dimension 4 (i.e., *Collective threat*) is the biggest with 655 instances. In addition, the dataset has an imbalance regarding *Stereotype* (1673 instance) vs. *Nonstereotype* (1962 instances). In general, all the labels have a similar distribution according to the length of the instances, but the nonstereotyped instances are slightly more consistent in length, with a smaller standard deviation. We take into account this distribution in the experimental settings (Section 5.5).

5.3.3 Evaluation of the Taxonomy

We asked nonexpert annotators for their judgement about the attitude expressed in the text. Concretely, each annotator had to say if this text expressed a pro-immigrant, an anti-immigrant, or an ambivalent attitude that annotators qualified as neutral. The purpose of this second task was to test if the theoretical value assigned to each category in terms of positive or negative stereotype was justified. Our aim was to analyse if there were any significant relations among categories and attitudes. For instance, we expect that given a text labelled with the 1st category (i.e., *Xenophobia's Victims*), there will be a significant high probability of being judged to express a pro-immigrant attitude.

To test the relationship among the categories of the taxonomy and the attitudes towards immigrants, we performed a chi-square test and a residual

Table 5.1: Statistics of the *StereoImmigrants* dataset. For each label, *Stereotype* and *Nonstereotype*, and for each category of the stereotype, we show the number of instances and the length based on tokens (ignoring punctuation marks).

			Length in Tokens		
		Instances	Min	Mean ± Standard Deviation	Max
Stereotype	**1. *Xenophobia's Victims***	186	6	50.55 ± 30.59	183
	2. *Suffering Victims*	557	7	47.32 ± 24.41	151
	3. *Economical Resources*	194	9	42.39 ± 22.31	128
	4. *Collective threat*	655	8	43.42 ± 23.28	162
	5. *Personal threat*	81	9	48.26 ± 25.56	149
	All dimensions joined	1673	6	45.62 ± 24.69	183
Nonstereotype		1962	3	36.00 ± 21.17	165
Total		3635	3	40.43 ± 23.35	183

analysis [179]. A residual is a difference between the observed and expected values for a cell. The larger the residual, the greater the contribution of the cell to the magnitude of the resulting chi-square obtained value. However, cells with the largest expected values also produce the largest raw residuals. To overcome that redundancy, a standardised or Pearson residual is calculated by dividing the raw residual by the square root of the expected value as an estimate of the raw residual's standard deviation. If the standardised residual is beyond the range of ±2.58 that cell can be considered to be a major contributor to the chi-square significance.

Results confirm a significant relation (Pearson $\chi^2 = 3828.24$, $df = 8$, $p < 0.001$; see Table 5.2) of the taxonomy's categories and the positive, neutral, or negative evaluation. The residual analysis in Table 5.2 shows that category 1 (*Xenophobia's Victims*) and category 2 (*Suffering Victims*) are significantly associated with positive attitudes, category 3 (i.e., *Economical Resource*) is significantly associated with a neutral attitude, and categories 4 and 5 (i.e., *Collective Threat* and *Personal Threat*) are significantly associated with a negative attitude.

Taking into account these results we consider it appropriate to use categories 1 and 2 as a supracategory that we named *Victims* and categories 4 and 5 in a supracategory *Threat*. These supracategories will be used to evaluate the effectiveness of the models at the automatic classification of stereotypes about immigrants. category 3, that we qualify also as supra-category named *Resources*, evaluated as neutral, will be left out of the experiments because of the small number of instances. Figure 5.3 summarises the taxonomy at its different levels.

Table 5.2: The relation among categories of the taxonomy and attitudes expressed in the texts of the dataset. Chi-square test with adjusted standardised residuals.

		Attitudes towards Immigrants			
Taxonomy Categories		**Pro-Immigrant**	**Anti-Immigrant**	**Neutral**	**Total**
Xenophobia's Victims	Obs	916	90	143	1149
	Adj. Res	**25.4**	−23.6	−3.0	
Suffering Victims	Obs	2002	414	363	2779
	Adj. Res	**34.8**	−32.3	−4.2	
Economical resource	Obs	482	187	259	928
	Adj. Res	4.4	−12.7	**11.1**	
Collective threat	Obs	339	2406	508	3253
	Adj. Res	−50.5	**51.2**	0.3	
Personal threat	Obs	108	268	45	421
	Adj. Res	−8.2	**10.4**	−2.8	
	Total (Obs)	3847	3365	1318	8530

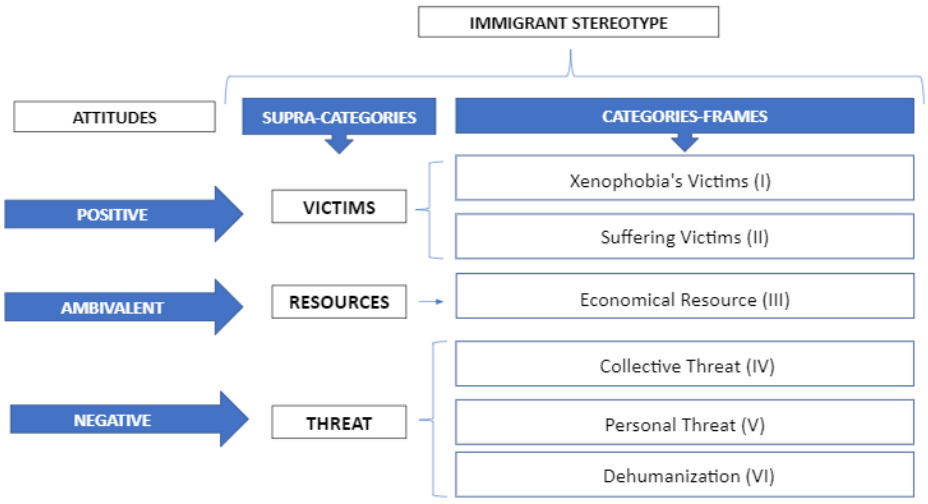


Figure 5.3: Explanatory taxonomy scheme.

5.4 Models

In this section, we briefly present the state-of-the-art models we have used for our experiments, which have been trained with huge general language datasets of pretrained systems based on Bidirectional Encoder Representations from Transformers (BERT).

BERT is a language representation model designed to pretrain deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers [47]. That is, BERT generates a representation of each word that is based on the other words in the sentence. This allows the model to capture complex patterns in the texts to study stereotypes, going beyond merely the use of words and capturing semantic and syntactic patterns in the same representation. BERT has also an attention mechanism that distinguishes if a word is attended by the model. These attention weights could be used also to give more insights about the results of the model and be used as a tool to support the work of human experts. Some important aspects of BERT include the pretraining, the fine-tuning, and the capability to be adapted to many types of Natural Language Processing (NLP) tasks like text classification.

To classify text in Spanish, we have used two monolingual models (BETO and SpanBERTa) and two multilingual models (MBERT and XLM-RoBERTa) briefly described below:

M-BERT: The Multilingual BERT is pretrained on the concatenation of monolingual Wikipedia datasets from 104 languages, showing good performance in many cross-lingual scenarios [47, 150, 212].

XLM-RoBERTa: It was trained on 2.5TB of newly created clean CommonCrawl data in 100 languages. It provides strong gains over previously released multilingual models like M-BERT on downstream tasks like classification, sequence labelling, and question answering. In [31], it was reported with better results to the one obtained by fine-tuning with Spanish data only.

BETO: This is a recent BERT model trained on a big Spanish dataset [31]. It has been compared with multilingual models obtaining better or competitive results [207]. BETO was trained using 12 self-attention layers with 16 attention heads each and 1024 as hidden sizes. It was trained using the data from Wikipedia and all of the sources of the OPUS Project [204]. BETO also was ranked in a better place than Logistic Regression in the prediction of aggressive tweets [30].

SpanBERTa: SpanBERTa (https://skimai.com/roberta-language-model-for-spanish/), accessed on April, 2021, was trained on 18 GB of the OSCAR's Spanish corpus (https://oscar-corpus.com/), accessed on April, 2021, following the RoBERTa's training schema [112]. It is built on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

Moreover, we use as baselines classical machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB).

5.5 Experimental Settings

We evaluate each model on two tasks. First, given a text about immigration, to predict whether or not it contains a stereotype about immigrants; second, given a text already known that is reflecting a stereotype, to detect if the stereotype corresponds to see immigrants as *victims* or *threat*.

Experiment I: Stereotype vs. Nonstereotype This experiment is very relevant for us because we have annotated not attributes that speakers assign to a group but narratives about the group that in an implicit way convey a stereotyped vision of the collective. Due to negative examples also being sentences from members of the Parliament speaking about immigration, we want to see if the models detect the subtle difference that consists in approaching the issue without personifying the problem in one group, i.e., immigrants as a social category.

Experiment II: Victims vs. Threat With this experiment we tried to see if the model can detect which dimension of stereotype about immigrants has been used in the political discourse. One common rhetorical strategy used by politicians that present immigrants as a threat to the majority group is to dedicate a part of their speech to recognise the suffering of the minority. However, these claims of compassion are framed by a discourse that clearly presents immigration as a problem and migrants as a threat. We are interested to see if a model is able to distinguish the deeper meaning of that statement as the human annotators did.

We apply a 10-fold cross-validation procedure and report our results in terms of *accuracy*. For the execution of each model, we balance the two labelled classes by randomly removing examples from the more populated class. Therefore, in Experiment I we use 1673 examples per label and in Experiment II we use 736 examples per label.

We use two monolingual Spanish transformer models: BETO cased and SpanBERTa; and two multilingual transformer models: MBERT (*bert-base-multilingual-cased*) and XLM-RoBERTa (*xlm-roberta-base*). We search the following hyperparameter grid: *learning rate* $\in \{0.01, 3e-5\}$; the *batch size* $\in \{16, 32\}$; and the optimizer $\in \{adam, rmsprop\}$. Moreover, we apply a dropout value of 0.3 to the last dense layer. Dropout is aimed at reducing overfitting by dropping a percentage of random units at each weight update.

Besides the length of the texts not being restricted to predicting stereotypes in general, we have to select a value for the max_length hyperparameter to use the transformer models. According to the characteristics of our data, the mean of the lengths of the instances is approximately 40 tokens (see the last row of Table 5.1), with a standard deviation of around 20 tokens. Taking this in mind and evaluating the number of instances with a length greater than 40 tokens, we finally select 60 as the value for the hyperparameter max_length in order not to lose too many instances. Accordingly, our transformer models expect an input text of around 60 tokens. In the case of longer texts, only the first 60 will be used while the rest is truncated.

Furthermore, we use the *sklearn* implementation of the four classical machine learning models. All the parameters were taken by default, except for LR in which we use the *newton-cg* optimization method. For SVM, we employ specifically the *LinearSVC* implementation, which uses a linear kernel and has more flexibility in the choice of penalties and loss functions. The number of trees in the forest of the RF classifier is 100 (the one by default). The four models were evaluated with the bag of words representation, using the *tfidf* term weighting. We tested with unigrams, bigrams, and trigrams of words, but unigrams allow for obtaining slightly better results. Stopwords and punctuation marks were removed in a preprocessing step.

5.6 Results and Discussion

In this section, we present the results of our two preliminary experiments. The optimal hyperparameter configuration for all the models is the following: $learning\_rate = 3e-5$, $optimizer = adam$, and $batch\_size = 32$. In general terms, we observe that the best performances were consistently achieved by BETO and M-BERT models. It is not surprising that M-BERT obtained better results than SpanBERTa, being the latter pretrained specifically for Spanish: a similar comparison between a multilingual model and a monolingual model was reported in [31]. Another general observation is that either for *Stereotypes vs. Nonstereotypes* and *Victims vs. Threat*, BETO seems to capture more complex patterns than the classical machine learning models, which are based on the bag of words representation.

The next subsections discuss the results of Experiment I and Experiment II in more detail.

Table 5.3: Accuracy achieved by each model in Experiment I on Stereotype vs. Nonstereotype. We indicate the p-value of the Mann–Whitney U test regarding the alternative hypothesis that the results of BETO are the highest compared to the other transformers. With * we indicate when accuracy is significantly lower than the result of BETO. The hypothesis is accepted with $p < 0.05$ except for XLM-RoBERTa model.

BETO	**SpanBERTa**	**M-BERT**	**XLM-RoBERTa**
0.861 ± 0.016	0.766 *±0.021 $_{p=0.00018}$	0.829 *±0.022 $_{p=0.00736}$	0.780 ±0.105 $_{p=0.06057}$
LR	**SVM**	**NB**	**RF**
0.82	0.81	0.73	0.81

5.6.1 Experiment I: Stereotype vs. Nonstereotype

In this section, we present the results concerning the identification of stereotype about immigrants. In Table 5.3 we can appreciate that all the models obtained an *accuracy* above 0.73. The highest result was obtained by BETO with 0.861 of *accuracy* and a standard deviation of 0.016. This performance was significantly better than the one of SpanBERTa and M-BERT for $p < 0.05$ using the Mann–Withney U Test. We use this test, also known as Wilcoxon–Mann–Whitney test (a nonparametric alternative to the Student's t-test), considering that we randomly removed instances from the most populated class each time. Therefore, we assume independence among the results of the models. We see that LR, SVM, and RF obtained a higher accuracy compared to SpanBERTa and XLM-RoBERTa. We believe that this could be associated with the size of the dataset since more data could have had an impact on deep learning models achieving better results [124].

Besides the fact that stereotypes involve more than the presence of specific words, word n-grams with the highest Pointwise Mutual Information (PMI) (PMI makes it possible to see the most relevant features (i.e., words, n-grams of words) for each topic and is computed as $PMI(L, w) = \log \frac{p(L,w)}{p(L)p(w)}$. Where $p(L, w)$ is the probability of a feature to appear in a text labeled as L, $p(L)$ is the probability of a label (we assume the label distribution to be uniform), and $p(w)$ is the probability of w.) to each label allow us to see that nonstereotypical texts talk more about *ayudas* (help) to refugees and Africa (the country some immigrants come from, and it is mostly mentioned in the speeches we are working with), *acuerdos* (agreement) between countries, etc. While in stereotypical texts we find more commonly bigrams such as *inmigración ilegal* (illegal immigration), *inmigrantes irregulares* (irregular immigrants), and *regularización masiva* (massive regularization) among others that indirectly reflect problems associated to immigration (see Table 5.4).

Interestingly, it is not evident from observing the relevant n-grams why some of them are more related to the stereotypes about immigrants and oth-

Table 5.4: Bigrams and trigrams with highest mutual information with respect to Stereotype and Nonstereotype labels.

	N-Grams with Highest Mutual Information
Nonstereotype	*inmigrantes irregulares; señor rajoy; señor zapatero; países africanos; abordar asunto; abordar problema; absolutamente acuerdo; acción exterior; acción política; acoger personas; acogida refugiados; acogida temporal; acorde derechos; acuerdo materia; acuerdos gobierno; acuerdos marruecos; acuerdos mauritania; acuerdos readmisión; adquisición nacionalidad; afecta unión europea; agencia europea fronteras; aguas canarias; aguas territoriales; asilo inmigración; asunto preocupa; atención inmigración; autoridad moral; ayuda refugiado; ayuda África;...*
Stereotype	*efecto llamada; inmigrantes ilegales; política inmigración; inmigración irregular; unión europea; materia inmigración; derechos humanos; inmigrantes irregulares; consejo europeo; inmigración ilegal; drama humano; regularización masiva; islas canarias; seres humanos; política común; proceso regularización; inmigración delincuencia; llegada masiva; respeto derechos; situación irregular; economía sumergida; mujeres inmigrantes; centros acogida; orden expulsión; centros internamiento; costas canarias; miles personas; europea inmigración; política migratoria; política exterior; respeto derechos humanos; menores acompañados; acogida canarias; drama humanitario; empresarios sindicatos; menores inmigrantes; crecimiento económico; acogida inmigrantes;...*

ers are not. This confirms that for the study of stereotypes about immigrants we have to go beyond the representative keywords that could define the social group. In other words, we should not rely only on intuitive words to base the measuring of bias in the case of stereotypes. In this sense, automatic approaches can detect other patterns that escape human detection.

We also confirm that the detection of stereotypes in this work, concerning immigrants, is not about characterising two opposite social groups but the immigrant group only. The nonstereotypical texts are neutral in this sense referring only to the topic of immigration without stereotyping at all.

In Table 5.5, we present the confusion matrix of BETO when obtaining the results shown in the Table 5.3. The model was similarly effective at predicting stereotypes and nonstereotypes, with a bit more confusion with the label Nonstereotype. Table 5.6 shows some examples where BETO misclassified the texts and their predictions.

Table 5.5: Confusion matrix of BETO in Experiment I on Stereotype vs. Nonstereotype.

	Predicted Labels	
	Stereotype	***Nonstereotype***
Stereotype	1451	222
Nonstereotype	240	1433

Table 5.6: Examples of texts correctly classified and misclassified by BETO in the Experiment I on Stereotype vs. Nonstereotype. The true label is indicated in each example. For the stereotypes, we indicate the dimension to which each sentence belongs.

Classified Examples with the Right Label
1. *Nos gustaría que lo acompañara de política migratoria real también.* (Nonstereotype) (We would like that actual immigration policy was considered as well.) 2. *Señorías, la política de integración es la gran asignatura pendiente.* (Nonstereotype) (Ladies and gentlemen, integration policy is the great pending issue.) 3. *Es decir, se tiene que cambiar la política de inmigración del Gobierno.* (Nonstereotype) (In other words, the government's immigration policy has to be changed.) 4. *El secretario de empleo dijo: España seguirá necesitando inmigrantes.* (S: Economical Resource) (The employment secretary said: Spain will continue needing immigrants.) 5. *En lo que va de año han llegado a Canarias más de 3500 personas en pateras.* (S: Threat) (So far this year, more than 3500 people have arrived in the Canary Islands in boats.) 6. *El problema, señora vicepresidenta, está en que en el cuerno de África mueren todas las semanas 40,000 niños por falta de nutrición.* (S: Victims) (The problem, Vice President, is that 40,000 children die every week in the Horn of Africa due to lack of nutrition.)
Misclassified Examples
1. *Entendemos que España puede jugar un papel destacado en cuanto a este problema, pero Europa será más creíble si afronta problemas reales que los ciudadanos perciban.* (Nonstereotype) (We understand that Spain can play a leading role in this problem, but Europe will be more credible if it faces real problems that citizens perceive.) 2. *Evidentemente nos encontramos ante una situación compleja, la relativa a las remesas en un momento en el que la política migratoria ha adquirido una gran dimensión.* (Nonstereotype) (Obviously we face with a complex situation, relating to remittances at a time when migration policy has acquired a great dimension.) 3. *Desde que aprobamos en 1985 la Ley de extranjería, de los derechos y obligaciones de los extranjeros en España, ha mantenido una línea congruente.* (Nonstereotype) (Since we approved in 1985 the Law on foreigners, on the rights and obligations of foreigners in Spain, it has maintained a congruent line.) 4. *El asunto de la inmigración requiere medidas de control pero fundamentalmente –y lo apuntaba usted ayer– medidas de solidaridad y este y este es un reto europeo.* (S: Victims) (The issue of immigration requires control measures but fundamentally—and you pointed this out yesterday–solidarity measures and this is a European challenge.) 5. *Decían que lo que pasaba en España era un coladero para los distintos países de la Unión Europea y a ustedes no les importó lo más mínimo. En aquella época el ministro.* (S: Threat) (They said that what was happening in Spain was a drain for the different countries of the European Union and you did not care at all. At that time the minister.) 6. *Por tanto, se abre un camino esperanzador, y yo solamente les deseo éxitos por el bien del conjunto de los trabajadores inmigrantes, por el bien de la política en el Estado.* (S: Economical Resource) (Therefore, a hopeful path opens, and I only wish you success for the good of all immigrant workers, for the good of politics in the State)

5.6.2 Experiment II: Victims vs. Threat

In Table 5.7, we can see all the performances are above the 0.70 of *accuracy*. The results are, in general, smaller than in Experiment I; this could be because the size of the training set for the current experiment is smaller or due to the difficulty that we have already mentioned in Section 5.5, describing the scenario of Experiment II: as some humans annotators reported, one of the common rhetorical strategies in political discourse is to precede any critical statement towards the immigrant collective with an expression of compassion towards the human drama that it represents in order not to be accused of xenophobia. For instance, the speaker is going to say that Spain could not admit more immigrants (threat) but she starts speaking about how many people died trying to arrive (*Victims*). Methodologically, we decided to assign only one label per sentence, and perhaps it would have been more effective to annotate different syntagmas within the same sentence, at least in those sentences in which this discursive strategy was developed.

Table 5.7: Highest accuracy achieved by each model in Experiment II on Victims vs. Threat. We indicate the p-value of the Mann–Whitney U test regarding the alternative hypothesis that the results of BETO are the highest compared to the other transformers. With * we indicate when accuracy is significantly lower than the result of BETO. The hypothesis is accepted with $p < 0.05$ only for SpanBERTa model.

BETO	**SpanBERTa**	**M-BERT**	**XLM-RoBERTa**
0.834 ± 0.034	0.704 *± 0.064 $_{p=0.00024}$	0.809 ± 0.022 $_{p=0.4965}$	0.785 ±0.070 $_{p=0.70394}$
LR	**SVM**	**NB**	**RF**
0.79	0.78	0.72	0.77

Similar to the previous experiment, the highest accuracy was obtained by BETO, but this time with 0.834 of accuracy and a standard deviation of 0.034. This performance was significantly better than the one of SpanBERTa for $p < 0.05$ using the Mann–Whitney U Test.

In Table 5.8, we show some of the n-grams (without including stopwords) more relevant for each label, for example: *atención humanitaria* (humanitarian care) , *atención sanitaria* (healthcare), *acogida personas* (welcome of people). These relevant n-grams allow us to figure out that the phrases are more likely to reflect the needs and pain of immigrants when they are seen as *victims*.

In Table 5.9, we present the confusion matrix of BETO at obtaining its result from the Table 5.7. We can see that BETO is almost equally effective at detecting *Victims* and *Threat* dimensions. In Table 5.10 we show some texts that were classified correctly and wrongly, respectively.

As we see in the misclassified examples 5 and 6 in Table 5.10, speak-

Table 5.8: Bigrams and trigrams with highest mutual information with respect to each label.

	N-Grams with More Mutual Information
Victims	*ley extranjería; ceuta melilla; guardia civil; política inmigración; presión migratoria; partido popular; acoger personas; acogida canarias; acogida inmigrantes; acogida integración; acogida personas; acogida temporal; acuerdos bilaterales; administración justicia; aeropuertos fronteras; fronteras terrestres; aflorar irregulares; afrontar problema; aguas canarias; aguas territoriales; amnistía internacional; apoyo; aquellas personas; aquellos inmigrantes; aquellos países; archipiélago canario; asuntos sociales; atención humanitaria; atención sanitaria;...*
Threat	*inmigración irregular; derechos humanos; regularización masiva; inmigración ilegal; inmigrantes ilegales; proceso regularización; efecto llamada; inmigrantes irregulares; señor caldera; asilo refugio; mujeres inmigrantes; inmigración delincuencia; personas muerto; control inmigración; derecho asilo refugio; mauritania senegal; canaria nueva; derecho asilo; inmigración problema; trafican seres humanos; principal problema; inmigración clandestina;...*

ers mention that thousands of people have arrived, or specifically the term *avalanchas* (avalanches of people) but also refer, in the same sentence, to some words or phrases that are in the semantic field of compassion and victims such as *sufrimiento* (suffering) and *han dejado su vida en el intento* (have lost their lives in the attempt). We could think that fear is an emotion stronger than compassion, so humans give more weight to the part of the sentence that generates fear than to the part that generates compassion and consider that in this *frame* migrants are presented as a *threat*.

To understand better the confusion matrix shown in Tables 5.9 and 5.10, we explore the hypothesis that some parties perform a rhetorical strategy to avoid being tagged as xenophobic that consists in mentioning some expression of compassion just before or after they present immigrants as a threat, a strategy that does not convince humans' annotators but confuses the performance of transformers. As we have seen in Table 5.8, the presence of the n-grams *personas muertas* (death people) and *trafican seres humanos* (traffic human beings) in the dimension of *Threat* could be indicative of this rhetorical strategy that we have mentioned above: "appealing to pity and misfortune just before presenting the immigrant collective as a threat".

To explore this hypothesis, we did an analysis using the word "party"

Table 5.9: Confusion matrix of BETO in Experiment II on Victims vs. Threat.

	Predicted Labels	
	Victim	***Threat***
Victim	611	125
Threat	119	617

in the dataset, which indicates the party that uttered the sentence. We have 12 different parties, but we keep only those parties that have more than one hundred sentences in the dataset and we created a new category for the rest (*Other Parties*). We performed a chi-square test and a residual analysis to identify if the disagreement between human annotators and the model has any relation with the rhetorical strategy of a party. We found a significant relation (Pearson X2 = 36,979, df = 8, $p < 0.000$, see Table 5.10) between parties and the type of confusion, see Table 5.11. As it was mentioned in Section 5.5, for this experiment we balanced the classes to have 736 examples per class and, therefore, in this analysis we used a total of 1472 labelled examples.

With the general category, *Other Parties* and with "Coalición Canaria", a regional party from the Canary Islands, very concerned about immigration, we did not find any significant differences in the type (One type of confusion is when the annotators label is *Victims* and BETO predicted it as *Threat*, and the other type is the opposite, the annotators label is *Threat*, but BETO predicted it as *Victims*.) of confusion between humans and the model (sometimes humans label *Victim* and the model predicts *Threat* and the opposite) but in case of the right win party Partido Popular (PP) humans label more often *Threat* when the model labels is *Victim*, whereas with the left win party Izquierda Unida (IU) the type of confusion is the opposite: humans label is *Victim* and the model labels predicts *Threat*. With PSOE, the socialist center-left party, the type of confusion goes in the same direction as with IU but is not statistically significant.

This result leads us to think that the model's confusion is based on the fact that politics use the same words for different purposes trying to avoid the label of xenophobic. This rhetorical strategy could be detected by humans that make an inference about the intentionality of the speaker, but computational models have more difficulties to detect it.

There is a great deal of research about how the human communication process occurs and which role the inference of the speaker's intentionality plays [88, 89, 90]. We think that one interesting approach to a further exploration of the computational linguistic difficulties is the one that Watzlawick and colleague suggest in their Pragmatic of Human Communication Theory [213]. They suggest that the study of human communication can be subdi-

Table 5.10: Examples of texts correctly classified and misclassified by BETO in Experiment II on Victims vs. Threat. The annotators label is indicated in each example.

Classified examples with the right label
1. *¿Por qué ha muerto una persona joven?* (*Victims*) (Why did a young person die?)
2. *Derechos de ciudadanía para los inmigrantes.*(*Victims*) (Citizenship rights for immigrants.)
3. *Esta no es la forma de enfrentarse con un problema que requiere, sobre todo, grandes dosis de solidaridad.* (*Victims*) (This is not the way to deal with a problem that requires, above all, large doses of solidarity.)
4. *Hay en España más ciudadanos irregulares que nunca.* (*Threat*) (There are more irregular citizens in Spain than ever.)
5. *España hoy está desbordada con la inmigración ilegal.* (*Threat*) (Spain today is overwhelmed with illegal immigration.)
6. *Hay más llegadas de inmigrantes irregulares que nunca.* (*Threat*) (There are more arrivals of irregular immigrants than ever.)
Misclassified examples
1. *Por cierto, el Gobierno debería explicarnos cuántos inmigrantes se fugaron este fin de semana del centro de Las Raíces, si fueron veinte, como dice el delegado del Gobierno, o si fueron cien, como afirman fuentes policiales.* (Threat) (By the way, the Government should explain to us how many immigrants escaped this weekend from the center of Las Raices, if there were twenty, as the Government delegate says, or if there were a hundred, as stated by police sources.)
2. *No queremos olvidar la operación Melilla, la expulsión de los 103 ciudadanos.* (*Victims*) (We do not want to forget the Melilla operation, the expulsion of the 103 citizens.)
3. *Por tanto, apostamos por una política de retorno, de repatriación humanitaria.* (*Threat*) (Therefore, we are committed to a policy of return, of humanitarian repatriation.)
4. *Esto es un escándalo, esto son más trabas a los migrantes cuando ya se encuentran dentro.* (*Victims*) (This is a scandal, these are more obstacles to migrants when they are already inside.)
5. *Ya son 25.000 los inmigrantes llegados a Canarias en lo que va de año y se cuentan por miles los que han dejado su vida en el intento.* (*Threat*) (There are already 25,000 immigrants who have arrived in the Canary Islands so far this year and there are thousands who have lost their lives in the attempt.)
6. *¿Por qué en tres meses no han tomado ninguna de las medidas propuestas para evitar las avalanchas que han generado tanto sufrimiento?* (Threat) (Why in three months have they not taken any of the measures proposed to avoid the avalanches that have generated so much suffering?)

vided into three areas: (i) *syntactic*—problems of transmitting information (a matter of mathematical logic), (ii) *semantic*—meaning of communication (a matter of philosophy of science), and (iii) *pragmatic*—communication affecting behaviour (a matter of psychology). While a clear conceptual separation of the three areas is possible, they are nevertheless interdependent. The same act of communication can express a content (then the question is "what" is being said and, therefore, we would be in the area of syntactic or semantic) but also the same act of communication can express a personal or a social relationship (the question is "how" is being said and, therefore, we would be in the psychology area).

Following this reasoning, Watzlawick said that humans communicate both digitally and analogically. *Digital concept* refers to humans convey-

Table 5.11: Relation between the type of confusion and political parties. Chi-square test with adjusted standardised residual.

			Annotators Say *Victims* But BETO Predicts *Threat*	Annotators Say *Threat* but BETO Predicts *Victims*	Annotators and BETO Agreement	Total
Parties	**Coalición Canaria**	**Obs.**	14	10	127	151
		Adj. Res.	0.4	−0.7	.2	
	IU	**Obs.**	22	2	130	154
		Adj. Res.	2.7	−3.3	0.3	
	PP	**Obs.**	11	40	330	381
		Adj. Res.	−4.6	2.0	1.9	
	PSOE	**Obs.**	42	29	324	395
		Adj. Res.	1.8	−0.6	−0.9	
	Other Parties	**Obs.**	36	38	317	391
		Adj. Res.	0.6	1.4	−1.5	
Total (Obs.)			125	119	1228	1472

ing meanings by using words (syntactic and semantics) and *analogical concept* refers to when humans convey relational content. For the analogical level, Watzlawick mentions nonverbal communication, posture, gesture, facial expression, voice inflection, rhythm, and cadence of words, etc. From a psychosocial point of view, we can reinterpret this definition of the analogical level of communication that Watzlawick identifies [213], in the broadest sense of the ability that allows humans to capture the level of the social relations that is behind the words. Using this human ability, people infer the speaker's intentionality in a sentence. For instance, when one speaker say "It is a humanitarian drama, more than 300 hundred boats had arrived at the Canarian Island this summer", the reader infers that for this speaker, the important part of the message is the second one, that means: "Spain can not accept more immigrant". This kind of inferences about the intention of the speakers is a natural cognition activity for humans, but it is more difficult for computational models.

5.7 Conclusions and Future Work

In order to advance in the study of stereotypes, for instance about immigrants, social sciences need to complement the classical paradigm that focuses on how a group is defined in terms of personal attributes with a new paradigm that emphasises the *frames*, that is, the narrative scenarios, in which a group (e.g. immigrants) is mentioned. In this work, we developed a taxonomy to identify stereotypes about immigrants not from the attributes that are assigned to the group but from the narrative scenarios in which the speaker places the group (**RQ1**). This fine grained taxonomy allows for the study of stereotypes about immigrants and covers the whole spectrum of the stereotype: from positive images of the group (as equals to the nationals or victims of xenophobia) to more negative images (a threat to the nationals or a less human people).

Based on psychosocial research on prejudice and stereotypes, this language-independent taxonomy is a new conceptual instrument with two objectives: (i) to provide computational linguistics with a new conceptual tool to detect and mitigate social bias in datasets, specifically, stereotypes about immigrants; and (ii) to strengthen the collaboration between social sciences and computational linguistics to understand better how stereotypes are generated in the context of public discourse.

We have validated our taxonomy considering how each category is related with some attitude (pro-immigrant, anti-immigrant, or neutral). We have identified two opposite supracategories of the stereotypes about immigrants: one that presents the minority as *victim* and the other that presents the minority as a *threat*. We annotated political speeches of the ParlSpeech V2 dataset, focusing on the speeches from 1996 to 1998, from 2006 to 2008, and from 2016 to 2018. The resulting *StereoImmigrants* stereotype-annotated dataset was created relying on the new taxonomy (**RQ2**). The dataset will be made public to the research community to foster the research on stereotypes about immigrants. *StereoImmigrants* was used to carry out some preliminary experiments using state-of-the-art transformer models and classical machine learning models. We obtained results between 0.70 and 0.86 of accuracy in the two experiments we performed: Stereotype vs. Nonstereotype and Victims vs. Threat. The best performance was obtained by BETO, a monolingual Spanish transformer model, suggesting that this model could be capturing a richer representation of stereotypes and their dimensions, than the classical machine learning models do (**RQ3**). We also point up that with these preliminary experiments we prove the existence of social bias, in particular stereotypes about immigrants, in political speeches, and the effectiveness of automatically detecting them.

From the most relevant n-grams from the examples labelled as *Stereotype*, we confirmed that they are not trivially associated with the immigrant group (they are not always biased attributes) and, therefore, we should not rely only

on a set of keywords to represent it. Additionally, we confirmed that in the study of stereotypes about immigrants, we have to consider not to define two opposite social groups, since the nonstereotypical texts are neutral phrases talking about immigration in general without stereotyping at all.

We have analysed the confusion matrices considering metadata from our *StereoImmigrants* dataset, in particular, the political party the speaker belongs to. This analysis has shown that the confusion could be explained because certain rhetorical strategies are particularly difficult to infer for transformers and not for humans. When speakers use the same words for different purposes, humans elaborate the different meanings of the sentence making an inference about the intention of the speaker, but transformers have difficulties inferring the rhetorical strategies. More work is necessary in this direction.

The taxonomy, although used to label text in Spanish, is applicable to other languages because it classifies into categories the different *frames* in which immigrants are placed. These *frames* are common in Western cultures. These categories express the dimensions of the stereotype about immigrants on the north area of a north-south axis of economical inequality. Moreover, we aim to apply this taxonomy on the Europarl corpus, as well as on texts of different genres like newspaper datasets.

As future work, we could test if these two big dimensions of immigrant stereotype (*Victims* and *Threat*) could be applied to other minorities' stereotypes as feminist or LGTBI people. In fact, these minorities are presented also as a *threat*, for instance, when feminist women are presented in scenarios that emphasise conflict and then are defined as *feminazis*. In addition, feminist women are presented as "victims" when the narrative context emphasises gender violence. A general idea for future work will be that the study of minorities' stereotype needs these two dimensions (*Victims* vs. *Threat*) to complement the well established Stereotype Content Model [57] that proposes *warmth* and *competence* as two fundamental dimensions of stereotypes.

Furthermore, we plan to analyse how social bias (in particular stereotypes) is reflected in the attention values of the transformer layers, in order to facilitate the explainability of the results and a further debiasing process. Moreover, we think that it will be interesting to enrich the dataset with more examples of each stereotype category for evaluating the multiclass classification by using the transformer models (e.g. BETO).

5.8 Taxonomy: Categories and Frames

Annotators should use the *frames* (labels in two digits) to label the texts, if it is possible; in other case, the label should be the category (label in one digit). For example, if the annotator recognises that a target text is saying that immigrants have the same rights as the nationals, she should label

the text using the *frame 1.1* (from category 1). However, if the annotator identifies that the text belongs to category 1, but she cannot specify the *frame*, then the label to put should be *1*.

If one text contains fragments that correspond to different *frames* from the same category, the label should be at one digit level: the one corresponding to the category.

If one text contains fragments that correspond to different *frames* from different categories, we ask the annotators to choose the most important or to discard the sentence.

Category 1: Xenophobia's Victims

1.1 With the same rights and with the same obligations. They are named as *ciudadanos*(citizens), *nueva ciudadanía* (new citizenship), etc.

1.2 They are presented doing a simile with the Spaniards who emigrated.

1.3 It is suggested that the problem is the racist or xenophobic attitudes of people.

1.4 It is claimed that immigration topic is used as an electoral or partisan weapon and that this is not right. Some party is accused of being racist and/or xenophobic. The rise of racist or xenophobic parties is seen as a problem.

1.5 It is stated that immigration is not a problem for coexistence. The population is not concerned about the presence of immigrants and the problem is xenophobia or racism.

1.6 Immigration is considered to bring cultural diversity, pluralism, etc. and that is positive for the country.

Category 2: Suffering Victims

2.1 Victims of suffering and poverty. It is argued that poverty and suffering in their countries of origin is the cause of the immigration. In addition, they are victims of suffering once they are here.

2.2 Victims of injustice and discrimination. Victims of labour exploitation and mafias. It is reported that they do not respect human rights in the treatment of immigrants or it is stated that they have to be respected.

2.3 Solidarity is required or manifested in the face of immigrant problems.

2.4 It is suggested that they die trying to get there (Spain in this work), for example, there is talk about the rescues.

Category 3: Economical Resource

3.1 They do the jobs that the Spanish do not want to do. They support the black economy. They are seen as workers in a situation of vulnerability, with special difficulties.

3.2 They bring economic prosperity: they pay taxes, send remittances abroad, etc.

3.3 They solve the problem of lack of population.

3.4 They propose measures to hire immigrant workers in their countries of origin: they must be approved, etc.

3.5 The entry of immigrants must be regulated according to the needs of the labour market.

Category 4: Collective Threat

4.1 They come in droves and create a situation of chaos. It could be mentioned *avalancha* (avalanche), *falta de control* (lack of control), *llegada a las costas* (arrival at the coast), and so on.

4.2 The problem is that they are illegal. They refer to them as *ilegales* (illegal) or *irregulares* (irregular) or using the category *inmigrantes* (immigrants) or *inmigracion* (immigration). It could be mentioned *repatriaciones* (repatriations), *devoluciones* (returns), or *expulsiones* (expulsions).

4.3 It is stated that immigration is a problem for the host society, causing imbalances in coexistence of the group.

Category 5: Personal Threat

5.1 It is argued that immigrants compete with the country's population for resources such as work, health services, and education. Immigration remains as a problem with regard to the use of these resources.

5.2 Immigrants are reported to bring diseases or are referred to as carriers of new diseases.

5.3 Immigration is associated with crime.

Category 6: Dehumanisation

6.1 They do not know how to live as human beings do.

6.2 They behave like animals.

6.3 Their deaths are not our problem: they come because they want to.

5.9 Keywords Used to Filter Immigration-Related Speeches

The keywords shown in Table 5.12 were used to discard those speeches that were not talking about immigration as a central topic. These keywords were carefully defined by a social psychologist who payed attention to the important historical events that occurred during the periods of the speeches.

Table 5.12: Keywords used to filter relevant speeches.

anti-inmigrante	*deportado*	*inmigración*	*pateras*
anti-inmigrantes	*deportados*	*inmigrante*	*permisos de residencia*
asilada	*deportar*	*inmigrantes*	*polizones*
asiladas	*desheredados de la tierra*	*islamofobia*	*racismo*
asilado	*devolución*	*migrantes*	*racista*
asilados	*efecto llamada*	*migratoria*	*refugiada*
centro de acogida	*efecto salida*	*migratorias*	*refugiadas*
centros de acogida	*emigrantes*	*migratorio*	*refugiado*
ciudadanía inmigrada	*etnocentrismo*	*multiculturalismo*	*refugiados*
ciudadano emergente	*expatriada*	*nativismo*	*repatriación*
ciudadanos emergentes	*expatriadas*	*nuevas ciudadanas*	*schengen*
colonialismo	*expatriado*	*nuevos ciudadanos*	*sociedad de acogida*
deportación	*expatriados*	*países de recepción*	*xenófoba*
deportada	*extranjería*	*países emisores*	*xenofobia*
deportadas	*indocumentados*	*países en tránsito*	*xenófobo*

Chapter 6

Masking and BERT-based Models for Stereotype Identification

In this chapter of the thesis we present the results of using the masking technique for detecting immigrant stereotypes. We compare this technique with BETO, a BERT-based model for Spanish in terms of F-measure and local explanations that both approaches could give. We also make further experiments to estimate the results of an ideal ensemble of these different approaches. The experiments with the attention scores of the transformers suggest new strategies to obtain cues of biased terms, and also how neutral words can be contextualized in different forms depending on *partisanship*.

The work presented in this chapter was published in the following paper:

- **Sánchez-Junquera J.**, Rosso P., Montes-y-Gómez M., Chulvi B. (2021) Masking and BERT-based Models for Stereotype Identification. *In: Procesamiento del Lenguaje Natural (SEPLN)*, num. 67, pp. 83-94. Best paper award.

Abstract

Stereotypes about immigrants are a type of social bias increasingly present in the human interaction in social networks and political speeches. This challenging task is being studied by computational linguistics because of the rise of hate messages, offensive language, and discrimination that many people receive. In this work, we propose to identify stereotypes about immigrants using two different explainable approaches: a deep learning model based on Transformers; and a text masking technique that has been recognized by its capabilities to deliver good and human-understandable results. Finally, we show the suitability of the two models for the task and offer some examples of their advantages in terms of explainability.

Keywords: social bias, immigrant stereotypes, BETO, masking technique.

6.1 Introduction

Nowadays, social media, political speeches, newspapers, among others, have a strong impact on how people perceive reality. Very often, the information consumers are not aware of how biased is what they are exposed to. To mitigate this situation, many computational linguistics efforts have been made to detect social bias such as gender and racial biases [20, 46, 70, 110]. The immigrant stereotype is another type of social bias that is present when a message about immigrants disregards the great diversity of this group of people and highlights a small set of their characteristics. This process of homogenization of a whole group of people is at the very heart of the stereotype concept [199]. As [111] said in his seminal work about stereotypes, stereotyping, as a cognitive process, occurs because "we do not first see and then define, we define first and then see". In short, we can say that a stereotype is being used in language when a whole group of people, itself very diverse, is represented by appealing to a few characteristics.

Unfortunately, the use of stereotypes promotes undesirable behaviours among people from different nationalities; an example is the violence against Asian Americans that have taken place recently [201]. Moreover, political analysts have associated the success of anti-immigration parties with the even more negative attitudes to the immigration phenomenon [42]. These stereotypes have received little attention to be automatically identified, despite the harmful consequences that prejudices and attitudes, in many cases negative, may have.

There have been some works related to the problem of immigrant stereotypes identification [173], but they are mainly focused on the expressions of hate speech; or social bias in general that involves racism [60, 133]. However, it is necessary to have a whole view of immigrant stereotypes, taking into account both positive and negative beliefs, and also the variability in which stereotypes are reflected in texts. This more refined analysis of stereotypes

would make it possible to detect them not only in clearly dogmatic or violent messages, but also in other more formal and subtle texts such as news, institutional statements, or political representatives' speeches in parliamentary debates.

Similar to applications of healthcare, security, and social analysis, in this task is not enough to achieve high results, but it is also mandatory that results could be understood or interpreted by human experts on the domain of study (e.g. social psychologists). Taking into account these two aspects, performance and explainability, the objective of this work is to compare two approaches diametrically opposite to each other in the text classification state of the art. On the one hand, a transformer-based model, which has shown outstanding performance, but high complexity and poor explainability; and, on the other hand, a masking-based model, which requires fewer computational-resources and showed a good performance in related tasks like profiling.

We aim to find explainable predictions of immigrant stereotypes with BETO by using its attention mechanism. In this way, we derive the explanation by investigating the importance scores of different features used to output the final prediction. With the other approach that we use, the masking technique [188, 198], it is possible to know what are the most important words that the model preferred to highlight. We compare these approaches using a dataset of texts in Spanish, which contains annotated fragments of political speeches from Spanish Congress of Deputies.

The research questions aim to answer in this work are:

RQ1: Is the transformer more effective than the masking technique at identifying stereotypes about immigrants?

RQ2: Is it possible to obtain local explanations on the predictions of the models, to allow human interpretability about the immigrant stereotypes?

The rest of the paper is organized as follow: Section 6.2 presents related work concerning the immigration stereotype detection and the models that we propose. Section 5.4 describes the two models, and Section 6.4 the dataset used in the experiments. Sections 6.5 and 6.6 contain the experimental settings and the discussion about the results. Finally, we conclude the work in Section 6.7 where we mention also future directions.

6.2 Related Work

6.2.1 Immigrant Stereotype Detection

There have been attempts to study stereotypes from a computational point of view, such as gender, racial, religion, and ethnic bias detection do [20, 70,

110]. Those works predefine two opposite categories (e.g. men vs. women) and use word embeddings to detect the words that tend to be more associated with one of the categories than with the other. In [133], the authors propose two different level tests for measuring bias. First, the intra-sentence test, with a sentence describing the target group and a set of three attributes that correspond to a stereotype, an anti-stereotype, and a neutral option. Second, the inter-sentence test, with a sentence containing the target group; a sentence containing a stereotypical attribute of the target group; another sentence with an anti-stereotypical attribute; and lastly, a neutral sentence. These tests are similar to the idea of [46] that consist in using natural language inference to measure entailment, contradiction, or neutral inferences to quantify the bias. To evaluate their proposal, in [133], the authors collected a dataset (StereoSet) for measuring bias related to gender, profession, race, and religion domains.

On the other hand, stereotypes are not always the (explicit) association of *words* (seen as attributes or characteristics) from two opposite social groups, like women vs. men in the context of gender bias. Such is the case of immigrant stereotypes, in which sentences like *¿Por qué ha muerto una persona joven?* (Why did a young person die?) do not contain an attribute of the immigrant group although from its context[1] it is possible to conclude that here immigrants are placed as victims of suffering. Also, it is not clear the representative word of the social group, since *persona joven* (young person) is neutral to immigrants and non-immigrants.

Other works have built annotated data to foster the development of supervised approaches. In [173], was presented an Italian corpus focused on hate speech against immigrants, which includes annotations about whether a tweet is a stereotype or not. This corpus was used in the HaSpeeDe shared task at EVALITA 2020 [116]. Most participant teams only adapted their hate speech models to the stereotype identification task, thus, representing (and reducing) stereotypes to characteristics of hate speech. One of the conclusions was that the immigration stereotype appeared as a more subtle phenomenon, which also needs to be approached as non-hurtful text. Additionally, in [133], it was proposed a dataset that includes the domain of racism (additionally to gender, religion, and profession). Although this dataset does not focus on the study of stereotypes about immigrants, its authors reported the word "immigrate" as one of the most relevant keywords that characterized the racism domain.

6.2.2 On the Explainability of AI models

Since eXplainable Artificial Intelligence (XAI) systems have become an integral part of many real-world applications, there is an increasing number of

[1] Fragment of a political speech from a Popular Parliamentary Group politician in 2006. The speaker is mentioning some of the conditions of immigrants in Spain in that period.

XAI approaches [84] including white and black boxes. The first group, which includes decision trees, hidden Markov models, logistic regressions, and other machine learning algorithms, are inherently explainable; whereas, the second group, which includes deep learning models, are less explainable [40]. XAI has been characterized according to different aspects, for example, (i) by the the level of the explainability, for each single prediction (*local explanation*) or the model's prediction process as a whole (*global explanation*); (ii) and if the explanation requires post-processing (*post-hoc*) or not (*self-explaining*).

XAI has also been characterized in accordance to the source of the explanations, for example: (i) *surrogate models*, in which the model predictions are explained by learning a second model as a proxy, such is the case of LIME [162]; (ii) *example-driven*, in which the prediction of an input instance is explained by identifying other (labeled) instances that are semantically similar [35]; (iii) *attention layers*, which appeal to human intuition and help to indicate where the neural network model is "focusing"; and (iv) *feature importance*, in which the relevance scores of different features are used to output the final prediction [40].

Taking into account this characterization, we frame our approach in the *self-explaining* scope, and consider two different models to obtain *local explanations* of the predicted texts. In this sense, we use the *attention layers* which have been commonly applied by local self-explaining models [19, 132]. For example, in [120] the attention weights were used to compare the posts' segments on which the labeling decision was based, highlighting the tokens that the models found the most relevant. Similarly, in [33] the authors used datasets for tasks like dependency parsing, to evaluate attention heads of BERT, and found relevant linguistic knowledge in the hidden states and attention maps, such as direct objects of verbs, determiners of nouns, and objects of prepositions. Finally, in [86] attention was used to prove that some swear words are inherently offensive, whilst others are not, since their interpretation depends on their context.

The other self-explaining model that we use to obtain the local explanations, is a masking technique which can be described as a white box. In this case, the explainable strategy is based on the *feature importance* idea, by measuring and observing the relevant words used in its masking process. The masking technique used in this work incorporates an additional way to explain decisions [77, 188]. It allows highlighting content and style information from texts, by masking a predefined and task-oriented set of irrelevant words.

6.3 Models

In this section, we briefly describe the two models that we use in our experiments.

Original text
la inmigración sigue siendo hoy - lo confirman los últimos sondeos del CIS - el principal problema que preocupa a los ciudadanos del estado
(Immigration is still today - confirmed by the latest CIS polls - the main problem that worries the citizens of the state)
Masking stopwords
*** inmigración sigue siendo hoy ** confirman *** últimos sondeos *** cis ** principal problema *** preocupa * *** ciudadanos ** *** ******
Masking the non-stopwords
*la *********** ***** ***** *** lo ******** los ******* ******* del *** el ******** ******** que ******** los ********** del estado*

Figure 6.1: Example of masking the stopwords, or keeping only the stopwords unmasked.

BETO: it is based on BERT, but it was pre-trained exclusively on a big Spanish dataset [31]. The framework of BETO consists of two steps: pre-training and fine-tuning, similar to BERT [47]. For the pre-training, the collected data included Wikipedia and other Spanish sources such as United Nations and Government journals, TED Talks, Subtitles, News Stories among others. The model has 12 self-attention layers with 16 attention-heads each, and uses 1024 as hidden size, with a total of 110M parameters. The vocabulary contains 32K tokens.

For fine-tuning, the model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream task, which in our case is a stereotype-annotated dataset (see Section 6.4). The first token of every sequence is always a special classification token ([CLS]), which is used as the aggregate sequence representation for classification tasks. In our work, we add to the [CLS] representation two dense layers and a Softmax function to obtain the binary classification.

The masking technique: it consists of transforming the original texts to a distorted form where the textual structure is maintained while irrelevant words are masked, i.e., replaced by a neutral symbol. The irrelevant terms are task-dependent and have to be defined in advance, following some frequency criteria or the expert's intuition.

The masking technique replaces each term t of the original text by a sequence of *. The length of the sequence is determined by the number of characters that t contains. One example of this is shown in Figure 6.1 considering the Spanish stopwords. In Section 6.5.1, we explain which are the relevant words that we considered better to mask.

After all texts are distorted by the masking technique, we use a traditional classifier to be compared with BETO. In our experiments, we use Logistic Regression (LR) classifier which has been used before to be compared with BERT [4].

6.4 Dataset

We use the StereoImmigrants dataset[2] for identifying stereotypes about immigrants [197]. In this previous work we collected texts on immigrant stereotypes from the political speeches of the ParlSpeech V2 dataset [158]; from where we also extracted the negative examples (labeled as Non-stereotypes). These texts are extracted from the speeches of the Spanish Congress of Deputies (*Congreso de los Diputados*), and are written in Spanish.

In the construction of StereoImmigrants [197], we proposed a new approach to the study of immigrant stereotyping elaborating a taxonomy to annotate the corpus that covers the whole spectrum of beliefs that make up the immigrant stereotype. The novelty of this taxonomy and this annotation process is that the work has not focused on the characteristics attributed to the group but on the narrative contexts in which the immigrant group is repetitively situated in the public discourses of politicians. To do this, the authors applied the frame theory –a social psychology theory– to the study of stereotypes. The frame theory allows us to show that politicians in their speeches create and recreate different *frames* [177], i.e. different scenarios, where they place the group. The result of this rhetorical activity of framing ends with the creation of a stereotype: a diverse group is seen only with the characteristics of the main actor in a particular scenario.

In [197], we identify different frames used to speak about immigrants that could be classified in one of the following categories: (i) present the immigrants as equals to the majority but the target of xenophobia (i.e., they must have the same rights and same duties but are discriminated), (ii) as victims (e.g. they are people suffering from poverty or labor exploitation), (iii) as an economic resource (i.e., they are workers that contribute to economic development), (iv) as a threat for the group (i.e., they are the cause of disorder because they are illegal, too many, and introduce unbalances in societies), or (v) as a threat for the individual (i.e., they are competitors for limited resources or a danger to personal welfare and safety). In the construction of the StereoImmigrants dataset, an expert in prejudice from the social psychology area annotated manually the sentences at the finest granularity of the taxonomy and selected also negatives examples where politicians speak about immigration but do not refer, explicitly or implicitly, to the people that integrates the group “immigrants". After this expert annotation, five non-experts annotators read the label assigned by the expert to each sentence and decided if they agreed with it or considered that another label from the taxonomy was better suited for this sentence. The dataset only contains sentences where at least three annotators agreed on the same category.

In [197], attending to a second annotation of the attitudes that each sentence expresses, we proposed two supra-categories of the stereotypes an-

[2]https://github.com/jjsjunquera/StereoImmigrants.

notated as Victims or Threat, where the categories (i) and (ii) belong to the Victims supra-category, and (iv) and (v) belong to the Threat supra-category. Table 6.1 shows the distribution per label of the dataset.

Table 6.1: Distribution of texts per label and the average length (with standard deviation) of their instances. The texts labeled as Victims or Threat are a subset of the texts labeled as Stereotype.

Label	**Length**	**Texts**	
Stereotype	45.62 ± 24.69	1673	3635
Non-stereotype	36.00 ± 21.17	1962	
Victims	48.93 ± 27.5	743	1479
Threat	45.84 ± 24.42	736	

Table 6.2 shows examples of Non-stereotypes and Stereotypes labels. The Stereotypes examples specify if they were labeled as Victims or Threat. From these examples, it is possible to see that the dataset contains stereotypes that are not merely the association of attributes or characteristics to the group, but texts which reflect biased representations of the group (i.e., how the immigrants are indirectly perceived or associated with specific situations and social issues).

Table 6.2: Examples from each label of the dataset.

Non-stereotype
No nos vale que se contabilice todo lo que se dedica a inmigración porque no estamos hablando de lo mismo. (We are not worth accounting for everything that is dedicated to immigration because we are not talking about the same thing.)
El Gobierno está desbordado por la inmigración, por su política improvisada, irresponsable, descoordinada y unilateral. (The Government is overwhelmed by immigration, by its improvised, irresponsible, uncoordinated and unilateral policy.)
Stereotype: Victims
¿Por qué ha muerto una persona joven? (Why did a young person die?)
Hay una situación de desamparo en muchas personas a la que necesitamos dar una solución. (There is a helplessness situation in many people to which we need to provide a solution.)
Stereotype: Threat
España hoy está desbordada con la inmigración ilegal. (Spain today is overwhelmed with illegal immigration.)
Esta alarmante situación, agravada por la incapacidad del Gobierno socialista, ha producido el colapso, el desbordamiento de los servicios humanitarios, judiciales y policiales que han generado una gran alarma social. (This alarming situation, aggravated by the incapacity of the socialist government, has produced the collapse, the overflow of humanitarian, judicial and police services that have generated great social alarm.)

6.5 Experimental Settings

We applied a 10-fold cross-validation procedure and reported our results in terms of *F-measure*.

For BETO, we searched the following hyperparameter grid to obtain the results of BETO: *learning rate* $\in \{0.01, \mathbf{3e\text{-}5}\}$; the *batch size* $\in \{16, \mathbf{34}\}$; and the optimizer$\in \{\mathbf{adam}, rmsprop\}$ (in bold we highlighted the optimal hyperparameter values). Moreover, we applied a dropout value of 0.3 to the last dense layer. We have selected a value of 180 for the max_length hyperparameter according to the maximum length of all the texts in the dataset. The model was finetuned for 10 epochs on the training data for each task.

For the masking-based approach, we used the *sklearn* implementation of the LR classifier. All the parameters were taken by default, except for the optimization method: we selected *newton-cg*. The model used the bag of words representation, using the *tfidf* term weighting. We tested with unigrams, bigrams, and trigrams of words; and with characters n-grams (n$\in \{3, 4, 5, 6\}$) obtaining the better results with **character 4-grams**. When we used LR with the original texts, unigrams of words achieved better results than character n-grams.

6.5.1 Unmasking Stereotypes

Related works on social bias detection have found a list of words that tend to be associated with one of two opposite social groups (e.g. female vs. male, Asian vs Hispanic people) [20, 70]. In the immigrant stereotypes case, it is particularly difficult to define two opposite groups and consequently to find such biased words. In this paper, we use the dataset described in Section 6.4 to find which could be the most relevant terms to be used in the masking process.

Intuitively, in the immigrant stereotypes' context, the relevant words could be content-related, although style-related terms like function words could play also an interesting role. After preliminary experiments, we found higher results by masking the words out of the following lists: (i) the words with higher relative frequency ($RelFreq$), i.e., the k words with a frequency in one class remarkably higher than its frequency in the opposite class; and (ii) the k words with the highest absolute frequency ($AbsFreq$) in all the collection, excluding stopwords (i.e., stopwords were masked). In our experiments we achieved better results with $k = 1000$.

Each list was computed using the corresponding set of texts depending on the classification task: Stereotype vs. Non-stereotype, or Victims vs. Threat. The information that is kept unmasked corresponds to the content-related words.

Table 6.3: F-measure in both classification tasks: Stereotype vs. Non-stereotype (S/N), and Victims vs. Threat (V/T).

	S/N	**V/T**
Original text	0.82	0.79
Masking Technique with $AbsFreq$	0.79	0.75
Masking Technique with $RelFreq$	0.84	0.81
BETO	**0.86**	**0.83**

6.6 Results and Discussion

We report the results of the models in Table 6.3. It is possible to see high results of LR with the original texts. However, we observe that masking the terms out of the list $RelFreq$ is slightly better than using the original text. These results suggest that the masking technique improves the quality of the stereotype detection and its dimensions.

In comparison with $AbsFreq$, maintaining unmasked the $RelFreq$ words helps to ignore more words that are less discriminative for classification tasks. This could be explained because $AbsFreq$ includes words similarly frequent in both classes, which could not help at predicting immigrant stereotypes: *países* (countries), *gobierno* (government), *señor* (mister), *partido* (party); or at identifying the immigrant-stereotype dimension: *fronteras* (frontiers), *política* (politic), *seguridad* (security), *grupo* (group). Table 6.4 shows examples of words included in $RelFreq$ that are indeed reflecting some bias according to the category. For example, it is not surprising to find words like *derechos* (rights), *humanos* (human[3] or humans), *pobreza* (poverty), *muerto* (dead), and *hambre* (hunger) more associated to immigrants seen as victims; and words like *irregular* (irregular), *ilegal* (illegal), *regularización* (regularization), *masiva*(massive), and *problema* (problem), more used in speeches where immigrants are seen as collective or personal threat.

BETO achieves the highest results in both classification tasks (**RQ1**). This is not surprising because the transformer-based models are known for their properties at capturing semantic and syntactic information, and richer patterns in which the context of the words are taken into account. However, we do not observe a significant difference between the results of such a resource-hungry model, and the combination of the masking technique with the traditional LR classifier. Considering the computational capabilities that BETO demands, and the less complexity of the masking technique, the latter shows a better trade-off between effectiveness and efficiency than the latter.

[3]It refers to the adjective: *human* rights.

Table 6.4: Examples of the relevant words that were not masked, considering the list $RelFreq$ in each classification task.

Stereotype	**Non-stereotype**	**Victims**	**Threat**
personas	política	derechos	inmigración
canarias	europea	personas	canarias
derechos	unión	humanos	gobierno
problema	materia	derecho	irregular
país	grupo	mujeres	ilegales
irregular	políticas	países	irregulares
situación	consejo	pobreza	regularización
regularización	cooperación	integración	ilegal
ilegales	gobierno	mundo	españa
humanos	moción	vida	problema
ciudadanos	europeo	solidaridad	proceso
irregulares	ley	asilo	masiva
efecto	parlamentario	condiciones	llegado
origen	comisión	millones	aeropuertos
centros	cámara	muerto	ministro
ilegal	desarrollo	refugiados	control
drama	consenso	social	efecto
acogida	subcomisión	miseria	pateras
llamada	socialista	internacional	llamada
menores	común	ciudadanos	medidas
vida	temas	xenofobia	llegada
mafias	tema	hambre	inmigrantes
llegada	asuntos	viven	marruecos
masiva	grupos	emigrantes	cayucos
extranjeros	emigrantes	muerte	presión

6.6.1 Discriminating Words

Motivated by the similar results of BETO and the masking technique, we wanted to observe and compare what portions of the texts they could be focusing on. For this purpose, we looked at the last layer of BETO and computed the average of the attention heads. Therefore, for each text, we had an attention matrix from which we could compute the attention that the transformer gave to each word in that texts. Figure 6.2 shows examples of texts where the two models agreed on the right label.

From the figure, it is possible to see what words were relevant for both approaches. Although some of the relevant words are function words (e.g. *para*, *muy*) and are not too informative at first glance for human interpretation, we can observe that some content-related words can be helpful for expert's analysis. For instance, the text labeled as Stereotype has as relevant words *fenómeno* (phenomenon), *inmigración* (immigration), *problema* (problem), *terrorismo* (terrorism), *paro* (unemployment), among others. The text labeled as Victims contains *desamparo* (abandonment), *personas* (people), *necesitamos dar una solución* (we need to give a solution), reflecting how immigrants were seen as people more than their illegal status (e.g. see

Stereotype:

BETO:
pues bien , el fenómeno de la inmigración es hoy sin ninguna duda , no solamente por las encuestas del cis sino por distintos diagnósticos de la opinión pública , un asunto que preocupa más a los ciudadanos que los problemas del terrorismo y del paro .
Masking:
**** bien ** fenómeno ** ** inmigración ** hoy *** ninguna **** ** ********* *** *** encuestas *** *** **** *** ********* ********** ** ** opinión pública ** asunto *** preocupa *** *** ciudadanos *** *** problemas *** ********** * *** ****

Non-stereotype:

BETO:
cuando se habla de inmigración , de qué está hablando el grupo parlamentario de esquerra ?
Masking:
****** ** habla ** inmigración ** *** **** ******** ** grupo parlamentario ** esquerra

Victims:

BETO: hay una situación de desamparo en muchas personas a la que necesitamos dar una solución .
Masking: *** *** ********* ** ********* ** muchas personas ** *** ********** dar *** solución

Threat:

BETO: el tiempo nos ha dado la razón , se ha convertido en un problema muy serio , en un problema muy importante tanto para europa como para españa .
Masking: ** tiempo *** ** **** ** ***** ** ** convertido ** ** problema *** ***** ** ** problema *** ********** ***** **** europa **** **** españa

Figure 6.2: Examples of attention visualization and masking transformation over the same texts. These examples were correctly classified by both models. The more intense the color, the greater is the weight of attention given by the model.

Tables 6.4), and the target of problems that need solutions. Moreover, in the example of Threat, some of the words and phrases receiving more importance (such as, *problema muy serio*, *problema muy importante*) reflect how immigrants were seen as a problem to the continent and the country, but not the country where immigrants come from.

Table 6.5 presents some of the words with the highest attention scores in only the true positive predictions of each class. Therefore, these words could be among the most discriminative for stereotype identification.

We contrasted the list of words with more attention on the BETO true positive predictions, with the $RelFreq$ words used by the masking technique as more discriminative for each class. Table 6.6 shows the percentage of

Table 6.5: Words whose attention scores are the highest only on the true positive predictions of BETO in each class.

Stereotype	Non-stereotype	Victims	Threat
drama	*consejo*	*derecho*	*masiva*
llegada	*asuntos*	*esclavitud*	*pateras*
ilegales	*temas*	*mujeres*	*llamada*
efecto	*comparecen*	*asilo*	*avalancha*
irregulares	*producen*	*refugiados*	*aeropuertos*
llamada	*recibir*	*pobreza*	*trasladar*
costas	*acuerdos*	*xenofobia*	*llegada*
expulsiones	*esfuerzo*	*muerto*	*zapatero*
trabajadores	*diálogo*	*sistema*	*caldera*
pateras	*pacto*	*devoluciones*	*alarma*
xenofobia	*congreso*	*miseria*	*ayudas*
mafia	*necesidad*	*desgracia*	*ilegalmente*
condiciones	*cumbre*	*difícil*	*afrontar*
legalidad	*proyecto*	*grupos*	*judiciales*
dinero	*zapatero*	*racismo*	*capacidad*
avalancha	*enmiendas*	*hambre*	*archipiélago*
vienen	*miembros*	*refugio*	*delincuencia*
peninsula	*colabora*	*persona*	*grave*
muertes	*conferencia*	*situaciones*	*congreso*
miles	*gobiernos*	*explotación*	*tropicales*
humanitaria	*exterior*	*denuncia*	*fallecido*
coladero	*importantes*	*muerte*	*coladero*
preocupa	*acción*	*democracia*	*oleada*

Table 6.6: The percentage of $RelFreq$ words that are in the top of words with the highest attention.

Attention Ranking	**% from a total of unmasked words**	
	S/N	V/T
top 10	48.96%	35.89%
top 20	78.92%	65.11%
top 30	95.84%	82.98%

$RelFreq$ words (which were not masked) that were present in the top of the ranking as *more discriminative* from BETO. In the top 30 of the ranking, we found the vast majority of the not masked words. This suggests that the two approaches have seen similar cues.

For now, we have seen that BETO and the masking technique achieved similar results and have an intersection in the discriminative words they focused on in the texts (which in fact answer **RQ2**), despite one of them is a resource-hungry model and the other requires less computational resources. We do not think that BETO should not be used because of its complexity: one of the differences we should highlight is that for the masking technique the list of words should be predefined with some limitations and algorithm bias that this could imply. However, BETO learns by itself to score the words gradually, instead of giving a binary score like in the masking technique (to

mask or keep unmasked). Therefore, we can apply from the transformer a comparison of the importance of the different words (like it was visualized in Figure 6.2), which was automatically learned from the context of the words in the texts. In the next sections, we confirm the advantages of both models by analyzing the results of an ideal ensemble and other utilities of the attention mechanisms.

6.6.2 An Ideal Ensemble

We have seen the results achieved by the proposed models and the intersection set of words they focus on in the texts. Therefore, one could think that these models are classifying correctly the same texts. In this section, we report that the models are misclassifying different instances in general.

Table 6.7 shows the misclassified instances of LR with the masked texts and BETO in each classification task. The models have good performances, so it is licit to think in an ideal ensemble that could wisely combine their predictions. The resultant ensemble will miss only the texts where both models are wrong: 272 texts at distinguishing Stereotype vs. Non-stereotype, which means that 92.5% of the 3630 texts will be correctly classified. A similar analysis can be done in the Victims vs. Threat classification task, which will result in 92.2% of the 1477 texts that will be potentially correctly classified.

Table 6.7: Misclassified instances and the performance of an ideal ensemble for Stereotype vs. Non-stereotype and Victims vs. Threat tasks.

	S/N	V/T
Total of instances	3630	1477
Misclassified by LR	624	263
Misclassified by BETO	518	259
Misclassified by both	272	115
Well predicted by an ideal ensemble	92.5%	92.2%

6.6.3 Relations with the Highest Attention Scores

Another advantage of the attention mechanism is the relations between the non-discriminative words and other words from each class. We could find noisy features similarly present in the opposite classes. One of the words with the highest attention scores in our dataset is *inmigración* (immigration); since we found its scores high in the two opposite classes, we did not count it as *discriminative* by BETO. However, we think that as the heads have the attention that each word gives to the others in the texts, we can observe how the "noisy" words are used in the opposite classes, by looking at the relations

with their context. We hypothesize that the immigration-related words are used in different contexts in the opposite classes.

Table 6.8 shows an example of the words whose relation with *inmigración* are the most scored in each class. We omitted the ones in the Non-stereotype class due to they are not informative. Interestingly, the words associated with *inmigración* are also describing differently the Stereotype, Victims, and Threat classes. For example, with this strategy we observe that words like *criminal*, and *enfermedades* (diseases) are now in the top of discriminating words of the Threat category (in contrast to Table 6.5). We conclude that the attention mechanism should be exploited in the future in this sense. Probably the attention scores could be a source of interesting cues not only in terms of biased words from $RelFreq$ list or the ones shown in Table 6.5, but also concerning the forms in which neutral terms are contextualized.

Table 6.8: Words with the highest attention scores in relation to *inmigración* (immigration).

Stereotype	Victim	Threat
muertes	discriminación	llega
saturado	colectivos	nuevo
miseria	mujeres	delincuencia
pobres	consenso	aeropuertos
policiales	dentro	procedente
dramaticos	refugiados	zapatero
descontrol	familias	saturado
humanitario	educativo	policiales
costas	planteamos	retención
avalancha	miseria	madrid
garantías	pobreza	enfermedades
delincuencia	reto	congreso
tráfico	voto	evolución
devueltos	voluntad	entran
explotación	especificas	francia
llamada	pobreza	tropicales
alarman	iniciativa	criminal
ilegales	enmienda	aeropuerto
pateras	podían	intentos
expulsión	saben	coladero

6.7 Conclusion and Future Work

This work is a contribution to the immigrant stereotype identification problem. The particularities of the immigration phenomenon make this bias detection task differs from other kinds of bias that have received much more

attention (e.g. gender bias). We addressed two classification tasks, the Stereotype vs. Non-stereotype detection, and Victims vs. Threat dimensions identification using an annotated dataset in Spanish. We proposed two different models: BETO, a resource-hungry model which demands strong computational capabilities; and a masking technique, a less complex approach that transforms the texts to be used by a traditional classifier. We demonstrate that both approaches are suitable for immigrant stereotype identification; and interestingly, the masking technique achieves almost the same results of BETO, despite its simplicity (RQ1).

We developed a comparison between the attention mechanism of BETO, and the list of relevant terms that the masking technique uses. These two different approaches focused on similar portions of the texts. Specifically, the majority of the relevant words maintained unmasked are at the top of the words that BETO gave the highest attention. Furthermore, with these models it is possible to highlight some stereotype cues that could be considered as *local explanations* for further studies about immigrant stereotypes (RQ2).

On the basis of the reported results, we conclude that both models are effective at identifying the immigrant stereotypes, and could be combined to build an ideal ensemble that overcomes the results of each one. We also point out that BETO can help to investigate with more detail the bias towards immigrants with the attention mechanisms. For these reasons, we think we cannot rule out the use of either model.

To our knowledge, this is the first work on immigrant stereotypes identification that compares deep learning with traditional machine learning approaches paying special attention to the explicability of the models in this task. However, more work is necessary to explore more deeply the advantages of the attention mechanisms in this sense. In future work, we plan to combine the two approaches to increase the performance; and to use discriminative words to find debiasing strategies to mitigate the immigrant stereotypes in social media and political speeches.

Chapter 7

Discussion of the Results

7.1 Introduction

As we mentioned in Section 1.1, to pursue their own political goals, politicians manage to influence the way people think by using specific linguistic means that often present **deceptive** behaviour. As a result, political messages give visibility to a part of reality in which specific social groups are **stereotyped**, and consequently biased conditions of **partisanship** or **hyperpartisanship** derive from. This research focused on these three axes that have been addressed throughout the preceding chapters of the thesis: deception detection, hyperpartisanship detection in political news, and immigrant stereotype identification in partisan interventions.

Not only because of the presence of deceptive language in political discourses, as we have referenced in Section 1.2, but more importantly, the difficulty of the deception detection task by itself, we have started our experiments with data annotated for this task. Experiments of Chapter 2 prove that the masking technique that we propose is effective at distinguishing the intention of deceiving from genuine narratives. Even when there are no universal cues for deception, our experiments show evidence that the masking technique has a good versatility (e.g. less domain dependency) because it can capture relevant patterns: in cross-domain scenarios, in domains where non-factual information is given, and in contexts where deceiving could have psychological implications to the deceiver (see Table 2.6). All of these serve as a solid base to use this technique for partisan and social biases detection, which we presented in Chapter 3 and Chapter 6, where we compare its results with transformers. Since we missed experiments with transformers for deception detection, we add in Section 7.2 some further experiments about this.

As we mentioned, we use the masking technique to detect partisan bias, in particular, hyperpartisanship in political news. The results described in Chapter 3 show that the masking technique can focus on the topic or the

style. However, the transformer-based models, which are resource-hungrier than the masking technique, are also more effective at capturing patterns where style- and topic-related information were kept in the texts. As it is possible to see in Figure 3.1, transformers' attentions score could help to visualize parts of the text that can be useful for explainable results. However, the masking technique can help to analyse in detail some specific aspects depending on the approach followed in selecting the terms to be masked. In addition, it requires less computational resources than transformers (Table 3.5 shows examples where the masks can indicate relevant features used by the masking technique). Section 7.3 describes additional experiments that we have done to evaluate how sensitive the results of the masking technique are to the variation of its hyperparameters k (number of words considered to be masked) and n (number of terms included in the sequence of n-grams).

The other type of bias that we consider essential to detect is concerning more complex social phenomena that are often embedded in partisan narratives, such as in social media or political debates. As we will describe in Section 7.4, a study of the content sponsored on Facebook by the main statewide political parties in Spain during the two General Elections held in 2019 shows that important topics are alluded to by some keywords that each party uses following different marketing strategies. Advertising strategies implemented by each party reflect the effort and economic resources that each one invests in order to highlight their political goals. Being interested in a more accurate study of the topics, instead of relevant keywords to detect partisan or social bias, we focused on those topics present in the political manifestos and analysed how they were addressed on Twitter in the context of the 10N Spanish Election.

As we described in Chapter 4, we found differences in the topics that each party considered important to talk about and differences in the sentiment and emotions that each one expressed. Surprisingly, we observed that the immigration topic was almost only mentioned by VOX (see Table 4.4b) and that either the expressed sentiment and emotions detected (see Table 4.1 and Table 4.2) suggested that the right-wing party used this topic to trigger specific reactions and gain popularity prior to the election. Therefore, in Chapter 5 we went deeper in analysing how the immigration topic has been talked about across different years in the Congress of Deputies, where the partisanship of each speaker is already known.

In Chapter 5, we described the work that we have done in order to: (i) obtain the speeches related to immigration; (ii) annotate the parts of texts where the speaker reflected a specific image of how he/she or his/her party perceives immigrants; and (iii) evaluate if traditional machine learning and state-of-the-art transformer models can effectively distinguish the presence of immigrant stereotypes in the annotated dataset. The proposed annotated dataset offers a challenge to the classifiers because the labeling follows a novel taxonomy (proposed in this thesis) that conceives stereotyping as a process

of framing (the selection of particular aspects of an issue that makes them salient in communicating a message). Our experiments proved that stereotypes could be detected automatically for the most part, both by traditional classifiers and by transformers. However, there were some disagreements regarding the manual annotation, which were correlated to the rhetorical strategy of each political party. In Section 7.5 we add further experiments in which we analyse in more detail the communication style of each party.

Taking into account the effectiveness of the models, we applied the masking technique over this dataset, as we described in Chapter 6, to identify explainable results and to improve the results of traditional classifiers. The masking technique outperformed traditional classifiers and obtained competitive results compared with transformers, which is very interesting because of its fewer requirements of computational resources. More surprisingly is the fact that both approaches are able to capture similar parts of texts detected as relevant for their predictions (see Figure 6.2).

7.2 Transformers for Deception Detection

Throughout this thesis, we have reported results of two diametrically opposite approaches: the masking technique and transformers. Both approaches have been compared in terms of their predictions, and we have also shown examples of *local explanations* i.e., examples where the level of explainability is for every single prediction instead of explaining the model's prediction process as a whole, Section 6.2.2. However, the results of the transformers were only shown for hyperpartisanship detection in political news and for stereotype identification. In this section, we show precisely local explanations of predicted deceptive texts, being possible to see how different they are visualized according to the attention scores that the model learned for each text.

The work that we have described in Chapter 2 focused on cross-domain deception detection. To propose a method that improves the cross-domain classification performance, we take advantage of some unlabeled instances of the target domain in the training process. In that way, we are able to see the vocabulary used in both domains and apply domain-specific terms filtering with the objective of removing/masking some noisy features. Since the experiments in which we have applied transformers so far have been set in in-domain scenarios, we do not find it fair to compare the predictions of the masking technique reported in Chapter 2 with transformers.

In the rest of this section, we describe the experiments that we have done with a similar cross-domain configuration and show some examples of local explanations. Note that the size of the datasets of controversial opinions and fake reviews is too small to apply deep learning; therefore, we use (as in Chapter 2) all the texts from one domain for training and all the texts from

another domain for testing.

7.2.1 Experimental Setup

For these experiments, we follow the same experimental setup described in Chapter 3 regarding the transformers-based models (the texts are also written in English). We approached the hyperparameter tuning also by grid search: *learning rate* $\in \{3e-03, 3e-04, 3e-05, 0.01\}$; the *batch size* $\in \{16, 32\}$; and for the optimizer the options were *adam* and *rmsprop*. We have selected a value for the *max_length* hyperparameter that covers completely the length of the opinions (Table 2.5). The transformer-based models that we use are BERT, M-BERT, and XLM-RoBERTa.

7.2.2 Results

Table 7.1 shows the results of the best predictions of each combination of domains. The source and the target domains were both from the same genre of corpora (reviews or controversial opinions). The table also indicates the hyperparameters used for achieving the corresponding F_1 score; in all the cases, the best score was achieved with a *batch size* of 16. Even if we used the same combinations of domains, these results should not be compared with those of Table 2.6, because, there, the classifiers needed information from the target, but the transformers, in these experiments, only used the information from the source domain. This difference is due to the current approach of the transformers, and it is not needed to select a common space between the two domains.

Table 7.1: F_1 score obtained with the transformers over the same data of Table 2.6.

Source	Target	F_1	Model	Optimizer
Hotel	Restaurant	0.84	BERT	adam
	Doctor	0.78	BERT	rmsprop
Restaurant	Hotel	0.77	BERT	adam
	Doctor	0.74	BERT	adam
Doctor	Restaurant	0.72	BERT	rmsprop
	Hotel	0.70	MBERT	adam
Abortion	Best Friend	0.70	BERT	adam
	Death Penalty	0.71	BERT	adam
Death Penalty	Abortion	0.75	BERT	rmsprop
	Best Friend	0.71	BERT	adam
Best Friend	Abortion	0.61	BERT	adam
	Death Penalty	0.60	BERT	adam

It is possible to note that in these experiments, like with the masking technique, the transformers achieved the best results for the combination of *Hotel* $\rightarrow$ *Restaurant*. The combination with the worst predictions was using the Best Friend corpus as the source domain. In general, the more effective model was BERT, with *adam* optimizer.

7.2.3 Deceptive Examples Visualized Using Attention Scores

Based on the previous results, we find it interesting to take a look at the true positive deceptive reviews on restaurants when the source domain is Hotel. In Table 2.7 we showed examples of some deceptive texts from Hotel and Restaurant domains, and we highlighted in yellow relevant char 4-grams for both. We depict in Figure 7.1 how they are visualized considering the attention scores learned by BERT.

on my next visit	is my next target visit
I will be back again	i will definitely be
but I was pleasantly	i was thirsty before
to anyone looking for	to anyone . try it and
I would recommend this	i highly recommend

Figure 7.1: Same relevant deceptive features highlighted in yellow in Table 2.7 (left), visualized considering the attention scores. The more intense the red, the greater the weight of attention (right).

Figure 7.1 shows on the right short text fragments of true positive deceptive opinions on restaurants. The terms are highlighted in red: the more intense the red, the greater the attention score. The figure also includes, on the left, part of Table 2.7 with fragments of deceptive restaurant reviews. We can see that the relevant char 4-grams (in yellow) are part of some terms that received high attention scores.

Figure 7.2 shows eight examples where it is possible to see that some terms have almost no score while others are very highlighted. It is the case of *I visited*, *this restaurant*, *a big fan of*, *experience*, and *love love love*. Surprisingly, the model learned that when the review has the reference to the restaurant (that is, in fact, redundant information), the expression has a very high score. The same happens when the deceiver emphasizes the *experience* that they have lived there. Moreover, thanks to the attention scores, it can be analysed several expressions that could be frequently used, perhaps in an exaggerated way, such as *a big fan of*. In the last example of the figure, we included the complete opinion in order to see the resultant attention scores of each term. It is important to comment that each term has different scores in each of its occurrences because it depends on the context in which it was used. For example, the deceiver repeated three times the word *love* consecutively, and we observe that each occurrence received a higher score than the precedent.

Based on the results presented in this thesis and complemented with this section, we can appreciate that the transformers provide some facilities for human experts. In particular, when the experts analyze words or expressions that may be relevant according to the context.

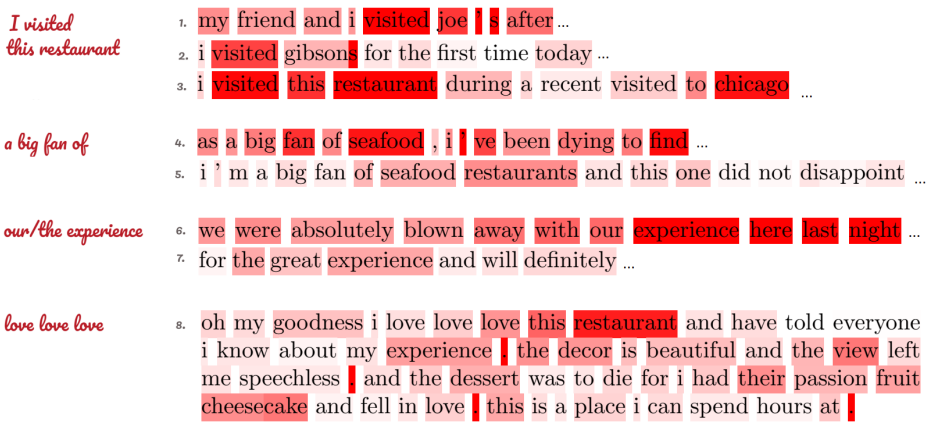


Figure 7.2: Text fragments of deceptive reviews. The last example is a whole opinion.

7.3 Robustness of the Masking Technique in the Hyperpartisan News Detection

In Chapter 3 we proposed the use of the masking technique to detect hyperpartisan news. We observed that this technique allows us to create a style-based model and also a topic-based model. Actually, these are examples of the flexibility that this technique offers. In this section, we describe additional experiments to measure the robustness of the two approaches, the one based on the style and the one based on the topic.

With the goals of: (i) understanding the robustness of the approaches to different parameter values; and (ii) determining if it is possible to overcome the $F_1 = 0.70$ from the baseline model, we vary the values of k and n (used in Chapter 3) and evaluate the macro F_1 using SVM.

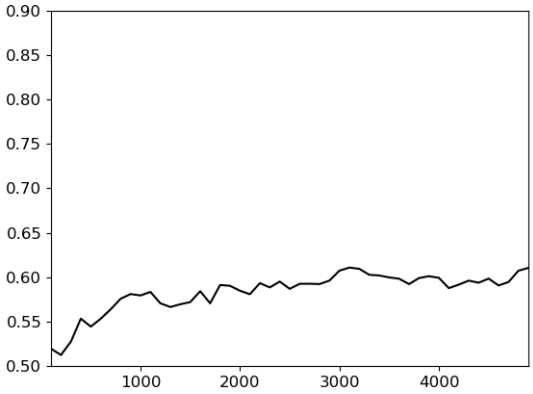
Figures 7.3 shows the results of the variation of $k \in \{100, 200, ..., 5000\}$. When $k > 5000$, we clearly can see that the topic-related model, in which the k most frequent terms are masked, is decreasing the performance. This could be explained by the fact that relevant topic-related terms start to be masked too. However, a different behaviour is seen in the style-related model, in which we tried to maintain only the style-related words without masking them. In this model, the higher is k, the better is the performance. This confirms that for the used dataset, taking into account only style-related information is not good, and also observing topic-related information benefits the classification. When k tends to the vocabulary size, the style-related model tends to behave like the baseline model, which we already saw in Table 3.3 that achieves the best results.

From this experiment, we conclude that: (i) the topic-related model is less sensitive than the style-related model when $k < 500$, i.e., the k most

frequent terms are style-related ones; and (*ii*) when we vary the value of k, both models achieve worse results than our baseline (based on the same text representation with the character n-grams features, but without masking any word).

On the other hand, the results of extracting character 5-grams are higher than extracting smaller n-grams, as can be seen in Figures 7.4. These results confirm that the performance of our approach overcomes the models proposed in [152] because of the length of the n-grams[1].

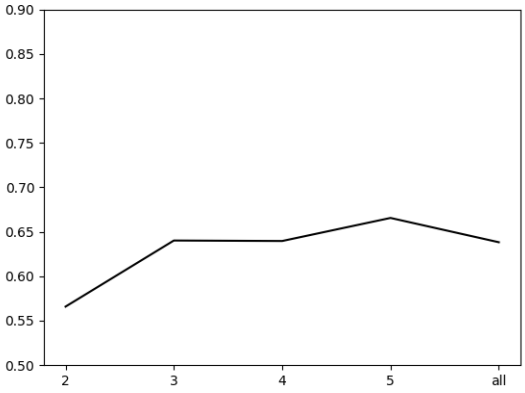
(a) Varying k values and masking the most frequent words: topic-based model.



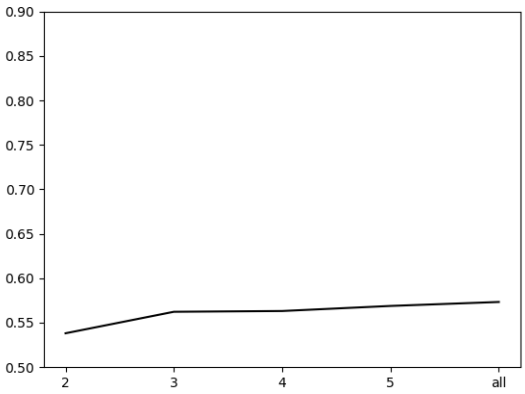
(b) Varying k values and maintaining without masking the most frequent words: style-based-model.

Figure 7.3: Macro F_1 results of the proposed masking technique. We set n=5 for comparing results of different values of k.

[1] In [152] the authors used $n \in [1, 3]$.



(a) Varying n values and masking the 500 most frequent words.



(b) Varying n values and maintaining without masking the 500 most frequent words.

Figure 7.4: Macro F_1 results of the proposed masking technique. We denote with *all* the representation that combines all the used n-grams.

7.4 Political Speech and Advertising

Partisan bias is often present in information spread in favor of specific political positions or when the candidates of electoral campaigns are pursuing popularity. In hyperpartisan news close to the U.S. presidential election of 2016, we have found discriminative patterns using a topic-based approach (Chapter 3), which suggests that each partisan position considers to highlight differently the topics that are important to the people. Relevant research conducted in [63] revealed that in Facebook ads, compared to television, there is significantly less attack of the opponent, less discussion, and they

increased their partisanship. The current section describes a study of the political contents of Facebook ads under a marketing and communication lens [143]. It focuses on the topics conveyed on the ads, the candidate presence, and the ideological issues.

We focus on the content of 14,684 Facebook ads sponsored by the five main statewide political parties in Spain (PSOE, PP, Unidas Podemos, Ciudadanos, and Vox) during the campaign for the April 28 and November 10 General Elections held in 2019. The same dataset has been used recently to study disinformation and communication strategies [27, 28]. As the Unidas Podemos coalition as such did not have Facebook advertiser account, we included in the corpus the ads promoted individually by the two parties of the coalition: Podemos and IU. We considered the textual message of an ad as the combination of two elements: the content written in text format and the eventual content written in the ad image. In this way, we distinguished 1,754 unique textual messages in the main corpus (see Table 7.2).

Table 7.2: Number of ads and number of different textual messages by party and by General Election.

	April 28th				**November 10th**				**Total 2019**	
	Ads	**Textual messages**	**%**		**Ads**	**Textual messages**	**%**		**Ads**	**Textual messages**
PSOE	336	146	43.45		285	244	85.61		621	390
PP	3,609	766	21.22		908	21	2.31		4,517	787
Ciudadanos	6,098	217	3.56		2,462	44	1.79		8,560	261
Podemos & IU	392	254	64.80		550	47	8.55		942	301
Vox	0	0			44	15	34.05		44	15
Total	10,435	1,383	13.25		4,249	371	8.73		14,684	1,754

7.4.1 Keyphrase Extraction

To analyse the topics in each campaign, we used the Keyphrase Digger (KD) algorithm proposed in [126]. KD is an automatic keyphrases detector that combines statistical procedures with linguistic information. It provides a list of keyphrases meant to capture the main important concepts discussed in a given text [205]. A keyphrase is usually a compound of one word, but it can also include two or more terms. The algorithm assesses the relevance of each keyphrase considering the whole text. In the case of the KD algorithm, keyphrases can be evaluated inside a set of documents. Furthermore, KD is ready to be used for English, Italian, and German texts but also provides an easy way to set up new languages. We made use of this functionality for analyzing the ad textual message corpus, being our research the first time KD has been implemented for Spanish texts, as far as we know.

After the tokenization and PoS tagging, KD extracts the keywords that match with one of the predefined patterns considered by an expert in the field. Patterns must be specified by the PoS tagging composition. For exam-

ple, to extract the bigram *banda terrorista* (terrorist gang), the PoS pattern *noun + adjective* must have been defined. This implies that all the bigrams whose PoS tags match this pattern will be extracted.

The algorithm computes a relevance score for every combination that matches the selected PoS patterns. Some of the parameters that the algorithm considers to compute the scores are the position of the first occurrence, the number of keyphrases to assess, and the function for performing the text vectorization. This function depends on the way the texts are assembled. In our case, we considered each ad's textual message as one document. Hence, we had two electoral campaigns and five political parties. However, as Vox did not sponsor Facebook Ads in the first campaign, we ultimately had nine sets of textual messages.

The PoS patterns selected for our analysis were: *noun, verb, adjective, proper noun, adjective+noun, proper noun+adjective, noun+verb,* and *proper noun+verb*. In the case of the bigram patterns, the algorithm considers equally the different order of both words. Finally, we set the algorithm to extract the 200 keyphrases more relevant for each set of Facebook Ads. When revising the results after running the KD algorithm for the first time, we missed some bigrams important for our research (e.g. *Unidas Podemos*). To solve these absents, we included the following PoS patterns in the algorithm: *verb+adjective* and *auxiliar verb+adjective*. After that, we run the KD algorithm.

7.4.2 Results

There were two clear political blocs: PSOE and Unidas Podemos as the left-wing options, and PP and Vox as the right-wing options. Ciudadanos positioned themselves as a center party so that they might support political policies from both sides of the ideological spectrum. In the following paragraphs, we will discuss our observations about the most promoted keyphrases, summarized in Figure 7.5. Note that in the figure, there is only one wordcloud for VOX due to there are no ads for the first campaign.

Podemos & IU: In April the top keyphrase is *Unidas Podemos*; in November, the dissemination was concentrated on *#quientienequedormir*, (#whohastosleep). The content associated with this hashtag consists of a video in which Podemos claims to prioritize the welfare of the people. The hashtag is also related to Pedro Sánchez's declaration in which he stated that he would distance himself from Podemos' leader.

Podemos & IU are the parties that have promoted more policy issues among the top 50 keyphrases: 15 in the first campaign and 9 in the second. They are related to **job** market –*empleo* (employment), *digno* (dignified), *estable* (stable), *salario* (salary), *precariedad* (insecurity), *desempleo* (unemployment), *contrato temporal* (temporal contract)–;

social issues –*prestación* (compensation), *mujeres* (women), *familia* (family), *violencias machistas* (sexist violences)–; and **economic** issues –*banca pública* (public banking), *rescate bancario* (bank rescue)–.

In November, we find very few repetitions, such as women, along with a new pool of keyphrases –*recortes* (cuts), *madres* (mothers), *abuelas*, (grandmothers), *desaceleración* (deceleration), *vivienda digna* (dignified household), *alquiler*, (rental) and *cambio climático* (climatic change)–. We find one term related to national identity ranking high (14th position) as a response to Vox's leader: *patriotismo* (patriotism).

Something to note related to these parties is the difference between the two campaigns. In April, Pablo Iglesias (Unidas Podemos's leader) is located in the 23rd position of the most relevant keywords, whereas in November, he holds the 4th position. Furthermore, the second candidate Irene Montero is absent from the April list, but she holds the 16th position in November. There has been a shift in the strategy to promote the candidates more intensely for the second ballot.

PSOE: PSOE prioritized electoral slogans and calls to vote and gave more relevance to promote the party over programmatic issues. In the first campaign, PSOE launched two slogans: *la España que quieres* (the Spain that you want), and *haz que pase* (make it happen). In the November campaign, the electoral slogan was *ahora sí* (now yes), as it evoked the second round of the general elections. There is a strong presence of the term Spain in both campaigns: among the top 5.

In the first campaign, the most viewed keyphrase associated to policy was *política migratoria* (immigration policy). After this issue, and going beyond the top 20 keyphrases, we find terms promoted in April regarding **social** issues –*igualdad* (equality), *mujeres* (women), *reto demográfico* (demographic challenge)–; **education** –*educación* (education), *educación gratuita* (free schooling)–; and **job** issues –*salario mínimo* (minimum salary), *microcréditos* (microcredits)–.

In November, the policy issues present in the ads were related to the **job** market –*empleo* (job), *estabilidad* (stability), *salario mínimo* (minimum salary)–; and **social** issues –*igualdad* (equality), *jóvenes* (youth), *pensiones* (pensions), *mejores oportunidades* (best opportunities)–. In both campaigns these keyphrases were much less promoted than calls to vote or the electoral slogans.

Ciudadanos: Ciudadanos' ads enhanced the figures of their candidates in both campaigns. In April, the references to other opponents were more viewed than references to policy issues. It is the case of keyphrases associated to PSOE, such as *Sánchez*, *gobierno* (government), *PSOE* and *PP*.

In April, the policy keyphrases were related mainly to **social** issues –*familias* (families), *hijos* (sons-and-daughters)–; **territorial** issues –*nacionalistas* (nationalists), *separatistas* (separatists), *constitución* (constitution)–; and **job market** –*trabajo* (job), *autónomos* (self-employees), *emprendedores* (entrepreneurs)–.

In November, policy issues had higher visibility. The most viewed keyphrase associated to policy was *impuestos* (taxes). Among the top 50 keyphrases, we find: *autónomos* (self-employees), *seguridad social* (social security), *jubilación* (retirement), *pensión* (pension), *empresas*; along with other topics more **social** such as *hijos* (sons-and-daughters), *madre* (mother), *conciliación* (work-family-balance), and *familia* (family). Ciudadanos bet clearly for the national identity: Spain was the 2nd most viewed keyphrase in April and the 4th in November.

PP: This party also prioritized their slogan in both campaigns. In April, their claim was *valor seguro* (safe value), and in November it was *por todo lo que nos une* (for everything that unites us). The most viewed keyphrases corresponded to these mottos: *#valorseguro* (#safevalue), and *une* (it unites).

The wide majority of the top 20 keyphrases most viewed in April are related to the slogan, calling to vote, and the party, with the except of three keyphrases: *empleo* (employment), *Pedro Sánchez*, and *paro* (unemployment). The policy issues developed by PP in April were mainly related to **job** keyphrases –*autónomos* (self-employees), *emprendedores* (entrepreneurs), in addition to employment and unemployment–; **education** –*educación* (education), *modelo educativo* (educative model), *red pública* (public network)–; **social** –*mujeres* (women)– and **economic** –*pensiones* (pensions), *impuestos* (taxes)–.

In November, we find keyphrases regarded to **education** –*universidad* (university), *capacitación* (training)–; **job** –*trabajo* (job), *empleado* (employee), *autónomo* (self-employee)–; **social** –*vivienda* (household), *hipoteca* (mortgage), *alquiler* (rental)–; and *economic* –*pensiones* (pensions), *impuesto* (tax)–. In this latter campaign, all these issues had very little impact overall in terms of impressions. The whole keyphrase list contains all the Spanish provinces, suggesting they implemented a geographic targeting. Probably the most remarkable finding in the April campaign is the presence of the PSOE references over their own candidate: *Pedro Sánchez*, *Zapatero* (the previous Spanish president who was from PSOE), and *Moncloa* (term that refers to the Government as it is the name of the President official residence).

VOX: The keyphrase with the highest diffusion corresponds to *España* (Spain), more than twice the number of impressions of the name of the party.

We only find one keyphrase slightly related to political issues among the top 20: *programa económico* (economic program). The references to policy issues were scant in the whole list of results, whereas three party's candidates accumulated a great deal of visibility (*Santiago Abascal*, *Iván Espinosa*, and *Javier Ortega*). Figure 7.5 also shows two keyphrases containing contemptuous allusions to left-wing ideology: *dictadura progre* (progressive dictatorship), and *progres explicando* (progressive people explaining).

Figure 7.5: Most relevant keyphrases used by the parties in the 2019 campaigns.

We conclude this section by pointing up some results. Except for VOX,

the rest of the parties had two topics in common: **job market** and **women**. The use of keyphrases about the job market reflected the different ideological framing of each party. Some terms were common to the left-wing parties (minimum salary, stability), whereas PP and Ciudadanos employed other keyphrases with right-leaning connotations (self-employees, entrepreneurs). The issue of sexist violence was present in Podemos's corpus as evidence of issue ownership [10]. In Chapter 4, in which we described our study on the sentiments and emotions of the parties towards the more relevant topics of the second campaign, it is possible to see that Podemos also was the party more focused on feminism in general (see Table 4.4b). With respect to immigration we observed that only PSOE addressed it in the Facebook Ads. In Table 4.4b we reported that VOX was more concerned with that topic on Twitter.

Concerning with marketing strategy, it was a trend of the prevailing candidate or campaign issues over policy issues. PSOE and PP, promoted calls to vote, the party, and electoral slogans. In the case of PP, we found more presence of the PSOE's candidate rather than theirs. The candidates of Ciudadanos and Vox were more prominent in the Facebook sponsored content than policy issues.

Finally, it is evident that each party uses different communication strategies to gain its goals and focuses on important issues/topics that others do not. In the next section, we emphasise the role that partisanship could have in the rhetorical strategy used in parliamentary debates about immigration.

In summary, our findings suggest that the main general function of the campaigns has been mobilizing users, the candidates have been more salient than political issues, and the ad contents reflected the ideological positions of the main parties. These conclusions contribute to knowing better the possibilities of Facebook advertising, but they also help to enrich the picture of partisanship in the Spanish elections.

7.5 Analysis of Immigrant Stereotypes as a Rhetorical Strategy

As we have already showed in this thesis, politicians can fulfill their own political goals by attempting to shape people's thinking. Throughout the previous chapters, we have presented different works that underlyingly show that political texts (e.g. news, speeches) offer biased information that directs the reader's attention to think according to the interests of the author or speaker. When we described our study of the communication strategies of five Spanish political parties in Chapter 4, we pointed out how emotions expressed on each topic differed according to partisanship; and that the far-right leaders were almost the only ones addressing immigration, unlike leaders of the other parties. We also proved that the masking technique

and the BERT-based model effectively detected partisan and social biases, particularly hyperpartisanship (Chapter 3) and identified immigrant stereotypes in partisan interventions (Chapters 5 and 6), respectively. In addition, in the previous section, we highlighted the effort that each party makes at sponsoring Facebook ads to mobilize users and gain votes, and also how the ads' content reflects their ideologies.

The StereoImmigrant dataset that we have used in Chapter 5 to detect immigrant stereotypes has been created extracting interventions of politicians from different ideologies. The speakers expressed from what angle their partisan vision is capable of focusing on one of the frames in which immigrants can be stereotyped.

In this section, we complement our study on detecting partisan and social biases by analyzing: (i) the correspondence among ideological positions and the use of different dimensions of immigrant stereotypes; (ii) and if politicians use language differently when they refer to different dimensions of the stereotypical image of immigrants.

7.5.1 Annotation at Speech Level

In Section 5.3.2 we explained the annotation process of the StereoImmigrant dataset. In that study, we were interested in having reliable examples of phrases alluding to one or other frames regarding immigration. Therefore, we split the speeches to annotate those sentences about immigration mainly under one principal frame or category. However, to better analyse the rhetorical strategy of speakers and observe its relation to social bias, we will consider the complete speeches. For the present analysis, we count 475 speeches pronounced by 143 politicians (63% men).

We annotated the speeches considering three dimensions of stereotyping (i.e., Victim, Threat, and Ambivalent):

- **Victim:** Discourses where there are only sentences annotated with the victim dimension of the stereotype
- **Threat:** Discourses where there are only sentences annotated with the threat dimension of the stereotype.
- **Ambivalent:** Discourses that contain at least one sentence annotated with the victim dimension and at least one annotated with the threat dimension.

7.5.2 Construction of Indices

For this exploratory study, an expert in social psychology elaborated the indices listed below.[2] The indices are based on the psycho-linguistic categories

[2]These indices are part of a next publication that the author is preparing with an interdisciplinary team led by the social psychologist Dra. Berta Chulvi Ferriols.

of LIWC2007 [146] and the POS tags (i.e., grammatical categories) recognized by SpaCy [186] using the *es_core_news_sm* package. We tokenized the speeches and computed a score for each psycho-linguistic and grammatical category. For example, for the psycho-linguistic category c_{LIWC}, the score is based on the number of words (and all their occurrences) belonging to that category. Later on, these scores are normalized by the amount of tokens from the speech but only counting those found in LIWC. This normalization relies on the fact that it is not the same that one speech's vocabulary has a high intersection with LIWC than another with less intersection. The scores of the POS tags are computed analogously, but since each token is assigned to a POS tag, the speech length is used to normalize scores for each POS tag. The psycho-linguistic indices are:

- **Victim vs. Threat stereotyping index:** Denoted as SI, it means the number of sentences that present immigrants as victims (V) less the sentences that present immigrants as a threat (T) in each parliamentary speech (i.e., $SI = V - T$). Positive scores indicate that the immigrant group is presented more as victims, and negative scores indicate that the immigrant group is presented more as a threat.

- **Political parties:** To resume the number of parties at the Spanish Parliament (15 in total), an ordinal variable has been coded that ranges from 1 (right) to 4 (left). Each speaker has been classified individually, attending the original position of the party and the agreement of two judges when the party presents some doubts. Right (117 speeches) is the position held by speakers from Partido Popular (PP), Moderate Right (106 speeches) are politicians from parties such as Partido Nacionalista Vasco (PNV) and other territorial parties with a liberal program but anchored in nationalist goals and, sometimes, giving support to the Moderate Left. Moderate left (150 speeches) is represented by Partido Socialista Obrero Español (PSOE). Left (102 speeches) are national parties – as Izquierda Unida (IU) and Podemos– and territorial parties (Esquerra Republicana de Catalunya o Compromís) that clearly present themselves at the left of PSOE. Due to some speeches in the corpus (30% of the speeches) have been pronounced by the same speaker, the code of each singular speaker is introduced as a covariant variable in the statistical analysis in order to avoid a confounding effect. The number of words and the number of speeches is also introduced as a covariant to control a possible effect of the amount of text because some groups have more words and more speeches. The number of words emerges as significant covariates in all the analyses but does not alter the conclusions of the analysis of variance that we present in the results.

- **In-group vs. Out-group index:** The GI index means the scores of

first-person plural pronouns and first-person plural verbs (W) less the scores of third-person plural pronouns and third-person plural verbs (T) in the speech. In this index $GI = W - T$, positive scores indicate that the speech talks more about the in-group, and negative scores indicate that the speech talks more about the out-group.

- **Analytic thinking index:** Applying LIWC2007, we compute the analytic thinking index elaborated in [145] to measure the use of a more analytic language or a more intuitive one. Initially labeled as the categorical-dynamic index (CDI), this pole of the psychological dimension was later renamed Analytic Thinking in LIWC2015. The CDI is computed as: $article + preposition - personal\ pronoun - impersonal\ pronoun - auxiliary\ verb - conjunction - adverb - negation$. Positive values in this index express more analytic thinking, and negative values more intuitive thinking.

- **Categorical vs. Narrative index:** Applying the POS tagging of SpaCy, we recognize the grammatical categories present in each speech and calculate the punctuation for each category as the percentage of this category over the total of words in each speech. Inspired by the previous research of Nisbett et al. (2001), we compute a categorical versus narrative index (CNI) as a simple algorithm: nouns + adjectives + prepositions -verbs-adverbs-personal pronouns. Positive values in this index express more categorical thinking, and negative values more narrative thinking.

- **Positive vs Negative emotion index**. Applying LWIC2007, we compute a new emotion index (EPN) with the scores of positive emotions (EP) less the scores of negative emotions (EN) in each speech. In this index (EPN=EP-EN), positive scores indicate that the speech is mainly expressing positive emotions and negative scores indicate the opposite.

- **Emotional language index.** Applying LWIC2007, we compute an emotional language index (ET=EP+EN) by adding the scores of positive emotions (EP) and negative emotions (EN) to measure how much the speech appeals to emotions.

7.5.3 Ideology and Immigrant Stereotypes

Table 7.3 shows the number of speeches of each dimension of stereotyping and the ideology of the speakers. There is a significant relation between the ideology of the politicians and the use of stereotypes about immigrants (Pearsonχ2= 51.399, df = 9, p $<$ 0.001). The residual analysis (included in Table 7.3) shows that **right-wing speakers see immigrants as threat**,

and **left-wing present significantly more the immigrants as victims**. No significant pattern is observed in moderate right and moderate left speakers.

Table 7.3: Number of speeches (Obs) in each dimension of stereotyping by the ideology of the speaker.

			Ambivalent	Threat	Economical Resource	Victim	Total
Ideology	**Right**	Obs	37	51	2	27	117
		Adj. Res	-0.2	4.9	-2.9	-2.7	
	Moderate Right	Obs	34	32	8	32	106
		Adj. Res	0	1.0	-0.2	-0.8	
	Moderate Left	Obs	56	31	16	47	150
		Adj. Res	1.6	-1.9	1.5	-0.70	
	Left	Obs	26	11	12	53	102
		Adj. Res	-1.6	-4.0	1.6	4.5	
Total (Obs)			153	125	38	159	475

Ideology, Stereotypes, and Language Style

We have obtained a significant Spearman[3] correlation (r=.120; p<.001) between **Victim vs. Threat index** and **In-group vs. Out-group index**: more the speeches present immigrants as victims more they use the in-group linguistic markers (first-person plural pronouns and first-person plural verbs), and the opposite: more immigrants are presented as a "threat" more the speech uses the out-group linguistic markers (third-person plural pronouns and third-person plural verbs).

This is the only significant correlation between the index of stereotyping and the indices that characterise the language. Moreover, we also find that the **In-group vs. Out-group index** presents a significant correlation with:

- The **Analytic Thinking index** ($r = .229; p < .001$), suggests that the use of in-group linguistic markers is related to a more analytic thinking style.
- The **Categorical vs. Narrative index** ($r = .227; p < .001$), suggesting that the use of in-group linguistic markers is related to a more categorical language.
- The **Positive vs. Negative emotion index** ($r = .180; p < .001$) suggests that the use of the in-group linguistic markers is related to expressing more positive emotions.

Out-group linguistic markers are related to a more intuitive thinking, a more narrative language style, and greater use of negative emotions.

[3]The Spearman correlation is used because of the presence of outliers in the data.

In addition, we performed a two factor ANOVA (dimension of stereotyping in the discourse vs ideology of the speaker) over the indices. Not significant interaction between the two variables is observed, but we observe a main effect of the ideology over the **In-group vs Out-group** ($F(9, 459) = 26, 124; p < .001$), the **Analytic Thinking** ($F(9, 459) = 4, 348; p < .005$), and the **Categorical vs Narrative** ($F(9, 459) = 5, 319; p < .001$). Therefore, this suggests that the rhetorical strategy varies depending on the ideology. We performed the t-test ($p < .05$) to compare groups, and we found that in the **In-group vs. Out-group index** the right differs significantly from the other three groups using more the out-group linguistics markers. The moderate right and the left-wing differ from the other two groups (right-wing and moderate left) by using more in-group linguistic markers. In the **Analytic Thinking index** the two moderate wings use more analytic thinking and differ significantly from the right and the left wings that do not differentiate between them. In the **Categorical vs. Narrative index** the only group that differs from the other three is the moderate right which is the one that uses more categorical language. We do not find any effect of the ideology on the two indices that measure the use of emotions.

We conclude this section by highlighting the relationship between ideologies and the dimensions of immigrant stereotypes. We also have seen that the more politicians present immigrants as victims, the more they use the in-group linguistic markers, and the more immigrants are presented as a threat, the more the politicians use the out-group linguistic markers. Therefore, ideology has an effect on how politicians perceive reality and how they communicate. This supports the idea that immigrant stereotypes are used in political conflicts as a rhetorical strategy.

7.6 Ethical Discussion

A NLP tool able to detect, identify, and predict if public figures are using deceptive language to manipulate people, is without a doubt a contribution to a world full of hyperpartisanship, prejudices, and toxic language in all its forms. However, researchers should think about the implications of their results and, at least, be aware of to what extent the intellectual challenge that motivated them to address conflictive issues at the same time allows to create a double-edged sword.

We have proposed some models that allow us to know from the input texts how biased the speaker is, whether the intention is to deceive and if stereotypes are employed in her/his speech. Because the accuracy levels of the models are relatively low, great care must be taken in using their results. Even though these scientific advances can be of great help, it is according to the individual's own choice, to count on them. Under no circumstances from our viewpoint, without prior consent, should these technological advances be

used to invade people's **privacy**.

Another ethical concern is about the use of transformers. We have presented the results of models based on word embeddings with accurate results, at least similar to or with better predictions than classical machine learning models. However, no matter the use of attention mechanisms, the transparency of these representations is still a big issue for which human experts are yet needed and not exempt from bias and misunderstanding. Therefore, the output of these models should only support human analysis, and not used directly for making decisions. In addition, it has to be considered that these models have an enormous impact on energy and cost that the training process requires [196].

Chapter 8

Conclusions and Future Work

8.1 Contributions

Partisan and social biases have lasting effects on people's polarization and lead us to even more harmful behaviour and the use of toxic language with undesirable outcomes. The work presented in this thesis addressed the detection of these biases that are present, for example, in political news and parliamentary debates. The automatic detection of these biases is a challenging task because, in many scenarios, they are not always false information but sometimes can contain prejudices, exaggerations, or omissions of specifics aspects of an issue. In contrast, other elements are visible to make salient what the politicians and also partisan journalists consider essential to achieve their goals.

The results we obtained in the experiments described throughout this thesis allow us to answer the research questions that we introduced in Section 1.5:

- **RQ1:** *Can **deceptive language** be detected employing the masking technique taking into account both content and style in cross-domain scenarios?*

  Our experiments in Chapter 2 showed that the masking technique could be effectively applied to transfer what the model learns from the source domain to distinguish deceptive from the truthful language in the target domain. We demonstrated that in deception detection, it is necessary to consider both content- and style-related words, i.e., what it is said and how it is said. We also evaluated this technique in domains where non-factual information is given, such as controversial opinions, which may be topics that politicians talk about, especially in electoral campaigns. Furthermore, the datasets used to evaluate the masking technique in deception detection, contain deceptive opinions with exaggerations (e.g. *the dessert was to die for*) that can be part

of partisan bias in political news (e.g. *Hillary is without a doubt, the worst and most despicable liar*).

- **RQ2:** *Can **hyperpartisanship in political news** be addressed from a deception detection perspective?*

  We demonstrated in Chapter 3 that the masking technique can be used effectively to detect hyperpartisanship with a topic-based approach (i.e., when the style-related words are masked) and that some relevant features are captured with a style-based approach (i.e., when the content-related words are masked). This versatility is due to the selection of the terms to be masked rely on the interest of the research, and that these terms being masked (and not removed) allows the classifiers to capture patterns in a modified text (e.g. ** imagination * assume * committing * * * fraud*) that nevertheless maintains a similar structure of the original text (e.g. *much imagination to assume he's committing some kind of fraud*), but without "noisy" terms. In this way, a classifier trained with n-grams can capture patterns that are combinations of specific content-related terms and any style-related word that was represented with the mask (e.g. *). With all this, we can conclude that partisan bias can be addressed with the masking technique; however, to obtain better predictions, the approach should consider both style and content for the masking, which we concluded after observing the results of our baseline model (where none term was masked).

- **RQ3:** *How can be approached the detection of social biases like **stereotypes** against immigrants in political speeches considering the manipulative strategies of this kind of narrative?*

  Immigrant stereotypes are social biases with a very complex theory behind them that makes it difficult to be addressed effectively with approaches that assume that two opposite social categories can be represented, as can be found in the literature in works on gender or racial bias. In Chapter 5, we addressed the detection of immigrant stereotypes when they are expressed in political debates, in particular, from different partisan ideologies. We have firstly proposed a fine-grained taxonomy that covers six main scenarios in which immigrants can be framed. We collected and filtered speeches related to immigration and proceeded to label them at the level of sentences or phrases reflecting (and perpetuating) any of the stereotypes. We demonstrated in Chapter 5 that immigrant stereotypes could be effectively identified, achieving above 0.83 of accuracy, showing that either state-of-the-art transformers or machine learning classifiers can help to distinguish how politicians speak about immigrants. Moreover, we proved that with the masking technique proposed in this thesis and evaluated in the detection of deceptive language and partisan bias, we also overcome the

machine learning classifiers (when they are used with the texts without applying any masking). At the same time, the masking technique achieves a similar F_1 to the transformers, which is very interesting considering the resource-hungry that the latter is. Furthermore, in Section 7.5 we described an additional study in which we found that the ideology of the politicians, and their partisan bias, affect how they perceive immigrants. In other words, the immigrant stereotypes have a strong relation with partisanship, and it makes sense to address the detection of partisan and social biases with the same approaches.

- **RQ4:** *Can the masking and transformer-based models* ***help human experts to further analyse*** *the above problems?*

  In Chapter 2, we showed how the masking technique could be used to see the relevant features that helped the model to predict deceptive language correctly; and we compared in Section 7.2.3 how some of the same examples were also relevant for the transformer. In Section 7.2.3 we added more examples where the attention scores serve to visualize some words or expressions (with high scores) that human experts can interpret. In Chapter 3 and Chapter 6, we showed how the masking technique and transformers help to compare, in the same sentences, those parts that each model found important in both partisanship detection (in hyperpartisan news) and social bias detection (immigrant stereotypes), respectively. We could observe, especially in the identification of immigrant stereotypes, that even when the two approaches have very different ways to predict, they agreed on some relevant expressions to support human experts' analysis. However, we found that the transformers can offer richer information in terms of explainability for two main reasons: (i) the attention scores are not binary, like the choice to mask or not a term in the masking technique, which can help to select and analyse the most relevant words or phrases (as we showed with the intensity of the color); and (ii) as we discussed in Section 6.6.3, the attention scores mechanism offers a way to find how non-discriminative words can be in fact used in both classes (e.g. Stereotype vs. Non-stereotype) but in different ways, probably depending on the context in which they appear. Despite this "richer" information from the transformers, their abilities to give explainable predictions are under discussion.

To sum up, we consider that the research questions have been successfully answered, besides some limitations that should be taken into account, such as the size of the annotated datasets, the undesirable but always present algorithm and annotation biases [9], and the complexity of social phenomena that make it difficult to extend this work to other social biases (e.g. stereotypes against LGBTIQ+ community, which involves more than two minority

groups) and force the construction of more taxonomies.

8.2 Future Work

In this thesis, we have addressed the partisan and social biases detection, particularly hyperpartisanship in the news, and the identification of immigrant stereotypes in partisan interventions. Although immigrants are currently one of the most affected social minorities, they are not the only ones. Recently, there have been homophobic political interventions from Vox, a right-wing political party, followed by an increased violence against the LGBTIQ+ community in Spain. We find the detection of stereotypes against this minority as one of the main directions for future work.

Considering that our StereoImmigrant dataset is annotated in terms of Victims and Threat supra-categories, one of the first attempts could apply transfer learning from one domain (e.g. immigration) to the other (e.g. LGBTIQ+). In addition, the positive and negative attitudes conceived in the proposed taxonomy (see Figure 5.3) can be considered in the detection of insults [48] and hate speech [8]. Furthermore, the dataset we used does not contain examples of dehumanization (probably because the politicians take care of the images of themselves in parliamentary speeches). Therefore, the future work also should include using our taxonomy to annotate examples of dehumanization concerning immigration or to adapt the current taxonomy (with the help of experts in social psychology) to the stereotypes about the LGBTIQ+ community. The authors of [123] analyzed dehumanizing language concerning homosexual people, and we think that the stereotypes about the two social categories, immigrants and LGBTIQ+ people, can have some dimensions in common. Moreover, we also find it interesting to extend this work to the stereotyping of Roma people [208], and to study the existence or not of any correlation with political ideologies.

Concerning XAI examples discussed in this thesis, a future direction is to keep working with attention scores learned by the transformers. For example, to explore how to use them to analyse *implicit bias* [114]. As we have shown in Section 6.6.3, some words can be seen as noisy at the first look at their attention scores. However, if we compute the attention scores of their relations with discriminatory terms, we can use them to obtain more interpretations of human analysis. In this sense, we find it interesting to combine the attention mechanism with the words considered in the masking technique to propose strategies to mitigate the bias when machine learning models are trained in datasets with stereotypes and, therefore, to reduce the *data bias* [9].

8.3 Research Publications

1. **Sánchez-Junquera J.**, Villaseñor-Pineda L., Montes-y-Gómez M., Rosso P., Stamatatos E. (2020) Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, vol. 135, pp.122-130 (**Impact Factor: 3.756 Q1**)

2. **Sánchez-Junquera J.**, Rosso P., Montes M., Ponzetto S. (2021) Masking and Transformer-based Models for Hyperpartisanship Detection in News. Proc. Int. Conf. on Recent Advances in Natural Language Processing, RANLP-2021, Bulgaria, September 1-4, pp. 1244-1251 (**CORE B**)

3. **Sánchez-Junquera J.**, Ponzetto S., Rosso P. (2020) A Twitter Political Corpus of the 2019 10N Spanish Election. Proc. 23rd Int. Conf. on Text, Speech and Dialogue, TSD-2020, Springer-Verlag, LNAI(12284), pp. 41-49.

4. **Sánchez-Junquera J.**, Chulvi B., Rosso P., Ponzetto S. (2021) How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Applied Science*, 11(8), 3610. (**Impact Factor: 2.679 Q2**)

5. **Sánchez-Junquera J.**, Rosso P., Montes-y-Gómez M., Chulvi B. (2021) Masking and BERT-based Models for Stereotype Identification. *In: Procesamiento del Lenguaje Natural (SEPLN)*, num. 67, pp. 83-94. **Best paper award.**

6. Rangel. F., Rosso P., Charfi A., Zaghouani W., Ghanem B., **Sánchez-Junquera J.** (2019) Overview of the Track on Author Profiling and Deception Detection in Arabic. In: Notebook Papers of FIRE 2019, FIRE-2019, Kolkata, India, December 12-15, CEUR Workshop Proceedings. CEUR-WS.org, vol. 2517, pp. 70-83

7. **Sánchez-Junquera J.** (2021) On the Detection of Political and Social Bias. Proc. of the Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T). vol. 3030, pp. 41-49.

8. Baviera Puig T., **Sánchez-Junquera J.**, Rosso P. (2022) Political Advertising on Social Media: Issues Sponsored on Facebook ads during the 2019 General Elections in Spain. Communication & Society. Vol. 35, págs. 39-49 (**Q3 in Communication**).

Bibliography

[1] Gavin Abercrombie, Federico Nanni, Riza Batista-Navarro, and Simone Paolo Ponzetto. Policy preference detection in parliamentary debate motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Saifuddin Ahmed, Vivian Hsueh Hua Chen, and Arul Indrasen Chib. Xenophobia in the time of a pandemic: Social media use, stereotypes, and prejudice against immigrants during the COVID-19 Crisis. *International Journal of Public Opinion Research*, 04 2021. edab014.

[3] Fareed Hameed Al-Hindawi and Nesaem Mehdi Al-Aadili. The pragmatics of deception in american presidential electoral speeches. *International Journal of English Linguistics*, 7(5):207, July 2017.

[4] Shivaji Alaparthi and Manit Mishra. Bert: a sentiment analysis odyssey. *Journal of Marketing Analytics*, pages 1–9, 2021.

[5] Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. The rise of partisanship and Super-Cooperators in the U.S. house of representatives. *PLoS ONE*, 10(4):e0123507, 2015.

[6] Talita Anthonio. Robust document representations for hyperpartisan and fake news detection. Master's thesis, University of the Basque Country UPV/EHU, 2019.

[7] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham, 2018. Springer International Publishing.

[8] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. A probabilis-

tic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, 173:114762, 2021.

[9] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

[10] Kevin K. Banda. The dynamics of campaign issue agendas. *State Politics & Policy Quarterly*, 13(4):446–470, September 2013.

[11] E. Barkan. *The guilt of nations: Restitution and negotiating historical injustices*. New York: Norton, 2000.

[12] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–31, 2021.

[13] Dan Barsever, Sameer Singh, and Emre Neftci. Building a better lie detector with bert: The difference between truth and lies. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020.

[14] Larry M Bartels. Beyond the running tally: Partisan bias in political perceptions. *Political behavior*, 24(2):117–150, 2002.

[15] Gregory Bateson. Ecology of mind. *Psychyatric Research Report*, 2, 1955.

[16] Hilary B Bergsieker, Lisa M Leslie, Vanessa S Constantine, and Susan T Fiske. Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, 102(6):1214–1238, 2012.

[17] Camiel J Beukeboom, J Forgas, O Vincze, and J Laszlo. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, 31:313–330, 2014.

[18] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert. *Cognitive Computation*, pages 1–11, 2021.

[19] Francesco Bodria, A. Panisson, A. Perotti, and Simone Piaggesi. Explainability methods for natural language processing: Applications to sentiment analysis. In *Symposium on Advanced Database Systems*, 2020.

[20] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to

homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[21] G. Breakwell and D.V. Canter. *Empirical approaches to social representations*. Oxford Science Publications. Clarendom Press, 1993.

[22] Rupert Brown. *Prejudice. Its Social Psychology*. Wiley-Blackwell, 2010.

[23] Jerome Bruner. *Acts of meaning*. Harvard University Press, 1990.

[24] John G Bullock, Alan S Gerber, Seth J Hill, and Gregory A Huber. Partisan bias in factual beliefs about politics. Technical report, National Bureau of Economic Research, 2013.

[25] Leticia C. Cagnina and Paolo Rosso. Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(Suppl. 2):151–174, 2017.

[26] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11), 2020.

[27] Dafne Calvo, Lorena Cano-Orón, and Tomás Baviera. Global spaces for local politics: An exploratory analysis of facebook ads in spanish election campaigns. *Social Sciences*, 10(7), 2021.

[28] Lorena Cano-Orón, Dafne Calvo, Guillermo López García, and Tomás Baviera. Disinformation in facebook ads in the 2019 spanish general election campaigns. *Media and Communication*, 9(1):217–228, March 2021.

[29] Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

[30] M Casavantes, R López, and LC González. Uach at mex-a3t 2020: Detecting aggressive tweets by incorporating author and message context. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*, 2020.

[31] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation

data. In *Practical ML for Developing Countries (PML4DC at ICLR 2020)*, 2020.

[32] CES. *La inmigración y el mercado de trabajo en España.* Colección Informes. Madrid: Consejo Económico y Social, 2004.

[33] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[34] Oana Cocarascu and Francesca Toni. Detecting deceptive reviews using argumentation. *Proceedings of the 1st International Workshop on AI for Privacy and Security - PrAISe '16*, pages 1–8, 2016.

[35] Danilo Croce, Daniele Rossini, and Roberto Basili. Auditing deep learning processes through kernel-based explanatory models. 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 4037–4046, Hong Kong, China, November 2019. Association for Computational Linguistics.

[36] André Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso. Team fernando-pessa at semeval-2019 task 4: Back to basics in hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 999–1003, 2019.

[37] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2020.

[38] Amy Cuddy, Susan Fiske, and Peter Glick. The bias map: Behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92:631–48, 05 2007.

[39] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics.

[40] Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu. Explainability for natural language processing.

In *Proceedings of the 27th ACM Special Interest Group on Knowledge Discovery and Data Mining Conference on Knowledge Discovery & Data Mining*, KDD '21, page 4033–4034, New York, NY, USA, 2021. Association for Computing Machinery.

[41] Lewis Davis and Sumit S Deole. Immigration and the rise of far-right parties in europe. *ifo DICE Report*, 15(4):10–15, 2017.

[42] James Dennison and Andrew Geddes. A rising tide? the salience of immigration and the rise of anti-immigration political parties in western europe. *The political quarterly*, 90(1):107–116, 2019.

[43] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological bulletin*, 129(1):74–118, 2003.

[44] Caroline Desombre, Mickaël Jury, Céline Bagès, and Célénie Brasselet. The distinct effect of multiple sources of stereotype threat. *The Journal of social psychology*, 159:1–14, 11 2018.

[45] Renard Muriel et al. Desombre Caroline, Jury Mickael. Validation factorielle d'une mesure des menaces du stéréotype en langue française. *L'Année psychologique*, 120(4):251–269, 2020.

[46] Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7659–7666, 2020.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, June 2-7, 2019, Volume 1*, pages 4171–4186. Association for Computational Linguistics, 2019.

[48] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3), sep 2012.

[49] John F Dovidio, Miles Hewstone, Peter Glick, and Victoria M Esses. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. *The SAGE handbook of prejudice, stereotyping and discrimination*, 80:3–28, 2010.

[50] Vlad Cristian Dumitru and Traian Rebedea. Fake and hyper-partisan news identification. In *International Conference on Human-Computer Interaction (RoCHI)*, pages 60–67, 2019.

[51] Alice H Eagly and Shelly Chaiken. *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers, 1993.

[52] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[53] Robert M Entman. Framing: Towards clarification of a fractured paradigm. *McQuail's Reader in Mass Communication Theory. London, California and New Delhi: Sage*, 2002.

[54] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377, 2020.

[55] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175, 2012.

[56] S. T. Fiske, A. J. C. Cuddy, and P. Glick. Universal dimensions of social perception: warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83, 2007.

[57] S. T. Fiske, A. J. C. Cuddy, P. Glick, and J. Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902, 2002.

[58] S.T. Fiske, J. Xu, A.C. Cuddy, and P Glick. (dis)respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, 55:473–489, 05 1999.

[59] Susan T. Fiske. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73, 2018.

[60] Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), 2018.

[61] Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, and Dirk Hovy. BERTective: Language models and contextual information for deception detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2699–2708, Online, April 2021. Association for Computational Linguistics.

[62] Knight Foundation. American views 2020: trust, media and democracy. Available at: https://knightfoundation.org/reports/american-views-2020-trust-media-and-democracy/, 2020.

[63] Erika Franklin Fowler, Michael Franz, Gregory Martin, Zachary Peskowitz, and Travis Ridout. Political advertising online and offline. *American Political Science Review*, 115(1):130–149, 2021.

[64] Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398, 2022.

[65] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. *arXiv preprint arXiv:2104.08790*, 2021.

[66] W.A. Gamson. *Talking politics*. New York: Cambridge University Press, 1992.

[67] William Gamson and Andre Modigliani. The changing culture of affirmative action. In Richard Braungart, editor, *Research in Political Sociology*, volume 3, pages 137–177. Jai Press, Inc, Greenwich, CT, 1987.

[68] William A Gamson, William Anthony Gamson Gamson, William Anthony Gamson, and William A Gamson. *Talking politics*. Cambridge university press, 1992.

[69] Wei Gao and Fabrizio Sebastiani. Tweet sentiment: From classification to quantification. In *2015 IEEE/ACM International Conference on ASONAM*, pages 97–104. IEEE, 2015.

[70] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[71] Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689, Online, April 2021. Association for Computational Linguistics.

[72] Goran Glavaš, Mladen Karan, and Ivan Vulić. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[73] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 18–23, Florence, Italy, July 2019. Association for Computational Linguistics.

[74] Peter Glick and Susan Fiske. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70:491–512, 03 1996.

[75] Erving Goffman. *Frame analysis: An essay on the organization of experience.* Nueva York: Harper & Row, 1974.

[76] Ana Granados, Manuel Cebrian, David Camacho, and Francisco de Borja Rodriguez. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102, 2010.

[77] Ana Granados, Manuel Cebrian, David Camacho, and Francisco de Borja Rodriguez. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102, 2011.

[78] Pär Anders Granhag and Leif A Strömwall. *The detection of deception in forensic contexts.* Cambridge University Press, 2004.

[79] Kelly M Greenhill. How trump manipulates the migration debate. *Foreign Affairs*, 5, 2018.

[80] N. Haslam, C. Loughnan, S. Reynolds, and S Wilson. Dehumanization: A new perspective. *Social and Personality Psychology Compass*, 1:409–422, 2007.

[81] L. Hemphill, A. Culotta, and M. Heston. #polarscores: Measuring partisanship using social media content. *Journal of Information Technology & Politics*, 13(4):365–377, 2016.

[82] We Are Social & DataReportal & Hootsuite. Global social network penetration rate as of january 2021, by region. Retrieved July 27, 2021, from https://www.statista.com/statistics/269615/social-network-penetration-by-region/, January 2021.

[83] Marjan Hosseinia. *Content and Stylistic Models for Authorship, Stance, and Hyperpartisan Detection*. PhD thesis, University of Houston, 2020.

[84] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021.

[85] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[86] Horacio Jesús Jarquín-Vásquez, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda. Not all swear words are used equal: Attention over word n-grams for abusive language identification. In Karina Mariela Figueroa Mora, Juan Anzurez Marín, Jaime Cerda, Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, and José Arturo Olvera-López, editors, *Pattern Recognition*, pages 282–292, Cham, 2020. Springer International Publishing.

[87] Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, 2019.

[88] Edward E Jones. The rocky road from acts to dispositions. *American Psychologist*, 34:107–117, 1979.

[89] Edward E Jones and Keith E Davis. From acts to dispositions: The attribution process in social perception. In *Advances in experimental social psychology*, volume 2, pages 219–266. New York: Academic Press, 1965.

[90] Edward E Jones and Victor A Harris. The attribution of attitudes. *Journal of Experimental Social Psychology*, 3:1–24, 1967.

[91] Kalman J Kaplan. On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77(5):361—-372, 1972.

[92] Douglas Kellner. Donald Trump and the politics of lying. In *Post-Truth, Fake News*, pages 89–100. Springer, 2018.

[93] Nicolas Kervyn, Susan Fiske, and Vincent Yzerbyt. Forecasting the primary dimension of social perception: Symbolic and realistic threats together predict warmth in the stereotype content model. *Social Psychology*, 46(1):36—-45, 2015.

[94] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, 2019.

[95] Donald R. Kinder. Opinion and action in the realm of politics. In D. T. Gilbert, S.T Fiske, and G. Lindzey, editors, *The handbook of Social Psychology*, volume 2, chapter 34, pages 778–867. N.J: Wiley., 1998.

[96] D. O. Kinder D. R.and Sears. Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology*, 40(3):414–431, 1981.

[97] Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. Detecting deception in political debates using acoustic and textual features. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 652–659, 2019.

[98] Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–515, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[99] Małgorzata Kossowska, Aneta Czernatowicz-Kukuczka, and Maciej Sekerdej. Many faces of dogmatism: Prejudice as a way of protecting certainty against value violators among dogmatic believers and atheists. *British Journal of Psychology*, 108(1):127–147, 2017.

[100] Lea Köstler and Ringo Ossewaarde. The making of ai society: Ai futures frames in german political and media discourses. *AI & society*, pages 1–15, 2021.

[101] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 313–322, 2019.

[102] Gretel Liz De la Peña Sarracén and P. Rosso. Aggressive analysis in twitter using a combination of models. In *IberLEF@SEPLN*, 2019.

[103] George Lakoff. *Don't think of an elephant! Know your values and frame the debate*. Chelsea Green Publishing, 2004.

[104] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138, 2020.

[105] Gerald Ki Wei Lee, Jun Choi ;Huang. Hyperpartisan news classification with elmo and bias feature. *Journal of Information Science & Engineering*, 37(5), 2021.

[106] Nayeon Lee, Zihan Liu, and Pascale Fung. Team yeon-zi at semeval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056, 2019.

[107] Tiane L. Lee and Susan T. Fiske. Not an outgroup, not yet an ingroup: Immigrants in the stereotype content model. *International Journal of Intercultural Relations*, 30(6):751–768, 2006. Special Issue: Attitudes towards Immigrants and Immigration.

[108] Jacques-Philippe Leyens, Paola M Paladino, Ramon Rodriguez-Torres, Jeroen Vaes, Stephanie Demoulin, Armando Rodriguez-Perez, and Ruth Gaunt. The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review*, 4(2):186–197, 2000.

[109] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)"*, pages 1566–1576, 2014.

[110] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics.

[111] Walter Lipmann. *Public Opinion*. New York:Harcourt Brace, 1922.

[112] Yinhan Liu and Myle Ott; Naman Goyal; Jingfei Du; Mandar Joshi; Danqi Chen; Omer Levy; Mike Lewis; Luke Zettlemoyer; Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

[113] Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, 36(1):123–155, 2 2011.

[114] Dermot Lynott, Michael Walsh, Tony McEnery, Louise Connell, Liam Cross, and Kerry O'Brien. Are you what you read? predicting implicit attitudes to immigration based on linguistic distributional cues from newspaper readership; a pre-registered study. *Frontiers in Psychology*, 10:842, 2019.

[115] Michael MacKuen and Courtney Brown. Political context and attitude change. *American Political Science Review*, 81(2):471–490, 1987.

[116] Sanguinetti Manuela, Comandini Gloria, Elisa Di Nuovo, Simona Frenda, Marco Antonio Stranisci, Cristina Bosco, Caselli Tommaso, Viviana Patti, Russo Irene, et al. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9. CEUR, 2020.

[117] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[118] Micol Marchetti-Bowick and Nathanael Chambers. Learning for microblogs with distant supervision: Political forecasting with twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612, Avignon, France, April 2012. Association for Computational Linguistics.

[119] Antonio Izquierdo Escribano María Concepción Carrasco Carpio Árbol académico, Carlos García Serrano Árbol académico. *Inmigración: mercado de trabajo y protección social en España.* Madrid: Consejo Económico y Social, 2003.

[120] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, 2021.

[121] John B. McConahay and Joseph C. Hough. Symbolic racism. *Journal of Social Issues*, 32(2):23–45, 1976.

[122] Julia Mendelsohn, Ceren Budak, and David Jurgens. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online, June 2021. Association for Computational Linguistics.

[123] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3, 2020.

[124] Vincent Menger, Floor Scheepers, and Marco Spruit. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6), 2018.

[125] Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[126] Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. Digging in the dirt: Extracting keyphrases from texts with KD. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 198–203. Accademia University Press, 2015.

[127] S. Moscovici. *Essai sur l'histoire humaine de la nature*. Paris: Flammarion, 1968.

[128] S. Moscovici. The coming era of representations. In J.P Codol and J.P. Leyens, editors, *Cognitive analysis of social behaviour*. Nijhoff, The Hague, 1982.

[129] S. Moscovici. The phenomenon of social representation. In R. Farr and S. Moscovici, editors, *Social Representations*, pages 3–69. Cambridge University Press, 1984.

[130] S. Moscovici and J.A. Pérez. A study of minorities as victims. *European Journal of Social Psychology*, 37:725–746, 2007.

[131] Serge Moscovici and Juan A. Pérez. A new representation of minorities as victims. In *Coping with Minority Status: Responses to Exclusion and Inclusion*, page 82–103. Cambridge University Press, 2009.

[132] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[133] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[134] Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. A second pandemic? analysis of fake news about COVID-19 vaccines in Qatar. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1010–1021, Held Online, September 2021. INCOMA Ltd.

[135] Federico Nanni, Goran Glavas, Simone Paolo Ponzetto, Sara Tonelli, Nicolo Conti, Ahmet Aker, Alessio Palmero Aprosio, Arnim Bleier, Benedetta Carlotti, Theresa Gessler, Tim Henrichsen, Dirk Hovy, Christian Kahmann, Mladen Karan, Akitaka Matsuo, Stefano Menini, Dong Nguyen, Andreas Niekler, Lisa Posch, Federico Vegetti, Zeerak Waseem, Tanya Whyte, and Nikoleta Yordanova. Findings from the hackathon on understanding euroscepticism through the lens of textual data. In *ParlaCLARIN 2018 Workshop Proceedings (LREC 2018)*. European Language Resources Association (ELRA), May 2018. 11th Edition of the Language Resources and Evaluation Conference, LREC 2018 ; Conference date: 07-05-2018 Through 12-05-2018.

[136] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 5 2003. PMID: 15272998.

[137] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[138] University of Oklahoma. Institute of Group Relations and Muzafer Sherif. *Intergroup conflict and cooperation: The Robbers Cave experiment*, volume 10. University Book Exchange Norman, OK, 1961.

[139] H Dan O'Hair and Michael J Cody. Deception. In *The dark side of interpersonal communication*, pages 181–214. Routledge, 1994.

[140] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[141] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 751–760, New York, NY, USA, 2010. ACM.

[142] Zhongdang Pan and Gerald Kosicki. Framing analysis: An approach to news discourse. *Political Communication*, 10:55–75, 01 1993.

[143] Baviera Puig Tomás; Sánchez-Junquera Javier; Rosso Paolo. Political advertising on social media: Issues sponsored on facebook ads during the 2019 general elections in spain. *Communication & Society.* [**in press**], 2022.

[144] Antonio Pascucci, Raffaele Manna, Vincenzo Masucci, and Johanna Monti. The role of computational stylometry in identifying (misogynistic) aggression in English social media texts. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 69–75, Marseille, France, May 2020. European Language Resources Association (ELRA).

[145] James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9(12):1–10, 12 2015.

[146] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007 the university of texas at austin. *Development*, 1(2):1–22, 2007.

[147] J.A. Perez, S. Moscovici, and B. Chulvi. The taboo against group contact: Hypothesis of gypsy ontologization. *British Journal of Social Psychology*, 46:249—-272, 2007.

[148] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 59–66, New York, NY, USA, 2015. ACM.

[149] Verónica Pérez-Rosas and Rada Mihalcea. Cross-cultural deception detection. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, 2014.

[150] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[151] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[152] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, 2018.

[153] Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131, 2019.

[154] Katherine Puddifoot. *How stereotypes deceive us.* Oxford University Press, 2021.

[155] J.A Pérez, S. Moscovici, and B. Chulvi. Nature and culture as principles for social classification. anchorage of social representations on ethnical minorities. *International Journal of Social Psychology*, 17(1):51–67, 2002.

[156] Verónica Pérez-Rosas and Rada Mihalcea. Experiments in open domain deception detection. In *Empirical Methods in Natural Language Processing*, pages 1120–1125. The Association for Computational Linguistics, 2015.

[157] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[158] Christian Rauh and Jan Schwalbach. The parlspeech v2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. *Harvard Dataverse*, 2020.

[159] David-Sven Reher, Luis Cortés, Fernando González, Miguel Requena, María Isabel Sánchez, Alberto Sanz, and Mikolaj Stanek. *Informe Encuesta Nacional de Inmigrantes (ENI – 2007)*. Instituto Nacional de estadística, 2008.

[160] Yafeng Ren and Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385-386:213–224, 2017.

[161] Ludovic Rheault and Christopher Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, pages 1–22, 2019.

[162] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[163] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

[164] Nazmul K Rony and Rahnuma Ahmed. Fake news conversation network in twitter: User type, emotional appeals and motives in network formation. *The Journal of Social Media in Society*, 10(1):121–139, 2021.

[165] Robert M Ross, David G Rand, and Gordon Pennycook. Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment & Decision Making*, 16(2), 2021.

[166] Paolo Rosso and Leticia C. Cagnina. Deception detection and opinion spam, chapter 8. In *A Practical Guide to Sentiment Analysis*, pages 155–171, Cham, 2017. Springer International Publishing.

[167] Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157, 2018.

[168] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*, 2017.

[169] Javier Sánchez-Junquera. Adaptación de dominio para la detección automática de textos engañosos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, 2018.

[170] Javier Sánchez-Junquera, Paolo Rosso, Manuel Montes-y Gómez, and Simone Paolo Ponzetto. Unmasking bias in news. *arXiv preprint arXiv:1906.04836*, 2019.

[171] Javier Sánchez-Junquera, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, and Paolo Rosso. Character n-grams for detecting deceptive controversial opinions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - Proceedings of the 9th International Conference of the CLEF Association. LNCS, vol. 11018, Springer-Verlag*, volume 11018 of *Lecture Notes in Computer Science*, pages 135–140, 2018.

[172] Javier Sánchez-Junquera, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, Paolo Rosso, and Efstathios Stamatatos. Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, 2020.

[173] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[174] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, 2019. Association for Computational Linguistics.

[175] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics.

[176] Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. GPolS: A contextual graph-based language model for analyzing parliamentary debates and political cohesion. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4847–4859, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[177] Dietram A. Scheufele. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122, 02 2006.

[178] Anamika Ashit Sen. *Hyperpartisanship in web searched articles*. PhD thesis, Virginia Tech, 2019.

[179] Donald Sharpe. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation*, 20, 2015.

[180] Chander Shekhar, Bhavya Bagla, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. Walk in wild: An ensemble approach for hostility detection in hindi posts. *arXiv preprint arXiv:2101.06004*, 2021.

[181] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[182] Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Trevino, and Juan Gordon. Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican International Conference on Artificial Intelligence*, pages 1–14. Springer, 2012.

[183] A. Sironi, C. Bauloz, and M. Emmanuel. *Glossary on Migration*. International Organization for Migration (IOM), 2019.

[184] Mahlon Brewster Smith. The open and closed mind. investigations into the nature of belief systems and personality systems. *Science*, 132(3420):142–143, 1960.

[185] Andrew D Spear. Resisting hyper-partisan silencing: Arendt on political persuasion through exemplification and truth-telling as action. *HannahArendt. Net*, 10(1), 2021.

[186] Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[187] Vertika Srivastava, Ankita Gupta, Divya Prakash, Sudeep Kumar Sahoo, RR Rohit, and Yeon Hyang Kim. Vernon-fenwick at semeval-2019 task 4: hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082, 2019.

[188] Efstathios Stamatatos. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149. Association for Computational Linguistics, 2017.

[189] Efstathios Stamatatos. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473, nov 2017.

[190] C. M. Steele and J. Aronson. Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology*, 69(3):797–811, 1995.

[191] S. Steele. *The content of our character*. New York: St. Martin´s Press, 1990.

[192] Andreas Steinmayr. Did the refugee crisis contribute to the recent rise of far-right parties in europe? *ifo DICE Report*, 15(4):24–27, 2017.

[193] Walter G. Stephan, Rolando Diaz-Loving, and Anne Duran. Integrated threat theory and intercultural attitudes: Mexico and the united states. *Journal of Cross-Cultural Psychology*, 31(2):240–249, 2000.

[194] Bozhidar Stevanoski and Sonja Gievska. Team Ned Leeds at SemEval-2019 task 4: Exploring language indicators of hyperpartisan reporting. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1026–1031, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[195] Preslav Nakov; Alan Ritter; Sara Rosenthal; Fabrizio Sebastiani; Veselin Stoyanov. Semeval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[196] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, Apr. 2020.

[197] Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Paolo Ponzetto. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8), 2021.

[198] Javier Sánchez-Junquera, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, Paolo Rosso, and Efstathios Stamatatos. Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, 135:122–130, 2020.

[199] Henri Tajfel, Anees A Sheikh, and Robert Charles Gardner. Content of stereotypes and the inference of similarity between members of stereotyped groups. *Acta Psychologica,*, 22(3):191–201, 1964.

[200] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. *European Conference on Information Retrieval*, pages 337–349, 2009.

[201] Hannah Tessler, Meera Choi, and Grace Kao. The anxiety of being asian american: Hate crimes and negative biases during the covid-19 pandemic. *American Journal of Criminal Justice*, 45(4):636–646, 2020.

[202] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[203] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335. Association for Computational Linguistics, July 2006.

[204] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *International Conference on Language Resources and Evaluation*, volume 2012, pages 2214–2218, 2012.

[205] Peter D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, May 2000.

[206] A Tversky and D Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[207] Pilar López Úbeda, Manuel Carlos Díaz-Galiano, L Alfonso Urena Lopez, M Teresa Martín-Valdivia, Teodoro Martín-Noguerol, and Antonio Luna. Transfer learning applied to text classification in spanish radiological reports. In *Proceedings of the LREC 2020 Workshop on*

Multilingual Biomedical Text Processing (MultiligualBIO 2020), pages 29–32, 2020.

[208] Paola Villano, Lara Fontanella, Sara Fontanella, and Marika Di Donato. Stereotyping roma people in italy: Irt models for ambivalent prejudice measurement. *International Journal of Intercultural Relations*, 57:30–41, 2017.

[209] Nikolai Vogler and Lisa Pearl. Using linguistically-defined specific details to detect deception across domains. *Natural Language Engineering*, 1(1):1–27, 2018.

[210] Aldert Vrij. *Detecting lies and deceit: the psychology of lying and implications for professional practice*. Wiley series in psychology of crime, policing and law. Wiley, 2000.

[211] Aldert Vrij. *Detecting lies and deceit: pitfalls and opportunities*. Wiley Series in the Psychology of Crime, Policing and Law. Wiley, 2008.

[212] Zihan Wang, Stephen Mayhew, Dan Roth, et al. Cross-lingual ability of multilingual bert: An empirical study. *International Conference on Learning Representations*, 2019.

[213] P Watzlawick, J.H Beavin, and D. Jackson. *Pragmatics of Human Communication : A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton & Company, Incorporated, W. W, 1967.

[214] Sandra R. Waxman. Racial awareness and bias begin early: Developmental entry points, challenges, and a call to action. *Perspectives on Psychological Science*, 16(5):893–902, 2021. PMID: 34498529.

[215] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[216] Heloisa Sturm Wilkerson, Martin J. Riedl, and Kelsey N. Whipple. Affective affordances: Exploring facebook reactions as emotional responses to hyperpartisan political news. *Digital Journalism*, 0(0):1–22, 2021.

[217] Qiong Wu, Songbo Tan, Miyi Duan, and Xueqi Cheng. A two-stage algorithm for domain adaptation with application to sentiment transfer problems. In *Information Retrieval Technology*, pages 443–453. Springer Berlin Heidelberg, 2010.

[218] Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online, September 2021. INCOMA Ltd.

[219] Mark P Zanna. On the nature of prejudice. *Canadian Psychology/Psychologie canadienne*, 35(1):11–23, 1994.

[220] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

[221] Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception. In Leonard Berkowitz, editor, *Advances in experimental social psychology*, volume 14 of *Advances in Experimental Social Psychology*, pages 1 – 59. Academic Press, 1981.