Universitat Politècnica de València

Departamento de Sistemas Informáticos y Computación

Tesis de Doctorado en Informática

Simona Frenda

Sarcasm and Implicitness in Abusive Language Detection: A Multilingual Perspective

Directores de Tesis
Viviana Patti
Università degli Studi di Torino, Italy
Paolo Rosso
Universitat Politècnica de València, Spain

Año Académico 2021-2022

Abril 2022

University of Turin

Doctoral School of Sciences and Innovative Technologies PhD Program in Computer Science
XXXIII cycle

PhD Dissertation
Simona Frenda

Sarcasm and Implicitness in Abusive Language Detection: A Multilingual Perspective

Advisors
Viviana Patti
Università degli Studi di Torino, Italy
Paolo Rosso
Universitat Politècnica de València, Spain

PhD Coordinator
Marco Grangetto

Academic Year 2021-2022

April 2022

To my family.
To Eduardo.

Acknowledgment

At the end of this work, I would like to thank all people who supported me during this new step of my professional and personal life. During this journey, various professional figures guided and inspired me. Thanks to Professor Viviana Patti and Professor Paolo Rosso, who patiently shaped the researcher that I am today. Thanks to Professor Manuel Montes-y-Gómez and Professor Rafael Valencia García, who from the beginning showed me how to look deeply and differently at data. Thanks to Professor Noriko Kando for hosting me at National Institute of Informatics in Tokyo, and introducing me to her fascinating culture. Thanks to Professor Mariona Taulé Delor for receiving me in Barcelona, and being always friendly and professionally available to help me. Thanks to reviewers, Professor Rafael Valencia García (Universidad de Murcia, Spain), Professor Vasudeva Varma (International Institute of Information Technology of Hyderabad, India) and Professor Els Lefever (Ghent University, Belgium), for their efforts in evaluating this dissertation[1] and for their insightful comments. Thanks to Professor Cristina Bosco and Dr. Valerio Basile and all the group of research in Turin who supported and encouraged my investigations, believing in me. And finally, I would like to thank all my colleagues for everything that they taught me and for everything that they will teach me!

This work was partially supported by various financial projects. Among them: the Spanish research project SomEMBED funded by Ministerio de Economía y Sostenibilidad (MINECO), the NII International Internship Program funded by JSPS KAKENHI, the Italian project M.EMO.RAI funded by RAI - Radiotelevisione Italiana Spa, the Italian project IhatePrejudice funded by Compagnia di San Paolo, and the European project "STERHEOTYPES" funded by Compagnia di San Paolo Foundation, Volkswagen Stiftung and Carlsberg Fondation.

[1] I started writing the first version of this dissertation in May 2021. I was able to submit the first draft in February 2022, and I later received the comments of reviewers in March 2022. The final (and current) version of the thesis has been registered in April 2022.

Abstract

The possibility to monitor hateful content online on the basis of what people write is becoming an important topic for several actors such as governments, ICT companies, and NGO's operators conducting active campaigns in response to the worrying rise of online abuse and hate speech. Hand in hand, abusive language detection turns into a task of growing interest in Natural Language Processing (NLP), especially when applied to the recognition of various forms of hatred in social media posts. Abusive language is a broad umbrella term which is commonly used for denoting different kinds of hostile user-generated contents that intimidate or incite to violence and hatred, targeting many vulnerable groups in social platforms. Such hateful contents are pervasive nowadays and can also be detected even in other kinds of texts, such as online newspapers.

The importance of understanding and automatically detecting abusive language is due to the observation of real manifestations of violent acts connected to negative behaviors online in its various forms, such as cyberbullying, racism, sexism, or homophobia. Various approaches have been proposed in the last years to support the identification and monitoring of these phenomena, but unfortunately, they are far from solving the problem due to the inner complexity of abusive language, and to the difficulties to detect its implicit forms.

In our doctoral investigation, we have studied the issues related to automatic identification of abusive language online, investigating various forms of hostility against women, immigrants and cultural minority communities in languages such as Italian, English, and Spanish. The multilingual frame allowed us to have a comparative setting to reflect on how hateful contents are expressed in distinct languages and how these different ways are transposed in the automated processing of the text. The analysis of the results of different methods of classification of hateful and non-hateful messages revealed important challenges that lie principally on the implicitness of some manifestations of abusive language expressed through the use of figurative devices (i.e., irony and sarcasm), recall of inner ideologies (i.e., sexist ideology) or cognitive schemas (i.e., stereotypes), and expression of unfavorable stance.

To face these challenges, in this work, we have proposed distinct solutions applicable also to different textual genres. We observed that, in particular, cognitive and creative aspects of abusive language are harder to infer automatically from texts. At the same time they are often recurrent elements, such in the case of sarcasm, a figurative device that tends to affect the accuracy of the systems. Indeed, for its peculiarities, sarcasm is apt to disguise hurtful messages, especially in short and informal texts such as the ones posted on Twitter. Our hypothesis is that information about the presence of sarcasm could help to improve the detection of hateful messages, even when they are camouflaged as sarcastic. In this line, it is interesting to study how the injection of linguistic knowledge into detection models can be useful to capture implicit levels of meaning.

According to rhetorical literature, sarcasm is considered as a specific form of irony. In line with our multilingual and linguistic approach, we studied the expression of irony in Italian and Spanish user-generated contents, revealing also the most common traits of ironic language. Therefore, we focused on Italian in order to validate our hypothesis

and to investigate the specific characteristics of sarcasm in delicate contexts such as the online debates on sensitive and social issues, like immigration.

In particular, we created novel resources that allowed us to examine deeply our hypothesis and develop specific approaches for the detection of two forms of abusive language in tweets and headlines: hate speech and stereotypes. Our idea is to fruitfully combine general knowledge from language models and linguistic information, obtained with specific linguistic features and the injection of ironic language recognition within a multi-task learning framework. The experimental results confirm that the awareness of sarcasm helps systems to retrieve correctly hate speech and stereotypes in social media texts, such as tweets. Moreover, linguistic features make the system sensible to stereotypes in both tweets and news headlines.

The corpora used in our experiments have been exploited as benchmark datasets within the EVALITA evaluation campaign for NLP tools for Italian, contributing to creating a new state of the art for these tasks in Italian. Moreover, the multidisciplinary and multilingual frame of our analyses allowed us to reflect on the boundaries between dimensions and topical focuses that often overlap in computational approaches to detect abusive language and related phenomena.

Abstract

La possibilità di monitorare i contenuti di odio online sulla base di ciò che le persone scrivono sta diventando un assunto importante per diversi soggetti, come governi, aziende ICT e operatori di ONG che mettono in atto campagne di sensibilizzazione in risposta al preoccupante aumento di abusi e dell'incitamento all'odio online. Di pari passo, la rilevazione automatica dei discorsi di odio è oggetto di crescente interesse nell'ambito del *Natural Language Processing* (NLP), soprattutto se volto all'identificazione di varie forme di odio e di *abusive language* nei post sui social media. *Abusive language* è un termine generico comunemente usato per indicare differenti contenuti ostili generati dagli utenti, che intimidiscono o incitano alla violenza e al disprezzo prendendo di mira i gruppi vulnerabili nei social network. Tali contenuti sono, oggigiorno, molto diffusi e possono essere rilevati anche in altri tipi di testi, come articoli e titoli di giornale online.

L'importanza di comprendere e rilevare automaticamente i discorsi di odio cresce di pari passo con l'incremento di manifestazioni di atti violenti collegati ai comportamenti abusivi online, come il *cyberbullying*, il razzismo, il sessismo e l'omofobia. Negli ultimi anni sono state proposte diverse tecniche per supportare l'identificazione e il monitoraggio di questi fenomeni, ma purtroppo gli approcci odierni sono lontani dal risolvere il problema a causa della complessità interna dei discorsi di odio e delle difficoltà nel rilevare le forme implicite.

Nella nostra ricerca di dottorato, abbiamo studiato le questioni relative all'identificazione automatica dell'incitamento all'odio online, indagando su varie forme di ostilità contro le donne, gli immigrati e le comunità culturali minoritarie, in lingue come l'italiano, l'inglese e lo spagnolo. La cornice multilingue ci ha permesso di avere un'impostazione comparativa per riflettere su come i discorsi di odio sono espressi nelle varie lingue, e su come tali espressioni devono essere rappresentate nel trattamento automatico del testo. L'analisi dei risultati dei vari metodi di classificazione dei messaggi in relazione alla presenza di *abusive language*, ha fatto emergere consistenti difficoltà legate principalmente alle manifestazioni più implicite dei discorsi di odio, riscontrate per esempio nei casi in cui vengono usate figure retoriche (come ironia e sarcasmo), quando si rafforzano delle ideologie (come l'ideologia sessista) o degli schemi cognitivi (come gli stereotipi), o ancora quando si esprimono posizioni contrarie a un tema di discussione.

Per affrontare queste difficoltà, abbiamo proposto soluzioni distinte e applicabili anche a diversi generi testuali. In particolare abbiamo osservato che gli aspetti cognitivi e creativi nei discorsi di odio sono più difficili da identificare automaticamente nei testi. Allo stesso tempo sono anche elementi molto ricorrenti, come nel caso del sarcasmo, un espediente retorico che tende a inficiare l'accuratezza dei sistemi. Infatti, per le sue peculiarità, il sarcasmo è adatto a mascherare i messaggi offensivi, soprattutto in testi molto brevi e informali come quelli pubblicati su Twitter. La nostra ipotesi è che informando il sistema sulla presenza del sarcasmo, si possa migliorare l'identificazione dei messaggi di odio, anche quando questi sono espressi in modo sarcastico. A tale scopo, risulta interessante studiare come l'introduzione di conoscenza linguistica nei modelli di *detection* possa essere utile per catturare i livelli più impliciti del significato.

Secondo la letteratura retorica, il sarcasmo è considerato come una forma particolare

di ironia. In linea con il nostro approccio linguistico e multilingue, abbiamo esaminato le espressioni ironiche nei contenuti generati dagli utenti in italiano e in spagnolo, rivelando i tratti più universali del linguaggio ironico. Su questa base, ci siamo concentrati sull'italiano per convalidare la nostra ipotesi e per indagare le caratteristiche specifiche del sarcasmo in contesti delicati, come i dibattiti online su temi sensibili e sociali, ad esempio, l'immigrazione.

Nello specifico, abbiamo creato nuove risorse che ci hanno permesso di approfondire la nostra ipotesi e sviluppare vari approcci per l'identificazione di due forme di *abusive language* nei tweet e nei titoli di giornale: i discorsi di odio (o *hate speech*) e gli stereotipi. La nostra idea è combinare fruttuosamente conoscenza generale dai *language model* e informazioni linguistiche, ottenute estraendo specifici elementi linguistici o con l'apprendimento simultaneo del riconoscimento del linguaggio ironico in un'architettura *multi-task*. I risultati sperimentali confermano che rendendo i sistemi consapevoli della presenza del sarcasmo si migliora il riconoscimento dei discorsi di odio e degli stereotipi nei testi provenienti dai social media, come i tweet. Mentre informandoli di specifici elementi linguistici i sistemi diventano più sensibili a identificare gli stereotipi sia nei tweet che nei titoli di giornale.

I corpora utilizzati nei nostri esperimenti sono stati proposti come dataset di riferimento per *shared task* in due edizioni di EVALITA, la campagna di valutazione degli strumenti di NLP per l'italiano, contribuendo a creare un nuovo stato dell'arte per questi *task* di *detection* in italiano. Inoltre, il quadro multidisciplinare e multilingue delle nostre analisi ci ha permesso di riflettere sui confini tra aspetti più generali e domini più specifici che spesso si sovrappongono negli approcci computazionali per identificare i discorsi di odio e i fenomeni correlati.

Resumen

La posibilidad de monitorear el contenido de odio en línea a partir de lo que escribe la gente se está convirtiendo en un asunto muy importante para varios actores, como gobiernos, empresas de TIC y profesionales de ONG's que implementan campañas de sensibilización en respuesta al preocupante aumento de los abusos y de la incitación al odio en línea. Al mismo tiempo, la detección automática del lenguaje abusivo (más conocido como *abusive language*) es un tema de creciente interés en el campo del Procesamiento del Lenguaje Natural (PLN), especialmente si el objetivo es identificar diversas formas de odio en las publicaciones de las redes sociales. El *abusive language* es un término genérico que se utiliza para definir los contenidos hostiles generados por usuarios, que intimidan o incitan a la violencia y al desprecio, dirigiéndose a grupos vulnerables en las redes sociales. Hoy en día, estos contenidos están muy extendidos, y se encuentran también en otros tipos de textos como los artículos y títulos de periódicos online.

La importancia de comprender y detectar automáticamente el discurso de odio se debe al aumento de las manifestaciones de actos violentos vinculados a conductas abusivas en línea, como el ciberacoso, el racismo, el sexismo y la homofobia. Se han implementado varios enfoques en los últimos años para apoyar la identificación y el monitoreo de estos fenómenos, lamentablemente estos están lejos de resolver el problema debido a la complejidad interna del lenguaje abusivo y las dificultades para detectar sus formas más implícitas.

En nuestra investigación de doctorado, hemos examinado las cuestiones relacionadas con la identificación automática del lenguaje abusivo en línea, investigando las diferentes maneras de hostilidad contra las mujeres, los inmigrantes y las comunidades culturales minoritarias, en idiomas como el italiano, el inglés y el español. El marco multilingüe nos ha permitido tener un enfoque comparativo para reflexionar sobre cómo se expresa el discurso de odio en varios idiomas, y cómo dichas expresiones se deben representar en el proceso automático del texto. El análisis de los resultados de los distintos métodos de clasificación de los mensajes en relación con la presencia del lenguaje abusivo, ha sacado a la luz algunas dificultades principalmente vinculadas a sus manifestaciones más implícitas. Por ejemplo, en los casos en que se utilizan figuras retóricas (como la ironía y el sarcasmo), cuando se fortalecen ideologías (como la ideología sexista) o esquemas cognitivos (como los estereotipos), o cuando se postulan contrarias a un tema de discusión.

Para abordar estas dificultades, hemos propuesto distintas soluciones que también se pueden aplicar a diferentes géneros textuales. En particular, hemos observado que los aspectos cognitivos y creativos del discurso del odio son más difíciles de deducir automáticamente de los textos. Al mismo tiempo, también son elementos muy recurrentes como el caso del sarcasmo un recurso retórico que tiende a socavar la precisión de los sistemas. De hecho, por sus peculiaridades, el sarcasmo es adecuado para enmascarar mensajes ofensivos, especialmente en textos muy breves e informales como los publicados en Twitter. Nuestra hipótesis es que al informar al sistema sobre la presencia del sarcasmo, se mejoraría la identificación de los mensajes de odio, incluso cuando estos están disfrazados de sarcásticos. Para ello, es interesante estudiar cómo la introducción

de conocimientos lingüísticos en modelos de detección puede ser útil para capturar los niveles de significado más implícitos.

Según la bibliografía retórica, el sarcasmo se considera una forma particular de ironía. De acuerdo con nuestro enfoque lingüístico y multilingüe, examinamos expresiones irónicas en contenido generado por usuarios en italiano y español, revelando los rasgos más universales del lenguaje irónico. Sobre esta base, nos hemos centrado en el italiano para validar nuestra hipótesis e investigar las características específicas del sarcasmo en contextos sensibles, como los debates en línea sobre temas sociales por ejemplo, la inmigración.

En concreto, hemos creado nuevos recursos que nos permitieron profundizar en nuestra hipótesis y desarrollar diversos enfoques para identificar dos maneras de lenguaje abusivo en tuits y títulos de periódicos: los discursos de odio (o *hate speech*) y los estereotipos. Nuestra idea es combinar de manera fructífera el conocimiento general de los modelos lingüísticos y la información lingüística obtenida mediante la extracción de elementos lingüísticos específicos o entrenando simultáneamente el sistema al reconocimiento del lenguaje irónico en una arquitectura multitarea. Los resultados experimentales confirman que hacer que los sistemas sean conscientes del sarcasmo mejora el reconocimiento del discurso de odio y los estereotipos en los textos de las redes sociales, como los tuits. Al informarles de elementos lingüísticos específicos, se vuelven más sensibles a la identificación de estereotipos tanto en los tuits como en los títulos de periódicos.

Los *corpora* utilizados en nuestros experimentos se propusieron como referencia para tareas compartidas en dos ediciones de EVALITA, la campaña de evaluación de herramientas de PLN para el italiano, ayudando a crear un nuevo estado del arte para estas tareas de detección en italiano. Además, el marco multidisciplinario y multilingüe de nuestros análisis nos permitió reflexionar también sobre los límites entre aspectos más generales y dominios más específicos que a menudo se superponen en los enfoques computacionales para identificar el discurso del odio y los fenómenos relacionados.

Resum

La possibilitat de monitorar el contingut d'odi en línia a partir del que escriu la gent s'està convertint en un assumpte molt important per a diversos actors, com ara governs, empreses de TIC i professionals d'ONG que implementen campanyes de sensibilització en resposta al preocupant augment dels abusos i de la incitació a l'odi en línia. Alhora, la detecció automàtica del llenguatge abusiu (més conegut com a *abusive language*) és un tema de creixent interès en el camp del Processament del Llenguatge Natural (PLN), especialment si l'objectiu és identificar diverses formes d'odi a les publicacions de les xarxes socials. L'*abusive language* és un terme genèric que s'utilitza per definir els continguts hostils generats per usuaris, que intimideixen o inciten a la violència i al menyspreu, adreçant-se a grups vulnerables a les xarxes socials. Avui dia, aquests continguts estan molt estesos, i es noten també en altres tipus de textos com els articles i títols de diaris en línia.

La importància de comprendre i detectar automàticament el discurs d'odi es deu a l'augment de les manifestacions d'actes violents vinculats a conductes abusives en línia, com ara el ciberassetjament, el racisme, el sexisme i l'homofòbia. S'han implementat diversos enfocaments en els darrers anys per donar suport a la identificació i monitorització d'aquests fenòmens, lamentablement aquests estan lluny de resoldre el problema a causa de la complexitat interna del llenguatge abusiu i les dificultats per detectar-ne les formes més implícites.

A la nostra investigació de doctorat, hem examinat les qüestions relacionades amb la identificació automàtica del llenguatge abusiu en línia, investigant les diferents maneres d'hostilitat contra les dones, els immigrants i les comunitats culturals minoritàries, en idiomes com l'italià, l'anglès i l'espanyol . El marc multilingüe ens ha permès tenir un enfocament comparatiu per reflexionar sobre com s'expressa el discurs d'odi en diversos idiomes, i com s'han de representar aquestes expressions en el procés automàtic del text. L'anàlisi dels resultats dels diferents mètodes de classificació dels missatges en relació amb la presència del llenguatge abusiu ha tret a la llum algunes dificultats principalment vinculades a les manifestacions més implícites. Per exemple, en els casos en què es fan servir figures retòriques (com la ironia i el sarcasme), quan s'enforteixen ideologies (com la ideologia sexista) o esquemes cognitius (com els estereotips), o quan es postulen contràries a un tema de discussió .

Per abordar aquestes dificultats, hem proposat diferents solucions que també es poden aplicar a diferents gèneres textuals. En particular, hem observat que els aspectes cognitius i creatius del discurs de l'odi són més difícils de deduir automàticament dels textos. Alhora, també són elements molt recurrents com el cas del sarcasme un recurs retòric que tendeix a soscavar la precisió dels sistemes. De fet, per les seves peculiaritats, el sarcasme és adequat per emmascarar missatges ofensius, especialment en textos molt breus i informals com els publicats a Twitter. La nostra hipòtesi és que en informar el sistema sobre la presència del sarcasme, es milloraria la identificació dels missatges d'odi, fins i tot quan aquests estan disfressats de sarcàstics. Per això, és interessant estudiar com la introducció de coneixements lingüístics en models de detecció pot ser útil per capturar els nivells de significat més implícits.

Segons la bibliografia retòrica, el sarcasme és considerat una forma particular d'ironia. D'acord amb el nostre enfocament lingüístic i multilingüe, examinem expressions iròniques en contingut generat per usuaris en italià i espanyol, tot revelant els trets més universals del llenguatge irònic. Sobre aquesta base, ens hem centrat en l'italià per validar la nostra hipòtesi i investigar les característiques específiques del sarcasme en contextos sensibles, com ara els debats en línia sobre temes socials per exemple, la immigració.

En concret, hem creat nous recursos que ens han permès aprofundir en la nostra hipòtesi i desenvolupar diversos enfocaments per identificar dues maneres de llenguatge abusiu en tuits i títols de diaris: el discurs d'odi (o *hate speech*) i els estereotips. La nostra idea és combinar de manera fructífera el coneixement general dels models lingüístics i la informació lingüística obtinguda mitjançant l'extracció d'elements lingüístics específics o entrenant simultàniament el sistema al reconeixement del llenguatge irònic en una arquitectura multitasca. Els resultats experimentals confirmen que fer que els sistemes siguin conscients del sarcasme millora el reconeixement del discurs d'odi i els estereotips als textos de les xarxes socials, com els tuits. En informar-los d'elements lingüístics específics, esdevenen més sensibles a la identificació d'estereotips tant als tuits com als títols de diaris.

Els *corpora* utilitzats en els nostres experiments es van proposar com a referència per a tasques compartides a dues edicions d'EVALITA, la campanya d'avaluació d'eines de PLN per a l'italià, ajudant a crear un nou estat de l'art per a aquestes tasques de detecció en italià. A més, el marc multidisciplinari i multilingüe de les nostres anàlisis ens ha permès reflexionar també sobre els límits entre aspectes més generals i dominis més específics que sovint se superposen als enfocaments computacionals per identificar el discurs de l'odi i els fenòmens relacionats.

List of Figures

List of Tables

List of abbreviations

AI: Artificial Intelligence
biLSTM: Bidirectional Long Short-Term Memory
BoW: Bag of Words
BoC: Bag of Characters
CGA: Course-Grained Annotation
CL: Computational Linguistics
CNN: Convolutional Neural Network
CRT: Critical Race Theory
DAL: Dictionary of Affect in Language
DL: Deep Learning
ECRI: European Commission against Racism and Intolerance
FN: False Negative
FP: False Positive
FRA: European Union Agency for Fundamental Rights
FT: Fine-Tuned
HC: Hard Cases
ID: Identifier
LM: Language Model
LR: Logistic Regression
LSTM: Long Short-Term Memory
MTL: Multi-Task Learning
NLP: Natural Language Processing
NUs: Nominal Utterances
PoS: Part-of-Speech
RBF: Radial Basis Function
RF: Random Forest
SA: Sentiment Analysis
SC: Simple Cases
SDA: Stance Detection towards the legalization of Abortion
SDF: Stance Detection towards Feminist movements
SEL: Spanish Emotion Lexicon
SVM: Support Vector Machine
TF-IDF: Term Frequency–Inverse Document Frequency
TN: True Negative
TP: True Positive
UBT: Unigrams, Bigrams and Trigrams
UD: Universal Dependencies

Contents

Chapter 1

Introduction

> The Internet is a reflection of our society and that mirror is going to be reflecting what we see. If we do not like what we see in that mirror the problem is not to fix the mirror, we have to fix society.
>
> Vinton Gray Cerf

Our historical period is characterized by deep transformations, especially social and technological. The new forms of interaction and communication supported by advanced devices bring along new legal issues underestimated until now. One of them concerns the spread and support online of hate against groups or individuals "on the ground of race, colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status" [ECRI Recommendation n. 15, 2015][1]. This social problem could not fail to involve one of the most important human rights: the freedom of expression.

The first clause of Art. 10 of the European Convention on Human Rights proclaims that the right to freedom of expression conveys the "freedom to hold opinions and to receive and impart information and ideas". But the linguistic act of *speech* is not merely expressive, indeed, it is intended to communicate, and, thus, may affect or harm others. As philosopher Onora O'Neill said in one of her interventions on this theme: "The nursery jingle 'sticks and stones may break my bones, but words can never hurt me' is palpably false"[2]. Words are not innocuous, and the necessity to use them responsibly emerges from the need to respect other rights, especially the right of living a life free of violence. As claimed in the second clause of Art. 10, "The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions, or penalties as are prescribed by law and are necessary in a democratic society".

[1] The ECRI is the European Commission against Racism and Intolerance and its General Policy Recommendation n. 15 on combating Hate Speech is available in several languages on https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15

[2] Onora O'Neill, "A Right to Offend?", The Guardian, 13 February 2006

Current technologies have realized a virtual public sphere that the initial enthusiasm saw as *democratic* [Barlow, 2001], but, very soon, its virtues converted into serious social risks. Internet does not seem to encourage tolerance of different perspectives as hoped, but it seems to be converted into 'echo chambers' of homogenized beliefs or ideas of individual and intolerant autocracies of thought [Bray and Cerf, 2019]. These echo chambers, supported by the global communication on social platforms, tend to reinforce social and cultural bias, such as racial or patriarchal stereotypes, also through hoaxes, without encountering other or opposite narratives. Therefore, this is where the internet technologies realize their paradox.

Especially in social media, the advantage of creating communities, making followers, and masking the identity, allows users to attack easily *outgroups* on the basis of perceived threads [Smith, 1993]. The hateful ideas are powered and spread also by the action of the trolls online that make hate viral. This dissemination, often untraceable, affects victims' life; as said before, words carrying insults and incitements to violence can hurt and even kill [Foxman and Wolf, 2013]. To prevent the consequences, every year European Commission monitors the conduct of social platforms on the basis of the Code of Conduct on countering illegal hate speech online signed in 2016. This document has the purpose to increase the cooperation between the IT Companies, civil society organizations and Member States authorities in Europe to enforce the legislation[3] that prohibits racist and xenophobic *hate crime* and hate speech.

Toxic messages affect social cohesion by reinforcing tensions between social groups, and, in particular, the victims' life. Some data about the relation between hate speech online and biased crimes come from the USA. Fulper et al. [2014] demonstrated the existence of a correlation between the number of rapes and the amount of misogynistic tweets per state in the USA, suggesting the fact that social media can be used as a social detector of violence.

Although a causal link between cyberharassment and hate crime is hard to demonstrate due to the difficulty to trace the texts encouraging the physical offence, the risk of crime is assessed by victim surveys collected by European Union Agency for Fundamental Rights (FRA)[4] and the systematic recording of crimes in the EU[5]. Moreover, the damages hardest to quantify are the effects on psychological and physical well-being of the victims. The exposure to harassments and *microaggressions* could provoke, in the long run, serious physical health issues such as cardiovascular disease [Calvin et al., 2003], and immediate complex mental health issues such as depression, distress, state of anxiety that might culminate in suicide [Hwang and Goto, 2008, Lambert et al., 2009, Nadal et al., 2014]

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008F0913&from=en

[4] https://fra.europa.eu/en/tools

[5] To clarify further, see the last study commissioned by the European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs about the online content regulation in the EU: https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)655135

especially in adolescents [Nikolaou, 2017]. We can remember the case of the Canadian teenager Amanda Todd suicide in 2012 because victim of cyberbulling[6].

One of the methodologies used to counter hate online concerns the *manual* elaboration of alternative narratives about positive impacts and values of civil society that incite the reflection on the object and the target of the discrimination. This activity aims principally to make people aware of the proliferation of toxic rhetoric and the violated human rights, and relies on the effort of several volunteers in Europe[7]. This technique could be supported by the *automatic* recognizer of hateful message. This system filtering these messages allows to monitor this phenomenon online[8], to create devices that arise the reflection on the characteristics of these messages in a scholar context[9] and, finally, to understand also how and where to intervene actively in the society. This technique of recognition is inherited by a more general approach of analysis of users' opinions and content online called *Sentiment Analysis and Opinion Mining* [Liu, 2012].

Text analysis studies focuses in general on the examination of the message to extract the sentiment, the emotion, the stance towards a targeted issue or person. However, the expression of negative or unfavorable opinions, sometimes, could involve a toxic language that both offends the targeted persons and foments a sentiment of social discord. This kind of opinions, stemmed principally from prejudices and social frustrations, needs to be stopped to avoid the spread of the plague of hate in our societies.

Nevertheless, *the detection of hate speech online* is a difficult task due to the complexity of natural language. Catching automatically the illocutionary force of words needs to involve the collaboration of "The Two Cultures" (literal and scientific one) as defined by C. P. Show[10]. This hybridity comes true in a specific branch of Artificial Intelligence (AI) called *Natural Language Processing* (NLP) that deals with the interaction between machines and humans exploiting linguistic and computer knowledge.

NLP approaches allow automatic systems to understand natural language, extracting relevant linguistic information from texts. With the help of new computational techniques of *Machine Learning*, this process of comprehension appears simplified, but a system that bases its knowledge only on stochastic processes cannot perceive the real meaning of the message. Language knowledge needs to be transferred into it, especially, to make it able to understand tortuous linguistic phenomenon such as the expression of toxic messages through implicit figurative expedients.

[6] http://di.unito.it/theguardian

[7] See No Hate Speech Youth Campaign promoted by the Council of Europe: https://www.coe.int/en/web/no-hate-campaign/home and the Task Force Hate Speech organized by Amnesty International: https://www.amnesty.it/entra-in-azione/task-force-attivismo/

[8] Some examples come from the Italian initiatives of *Mappa contro l'odio* (https://mappa.controlodio.it) and Vox (http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-4/)

[9] Another example is the device employed in Frenda et al. [2021]: https://didattica.controlodio.it/

[10] Snow and Collini [1998] with the introduction of Stefan Collini.

Technologies able to understand natural language and, thus, identify hate speech, could help to monitor and counter it responding to the ethic need to ensure the fundamental rights for all humans as well as the freedom of speech and the right to life without discrimination.

1.1 Natural Language Processing

Several years ago, the Turing test, one of the first tests that judges whether a machine is intelligent or not, evaluates a machine as intelligent if it possesses conversational abilities that are comparable to those of a human being. This implies that an intelligent machine must possess linguistic abilities of comprehension and production of natural language. From the 1950s, the interest in language especially in technical fields, such as Computer Science, began to increase stimulating, on the one hand, the imagination of the science fiction and, on the other one, the development of the first applications in the real world such as Machine Translation systems. Therefore, the need to make machines able to understand natural language becomes, from the very beginning, an essential prerequisite for this kind of technologies.

The disciplinary field intended to process automatically natural language is NLP. Since the common focus on the study of language, this term is, sometimes, used as a synonym of Computational Linguistics (CL). A shared interpretation is that CL relates more to the formalization of computational models of various linguistic and psycholinguistic phenomena. Whereas, NLP refers to the application of computational techniques that process linguistic data (speech or text) [Clark et al., 2013, Kurdi, 2016, Basile, 2020a]. Our investigation is not only motivated from a technological interest but from a scientific perspective focused on the study of the implicitness of language and especially of hate speech.

In spite of these terminological divisions, in order to develop a NLP technology it is necessary to adopt a complete perspective about language that intersects humanistic and technical studies such as Linguistics, Computer Science and Cognitive Psychology. This hybridization, indeed, helps researchers to design models that aim at replicating the complex mechanisms of comprehension that are spontaneous for humans. It is in this branch of NLP related to Natural Language Understanding (NLU) that our work is inserted in.

1.1.1 Abusive Language Detection

As said above, a support to counter hate online is its automatic recognition that helps to easily filter and analyze it. Although IT companies have already developed techniques that could be efficient enough, the majority of them rely principally on the feedback or warnings of their users.

Realizing an automatic process of detection of hateful content online is hard due to the complexity of natural language. Systems of detection, indeed, need to be equipped with linguistic intuition typical of humans by means of linguistic features or specific annotated data. Especially in a sensitive and important issue like toxic messages online, our linguistic intuition helps to understand the most implicit forms of hate speech that could be disguised as stance (Example 1), funny comment (Example 2 and 3) or simple banality that actually stems from negative stereotypes or prejudices (Example 3):

(1) *@USER @USER is a seven pound baby murderer. Too bad her Mom didn't have the same operation. #Baby #Democrats #Losers #SemST*[11]

(2) *Signore, hanno tutti diritto a una vita dignitosa, ma mettete un migrante sulla mia strada e io sarò Salvini. (Matteo 15, 83)*[12]

(3) *Un piatto di pasta e chiediamogli scusa per non essere anche noi musulmani. Magari così diventano nostri amichetti e non ci uccidono più.*[13]

As in Example 1, the expression of own stance in controversial social issues, such as legalization of abortion, could involve an offensive and aggressive tone, inciting sometimes also to violence against the opposite group (in this case women in favor of the legalization of abortion). The same intention of attack is expressed in Example 2, even if it could amuse the readers with the metonymic use of 'Salvini' to refer to his immigration policy. Whereas, Example 3 does not contain an expressed intention of attack, but it aims to underline, although in a sarcastic way, the stereotyped idea that "all Muslims are terrorists".

In accordance with Merriam-Webster Dictionary, a *stereotype* is

> a standardized mental picture that is held in common by members of a group and that represents an oversimplified opinion, prejudiced attitude, or uncritical judgment.

The proliferation of *oversimplified* and *uncritical* judgments about especially minorities causes the reinforcement of *outgroup homogeneity* perceived as different and, sometimes, in contrast with own ingroup [Fiske, 1998] generating offensive expressions, such as:

(4) *#DecretoSalvini esatto e' buono anche per gli immigrati regolari che si vogliono integrare sul serio. la nostra cultura millenaria fara loro del bene.*[14]

[11] Tweet extracted from the corpus released by the organizers of Task6 in SemEval 2016 about Stance Detection (StanceDataset) [Mohammad et al., 2016]. This tweet is annotated as unfavorable toward legalization of abortion.

[12] *Sir, everybody has the right to a dignified life, but if you put a migrant in my way, I will be Salvini. (Matthew 15, 83).* Tweet extracted from HaSpeeDe2020 [Sanguinetti et al., 2020].

[13] *A plate of pasta and let's apologize for not being Muslims too. Maybe then they become our friends and won't kill us anymore.* Tweet extracted from HaSpeeDe2020.

[14] *#DecretoSalvini exactly is even good for legal immigrants who want to integrate seriously. our millennial culture will do them good.* Tweet extracted from HaSpeeDe2020.

Although hate speech detection is a recent issue in communities of NLP and CL, the existing literature in this field is vast and not uniform. The subjective perception of the issue has, indeed, caused various interpretations of the term *hate speech* and certain vagueness in the use of related terms such as abusive, toxic, dangerous, offensive and aggressive language [Poletto et al., 2021, Vidgen and Derczynski, 2021].

A helpful reflection on the overlapping of *hate speech* with other terms is provided by Waseem et al. [2017]. They define similarities and differences between the various *subtasks* in abusive language detection taking into account two primary factors: the type of target and the degree to which it is explicit. The former, relying on sociological literature [Weber, 1968], furnishes an interesting distinction between *individual* or entity (for example specific community online) targeted by cyberbulling and trolling, and the generalized *other* or group [Wimmer, 2013] with certain ethnicity or protected characteristics targeted by racism, homophobia, or misogyny. The latter implies the linguistic and semiotic definitions of *denotation* (literal meaning) and *connotation* (sociocultural associations or assumptions) [Barthes and Lavers, 1993]. On the one hand abusive language could be unambiguous and explicit, on the other one it implies some connotations that are difficult to interpret as abusive for the lack of profanities or negativity, the use of ambiguous terms or rhetorical elements (i.e., sarcasm) that recall contextual knowledge (Example 2) or stereotypes (Example 3 and 4) making offensiveness indirect [Dinakar et al., 2011].

Following the typology delineated by Waseem et al. [2017], in this work we adopt the term *abusive language* as an umbrella term to enclose the variety of hateful discourses linguistically studied and computationally processed. Figure 1.1 from Poletto et al. [2021] shows perfectly our representative intuition of abusive and toxic language and relative concepts.

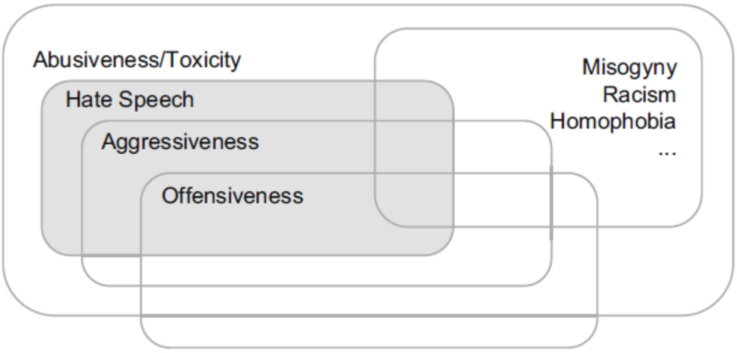


Figure 1.1 – Representation of Relations among different Abusive Language Subtasks.

With respect specifically to *hate speech*, one of the most complete definitions is provided by Sanguinetti et al. [2018d]: a content is considered hateful on the basis of its *action* and its *target*. The action is the illocutionary act of the utterance aimed to spread or justify hate, incite violence, or threat people's freedom, dignity, and safety. The target must be a protected group or an individual belonging to such a group, attacked for his/her individual characteristics.

This definition could be complemented with the definition proposed by Fortuna and Nunes [2018] that puts the emphasis on the linguistic style that makes hate speech explicit and implicit such as humour (recalling the factors underlined by Waseem et al. [2017]):

> Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used. [Fortuna and Nunes, 2018, p. 5]

The necessity to underline the possibility to express hate speech in a very implicit way using humorous exclamations or sarcastic utterances, is due to the harshness of some jokes. Kuipers and Van der Ent [2016] investigated on the seriousness of ethnic jokes. Some people, as well as scholars [Davies, 2011], consider the ethnic joke not serious, whereas others stress the importance of the context and, thus, the relation between stereotypes-based jokes with the social exclusion of the group targeted. This idea comes out by the possibility that jokes can reinforce negative stereotypes and foster, especially online, the spread of hateful discourse leading to serious consequences [Weaver, 2013]. In this study, and also in our point of view, the seriousness of humour relies principally on the rhetoric of jokes (when are related to hostility, exclusion, or hierarchies) and on the hurtfulness of their content (that could have psychological repercussions [Douglass et al., 2016]).

The existing surveys on abusive language detection [Schmidt and Wiegand, 2017, Fortuna and Nunes, 2018] underline the necessity to approach computationally the implicitness of toxic discourses, especially in cases where these discourses are disguised by sarcasm, euphemism, rhetorical questions, litotes, or where there is no explicit accusations, negative evaluations or insults. This kind of implicitness eludes the abusivity of language making its recognition hard for machines and even for humans [Wiegand et al., 2021].

1.1.2 Ironic Language Detection

The investigation on figurative language from a computational perspective takes root into the necessity to solve semantic disambiguation and, thus, make machines able to understand the real meaning of the verbal and written utterances. Various researchers have run into the limits of literal meaning in the automatic interpretation of: users' opinion about specific subjects (product, service, organization, or event) [Liu, 2012, Cambria et al., 2017]; users' stance about some topics or social issues, persons or political ideology [Küçük and Can, 2020]; sentiment dynamics between characters of plays [Kim and Klinger, 2018]; and recognition of abusive language [Nobata et al., 2016].

Figurative language is, as asserted by Quintilianus [2020], a type of language that moves away from the usual and conventional modes of expressions. One of the most recent rhetorical perspectives, led also by a more general semiotic interpretation, tends to distinguish between figures related to the *expression plane* and figures related to the *content plane* [Dubois et al., 1970]. In linguistics, the former concerns the morphological and syntactic manifestation of language; the latter covers the semantic and logic interpretation of language. Irony is placed among these last figures and, in particular, among *metalogic figures* that are the figures that modify the logic value of the utterance breaking the maxim of quality [Grice, 1975] and affecting the literal meaning [Garavelli, 1997].

Therefore, ironic language relies on the inference of information that can go beyond the lexical, syntactic and also semantic knowledge of the words. Looking at the following examples, we notice that to solve ironic interpretation we need to understand: the oxymoron between words *fondatore* (founder) and *affondatore* (sinker) (Example 5); the mechanisms of attenuation of highly positive concepts (stimulated by the adjective *solo*, the negation and the rhetorical question [Giora et al., 2015a]) and of the consequent polarity reversal [Bosco et al., 2013] (Example 6 and 7); the common reference of words *Maserati* and *Montecarlo* to richness and the knowledge of world related to the fact that the Ministry of Defense does not need this luxury (Example 8); and the analogy between *ultras* and *pregiudicati* (offenders) that evokes information external to the assertion [Sperber and Wilson, 1981] (Example 9):

(5) *Mario Monti: c'è il rischio... di trasformare l'Italia da Stato fondatore in Stato affondatore dell'Unione europea ! URL*[15]

(6) *La destinazione delle vacanze e' il solo problema che non riuscira' a risolvere l'ottimo concergie Mario Monti.*[16]

(7) *#la7 ma perche' Mario Monti non fa il premier? Che persona competente e per bene!*[17]

(8) *Il ministero della Difesa compra 19 Maserati. Prende corpo il piano di invasione di Montecarlo. [giga]*[18]

[15] *Mario Monti: there is a risk ... of transforming Italy from a founding state into a sinking state of the European Union ! URL.* Tweet extracted from IronITA2018 [Cignarella et al., 2018b]

[16] *The holiday destination is the only problem that the excellent concierge Mario Monti will not be able to solve..* Tweet extracted from IronITA2018.

[17] *#la7 but why isn't Mario Monti the prime minister? What a competent and good person!.* Tweet extracted from IronITA2018.

[18] *The Ministry of Defense buys 19 Maseratis. The Monte Carlo invasion plan takes shape. [giga].* Tweet extracted from IronITA2018.

(9) *All'Olimpico si dialoga con gli ultras. In Italia prima di decidere qualsiasi cosa va di moda consultare dei pregiudicati. [@USER]*[19]

Without the activation of these linguistic and cognitive mechanisms of reference, it is difficult to solve the ironic interpretations and to understand the presence of secondary meanings that could be opposite to the literal ones (such as in the case of Examples 6 and 7). A superficial interpretation of texts, especially in delicate issues such as abusive contexts, must be avoided.

Looking in particular at Example 6 and 7, we can observe a marked use of highly positive words to infer a very negative message. One of the most common figurative devices used to sugar-coat negative meanings is sarcasm [Liu, 2012, Wang, 2013]: a specific type of irony that aims to mock and scorn a victim. In particular, Lee and Katz [1998] demonstrates that, differently from other forms of irony, sarcasm is used to ridicule a specific target (i.e., Mario Monti in Example 6 and 7).

In spite of the amusement provoked by sarcastic texts, the ironic sharpness of sarcasm is perceived as offensive by victims. Bowes and Katz [2011], for example, noted that the targets of sarcastic utterances do not perceive these expressions as humorous, differently from their aggressors. This study seems to be in contrast with the line of some scholars that stress the 'muting the meaning' hypothesis that considers ironic language as a device to mute the negative meaning [Dews and Winner, 1995].

In this regard, Pexman and Olineck [2002] proposed a pragmatic analysis of ironic insults and ironic compliments and how they are perceived by social impression. Ironic insults are perceived as more polite whereas ironic compliments as more mocking and sarcastic: speakers tend to criticize someone lowering the social cost of doing so, and ironic language seems appropriate to cover the scorn.

This is observed especially in abusive context, such as debates online on immigration and presence in Italy of minorities like Roma and Muslim communities. In the following examples, we can notice that ironic language could in some cases lessen the tones but in the majority of cases enhances the negativity of the message [Colston, 1997] or create some kind of in-group identification [Bowes and Katz, 2011]:

(10) *@USER Mi hanno insegnato che non tutti i musulmani sono terroristi ma il 99% dei terroristi nel mondo sono musulmani.*[20]

(11) *Complimenti agli islamici x aver relegato le donne sotto un tendone alla loro festa. Una religione comprensiva, giusta, ecc Ma per favore.*[21]

[19] *At the Olimpico we talk with the ultras. In Italy, before deciding anything, it is fashionable to consult with offenders. [@USER].* Tweet extracted from IronITA2018.

[20] *@USER They taught me that not all Muslims are terrorists but 99% of terrorists in the world are Muslims.*. Tweet extracted from IronITA2018.

[21] *Congratulations to the Muslims for having relegated women to a marquee at their party. An understanding religion, just, etc. But please.*. Tweet extracted from IronITA2018.

(12) *Un pensiero di ringraziamento ogni mattina va sempre ai comunisti che ce li hanno portati fino a casa musulmani rom e delinquenti grazie*[22]

(13) *@USER il nuovo stile di vita invocato dalla Boldrini! che retrogradi noi non-musulmani....non accogliere questa ricchezza culturale!*[23]

1.2 Motivation and Objectives

To support the activities of counter hate speech online, such as awareness raising projects[24], the collaboration of systems of AI that quickly monitor big amounts of data in real time represents an important help to community.

However, the process of comprehension of the meaning of what people write is not easy. In social platforms, the users want that their opinions are understood by the majority of people [Turner, 2010], and, especially when some limitation of characters are established, the necessity of conciseness and openness encourages users to use creative devices to express clearly his/her thought, such as: visual elements (emoticons and emojis) to symbolize various paralinguistic expedients; graphic details (punctuation and capital letters) to convey the prosodic level of language; spontaneous language (dialectal forms, contractions, colloquialisms, abbreviations); and rhetorical expedients that, like metaphor, irony or euphemism, could express in a compact fashion a longer message. All these devices respond to the criteria of space-savings, functionality and efficacy typical of digital writings [Chiusaroli, 2017].

Especially in case of negative and hateful opinions, Sanguinetti et al. [2018d] noticed that users tend to be less explicit in their claims in order to limit their exposure. This implicitness, as seen above, materializes sometimes in the use of ironic language such as sarcasm:

(14) *Vabbè, come dice la Boldrini i nomadi sono una risorsa per tutti.... al pari degli immigrati africani e dell'est Eu... URL*[25]

Its ironic sharpness and its echoic function of recalling a meaning that is the opposite or an extension of the literal one, make sarcasm appropriate to lower tones without losing the hurtfulness of the message. Moreover, funny messages are more likely to be accepted and shared by the community, making the abuse viral.

[22] *Each morning, I would like to thank communists who bring home musulmans, roms and delinquents thanks.* Tweet extracted from IronITA2018.

[23] *@USER the new lifestyle invoked by Boldrini! how retrograde we non-Muslims .... not to welcome this cultural richness!.* Tweet extracted from IronITA2018.

[24] For example, the ones activated by NGOs, such as Amnesty International, or supported by the EU, like 'Silence Hate' in schools `https://www.amnesty.org/en/latest/education/2020/04/silence-hate-students-in-italy-use-art-to-create-a-campaign-against-online-hatred/`

[25] *Oh well, as Boldrini says, nomads are a resource for everyone .... like African and Eastern Eu immigrants ... URL.* Tweet extracted from HaSpeeDe2020.

However, understanding implicit meanings of natural language is in general a hard task, and systems trained to distinguish hate speech from non-hurtful are not aware of probable sarcastic meaning. In our experience, we noticed that even the novel computational techniques, that exploit pre-trained models to generalize better, show difficulties to interpret correctly these translations of meaning. Indeed, these techniques that are principally data driven, miss the linguistic intuition typical of humans. This intuition, indeed, activates complex semantic mechanisms to understand the intentional meaning of the message. Nevertheless, human intuition is strongly influenced by cultural background [Basile, 2020b] that pours also on the ground-truth datasets used to train and test automatic systems, creating biased models of language understanding. For this reason, computer technologies need to be informed adequately to acquire linguistic knowledge limiting as much as possible negative prejudices.

1.2.1 Research Questions

Considering this premise, our principal question is: *How do we provide machines with linguistic intuition to detect abusive language in indirect contexts?* To answer it, we elaborate specific research questions that help us to focus on specific steps of analysis and face single problems.

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?

RQ2 What is the role played by sarcasm in hateful messages online?

RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

1.2.2 Objectives

In our investigation, to answer the previous research questions, we propose to reach the following objectives:

1 Investigating the characteristics of implicit manifestations of hate speech and examining, in terms of performance, the techniques that could help systems to infer them, such as:

 1.1 using lexical resources that allow systems to consider specific words or sequence of words whose meaning reflects negative stereotypes and prejudices;

 1.2 using distributional semantics' metrics and models to capture semantic relations between words and documents even in different textual genres;

 1.3 using transformers-based techniques that make systems more sensitive to style and semantics of the informal writings and combining them with linguistic features;

1.4 using multi-task learning approaches to increase the acuteness of systems towards stereotyped biases.

2 Analyzing the role of ironic language in hateful texts:

2.1 creating the adequate corpora to investigate the issue;

2.2 observing the multilingual characteristics of irony;

2.3 distinguishing the linguistic and cognitive traits of sarcasm respect to other forms of irony;

2.4 validating, with experiments of classification, these traits in terms of features (such as statistical measures to identify sentiment polarity reversal, variations of emotions within the text revealing the ironic contrasts, or hurtful words).

3 Evaluating the benefits of ironic awareness in hate speech detection:

3.1 exploiting computational techniques that make systems aware of ironic language, such as the multi-task learning approach that enables systems of abusive language detection to acquire specific knowledge about ironic language;

3.2 measuring the significance of the obtained results in comparison to existing approaches and baseline models.

1.3 Thesis Contributions

Considering the issues faced in this thesis, below we delineate the principal theoretical and technical contributions of our work such as methodologies, resources, and benchmark datasets that could be used by other scholars to approach and deepen *the detection of abusive language.*

a We examined explicit and implicit forms of abusive language against women and immigrants in various languages such as Italian, Spanish, and English, noting that:

- the stereotypes about women, that enforce the patriarchal social order and the dehumanization of women, are perceived as expressions of discrimination, and they are often present in the attacks against women. Differently from misogyny, in racist context the stereotypes not always occur in hateful messages, especially in newspapers headlines.
- in debates on sensible social issues that involve a specific group of people, such as legalization of abortion or feminist manifestations, the expression of unfavorable stance often disguises abusive messages even in newspaper articles.
- hate speech against immigrants appears differently expressed in social media posts and news headlines. In particular, headlines tend to be characterized by a nominal syntactic structure that recalls the slogan's one.

  - the linguistic information about specific lexical knowledge and the awareness of the presence of stereotypes make systems, based on classical or deep learning architectures, significantly sensible to recognize correctly abusive language.

b Focusing on the difficulty of systems to detect abusive language when expressed with ironic devices, we investigated the characteristics of sarcasm as a specific form of irony, specifically examining its role in abusive context. We observed that:

  - some specific affective, rhetorical and pragmatic elements common in various languages (Italian, Spanish, and English) tend to trigger ironic interpretation in supervised approaches, such as negative emotions, hyperboles, oxymoron, and context shift.
  - sarcasm, especially in abusive context, appears to be characterized by offensive words, aggressive language and sentiment shifts, revealing to be apt to transmit hurtful messages.
  - informing systems with these characteristics, the performance in the recognition of irony and sarcasm in user-generated contents ameliorates, creating a reference for the following computational approaches in irony and sarcasm detection.

c Taking into account the role of ironic devices in abusive context, we made systems aware of the presence of irony and sarcasm exploiting a multi-task learning approach, and we noticed that:

  - training abusive detection systems, especially, on sarcasm recognition, the performance significantly improves in spontaneous texts such as tweets.

d To approach abusive language detection in Italian, English, and Spanish, we explored various computational techniques ranging from classical to deep learning based architectures, evaluating principally the contribution of data-driven approaches respect to more complex informed systems. We tried to translate the linguistic and cognitive mechanisms of comprehension of the intentional meaning of the message into formal and technical solutions. In particular, we designed:

  - systems based on classical and deep learning architectures, led mainly by specific lexical features. To this purpose, we manually created multilingual core-lexica (for Italian, English, and Spanish) that have been also extended automatically using GloVe [Pennington et al., 2014b] and TWITA [Basile and Novielli, 2014] embeddings. For the Mexican variant of Spanish, that is a low-resource language, we created also some lexica of implicit and explicit offensive and derogatory expressions.
  - systems that combine general knowledge, coming from various language models, and linguistic information, obtained with specific linguistic features, or

with the injection of ironic language recognition within a multi-task learning framework.

e We created the following resources.

  - With respect to the analysis of characteristics of explicit and implicit forms of the abusive language, we collaborated in the creation of the second edition of the HaSpeeDe benchmark corpus [Sanguinetti et al., 2020], called here HaSpeeDe2020, in occasion of EVALITA 2020 with the aim to encourage the detection of hate speech and stereotypes in Italian tweets and newspapers headlines against minorities such as Muslims, Romas and migrants. The annotation of this dataset was, successively, extended, taking into account other phenomena such as aggressiveness and ironic language. This additional annotation was useful to reflect on the implicitness of hateful comments and propose a computational approach to detect it.
  - About the examination of debates online on sensible social issues in newspapers online we created a dataset from GDELT[26]. In particular, we collected news articles in English about feminist movements related to events happened from the 1st of October to 31st of December in 2017 in Europe, Japan and USA, linked to the viral spread of #metoo in occasion of the legal case of sexual assault and harassment in the workplace risen against Harvey Weinstein. This corpus is called here GDELT-FM.
  - To support the investigation of the linguistic characteristics of sarcasm, we collaborated in the creation of the corpus for IronITA, called here IronITA2018. IronITA2018 is a benchmark Italian corpus [Cignarella et al., 2018b] released for EVALITA 2018 to stimulate the reflection on the peculiarities of sarcasm at a computational level and on the possible differences with the irony detection task. In particular, we gathered Italian tweets from Hate Speech Corpus (hsc) [Sanguinetti et al., 2018d] and twittirò corpus [Cignarella et al., 2017] in order to examine also the contribution of irony and sarcasm in two different contexts: political and abusive. To this purpose, we exploited the multi-source composition of this dataset and retrieved the original labels of the hsc and twittirò corpora.

1.4 Structure of the Thesis

The chapters of this work are grouped in three principal parts:

I Abusive Language Detection: Chapters 2 and 3;

II Irony and Sarcasm Detection: Chapters 4 and 5;

III Abusive and Ironic Language: Chapters 6 and 7.

[26] The acronym of Global Database of Events, Language, and Tone Project supported by Google Jigsaw: https://www.gdeltproject.org/

Chapter 1

The first chapter is the introductory section, where we describe the social problems related to the new technologies, introducing the issue of the *abusive language* and the difficulties to detect it automatically. We define also the main hypothesis of our work and its related research questions that will be answered in this thesis.

Chapter 2

In the second chapter, we define the concept of *abusive language*, looking at the juridical and linguistic theories. Moreover, we resume the state of the art from a computational perspective, focusing especially on the studies that approach misogyny, hate speech, stereotype, and aggressiveness detection that are the specific tasks addressed in this work. Finally, we define also the open challenge of abusive language detection about its implicit forms, reporting some of the few studies that recently addressed it.

Chapter 3

In the third chapter, we describe the linguistic, statistical and computational analysis performed on benchmark datasets to individuate the characteristics of the explicit and implicit manifestations of hate speech and the linguistic elements that make hateful content indirect and difficult to recognize automatically. In this section, we describe also various linguistic resources created manually and the designed approaches, based on classical and deep learning algorithms, that make systems able to infer indirect abusive messages (**RQ1**). Finally, we describe the second edition of the HaSpeeDe shared task organized at EVALITA 2020 on hate speech and stereotypes detection in Italian tweets and news headlines.

Chapter 4

In the fourth chapter, we define what is *ironic language*, looking at the linguistic theories stretching from pragmatic to cognitive studies. In addition, we introduce the state of the art on irony and sarcasm detection, focusing especially on studies that analyzed, linguistically and statistically, the peculiarities of sarcasm and on the emotions that move the expression of irony.

Chapter 5

In the fifth chapter, we propose statistical and computational analysis to individuate the characteristics of irony and sarcasm. We observe, in particular, linguistic traits of irony from a mono and multi-lingual perspective, emotional and aggressive language involved in the expression of irony and sarcasm especially when the topic of the text regards controversial issues such as the integration of cultural minorities (**RQ2**). In this chapter, we describe also our experience as organizers of the IronITA shared task at EVALITA 2018 on irony and sarcasm detection.

Chapter 6

Taking into account the findings emerged from previous chapters, in the sixth one, we propose a new computational approach that exploit the simultaneous learning from abusive and ironic language to detect hate speech in Italian tweets and news headlines. The results show an interesting improvement of the performance, especially in hate speech detection in tweets (**RQ3**).

Chapter 7

In the last chapter, we report the obtained results and the observations emerged from our analyses. We individuate also the remaining challenges that we want to address in further works, and we summarize the contributions to the NLP community in terms of findings, methodologies, resources, and publications.

Part I

Abusive Language Detection

Chapter 2

The Language of Hate

In this chapter, we draw a multidisciplinary background on abusive language and in particular on *hate speech*. This kind of speech that instigates in particular to violence and discrimination is prosecuted in various countries as offense (such as the propaganda against a specific group)[1], and, in general, the purpose of discrimination and hate in the offenses is considered as an aggravating circumstance[2]. The ECRI in its 15th General Policy Recommendation on Combating Hate Speech[3] has expressed the need of defining distinctly the offenses and taking effective actions against this kind of speech in each country of EU. Moreover, the ECRI has insisted on the necessity to recognize hate speech as a problem that affects our societies, reinforcing the social asymmetries and injustices. Indeed, the *language* is another place where the discrimination takes shape. In some speeches, specific social groups are positioned in a sort of social scale on the basis of their attitudes or characteristics [Bianchi, 2021]. This ranking makes the 'lower' groups dehumanized or not able to be considered similar to 'higher' groups. This process is supported also by negative *stereotypes* that collect uncritical judgments about the perceived *outgroup* and that easily could be assimilated by the majority of people [Fiske, 1998].

Adopting a multidisciplinary perspective, in the next sections we provide a linguistic and philosophic interpretation of hate speech starting from a problem, even juridical, of definition. On the basis of the typology delineated by Waseem et al. [2017] (see Section 1.1.1), we distinguish the various hateful discourses underlining the computational

[1] In Italy, the Reale's Law, Mancino's Law and the Law n.85 of 2006, reproduced in the Art. 604-*bis* of the Penal Code, criminalize the instigation and the act of racial discrimination. About the other forms of manifestations of hate based on disabilities, genre or sexual orientation, there are ongoing proposals on the possibility of extension of article 604-*bis* and -*ter* of the Penal Code. The need of specifying all these forms of manifestations of hate emerged also for Spanish Penal Code from the ECRI report of 2018: `https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/spain`.

[2] In Italy it is established by the Mancino's Law of 1993 and reproduced in the Art. 604-*ter* of the Penal Code only in case of racial, ethnic and religious hate. Whereas in the Art. 22 of Spanish Penal Code, the aggravating circumstances already include all the forms of hate.

[3] See footnote 1 in Section 1.

approaches more representative of the new technologies and the benchmark datasets created for Italian, Spanish, and English.

2.1 Theoretical Background

As seen in Section 1.1.1, deciding what is *hate speech* is difficult; and in the NLP and CL communities, as well as in the juridical one, an unambiguous definition is missing. The American legal scholar Dean Robert C. Post in Hare and Weinstein [2009] considers that to prosecute legally a hateful expression, it is essential that it is qualified as *extreme* intolerance or dislike. To justify a legal intervention, hate speech must be defined both in terms of expression of abhorrence and in terms of elements that underline the presence of *extreme hate*. Post roughly individuated two principal elements: the 'manner of speech' and the emphasis of causing a contingent harm against the victim. The former involves not only its content, but also its *style of presentation* that aims to degrade and insult. The latter takes into account the *harmful effects* of the speech.

Especially in the Recommendation n. R (97) 20 to Member States on "Hate Speech", the European Committee of Ministers, with the purpose of guaranteeing both the freedom of expression and respect of the human dignity and reputation, underlines that the courts should "bear in mind that specific instances of hate speech may be *so insulting* to individuals or groups as not to enjoy the level of protection afforded by Article 10 of the European Convention on Human Rights" (Principle 4). In this legal framework, they proposed a first definition of the term *hate speech* understood as covering:

> all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants, and people of immigrant origin. [Recommendation n. R (97) 20, 1997, p. 107]

This definition has been reconsidered by ECRI in the Recommendation n. 15[4] extending in particular the range of manifestations and types of hate underlining, always, that "the right to freedom of expression can and should be restricted in *extreme* cases" (such as the instances of incitement to hatred that involve intent, content, extent, probability of harm occurring, imminence and context). They consider as hate speech

> the advocacy, promotion or incitement, in any form, of the denigration, hatred, or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat with respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of 'race', color, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual

[4] See footnote 1 in Section 1.

> orientation and other personal characteristics or status. [Recommendation n. 15, 2015, p. 3]

In this recommendation, the specific *style of presentation* of hate speech (*harassment, insult, negative stereotyping, stigmatization or threat*) and its *harmful effects*, such as “to incite, or reasonably expected to have the effect of inciting others to commit, acts of violence, intimidation, hostility, or discrimination” (2015, p. 3), are better defined.

The term *hate speech* appears for the first time in the battlefield of the Critical Race Theory (CRT) movement at the end of the 80s’. This movement has aimed at challenging the ability of the American legal system affected by racist ideology to deliver social justice. One of the most important critiques of CRT was towards the First Amendment of United States Constitution, considered formal but not coherent with the reality. In particular, Delgado [1994] argues mainly three points:

1 The social ills, such as racism and sexism, are embedded in the set of conventions of the society and, thus, reflected in the *marketplace of ideas* guaranteed by the 1st Amendment leaving that who speaks against these ills appear incoherent and his/her speech less effective.

2 A perfect marketplace of ideas is based on the condition that social power and resources are equally distributed in the society. However, this condition is not fair and the prevalent system of ideas makes some people less credible than others.

3 Some speech should be viewed in terms of the harmful effects it causes, rather than valued only on the basis of being a speech. Indeed, for instance, the incessant depiction of a group as lazy or stupid constructs a *social reality* that is disadvantageous for the targeted group.

On these bases, CRT have required a hate speech regulation that has not been approved.

Looking at these augmentations from a linguistic and philosophical perspective, we can resume them saying that people do things with words, even hurt. Scholars, like Austin [1975], argued the presumed neutrality of language and of the context where the speech takes place. With language, we can create a biased reality that affects the power of some groups and, consequently, limits also the effectiveness of their speeches. This limitation is defined by Bianchi [2021] as *discursive injustice*. Bianchi [2021] distinguishes mainly two types of discursive injustice: the illocutionary distortion and the reduction to silence. In the first case, an individual of a penalized group does with his/her words something differently from what expected (for example, when an order of a female manager/boss is perceived as request by listeners). In the second, the speaker has not illocutionary force and his/her words have no effects (for example, when a woman says ‘no’ to sexual proposal and her refusal is not taken into account). The *illocutionary act* is defined by Austin [1975] as the act that speaker realizes with his/her speech: if the judge sentences a person as guilty, he/she will be guilty whether or not he/she committed the murder; the sentiment of frustration of the guilty or the joy of the family of the victim belong to the sphere of the *perlocutionary act*. The latter is related to all the extra-linguistic effects provoked by the speech:

(15) *@USER @USER @USER La grande differenza dei nostri antenati immigrati per lavoro, umili, onesti e grandi lavoratori. con i parassiti africani che oggi clandestinamente arrivano e vogliono essere mantenuti vita natural durante da noi Italiani.*[5]

(16) *Guai a voi se permetterete uno scempio del genere. Questi cialtroni non devono arrivare in Italia. Sig. Ministro, cacci via subito tutti gli islamici, abbatta tutte le moschee. Liberiamoci da questi terroristi e codardi che picchiano le donne perché dagli uomini le prenderebbero.*[6]

Therefore, defining immigrants as parasites (15) and Muslims as terrorists (16) the users produce a *normative effect*, related to the illocutionary act, and at the same time a *causal effect* related to the perlocutionary act [Bianchi, 2021]. The first helps to create, legitimate and reinforce the beliefs, the subordination and the stereotyped ideas about the outgroup, and the second relies on the production of a behavior of discrimination, harm, and the damage against the target. These effects do not change in the implicit expressions of abuse:

(17) *RT @USER call me sexist if you want, but I find female sportscasters really annoying.*[7]

(18) *I carabinieri hanno individuato come possibile spacciatore un 27enne del Marocco. La tipica #risorsa straniera, ammiro la madre! URL*[8]

Examples like (17) and (18) tend to propagate, on the one hand, the sexist ideology that justify the patriarchal social order, and on the other hand the stereotype that sees immigrants as criminals. Bianchi [2021] individuates especially two dimensions of abusive language: an *evident* dimension that consists in the 'verbal violence' that evokes the 'physical violence' appearing explicitly aggressive and offensive; and a dimension that could be called *of propaganda* that aims at attesting the social identity presenting some roles or assumptions as normal/conventional appearing not only aggressive but as a form of proselytism of negative idea (see for instance tweet 17). Especially this last

[5] *@USER @USER @USER The great difference of our immigrant ancestors for work, humble, honest and hard workers. with the African parasites that today clandestinely arrive and want to be kept by us Italians for life.* Tweet extracted from HaSpeeDe2020.

[6] *Woe to you if you allow such a mess. These scoundrels must not arrive in Italy. Mr. Minister, immediately drive out all the Muslims, tear down all the mosques. Let's get rid of these terrorists and cowards who beat women because they would be beaten by men.* Tweet extracted from HaSpeeDe2020.

[7] Tweet extracted from the NAACL_SRW_2016 corpus released by Waseem and Hovy [2016].

[8] *The Italian police have identified a 27-year-old from Morocco as a possible drug dealer. The typical foreign #resource, I admire the mother! URL.* Tweet extracted from HaSpeeDe2020.

dimension has been extensively studied by social and legal scholars, focusing especially on the illocutionary force of public speeches.

As said above, the illocutionary force of the speech have to be recognized by listeners to realize *happily* the illocutionary act of the speech (otherwise it is a case of discursive injustice). To this purpose, the speaker needs to have the authority (such as the judge) and to enunciate the speech in an adequate context (i.e., courthouse). These two variables are taken into account by Leader Maynard and Benesch [2016] as elements to establish the dangerousness of a speech. Benesch [2012] defines as *Dangerous Speech* every act of speech that "has a reasonable chance of catalyzing or amplifying violence by one group against another, given the circumstances in which it was made or disseminated". This *chance* materializes when the circumstances in which the speech takes place consist of: 1) a powerful speaker or source with a high degree of influence, 2) an audience that believe to be subject to a threat, 3) a social and historical context propitious for the violence, 4) the means of dissemination (radio, newspapers, social media, or a specific language), 5) the content of the speech that aims at the process of dehumanization, guilt attribution, threat construction, destruction of alternatives, creation of a new semantics of the violence conceived as admirable, linked to praiseworthy qualities and based on specific biased references that justify it [Leader Maynard and Benesch, 2016]. Their studies, based on the analysis of speeches disseminating ideologies that played a critical role in the realization of mass atrocity crimes, give us the possibility to reflect on the climate of violence that certain assertions could create when spread in social platforms or more credible media such as newspapers.

Therefore, the normative effect of abusive language, especially in its form *of propaganda*, brings with it an atmosphere of intolerance, and its dangerousness lies in the implicit legitimization or justification of prejudiced behaviors against the perceived outgroup (such as Examples 17 and 18).

2.1.1 Social and Traditional Media

Increasingly, social platforms along with traditional media are becoming dominant sources of information for the majority of people with the risk that they can come to believe in dangerous ideological claims; especially when they are 'trapped' into echo chambers fed by false and hateful messages[9]. A practical example is the spread of racial hoaxes that aimed "to circulate information that is an allegation of a threat posed against someone's health or safety associated to an individual or a group because of race, ethnicity, or religion" [Russell-Brown, 2009].

Like racial hoaxes, also some newspapers contribute to the dissemination of negative evaluations about specific categories of people. Their formal language and the need to

[9] A first reflection about the deep connection between fake news and hate speech is provided in Scamuzzi et al. [2021]

respect the deontological norms tend to make the references to the common stereotypes more implicit, as well as to the opinion of the journalists:

(19) *Il regno di immigrati e no global: “Ecco l’anticamera dell’inferno”*[10]

(20) *Calci e pugni alle auto parcheggiate in strada: arrestati due immigrati*[11]

In these headlines of articles published online in 2019, we notice that also traditional media could contain explicit (19) and implicit (20) toxic speeches reporting the same normative and causal effects of tweets showed as examples from (15) to (18). Therefore, even news could not be considered as ‘transparent’ messages, but like messages of users in social media could conceal prejudices feeding the social inequality. This easy sharing of stereotyped ideas, and thus also news and hoaxes containing conventional beliefs about a group or individual, is due to their fast assimilation. Fiske [1998] reported various psychological studies showing that people, in general, attend longer to stereotype confirming than disconfirming their idea on the perceived outgroup and ingroup.

As false and toxic speeches, journalists could express their stance in reporting the news. An interesting case that captured our attention is when the stance is in favor or against a social issue that involves a specific group of people such as women in themes related to legalization of abortion, gender violence and feminist movement. In fact, in social media communication, it is typical that some users use strong and offensive words to express their stance against a specific event that involves individuals or groups (such as Example 1). In traditional media, the expression of stance for deontological reasons appears to be moderate and implicit enough:

(21) headline: *Women need to free themselves from permanent victimhood*
text: *If there is one thing the reactions to the Harvey Weinstein accusations have confirmed, other than the common knowledge that human beings are corruptible and will sometimes try to exploit their position of superiority, it is feminism’s obsession with men in power. [...] At least one of the reasons for this is, quite logically, the reliance on the male villain for the rationalisation and validation of the position to which these women are clinging in order to avoid facing the most pressing issue for privileged women when it comes to the lack of gender equality: their own dependency issues. [...] The hysteria and irrationality of the reactions are revelatory of an obsession with male power. [...]*[12]

Especially in Van Dijk [1991], the role of the media in the reproduction and proliferation of toxic and in particular racist ideology, is well defined. The linguist Van Dijk noticed that some events related to ethnic stories or issue are described by newspapers reporting

[10] *The kingdom of immigrants and no global: “Here is the antechamber of hell”.* Tweet extracted from HaSpeeDe2020.

[11] *Kicks and punches to cars parked in the street: two immigrants arrested.* Tweet extracted from HaSpeeDe2020.

[12] Article extracted from GDELT-FM.

the perspective of dominant political group or other elites (such as the perspective of officials rather than the opinions of black people involved in the event). It seems that despite the neutral point of view theoretically adopted by journalists, they tend to report stories that align with their preconceptions about the outgroup. In Van Dijk [2000], he underlined that issues related to outgroups, such as immigrants, tend to assume a negative dimension. For example, immigration often is topicalized as a threat or in relation with crimes, drugs and so on. Whereas, many other topics that are also part of the ethnic affairs, such as the political and social situation in other countries or the everyday life in communities for migrants, occur much less in the news.

Media *discourse*, both in traditional and new form, is the main source of daily people's knowledge and its influence is very powerful. The authority of newspapers as credible source is exploited by users to confirm their fear and their beliefs in social media, like:

(22) *Un grave episodio che evidenzia i danni dell'#immigrazione senza freni che provocano anche allarme sanitario. Ogni giorno di più scopriamo quanto è pericoloso il business dell'#accoglienza!* ***https://t.co/FZZFGiRcJy***[13]

The communicative strategies employed by users, especially in social platforms, are various. As seen in Chapter 1, to respond to necessities of shortness and immediate comprehension limiting their exposure, users tend to make their messages implicit using for example ironic language (see, for instance, tweets from 10 to 14). The limitation of exposure, and the consequent sensation of impunity, is also facilitated by the possibility to make their identities anonymous using false profiles devoted to spread hate [Fox et al., 2015].

Specialists of politics and philosophy recently have tried to define the elements that make online hate speech *special* compared to offline hate speech. Brown [2018] commented some specific features of the Internet, finding in particular that the instantaneous and spontaneous nature of some platforms encourage users to write even hateful messages, along with the ease of use and relatively low cost of being online. Their diffusion is then supported by the amount of possible readers, and especially by the virality and the permanence of the abuse online [Foxman and Wolf, 2013]. The instantaneousness and spontaneity that characterize social media communication as well as the implicitness and formality of newspapers online motivate our interest of investigation towards these two particular media.

The study of the phenomenon of the spread of hatred and its consequences, in recent years, has involved various disciplines that together are trying to formalize and counter

[13] *A serious episode that highlights the damage of unrestrained #immigration which also causes health alarms. Every day, we discover how dangerous the #hospitality business is!* ***https://t.co/FZZFGiRcJy***. Tweet extracted from HaSpeeDe2020. Differently from other examples, we do not hide the link of this tweet to allow the reader to understand the discursive strategy employed by users to assess their beliefs about the outgroup.

the many facets that hatred acquires. Unlike other contexts, the hatred expressed online has an uncontrollable resonance: any content spread on social media can become viral even when you do not have many followers thanks to the continuous interaction. The widespread content has an eternal permanence on the web and if it is a content that harms someone, the victim may never feel safe again. Furthermore, especially social platforms, as seen above, alongside democratization, favor the *deresponsibility* of communication which brings with it a feeling of impunity also guaranteed by the possible anonymity and invisibility. This deresponsability is encouraged, also, by the difficulty of feeling empathy towards the victim, due to the fact that online communication takes place in front of a screen and not in front of the attacked person. The effects of this *online disinhibition* were observed by Lapidot-Lefler and Barak [2012] who studying in particular factors such as anonymity, invisibility and lack of eye-contact realize that if, on the one hand, together these factors lead to flaming behaviors, on the other one, the lack of eye-contact appears as the main contributor to the negative effects. And, as seen above, the consequences are varied and can easily lead to violent offenses that have repercussions in the lives of the victims. To prevent and avoid these consequences, the efforts of the NLP and CL communities, in recent years, have intensified, giving rise to various corpora and linguistic resources as well as an extensive series of models to automatically recognize abusive language online.

2.2 Open Challenge: Implicit Abuse

As seen in previous sections, abusive language could be expressed with different degrees of denotation: explicitly (*evident dimension*) and implicitly (*dimension of propaganda*). If the first dimension (when they contain an aggressive tone and offensive language like in Example 23) is easier to recognize automatically, the second one implies cognitive and creative aspects harder to be recognized even by humans. For example, they can involve stereotypes (Example 24), be devoid of explicit linguistic patterns against the target and express a prejudice that hurts (Example 25) [Breitfeller et al., 2019] or mask the abuse using ironic language (Example 26):

(23) *«Profughi in fuga dall'inferno», si sono dichiarati davanti alle autorità. Ma l'inferno lo hanno portato loro a casa nostra. Secondo gli accertamenti medici, Desirée è stata prima drogata con un mix di stupefacenti e poi è stata violentata per ore. URL*[14]

(24) *#BeatriceLorenzin del #Pd : "Vaccinate i vostri figli perchè gli immigrati (#risorseINPS) riportano malattie scomparse." Al governo sono consci degli enormi rischi sanitari che arrivano dall'Africa, eppure continuano a tenere i porti aperti. Incredibile. #CoronaVirus #RadioSavana URL*[15]

[14] *«Refugees fleeing from hell», they declared themselves before the authorities. But they brought hell into our house. According to medical findings, Desirée was first drugged with a mix of drugs and then raped for hours. URL.* Tweet extracted by HaSpeeDe2020.

[15] *#BeatriceLorenzin from #Pd: "Vaccinate your children because immigrants (#risorseINPS) report missing diseases." The government is aware of the enormous health risks coming from Africa, yet*

(25) *Ha ragione Salvini. L'Italia non é in grado di ospitare chi non invita. I migranti possono bussare alle nostre porte, ma noi non siamo obbligati ad aprirle a chiunque.*[16]

(26) *Miracolo a #Milano, la mendicante rom guarisce e torna a camminare. @utente @utente #RadioSavana #resilienza20 ?? #risorseINPS URL*[17]

Especially, the messages with a higher degree of connotation involve various linguistic and non-linguistic elements that could be recurrent but difficult to recognize automatically. In the examples above we find: the rhetoric of fear that justify intolerance towards immigrants (*But they brought hell into our house*), stereotypes and prejudices that depict immigrants as carriers of diseases (*enormous health risks coming from Africa*) or unwanted invaders (*Italy is unable to accommodate those who does not invite*), and figurative expedients such as sarcasm that tend to veil the offensiveness of the message. Current systems prove not to be able to recognize correctly hateful messages when intolerant discourses are not evident because they are disguised by rhetorical devices, or they do not contain explicit accusations or prejudices [Schmidt and Wiegand, 2017, Fortuna and Nunes, 2018]. Scholars like Nobata et al. [2016] and MacAvaney et al. [2019] assess the necessity to approach the implicitness of language, observing that these expedients elude the abusivity of language, making its automatic recognition hard.

In a very recent investigation, Wiegand et al. [2021] analyzing various benchmark datasets in English identified specific subtypes of implicit abuse: stereotypes, perpetrators, comparisons, dehumanization, euphemistic constructions, call-for-action, multimodal abuse, and all the phenomena that require world knowledge and inferences such as jokes, sarcasm and rhetorical questions. Some of these subtypes have been identified by scholars as problematic challenges in abusive language detection, demonstrating that only their explicit manifestations are identified by current classifiers (supervised and unsupervised).

For instance, Van Aken et al. [2018] in a detailed error analysis of an ensemble classifier's performance in a Wikipedia[18] and Twitter [Davidson et al., 2017] dataset, identified some of these subtypes as elements that make abusive language difficult to recognize, such as: lack of explicit offenses (such as swear words), idiosyncratic expressions, rhetorical questions, metaphorical and ironic language. As showed also by Wiegand et al. [2019], the performance of classifiers in presence of implicit abuse decreases considerably, with

they continue to keep the ports open. Incredible. #CoronaVirus #RadioSavana. Tweet extracted by HaSpeeDe2020.

[16] *Salvini is right. Italy is unable to accommodate those who does not invite. Migrants can knock on our doors, but we are not obliged to open them to anyone.* Tweet extracted by HaSpeeDe2020.

[17] *Miracle in #Milano, the Roma beggar gets well and starts walking again. @USER @USER #RadioSavana # resilienza20 ?? #risorseINPS URL.* Tweet extracted by HaSpeeDe2020.

[18] This dataset has been published by Google Jigsaw in December 2017 in the context of Toxic Comment Classification Challenge on Kaggle.

some exception regarding those cases where the sampling process introduces data bias in the training and test sets. These analyses that take into account the explicit and implicit portion of abusive documents are carried out looking at the vocabulary of the corpora: a document contains explicit abusive language if at least one word from the lexicon of abusive words [Wiegand et al., 2018] is included.

Caselli et al. [2020] employed a similar approach to analyze OLID/OffensEval dataset [Zampieri et al., 2019] in order to reflect about the notions of explicit/implicit and offensive/abusive. On this reflection, they proposed to apply a new annotation layer on OLID/OffensEval creating AbuseEval v1.0. As expected, the authors showed that the documents annotated as offensive in OffensEval overlap largely with the documents annotated as explicitly abusive in AbuseEval, and that the prediction of implicit abuse is more challenging than explicit one.

Coping with the implicitness is necessary to make systems able to understand these messages that have a strong abusive effect but very weak offensive forms. As seen in Section 1.1.2, linguistic studies underlined a double use of ironic language. On the one hand, irony helps the speaker to reduce the social cost of criticizing or insulting someone. On the other hand, especially sarcasm reinforces the aggression of the message, hurting the victims that perceive the sarcastic utterance as mock and not as humorous as their aggressor [Bowes and Katz, 2011, Pexman and Olineck, 2002].

In spite of the theoretical literature is clear on describing the implicitness of abusive language, the computational efforts that could support it are few. To our knowledge, only metaphorical and stereotyped information have been exploited for abusive language detection. Interestingly, Lemmens et al. [2021] proved the contribution of hateful metaphors as features for the identification of the type and target of hate speech in Dutch Facebook comments in models based on classical machine learning and transformers. Whereas Lavergne et al. [2020] exploit the multi-annotation proposed in HaSpeeDe2020 about the presence of hate speech and stereotype in tweets to train a multi-task learning-based model reaching the best score in hate speech detection in tweets.

Most of the proposed approaches are specific to detect explicit abusive language, even if in some cases scholars employ features oriented to capture less evident aspects.

2.3 Computational Approaches to Abusive Language Detection

Considering the purposes of our work (Section 1.2.2), in this section we describe the main works and the shared tasks organized in the last years on *Abusive Language Detection* focusing principally on targets such as women and immigrants, and on different types of abuses, such as hate speech, aggressiveness, and stereotype.

The studies on Abusive Language Detection and on its sub-tasks are various. The efforts of the NLP and CL communities are visible in the vast range of linguistic resources and datasets created to analyze the various facets of abusive language, even from a multilingual perspective. Nevertheless, English is the most represented language in NLP. But, as underlined by Bender [2019], English fails to represent all languages because the linguistic properties of English are not broadly shared. This is one of the reasons that encourage us to approach computationally various languages apart from English, such as Italian and Spanish. Another reason lies in the will of unmasking cognitive and linguistic processes, regardless of languages, implied in the expression and, also, in the comprehension of abusive language.

Bringing to light these processes could encourage the development of models not strictly depending on language or on data, often biased. The creation of annotated corpora is, indeed, a complex process. On the one hand, it is difficult to find qualified annotators who support the annotation of large amounts of data. On the other hand, despite the institutionally and conventionally accepted definitions of abusive language, the perception of hateful message still remains subjective and also dependent on one's own cultural background [Basile, 2020b], determining the creation of individual perspective-based corpora.

The *sub-tasks* of Abusive Language Detection approached in this work are:

- Misogyny and Sexism Detection;
- Hate Speech Detection against immigrants;
- Stereotypes Detection;
- Aggressiveness Detection.

2.3.1 Misogyny and Sexism Detection

One of the first corpora of abusive language targeting women was released by Waseem and Hovy [2016]. The authors created a corpus containing racist and sexist tweets annotated taking into account a list of criteria founded in CRT. In their experiments, they noticed that sexist and racist tweets are better identified if information about the gender of users is provided. With a technique of gender identification based on recovering pronouns, gender-specific nouns and names of users, Waseem and Hovy [2016] noticed that the majority of hateful and especially sexist tweets are written by men. However, *what is sexism?*

From the philosophical perspective of Manne [2017], sexism is seen as:

> the 'justificatory' branch of a patriarchal order, which consists in ideology that has the overall function of rationalizing and justifying patriarchal social relations [Manne, 2017, p. 79]

whereas, misogyny is:

> the 'law enforcement' branch of a patriarchal order, which has the overall function of policing and enforcing its governing norms and expectations. [Manne, 2017, p. 78]

Therefore, on the one hand sexist ideology consists in the beliefs, assumptions, or stereotypes that represent men and women as different; on the other hand, misogyny differentiates between good and bad women, punishing the latter or forcing them to back into an *order*. In this view, sometimes sexism could be used for misogynist purposes, whereas in other cases, it can appear alone. In any case, both phenomena support the patriarchal order that tends to subjugate or reduce the power of women respect to the power of men, encouraging a discrimination based on sexual genre.

The difference of meaning stated by Manne [2017] between misogyny and sexism is not intercepted in most of the computational works that approach misogyny and sexism detection. For example, Waseem and Hovy [2016] considered as sexist the tweets that are offensive towards women according to the established criteria that defined a tweet as *offensive*. However, they reported that some cases of disagreement among annotators are based specifically on the difference of opinion about sexism.

The broader use of 'sexism' allowed scholars to individuate some communicative strategies against women. For instance, Clarke and Grieve [2017] investigated the functional linguistic variations between racist and sexist classes proposed in the corpus of Waseem and Hovy [2016], discovering that tweets against women tend to be more interactive and attitudinal than racist ones, addressed principally to persuade and argue the discrimination reporting events. Computational experiments on this corpus have been reported by Gambäck and Sikdar [2017]. The authors proposed a deep learning system based on Convolutional Neural Network (CNN) that assigns each tweet to one of the four categories (racism, sexism, both racism and sexism and non-abusive language), comparing its performance (with a $f1$-score of 0.78) with the Logistic Regression classifier based on character n-grams (with a $f1$-score of 0.74) proposed by Waseem and Hovy [2016].

Specifically about misogyny, a first computational effort in terms of automatic detection is provided for English by Anzovino et al. [2018]. The authors compared the performance of different supervised approaches using word embeddings, stylistic and syntactic features. The results revealed that the best machine learning approach for misogyny classification is the linear Support Vector Machine (SVM) classifier, with an accuracy of 0.77. This work inspired most of our experiments presented in Section 3.1. Part of this first dataset was used to create the English version of the benchmark datasets released by the organizers of the AMI (Automatic Misogyny Identification) shared tasks at IberEval 2018[19] [Fersini et al., 2018b] and EVALITA 2018[20] [Fersini et al., 2018a]. AMI shared task in its edition

[19] https://sites.google.com/view/ibereval-2018

[20] http://www.evalita.it/2018/tasks

at IberEval 2018 and EVALITA 2018 aims at detecting misogyny in tweets in various languages (English, Spanish, and Italian). In particular, the organizers asked participants to detect firstly if the message is misogynistic (task A), and secondly to classify the target (individual or not), and the category of misogyny according to the classes proposed in Poland [2016] (task B):

- ***Stereotype & Objectification***: a widely held but fixed and oversimplified image or idea of a woman; description of women's physical appeal and/or comparisons to narrow standards.
- ***Dominance***: to assert the superiority of men over women to highlight gender inequality.
- ***Derailing***: to justify woman abuse, rejecting male responsibility; an attempt to disrupt the conversation in order to redirect women's conversations on something more comfortable for men.
- ***Sexual Harassment & Threats of Violence***: to describe actions as sexual advances, requests for sexual favors, harassment of a sexual nature; intent to physically assert power over women through threats of violence.
- ***Discredit***: slurring over women with no other larger intention. [Fersini et al., 2018b, p. 125]

The best performing systems in English (with an accuracy of 0.91) and Spanish (0.81) version of AMI at IberEval 2018 are based on the SVM classifier and a complex set of designed features aimed mainly to capture the style (such as link presence, swear word count), the lexicon or vocabulary of the texts (insults, hashtags, bag of words, bag of hashtags, bag of emojis), and the negative stereotypes behind the use of certain words or expressions. Similar supervised approaches have been used in AMI at EVALITA 2018. In particular, the best scored systems for Italian (0.84) are based on the Logistic Regression (LR) and SVM classifiers exploiting Term Frequency–Inverse Document Frequency (TF-IDF) of character n-grams (for better dealing with misspellings and capturing few stylistic aspects), Singular Value Decomposition and feature abstraction techniques. For English (0.70), the best systems used the LR classifier and ensemble models with vector representation that concatenates sentence embedding, TF-IDF, average word embeddings and a bag of n-gram representation. Looking at the reports of both editions of AMI [Fersini et al., 2018b,a], the results achieved in the task B are lower than the ones obtained in the task A (from macro F-score of 0.29 to 0.34). The organizers thought that this difference is due to the fact that there can be a high overlap between textual expressions of different misogyny categories, therefore even for annotators it is difficult to identify a specific category (especially for derailing and dominance classes). A confirmation of their hypothesis is reported in Lazzardi et al. [2021]. Regarding the target classification, systems reached higher scores than category identification, but lower than the task A (with a macro F-score of about 0.55). Indeed, systems can be easily misled by the presence of mentions that are not the target of the misogynous content.

As said before, the boundaries between *misogyny* and *sexism* in many CL and NLP works are blurred; and for this reason, it is common to find in literature categories of sexism similar to the ones identified in AMI. In particular, Sharifirad et al. [2018] distinguish in sexist tweets: indirect harassment (expressing stereotypes about women and superiority of male), information threat (when women are threatened), sexual harassment (containing also insults), and physical harassment (attacks about physical aspects). Differently from these mutually exclusive classes, Parikh et al. [2019] proposed 23 categories, keeping in mind the campaigns on gender issues and the potential policy implications. They annotated the instances of sexism collected from the Everyday Sexism Project website[21] and for the classification experiments considered only the most representative categories (14). The purpose of their model is to categorize automatically the sexist experiences on the web in order to assist social scientists and policymakers in their investigations. Therefore, their intuition about multi-label classification (a possible solution even for AMI) lies in the possibility that the reported experiences could involve different types of sexist discrimination, such as: pay gap, internalized oppression, body-shaming, menstruation, motherhood, religion, and role based discrimination, tone policing (like the reduction to silence), moral policing, rape, threats, and violence. Although most of these categories could be arisen from sexist ideology, others accomplish the function of misogyny.

Similar categories have been proposed by the organizers of the EXIST share task on sEXism Identification in Social neTworks at IberLEF 2021[22] in Spanish and English tweets [Rodríguez-Sánchez et al., 2021]. Differently from the AMI shared task, EXIST dataset covers sexism in a broad sense from misogyny to various sexist behaviors such as inequality, stereotypes, dominance, objectification and sexual violence. Similarly, García-Díaz et al. [2021] created Spanish MisoCorpus-2020, a collection of tweets annotated as misogynistic and non-misogynistic in Spanish. In particular, they split the entire corpus in three main subsets: (1) violence against relevant women, (2) European Spanish vs that of Latin America, and (3) discredit, dominance, sexual harassment and stereotype. The first group of tweet concerns the verbal violence towards women who cover a relevant social positions; the second one takes into account the cultural and linguistic differences among the variants of Spanish; and the third one collects tweets containing general aspects related to misogyny and sexism like EXIST dataset.

The organizers of the last edition of the AMI shared task at EVALITA 2020[23] proposed the detection of misogyny and aggressiveness in Italian tweets and a test about the fairness of the models in terms of unintended bias on a synthetic dataset. Differently from the previous editions, Fersini et al. [2020] wanted to test the 'ability' of the model to be fair and not biased by the presence of certain terms called identity terms [Nozza et al., 2019]. The results on the second task showed that debasing techniques based on the augmentation of negative class [Lees et al., 2020] and on specific lexica on misogynistic aspects [Attanasio and Pastor, 2020] could help systems to reduce biases. About the first

[21] https://everydaysexism.com

[22] http://nlp.uned.es/exist2021/

[23] http://www.evalita.it/2020/tasks

task on misogyny and aggressiveness detection, the best results (macro F-score of 0.74) are reached fine-tuning the BERT pre-trained model [Devlin et al., 2019b] adapting it to the challenge domain and using a transfer multilingual strategy and ensemble learning.

Another shared task that addresses the problem of hateful language against women is HatEval proposed in the framework of SemEval 2019[24] [Basile et al., 2019]. The organizers asked participants first, to detect hate speech against immigrants and women (task A), and then, classify if the hateful tweet contains aggressive language and what is the type of addressed target (individual or generic) (task B). The released dataset for this occasion is a multilingual corpus containing English and Spanish tweets[25]. The best scores especially for the first task in both languages have been obtained with approaches based on an SVM classifier exploiting sentence embeddings, bag-of-words, bag-of-characters, tweet embeddings, and various linguistic features such as lexicon of derogatory words and different types of n-grams (macro F-score of 0.65 in English and 0.73 in Spanish). About the task B, the best performing systems for both languages (macro F-score of 0.57 in English and 0.70 in Spanish) are based on the SVM and a combination of LR, Multinomial Naïve Bayes, Classifiers Chain and Majority Voting. The principal features include lexical and syntactic information weighted by means of TF and TF-IDF and external lexica to capture hate speech. Some techniques of pre-processing of data have been employed to convert slang and short forms. It is interesting to notice that even considering the application of very recent computational techniques such as deep learning algorithms or transformer-based systems, the best scores were obtained with classical machine learning approaches. Like in AMI, participant systems at HatEval found much harder to predict the aggressiveness and targets than just the presence of hate speech. This international interest about targets such as immigrants and women is justified by the impact that misogyny/sexism and racism have on the daily life of a society. Indeed, the continued exposition, especially, to their most subtle forms, such as racist and sexist humorous expressions, tends to create greater tolerance of abusive events, modifying their perceptions as norms and not as negative behaviors [Ford et al., 2001].

2.3.2 Hate Speech Detection against Immigrants

As regards the state of the art on the automatic detection of hate speech specifically against minorities, only few contributions have been provided. Most of the literature about abusive language, in general, focuses on the type of the language (aggressive[26], offensive [Zampieri et al., 2019, 2020], hateful [Nobata et al., 2016], toxic[27]) and not on the type of target. Looking at the efforts related to detect abusive language against minorities on the basis of race, color of skin, religion, culture and nationality, one of

[24] https://alt.qcri.org/semeval2019/

[25] The subsets of English and Spanish tweets against women come mostly from the AMI datasets provided in the two editions of IberEval 2018 and EVALITA 2018.

[26] We will provide information about computational approaches to aggressive language detection in Section 2.3.4.

[27] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

the benchmark datasets is provided by the already mentioned Waseem and Hovy [2016] reporting racist tweets. Another mentioned dataset reporting tweets against immigrants is the corpus of HatEval. In the overview of the shared task, Basile et al. [2019] reported the distribution of the annotated categories for each target, and it is clear how the hateful messages against immigrants tend to offend the generic group and not individuals like in tweets addressed to women. If we look at some examples reported until now of hateful messages addressed to minorities (such as Examples 2, 3, and others in Section 1.1.1) we can notice that even when the specific target is an individual, like in Example 18, users tend to generalize the negative opinion to all members of the outgroup ("*la tipica #risorsa straniera*"[28]). These (negative) generalizations reinforce or create the stereotypes about minorities. And, as seen above, this happens not only on social media but even on newspapers where sometimes the hatefulness of the message is explicitly expressed (see Example 19).

A specific attention to hate speech against minorities is showed in the two editions of HaSpeeDe at EVALITA 2018 [Bosco et al., 2018b] and 2020 [Sanguinetti et al., 2020]. In particular, taking into account this necessity to face abusive language even in newspapers, in the second edition of HaSpeeDe at EVALITA 2020, we asked participants to detect hate speech (task A) and stereotypes (task B) in tweets and news headlines, and, if the text is hateful, to identify nominal utterances (task C) that make the message abusive[29]. Differently from this second edition, the HaSpeeDe shared task at EVALITA 2018 [Bosco et al., 2018b] focused on the detection of hate speech on the two of the most used social media nowadays (Twitter in the task 1 and Facebook in the task 2) with the main aim of testing also the robustness of systems in a cross-domain context (task 3). Although both tasks 1 and 2 addressed the problem of hate speech detection on social media, hateful messages targeted, in the first task, immigrants and, in the second task, various victims (among them even women). As expected the systems proved to be more efficient on in-domain context (with a higher macro F-score of 10-20%), and in Facebook domain where the comments tend to be longer and more correct than in Twitter (macro F-score of 0.83 on Facebook and 0.80 on Twitter). A similar cross-domain challenge was proposed also in HaSpeeDe 2020 where the provided training set collects only tweets whereas the systems are evaluated on tweets and news headlines. The best performing systems in HaSpeeDe 2018 in all tasks are based manly on the multi-task learning technique, recurrent neural networks (like Long Short-Term Memory) and exploited lexica about polarity and subjectivity.

Both editions of HaSpeeDe showed to be very fruitful ventures, bringing to international level the computational attention on the Italian language. Among the participants in both shared tasks, various not-Italian teams brought their contribution to hate speech detection. Indeed, before the proposal of these shared tasks, very few scholars have been worked on Italian hate speech [Vigna et al., 2017, Pelosi et al., 2017, Musto et al., 2016].

[28] the typical foreing #resourse

[29] More details about the HaSpeeDe competition in 2020 see Section 3.2.1

Similar lack could be observed in Spanish. Recently, a novel shared task about abusive language against immigrants (DETOXIS) [Taulé et al., 2021] has been proposed at IberLEF 2021[30] and aims principally to detect toxicity in comments posted in response to different online news articles related to immigration. Differently from other mentioned shared tasks, DETOXIS focus on toxicity (not specifically hate speech) distinguishing different levels of toxicity contained in the messages. All of these efforts from a multilingual perspective encourage the investigation of abusive language even in countries, like Italy and Spain, that are suffering from daily racist and misogynist attacks.

2.3.3 Stereotypes Detection

Stereotype is a phenomenon strictly connected to hate speech. Some participants in the competitions of AMI exploited this aspect as feature to detect misogyny. However, *what stereotype actually concerns?* Merriam-Webster Dictionary defines stereotype as "a standardized mental picture that is held in common by members of a group and that represents an oversimplified opinion, prejudiced attitude, or uncritical judgment". This definition reflects the recent conceptualization of stereotype in social and cognitive psychology. Dovidio et al. [2010] distinguished mainly three forms of social bias toward a group and its members: prejudice, stereotypes and discrimination. If prejudice and discrimination are related to attitude and behavior, stereotypes regard the association and attribution of specific characteristics to a group.

The authors underlined that if early researchers stress the inflexibility and the fault of the process of stereotyping a group, more recent investigations emphasize the functional and dynamic aspect of stereotyping as a process of simplification of a complex environment. Therefore, stereotypes are defined as cognitive schemas used by humans to process information about others. But, if, on the one hand, a stereotype implies an amount of information about people, generating expectations, on the other hand, constrains, producing a readiness to perceive information that are consistent with it [Fiske, 1998]. This simplification of reality helps to reinforce and create discrimination.

In particular, Fiske et al. [2002] elaborated the Stereotype Content Model that individuates systematic principles that help to understand why different groups are stereotyped in a very similar way (for example the members of Roma groups, Muslims and immigrants that, although they have distinct ethno-religious roots, are equally perceived as thieves and criminals). This model proposes two dimensions of stereotypes: warmth/cold and competence/incompetence. When the outgroup is perceived as warm and competent, it elicits pride and admiration, but when the outgroup is perceived as cold and incompetent, it generates disgust and anger[31]. The latter reflects the general perception of immigrants, Muslims and Roma. These processes of simplification and perception

[30] `https://detoxisiberlef.wixsite.com/website`

[31] Fiske et al. [2002] individuated also other two combinations of these dimensions: when the outgroup is perceived as warm and incompetent and when is perceived as cold and competent eliciting respectively pity and jealousy.

take place especially in the language, considered as one of the most important means of transmission of stereotypes [Dovidio et al., 2010]. In a context of communication, indeed, people tend to focus on the traits viewed as most informative, because more distinctive of a group and because easy to assimilate. For these reasons, social and public discourses that involve stereotypes are easily shared and spread. Some authors underline the importance of the creation and recreation of *frames* especially in politic and social speeches where the feelings of unease are used to produce media content and effects (the *media logic*)[32]. The frames should be considered as cognitive structures or the story line that provides meaning to different events connecting them [Gamson and Modigliani, 1994, Kinder, 1998]. Thanks to frames, ingroup members (politics, users) can shape stereotypes without mentioning explicitly the attributes of the outgroup.

One of the most recent works that approaches stereotypes computationally looking at the framing process, is the one of Sánchez-Junquera et al. [2021]. In this study, the authors proposed a taxonomy of stereotypes about immigrants, and approached the problem of automatic classification of stereotypes focusing on the narrative frames that spread the stereotypes. The taxonomy involves six categories extracted from political speeches about immigrants that cover pro and anti-immigrant attitudes: xenophobia's victims, suffering victims, economic resource (seen as instruments), threat for the group (threat for the society), threat for the individual (seen as competitors), and dehumanization. Following this taxonomy, the authors proposed a new annotated dataset used to train transformer-based models in order to predict the presence of stereotypes against immigrants and then to classify if the stereotype sees immigrants as victims or threat.

Similarly, Fokkens et al. [2018] approached stereotypes detection extracting from the text the microportraits, that are the collections of information and description that a text provides about a target. Card et al. [2016] combined syntactic relations and labels applying a Dirichlet process in a Bayes model to extract small stories about individuals, showing that these descriptions are useful for identifying frames. Differently from stories, the microportraits are based on descriptions that involve labels assigned to an entity, property and role that the entity plays in a specific event [Fokkens et al., 2018]. This approach follows the Social Categories and Stereotypes Communication framework proposed by Beukeboom and Burgers [2019]. The authors of this framework focused on how, linguistically, stereotypes are shared through language. In particular, they individuated two dimensions: bias in labels and bias in the description of characteristics and behaviors. The former involves the terminological choices (the use of *aliens* rather than *refugees*); the latter relies on the selection of information about characteristics and behavior that are presented by media/speaker (i.e., the emotional personality of women, the extremist religious perspective of Muslims and so on). The efforts to counter these negative stereotypes tend also to switch the focus on other (and positive) aspects of

[32] Dixon and Williams [2015] showed the representation of various targets (such as Muslims, latinos, blacks and whites groups) on American media between 2008 and 2012, and interpreted their findings looking principally at the 'guard dog' media coverage theory, economic interest of media, ethnic blame discourse and philanthropy perspective of community.

these groups. Other common computational approaches focused mainly on measuring and quantifying social bias towards different groups (women, immigrants and so on). From this perspective the most effective technique relies on the word representation, such as word embedding [Bolukbasi et al., 2016], transformers [Card et al., 2016] and techniques of natural language inference [Dev et al., 2020].

The majority of computational efforts on stereotypes detection tends to focus specifically on political speeches and news. These efforts reveal a special attention on the implicit dimension of stereotypes, leaving unexplored their explicit manifestation even in hateful content. To explore both dimensions of expression of stereotypes about minorities, we organized the second edition of HaSpeeDe shared task at EVALITA 2020, asking participants to detect stereotypes in the task B in hateful and not hateful tweets (see Examples 23 and 22) and headlines (e.g., Examples 19 and 20). To define this task, we adopted a perspective that sees *stereotype* as an orthogonal dimension of abusive language which does not necessarily coexist with hate speech, in line with the explored psychologist and cognitive studies.

2.3.4 Aggressiveness Detection

The boundaries of what is defined as aggressive language are very fuzzy, and most of the time is associated with flaming [Lapidot-Lefler and Barak, 2012], venting and uncivil language [Rösner and Krämer, 2016]. Especially studies on social media communication talking about verbal aggression refer to *any behavior that uses words rather than physical attacks to do harm* [Rösner and Krämer, 2016] appearing as a kind of destructive form of communication that includes, in particular, "hostile words and expressions, swear words and derogatory names, direct and indirect threats, use of letters, symbols and punctuation marks conveying hostility or aggression, and insulting, sarcastic, teasing, negative, or cynical comments" [Lapidot-Lefler and Barak, 2012].

The strong relation with other computational tasks such as hate speech [Burnap and Williams, 2015], cyberbullying Dinakar et al. [2011], flames [Spertus, 1997] and offensive language [Razavi et al., 2010] detection makes the extraction of aggressiveness difficult as individual dimension of verbal hate. Looking at the different aspects of abusive language online, Sanguinetti et al. [2018d] considered a message as aggressive on the basis of "the user intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target". The attention on the *intention* and the *aim* of the aggressive message is underlined also in the definition adopted by Carmona et al. [2018]. In particular, Carmona et al. [2018] considered a message as aggressive "if the purpose is to humiliate, belittle, discredit a person or group of people using rude words or pejorative language". This definition has been used to annotate the dataset MEX-A3T2018 that collected Mexican Spanish tweets, released in occasion of the first edition of the Aggressiveness Detection shared task proposed at IberEval 2018 in the forum of MEX-A3T[33]. To our knowledge, this shared task was the first that focused on aggressiveness, especially

[33] https://sites.google.com/view/mex-a3t2018/home?authuser=0

in a language different from English. The attention on Mexican variation of Spanish is due also to the important role that various linguistic contexts play in the individuation of aggressive, and more general abusive, messages. Indeed, distinctive lexica, syntactic characteristics and specific conventional meanings could lead to different interpretations or incomprehension:

(27) *#UnMexicanoPodría hacer luchona a cualquier mujer*[34]

(28) *#DosDeCadaTres mentadas de madre llevan tu nombre bordado en oro*[35]

To annotate this dataset Carmona et al. [2018] considered the provided definition of aggressive message and some rules that help the annotators to identify some specific characteristics of aggressiveness such as offensive nicknames, jokes, derogatory adjectives and rudeness. In particular, texts reporting quotes, pornography and prostitution content, self-attacking intention and abuses towards objects are not considered aggressive. This schema of annotation has been exploited also to annotate the dataset released for the second edition of the Aggressiveness Detection shared task in the forum MEX-A3T at IberLEF 2019[36] [Aragón et al., 2019]. A more specific definition of aggressive language is provided in the schema of annotation proposed for the dataset released in the third edition of MEX-A3T at IberLEF 2020[37] [Aragón et al., 2020]. The results in the previous editions, indeed, showed the need to distinguish the aggressive language specifically from offensive and vulgar language. In this line, Díaz-Torres et al. [2020] defined an annotation diagram for categorizing abusive language: a tweet could be vulgar, when involves, i.e., profanity or sexual connotations (independently of the presence of the target), could be aggressive if a target is involved and the message aims to hurt or incite to violence, and it could be even offensive when contains pejorative, derogatory or negative intensifiers of a term to refer to the target humiliating and insulting him/her. The application of this new methodology of annotation lead to better identification of aggressive tweets, passing from an F-score of 0.49/0.48 in 2018/2019 to 0.80 in 2020.

The best performing systems in the three editions of the aggressiveness detection shared task of MEX-A3T reflect the computational trend seen in HatEval 2019 and AMI 2020 (Section 2.3.1). Specifically in MEX-A3T 2018 and 2019, the best scoring systems adopted approaches based on model selection and the classical SVM exploiting as features tailor-made lexica, character n-grams and word embeddings representations. It is interesting to notice that the system best scored in 2019 uses a simpler model than the one employed in 2018, reaching very similar results. In the competition of 2020, the best performing model is based on a majority and weighted voting technique applied to an ensemble of different BETO models (BERT models trained in Spanish) exploiting the adversarial data augmentation.

[34] *#AMexicanCould make any woman fighter.* The term *fighter* could have an offensive meaning sometimes. Tweet extracted from MEX-A3T2018 dataset.

[35] *#EveryTwoForThree fuck, they bring your name embroidered in gold.* Tweet extracted from MEX-A3T2018.

[36] https://sites.google.com/view/mex-a3t2019

[37] https://sites.google.com/view/mex-a3t/home?authuser=0

2.4 Conclusions

Observing the efforts of the NLP community, we can notice that during the last years, the attention on abusive language detection has been increased. These efforts concern general and specific categories of abusive language, encouraging also the investigation of particular characteristics or crossed phenomena. Even in the definition of the terms used in the world of abusive language, we can notice a sharpening that is the result of deeper analyses and multidisciplinary collaborations. Although the evident improvement at the theoretical and computational level, the attention on implicit manifestation of abusive language is scarce.

The problem is clear and marked also by the organizers of the mentioned shared tasks. For instance, the organizers of HatEval [Basile et al., 2019] provided an interesting error analysis, which brings to light the necessity to address: 1) the contextual use of swear words (i.e., bitch) that could be used with not offensive purposes; and 2) the humorous intention especially in wordplay not recognized by the majority of the systems in both considered languages. Similar permanent errors are found in misclassified tweets in all the editions of MEX-A3T aggressiveness detection shared task: ironic comments, implicit offenses (not expressed with vulgar words), and the use of indirect discourse that contains derogatory words but is a description of an event and not an attack [Carmona et al., 2018, Aragón et al., 2019, 2020]. In misogyny detection, the main errors regard the case of tweets where hate speech is not addressed to women, or that contain general aggressive expressions without a specific target [Fersini et al., 2020]. This problem, that reflects also on the general lower results in target identification in AMI and HatEval shared tasks, underlines the necessity to focus on the victim that in real-world application could appear fundamental. Other errors are related to the lack of dedicated approaches for specific problems. For instance, as showed by the organizers of AMI 2018 [Fersini et al., 2018b,a], participating teams did not approach specifically the detection of individual misogynistic categories, leading to the problem of not detected categories. Or in HaSpeeDe 2020, Sanguinetti et al. [2020] reported that the teams employed the same approach to detect hate speech and stereotypes, treating these phenomena as the same.

In general, the used approaches are supervised. Only early works on abusive language, such as Spertus [1997], approached the problem with an unsupervised system based on rules and lists of words. From the advent of machine learning, the majority of models have become dependent on the data used for their training. Therefore, the research community tends to provide more and more balanced dataset annotated by experts. But, despite the effort, data-driven systems tend to reflect the perspective (such as bias and cultural background) of small groups [Akhtar et al., 2021]. To prevent the consequences, supervised systems are often supported by lexica and elaborated linguistic features that capture syntactic and stylistic patterns. Another recent solution relies on the application of the transfer learning technique, especially fine-tuning BERT models, that could help to adopt an extended perspective.

Taking into account these main problems and lacks in the current literature, in the next chapter we suggest some proposals and describe the experiments and results obtained using specific approaches in multilingual and multi-genre contexts.

Chapter 3

Detecting Different Forms of Abusive Language Online

In this chapter, we describe the experimental techniques and approaches employed to overcome the principal difficulties noted in the current literature, especially the need to address implicit abuses. In particular, we address the first research question of this thesis:

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?

To this purpose, we analyze the characteristics of hateful texts and explore some linguistic and computational approaches that allow systems to perceive indirect interpretations.

3.1 Misogyny Online

An interesting episode that could help to highlight the difference between misogyny and sexism is the speech of the prime minister of Australia, Julia Gillard, in October 2012, about the sexist and misogynistic behavior of the leader of the opposition, Tony Abbott. In particular, as underlined by Manne [2017], she defined some behaviors and expressions as sexist (*If it's true that men have more power generally speaking than women, is that a bad thing?*, *What the women of Australia need to understand, as they do the ironing...* or defining abortion as *the easy way out* when he was health minister), and, only when he used offensive explicit and implicit remarks against her, she started talking about misogyny (*If the Prime Minister wants to, politically speaking, make an honest woman of herself...*, or describing her as *a man's b***h* or *witch*). It is clear that both behaviors offend women, but the ones described firstly are expression of an *ideology* that rationalize and justify the men's power and the patriarchal order; the others tend to express a sort of enforcement and punishment of women, as well as hatred and explicit insults[1]. Therefore, although both can appear individually, they tend to support patriarchal order.

[1] The resonance of Gillard's speech was such that some dictionaries, such as the Macquarie Dictionary, updated the definition of misogyny.

Focusing on misogyny, Manne [2017] individuated three important characteristics:

- misogyny involves even third-personal indignation: hostility is shown even towards women *who are held to wrong others*, like unborn children;
- misogyny comprises social and institutional practices and structures that can support forms of hostility against women;
- misogyny and racism are connected, and appear inseparables in cases of hostility against *non-white* women [Crenshaw, 1991, Hancock, 2007].

In our work, we touch especially the first and second characteristics, and taking into account the difficulties emerged from the existing approaches (see Section 2.3.1), we focus mainly on:

1) the automatic detection of misogyny in multilingual texts (English, Spanish, and Italian tweets) looking at the different types of misogynistic attacks more common in each language (Section 3.1.1);

2) the differences and analogies between sexist and misogynistic tweets from the automatic language processing perspective, observing principally the social and conventional biases (Section 3.1.2);

3) the expression of misogyny as third-personal indignation or as result of sexist attacks, and its coexistence with stance on Twitter and newspapers (Section 3.1.3).

3.1.1 Multilingual Misogyny Detection

Focusing on the first point, in this section, we describe two simple but efficient approaches used in our participation at the AMI competitions, organized in the framework of IberEval and EVALITA 2018, aiming at detecting misogyny in English, Italian and Spanish tweets. As described in Section 2.3.1, Fersini et al. [2018a,b] organized in 2018 two shared tasks

in occasion of IberEval and EVALITA 2018, asking participants to detect firstly whether a tweet is misogynistic or not (task A), and secondly, in case of misogynistic tweet, to define the category of misogyny (task B1) and the type of target (individual or group) (task B2). The automation of these tasks could constitute an important support to policing and monitoring activities, especially in contexts, such as in the video games online, where the social dominance and masculine norms are strong and women need to comply with them [Fox and Tang, 2014].

The organizers released a total of four multilingual datasets: two at IberEval containing English and Spanish tweets, and two at EVALITA containing English and Italian tweets. Tables 3.1 and 3.2 report the distribution of labels for the tasks A and B in all datasets. The tweets were collected using keywords and hashtags regarding harassment and attacks against women in each language, as described in Anzovino et al. [2018]; some examples from all datasets are provided in Table 3.3. On the basis of the collected corpora, Jane's intuition [Jane, 2014] seems to be confirmed across language: the hostility against women is often expressed using sexually explicit language, suggesting various acts connected to the sexual sphere that could be carried out as forms of correction to bring women back into line, and insults about their physical appearance.

	Task A		**Task B**		
	English	Spanish		English	Spanish
misogynistic	1,568/283	1,649/415	stereotype		
			& objectification	137/72	151/17
			dominance	49/28	302/54
			derailing	29/28	20/6
			sexual harassment		
			& threats of violence	410/32	198/51
			discredit	943/123	978/287
			active (individual)	942/104	1,455/370
			passive (generic)	626/179	194/45
non-misogynistic	1,683/443	1,658/416			
Total	**3,251/726**	**3,307/831**			

Table 3.1 – Label Distribution of the AMI IberEval Dataset (Training/Test).

	Task A		**Task B**		
	English	Italian		English	Italian
misogynistic	1,785/460	1,828/512	stereotype		
			& objectification	179/140	668/175
			dominance	148/124	71/61
			derailing	92/11	24/2
			sexual harassment		
			& threats of violence	352/44	431/170
			discredit	1,014/141	634/104
			active (individual)	1,058/401	1,721/446
			passive (generic)	727/59	96/66
non-misogynistic	2,215/540	2,172/488			
Total	**4,000/1,000**	**4,000/1,000**			

Table 3.2 – Label Distribution of the AMI EVALITA Dataset (Training/Test).

Inspired by this attempt to categorize different misogynistic behaviors, we tried to explore the different aspects of this form of hate speech, comparing them in the different languages. As we can see in the distribution of labels in Tables 3.1 and 3.2, in Spanish the users tend to use terms that transmit a will of male dominance on women, whereas verbal harassment and threats are typical in English and Italian tweets as well as the intention to objectify the image of women or describing her in a stereotyped way.

Linguistic Features To better understand these linguistic differences, in both competitions we proposed an approach based mainly on a set of lexica that reflect the specific topics of misogynistic attacks: sexuality, profanity, femininity, human body and generic

category	target	text
Stereotype	passive	*Women play your role..#gobacktothekitchen*
Objectification	passive	*Big te**ts and no brain.this is what I want as wife!*
Dominance	active	*USER Cállate y hazme la cena.* → *USER Shut up and make me dinner.*
Derailing	active	*yo creoque mereces un castigo por z***a no? *con el cigarro en la mano* (md) URL* → *I think you deserve a punishment for b***h, right? *with cigar in hand* (md) URL*
Sexual Harassment	active	*@USER ciao Selena buon pm quanto 6 figa e sexy ti scoperei un abbraccio e un bacio* → *@USER hello Selena good afternoon how hot and sexy you are I would f*** you a hug and a kiss*
Threats of Violence	active	*Stupid bi**h I'll put you down on the floor and I'll rape you! You should be scared!*
Discredit	active	*@USER un vomito più che rigurgito, questo comunista assieme alla culona in6 anni è riuscito a quintuplicare il numero di poveri in italia* → *@USER vomiting rather than regurgitation, this communist together with the big ass in 6 years has managed to quintuple the number of poor people in Italy*

Table 3.3 – Examples Extracted from the AMI Datasets.

stereotypes about women. Along with them, we used also lists of hashtags, abbreviations, and slangs considering the social media context (see Table 3.4). The majority of these lexica were built mainly considering the most relevant words for each category using a process of extraction based on Information Gain (IG). The main advantages of using this kind of lexicon-based approaches can be summarized as follows: 1) the possibility to reduce the bias of the systems, as shown in the results of the third edition of AMI [Fersini et al., 2020]; 2) the opportunity to analyze the language used by users to offend or to express hostility against women [Hewitt et al., 2016]; 3) the fact that lexica represent the linguistic cues that allow to discover the stereotypic inferences in texts [Beukeboom and Burgers, 2019].

3.1.1.1 AMI-IberEval 2018

Taking into account that the collected English and Spanish tweets are not geolocalized, one of the challenges is to cope with the linguistic variations. To this purpose, we designed specific linguistic features aimed to capture the style and variations by means of character n-grams, sentiment and emotional information.
The additional features are:

category	**description**
Sexuality	One of the most frequent topics in misogynistic tweets is the sexuality (*orgasm*, *orgy*, *pussy*, *concha*) and in particular the desire of domination in sexual way, especially in English (*rape*, *pimp*, *slave*).
Profanity	We collected the general derogatory words, excluding common words, like *fuck* or *puta*, which could be used also without offensive purposes [Clarke and Grieve, 2017, Hewitt et al., 2016]. Especially for Spanish, this lexicon gathers several vulgarities from different variants. Some examples are: *motherfucker*, *slut* and *scum*.
Femininity	In order to identify women as target, we collected personal pronouns or possessive adjectives (such as *she*, *her*, *herself*), common words used to refer to women (*girl*, *mother*) even in negative way (*gallina*, *blonde*) and also offensive words towards women (such as *barbie*, *hooker* or $non - male$).
Human body	We collected a set of terms about feminine body (*gambe*, *pies*) even with negative connotation (such as *holes*, *throat*, *boobs*).
Hashtags	As in previous works [Fox et al., 2015], we took into consideration the hashtags used as referents for shared concepts by online communities, such as #*todasputas*, #*womensuck*, #*ihatefemales* or #*bitchesstink*.
Abbreviations	This list contains vulgar abbreviations typical of Internet slangs found in the data, such as: *idgaf*, *smh*, *hdp*, *wtf* or *stfu*.
Stereotypes	This last list embraces various terms related to the stereotypes or myths about women, like technology, cooking or taking care of children.

Table 3.4 – Lexica for Misogyny Detection.

- *Sentiment lexica* As previous studies about abusive language [Dinakar et al., 2012, Gitari et al., 2015], we used sentiment analysis. For English, we used SentiWordNet [Baccianella et al., 2010] and SentiStrength [Thelwall et al., 2010] specific for informal language on social platforms. For Spanish, we used ElhPolar dictionary [Urizar and Roncal, 2013]. In both languages, sentiment analysis increased the accuracy of the system, confirming our intuition that hateful expressions largely exhibit a negative polarity.

- *Affective lexica* Finally, specifically for the second task in Spanish, we used Spanish Emotion Lexicon (SEL) [Sidorov et al., 2012, Rangel et al., 2014], in order to understand the impact of the emotions on specific misogynistic classes.

These sets of features are experimented employing the SVM classifier and the ensemble technique. The data were preprocessed, deleting emoticons and urls to leave the most informative elements for the lexical approach. Then, preprocessed tweets have been represented as a vector composed of: all specific topic features (set of lexica) pondered with IG, and character n-grams weighted with TF-IDF. In addition, we used also the FreeLing lemmatizer provided by Carreras et al. [2004] to face the inflectional morphology of Spanish language, and the Porter stemmer by Natural Language Toolkit (NLTK)[2] for English[3].

Experiments and Results Although the evaluation measures adopted by the organizers [Fersini et al., 2018b] are different for each task (respectively accuracy for the task A and $f1$-macro for the task B), we carried out a similar set of experiments for both tasks. First, our efforts focused on finding the best classifier for each task, and the SVM proved to be the best like in the previous work of Anzovino et al. [2018]. In particular, we employed a linear kernel and, for each task, different values of C and *Gamma* parameters were chosen. We experimented also with an ensemble technique based on the majority voting methodology, looking at the predictions obtained from the three best performing classifiers: SVM, RF and Gradient Boosting.

In this experimental phase, to evaluate our approach we used K-Fold Stratified method with K=5, and as baseline we used character n-grams for each classification task with the best n-gram length (from 3 to 5 grams). The experimental results are reported in Table 3.5. In this step, we also selected the best contributing features for each task and for each language. Although the distribution of labels in Table 3.1 suggests some differences in the expression of hostility against women between English and Spanish, the experimental results indicate that the classifiers need to be informed lexically with all the created lexica. Selected features are shown in Table 3.6.

approach	**Task A**		**Task B1**		**Task B2**	
	accuracy		$f1$-macro		$f1$-macro	
	En	*Sp*	*En*	*Sp*	*En*	*Sp*
char n-grams - SVM (Baseline)	0.762	0.778	0.251	0.360	0.711	0.690
Lexica - SVM	0.790	0.788	0.270	0.381	0.520	0.545
Sentiment+SEL - SVM	0.743	0.758	0.209	0.295	0.501	0.516
Selected Features - SVM	0.782	0.790	**0.284**	0.408	0.713	**0.693**
Selected Features - Ensemble	**0.801**	**0.794**	0.284	**0.414**	**0.714**	0.672

Table 3.5 – Experimental Results for each Task of AMI at IberEval 2018.

Among all results reported in Table 3.5, we can notice that, in general, lexical features achieved higher results compared to the others. We noticed, in particular, that without the application of the lemmatizer the accuracy score for Spanish decreased to 76.11.

[2] https://www.nltk.org/

[3] In our experiments, we noticed that the best solution for misogyny detection in English was the use of a stemmer; in fact, the application of lemmatization reduced the accuracy.

feature	**Task A**		**Task B1**		**Task B2**	
	En	*Sp*	*En*	*Sp*	*En*	*Sp*
Sexuality	v	v	v	v		
Profanity	v	v	v	v		
Femininity	v	v	v	v		
Human body	v	v	v	v		
Hashtags	v	v	v	v		
Abbreviations	v	v	v	v		
Stereotypes			v	v		
Sentiment	v	v	v	v		
Emotions				v		
Character n-grams	v	v	v	v	v	v

Table 3.6 – Selected Features for each Task of AMI at IberEval 2018.

The evaluation on the test set provided by the organizers confirmed our experimental results. Tables 3.7 and 3.8 report results obtained in the competition compared with the provided baseline and the first ranked system. The organizers provided baseline scores using a model based on an SVM classifier with linear kernel trained on the unigram representation of tweets. As we can see, in both tasks our approach has reached a good performance. Especially, in the task B[4], we achieved the highest results for English[5].

approach	**accuracy**		**team rank**		**run rank**	
	En	*Sp*	*En*	*Sp*	*En*	*Sp*
Pamungkas et al. [2018]	***0.913***	***0.815***	*1*	*1*	*1*	*1*
Selected Features - Ensemble	0.870	0.813	2	3	5	3
Selected Features - SVM	0.862	0.805	2	3	7	10
AMI-Baseline	*0.783*	*0.768*			*15*	*18*

Table 3.7 – Results in Official Ranking for the Task A of AMI at IberEval 2018.

Performing the error analysis of the best models, we noticed that one of the main problems in both languages was the use of humorous or sarcastic expressions in the tweets. As confirmed by Boxer and Ford [2010], humorous utterances are common in misogynistic and sexist speeches:

(29) *¿Cuál es la peor desgracia para una mujer? Parir un varón, porque después de tener un cerebro dentro durante 9 meses, van y se lo sacan*[6]

[4] The organizers provided the results for misogynistic categories and target classification in terms of average of the $f1$-macro.

[5] Although the submitted runs are various (few changes distinguish them), in these tables we have preferred reporting only the runs obtained using different approaches.

[6] *What is the worst misfortune for a woman? Give birth to a boy, because after having a brain inside for 9 months, they go and take it out*

approach	accuracy		team rank		run rank	
	En	*Sp*	*En*	*Sp*	*En*	*Sp*
Pamungkas et al. [2018]	*0.369*	***0.446***	*2*	*1*	*6*	*1*
Selected Features - Ensemble	0.402	0.441	1	2	5	4
Selected Features - SVM	**0.442**	0.427	1	2	1	9
AMI-Baseline	*0.337*	*0.409*			*16*	*14*

Table 3.8 – Results in Official Ranking for the Task B of AMI at IberEval 2018.

(30) *What's the difference between a blonde and a washing machine? A washing machine won't follow you around all day after you drop a load in it*

Moreover, by means of the analysis based on IG, we noticed that, in both tasks, sexual language tends to be used especially in misogynistic tweets in English, and the profanities or vulgarities in misogynistic ones in Spanish. One of the points that surprised us is the fact that in the analysis of affective features in Spanish, *joy* is the principal emotion that provokes misogynistic speech.

3.1.1.2 AMI-EVALITA 2018

Considering the encouraging results obtained with the lexicon-based approach in Spanish and English languages, we re-proposed a similar approach for Italian tweets and the new collection of English tweets released by the organizers in the second edition of AMI at EVALITA 2018 [Fersini et al., 2018a]. In particular, we propose a comparison between two lexicon-based approaches:

- the Manually-Modeled Lexica (**MML**) approach, similar to the previous work, is based on topic, linguistic and stylistic information captured by means of the manually-modeled lexica and n-grams of words and characters;
- the Automatically-Enriched Lexica (**AEL**) approach involves the automatically extended versions of the original lexica.

The purpose is dealing with the continuous variation of languages on social platforms. To do that, we enriched the created lexica (see Table 3.4) considering the contextual similarity of lexica by the use of the pre-trained word embeddings. This technique helps the system to consider also new terms in multilingual contexts relative to the topic information of the original lexica. It could be considered as a good methodology to upgrade automatically the existing multilingual list of words used to block offensive contents in real applications.

Manually-Modeled Lexica Approach

Features for English For the detection of misogyny in English tweets, we employed the manually-modeled lexica concerning mainly sexuality, profanity, femininity and human body (see Table 3.4). These lexica contain also slang expressions, hashtags, and abbreviations. Experimenting with various n-grams of words in both tasks, in our final systems we employed: bigrams for the first task and the combination of unigrams, bigrams and trigrams (hence defined as UBT) for the second task. Moreover, the bag of characters (BoC) in a range from 1 to 7 grams is employed to manage misspellings and to capture stylistic aspects of writing online. In order to perform the experiments, each tweet is represented as a vector. The presence of words in each lexicon is pondered with IG, and character and word n-grams are weighted with TF-IDF. In addition, considering the fact that in the previous work several misclassified misogynistic tweets were ironic or sarcastic, we tried to analyze the impact of irony on misogyny detection in English. In particular, inspired by Barbieri and Saggion [2014], we calculated the imbalance of the sentiment polarities (positive and negative) in each tweet using SentiWordNet provided by Baccianella et al. [2010]. For each degree of imbalance, we associated a weight used in the vectorial representation of the tweets. Despite our hypothesis is well funded, we obtained lower results for the runs that contain sentiment imbalance as feature (see Table 3.11).

Features for Italian For the Italian language, we selected specific categories of hurtful words extracted from HurtLex[7] [Bassignana et al., 2018] that is a multilingual lexicon of hateful words created from the Italian lexicon “Le Parole per Ferire” by Tullio de Mauro. The entries in the lexicon are categorized in 17 types of offenses (see Table 3.9). In particular, from this lexicon we employed the following categories for the Italian part: `AN` (*sanguisuga* or *pecora*), `ASF` (*fessa*), `ASM` (*verga*), `CDS` (*bastardo* or *spazzatura*), `OR` (*finocchio* or *rapa*), `PA` (*portinaia* or *impiegato*), `PR` (*bagascia* or *zoccolona*), `PS` (*negro* or *ostrogoto*), `QAS` (*parassita* or *dilettante*), `RE` (*stupro* or *violento*). Differently from English, the experiments reveal that: UBT is useful for both tasks and the best range for BoC is from 3 to 5 grams[8]. Indeed, in a morphological complex language like Italian the desinences of the words (such as the extracted n-grams “tona” or “ana”) contain relevant linguistic information. Whereas in English, longer sequences of characters could help to capture multiword expressions containing also pronouns, adjectives or prepositions, such as “ing at” or “ss bitc”.

Before the training phase, we preprocessed the data, deleting emoticons, emojis and URLs. Our experiments proved that emoticons and emojis are not relevant for this task. In order to achieve a good lexical match, we used the lemmatizer provided by the

[7] The multilingual lexicon is available on `http://hatespeech.di.unito.it/resources.html`. Our experiments on English suggested that the system informed with the manually-modeled lexica, achieves better performance in misogyny detection.

[8] The experiments are carried out using the Grid Search technique provided by the *scikit-learn* library `https://scikit-learn.org/stable/index.html`

category	length	description
PS	254	Ethnic Slurs
RCI	36	Location and Demonyms
PA	167	Profession and Occupation
DDP	496	Physical Disabilities and Diversity
DDF	80	Cognitive Disabilities and Diversity
DMC	657	Moral Behavior and Defect
IS	161	Words Related to Social and Economic advantages
OR	144	Words Related to Plants
AN	775	Words Related to Animals
ASM	303	Words Related to Male Genitalia
ASF	191	Words Related to Female Genitalia
PR	138	Words Related to Prostitution
OM	145	Words Related to Homosexuality
QAS	536	Descriptive Words with Potential Negative Connotations
CDS	2042	Derogatory Words
RE	391	Felonies and Words Related to Crime and Immoral Behavior
SVP	424	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 3.9 – HurtLex Categories.

NLTK for English, and the Snowball Stemmer for Italian. Like in Spanish, the use of a lemmatizer for Italian tweets hinders the match.

Automatically-Enriched Lexica Approach The second approach aims to deal with the dynamism of the informal language online, trying to capture new words relative to the context defined in each lexicon. In particular, we used the enriched versions of the original lexica (Table 3.4 and selected categories from Table 3.9), and stylistic and linguistic information captured by means of n-grams of words and characters like in the first approach. The method used to expand automatically a lexicon is based on the identification of new words looking at their contextual similarity with the original lexicon, exploiting a pre-trained word embedding considered for each language. The recovered new words, thus, are strongly related to the context defined in the original lexicon.

For its description, let us assume that $\mathcal{L} = \{l_1, \dots, l_m\}$ is the initial lexicon of m words, and $\mathcal{W} = \{(w_1, e(w_1)), \dots, (w_n, e(w_n))\}$ is the set of pre-trained word embeddings, where each pair represents a word and its corresponding embedding vector. The main steps are:

1. *Dictionary modeling* Firstly, we extract the embedding $e(l_i)$ for each word $l_i \in \mathcal{L}$; then, we compute the average of these vectors to obtain a vector describing the entire lexicon, $e(\mathcal{L})$. We named this vector *context embedding*.

2. *Dictionary expansion* Then, using the cosine similarity, we compare $e(\mathcal{L})$ against the embedding $e(w_i)$ of each $w_i \in \mathcal{W}$; and, extract the k most similar words to $e(\mathcal{L})$, defining the set $E_L = (w_1, \ldots, w_k)$. Finally, we insert the extracted words into the original lexicon to build the new lexicon, i.e., $\mathcal{L}_{\mathcal{E}} = \mathcal{L} \cup E_L$.

feature	**Task A**						**Task B1**		**Task B2**	
	En			*It*			*En*	*It*	*En*	*It*
	run1	*run2*	*run3*	*run1*	*run2*	*run3*[9]				
Sexuality	v	v	v							
Profanity	v	v	v							
Femininity	v	v	v							
Human body	v	v	v							
Hashtags	v	v	v							
Abbreviations	v	v	v							
Stereotypes	v	v	v							
Hurtlex				v	v	v				
AEL	v	v		v		v				
Sentiment Imbalance	v		v							
Unigrams				v	v	v	v	v	v	v
Bigrams	v	v	v	v	v	v	v	v	v	v
Trigrams				v	v	v	v	v	v	v
BoC	v	v	v	v	v	v		v	v	v

Table 3.10 – Selected Features for each Task of AMI at EVALITA 2018.

The exploited pre-trained word embeddings for each language are: *GloVe* embeddings trained on 2 billion tweets [Pennington et al., 2014a] for English, and word embeddings built on the TWITA corpus[10] for Italian [Basile and Novielli, 2014]. This expansion method is a parametric proposal and requires a value for k, the number of words that are going to extend the lexica. In this work, we experimented with $k = 1000$, 500 and 100.

Experiments and Results On the basis of our experiments we used:

- for the tasks A and B1 (category identification), an SVM classifier with the radial basis function kernel (RBF) and the following parameters: $C = 5$ and $\gamma = 0.1$ for English and $\gamma = 0.01$ for Italian;

- and an RF classifier that aggregates the votes from different decision trees to decide the final class of the tweet for the task B2 (target identification) in Italian. We chose this strategy to face the imbalance between the classes *active* and *passive* in the dataset (see Table 3.2).

[9] with $k = 100$

[10] http://valeriobasile.github.io/twita/about.html

In this phase we selected the most contributing features, shown in Table 3.10, and we tested our approaches employing the 10-fold cross-validation taking into account the measures proposed by the organizers: accuracy for the task A and a general average of $f1$-macro ($f1$-avg) for the task B. Looking at the results obtained in the competition in the task A, we noticed a discordance, especially in Italian, between the results obtained with 10-fold cross-validation and the test set provided by the organizers. Indeed, the AEL approach with the enriched lexica using k equal 100, performed an Accuracy of 0.880 whereas the evaluation on the test set reached 0.823.

approach	**accuracy**		**run rank**	
	En	*It*	*En*	*It*
First Ranked System	***0.704***[11]	***0.844***[12]	*1*	*1*
run1 (AEL)	0.592	0.824	21	9
run2	0.613 (AEL)	0.822 (MML)	17	12
run3	0.584 (MML)	0.823 (AEL)	25	11
AMI-Baseline	*0.605*	*0.830*	*19*	*7*

Table 3.11 – Results in Official Ranking for the Task A of AMI at EVALITA 2018.

approach	**$f1$-avg**		**run rank**	
	En	*It*	*En*	*It*
First Ranked System	***0.406***[13]	***0.501***[14]	*1*	*1*
run1	0.335	0.448	8	8
run2	0.344	0.446	6	10
run3	0.328	0.449	10	7
AMI-Baseline	*0.370*	*0.487*	*3*	*2*

Table 3.12 – Results in Official Ranking for the Task B of AMI at EVALITA 2018.

Observing Table 3.11, reporting the official results of the AMI task, only *run2* overcomes the baseline for the detection of misogyny in English, and for this run we used the AEL approach excluding the sentiment imbalance as a feature (Table 3.10). About the identification of misogyny in Italian, in general, the results are lower than the AMI-Baseline as well as the scores in the task B for both languages (see Table 3.12).

Despite the usefulness of lexica for a specific domain like misogyny, a lexicon-based approach proves to be insufficient for this task. Actually, as the error analysis confirms, misogyny, and in general abusive language, involves linguistic devices such as humor,

[11] Saha et al. [2018]

[12] Bakarov [2018]

[13] Ahluwalia et al. [2018]

[14] Basile and Rubagotti [2018]

exclamations typical of orality and contextual information that make the meaning transmitted by the tweet, implicit and less intuitive. Moreover, the low scores obtained also in the task B suggest the necessity to implement a dedicated approach for each misogynistic category, as seen in the proposed approach at AMI for IberEval 2018.

Carrying out the error analysis, we noticed in both datasets aspects similar to those observed in English and Spanish misclassified tweets, such as: the content of URL tends to affect the transmitted information in the tweet (Example 31); the swear words are often used as exclamations without the aim to offend (Example 32); and, despite the actual English corpus does not contain several jokes, Italian misclassified tweets involve various humorous utterances (Example 33).

(31) *Right! As they rape and butcher women and children !!!!!!* `https://t.co/maEhwuYQ8B`

(32) *Volevo dire alla Yamamay che tettona non sinonimo di curvy dato che di vita ha una 40, quindi confidence sta minchia.*[15]

(33) *@USER @USER A parte il fatto poi che culona inchiavabile "è il miglior giudizio politico sentito sulla Merkel negli ultimi anni??"*[16]

3.1.2 Patriarchal Culture Online

Considering the observations about the complexity and peculiarities of misogynistic tweets, emerged from these first experiments on misogyny detection in various languages, and inspired by the considerations about misogyny advanced by Manne [2017], we decided to investigate from a computational perspective the relation between sexism and misogyny looking particularly at the social or conventional practices/ways of thinking that support forms of hostility against women (such as sexist stereotypes). In particular, we were interested to understand:

a) if sexist utterances discriminating women could be indicative of misogynistic attitudes and, thus, be recognized as attacks towards women

and then,

b) if it is possible to approach their automatic recognition similarly as two sides of the same coin (patriarchal mentality).

To this purpose, we proposed a classifier of misogynistic and sexist tweets using the same set of features: stylistic features and lexica modeled on topics and stereotyped information (as described in Table 3.4). The used benchmark corpora include English misogynistic tweets released for AMI at IberEval and EVALITA [Fersini et al., 2018b,a]

[15] *I wanted to tell Yamamay that busty is not synonymous with curvy since she has a waist with a size of 40, so confidence is f***ing.*

[16] *Apart from the fact that unf***able big ass "is the best political opinion heard about Merkel in the recent years??"*

and English sexist tweets released by Waseem and Hovy [2016]. As described above, AMI corpora include misogynistic tweets collected using keywords and hashtags regarding harassment and attacks on women [Anzovino et al., 2018]. The sexist corpus, available online (NAACL_SRW_2016_tweets[17]), was recovered using the ids of tweets provided by Waseem and Hovy [2016] jointly with the labels: `sexist`, `racist` and `none`. In this work, sexist tweets contain “prejudice or discrimination based on sex” (Merriam-Webster Dictionary). Unfortunately, some tweets were no longer available[18]. Despite a balanced collection of positive and negative samples is not a faithful representation of the real world, an equal number of sexist and non-sexist tweets has been selected such as in the balanced AMI collections. Therefore, we collected in a corpus, called here SRW, all the available sexist tweets and a correspondent number of non-sexist tweets (from the `none` class). Table 3.13 shows the statistics of the datasets.

dataset	`misogynistic`	`non-misogynistic`	`sexist`	`non-sexist`	**total**
AMI_IberEval	1,568	1,683	–	–	3,251
AMI_EVALITA	1,785	2,215	–	–	4,000
SRW	–	–	2,503	2,503	5,006

Table 3.13 – Distribution of Labels in Misogynistic and Sexist Datasets.

3.1.2.1 Statistical Analysis

In order to better understand differences and similarities between these datasets, we performed an analysis comparing: the size of corpora, vocabulary, lexical richness and the average of words per tweet. The lexical richness of the collections of tweets is calculated by means of the Type-Token Ratio (TTR) that measures the variation of the lexicon in each corpus. Table 3.14 summarizes the resulting statistics.

	AMI_IberEval	**AMI_EVALITA**	**SRW**
Number of tweets	3,251	4,000	5,006
Vocabulary	9,158	10,532	11,966
Number of tokens	55,431	68,573	79,138
Type-token ratio	16.52%	15.36%	15.12%
Average of words	17.05	17.14	15.81

Table 3.14 – Statistics of Datasets.

For this analysis, every symbol and punctuation is cleaned off, as well as the urls. Considering the important role played by hashtags and mentions (@USER) in the tweet context, they have been taken into account as tokens. The results of the analysis reveal that the corpora have a similar percentage of lexical diversity, with a soft variation of words per

[17] https://github.com/ZeerakW/hatespeech

[18] We used Twitter API for Python to download the tweets by ids.

tweets. Despite the different size of the corpora, the similar TTR value suggests that the users could use a similar, and probably informal lexicon in both contexts. To better understand the analogies between positive (i.e., `misogynistic` and `sexist`) and negative (i.e., `non-misogynistic` and `non-sexist`) classes in our corpora, a lexical and stylistic analysis was carried out. To consider also the level of offensiveness of each corpus, we used an available online lexicon of English swear words[19]. Tables 3.15 and 3.16 report the obtained values.

	AMI_IberEval	**AMI_EVALITA**	**SRW**
Number of tweets	1,568	1,785	2,503
Vocabulary	5,155	5,932	7,846
Number of tokens	27,477	31,535	44,803
Type-token ratio	18.76%	18.81%	17.51%
Average of words	17.52	17.67	17.90
Swear words	3,176	3,587	1,261
Feminine pronouns	353	344	251
Masculine pronouns	111	80	66

Table 3.15 – Statistics for Positive Classes in each Dataset.

	AMI_IberEval	**AMI_EVALITA**	**SRW**
Number of tweets	1,683	2,215	2,503
Vocabulary	6,228	7,133	6,633
Number of tokens	27,954	37,038	34,335
Type-token ratio	22.28%	19.26%	19.32%
Average of words	16.61	16.72	13.72
Swear words	2,087	2,251	600
Feminine pronouns	158	150	88
Masculine pronouns	157	159	117

Table 3.16 – Statistics for Negative Classes in each Dataset.

Looking at these tables, despite the different number of tweets in the corpora, the TTR and the number of words per positive samples is very similar, differently from the values of negative samples. Focusing on the differences between the values obtained for positive and negative classes, we can see that a richer lexicon is used in non-misogynistic and non-sexist tweets. This factor could be due to the fact that misogynistic and sexist texts mainly contain a substantial number of insults and profanities, since the misogynistic and sexist tweets often aim at offending or hurting the target. Moreover, a very simple investigation about the presence of masculine pronouns ("he", "his") and feminine pronouns ("she", "her") was carried out. As Tables 3.15 and 3.16 show, the positive samples

[19] https://www.cs.cmu.edu/~biglou/resources/

are principally focused on women, respect to the negative ones. The obtained values indicate that also sexist tweets talk more about women than men. Finally, the average of words per tweets reveals that misogynistic and sexist messages are longer than non-misogynistic and non-sexist ones. Indeed, the user tends to justify the negativity of his opinion or underline that his statement is not misogynistic or sexist, such as:

(34) *Because femininity is so horrible! @USER I'm not sexist but if a dude cries because of a girl in a wedding dress then he has a vagina*[20]

Some important analogies emerged from this analysis, and they can also be inferred observing some positive examples, as reported in Table 3.17. It is evident that they tend to target women with the purpose to discredit them and underline male superiority.

corpora	**text**
AMI_IberEval	*What do you call a women that has a brain? Pregnant with a baby boy.* *What's worse than a girl who gives rough handjobs? A feminist.*
AMI_EVALITA	*@USER no I said hope. I hope you women learn your place! #SitDownInTheKitchen* *Who makes the sandwiches at a feminist rally?*
SRW	*RT @USER Dont ever let women drive, they'll break your arm! #notsexist* *Are you even a real person? @USER I'm not sexist. But Men are superior to women.*

Table 3.17 – Some Positive Examples from the Datasets.

3.1.2.2 Experiments and Results

The previous analyses highlight that sexist and misogynistic corpora are similar enough. Especially, the fact that sexist tweets are more focused on women than men supports the idea that sexist messages in the majority of the cases tend to discriminate women. Despite the corpora that we used are representative of a little part of real big data on the web, they help to understand the aspects of hate speech against women. The problem of the identification of sexist (presumably against women) and misogynistic tweets is addressed as a classification task. On the basis of the performance obtained in previous works (Section 2.3.1), we employed an SVM classifier informed with linguistic features to detect misogyny in English. In particular, we used the SVM with an RBF kernel and parameters of $C = 5$ and $gamma = 0.1$; and a set of features that involves: modeled lexica about specific aspects of online hate speech against women (see Table 3.4) and stylistic features captured by means of n-grams of words and characters.

[20] Tweet extracted from SRW corpus.

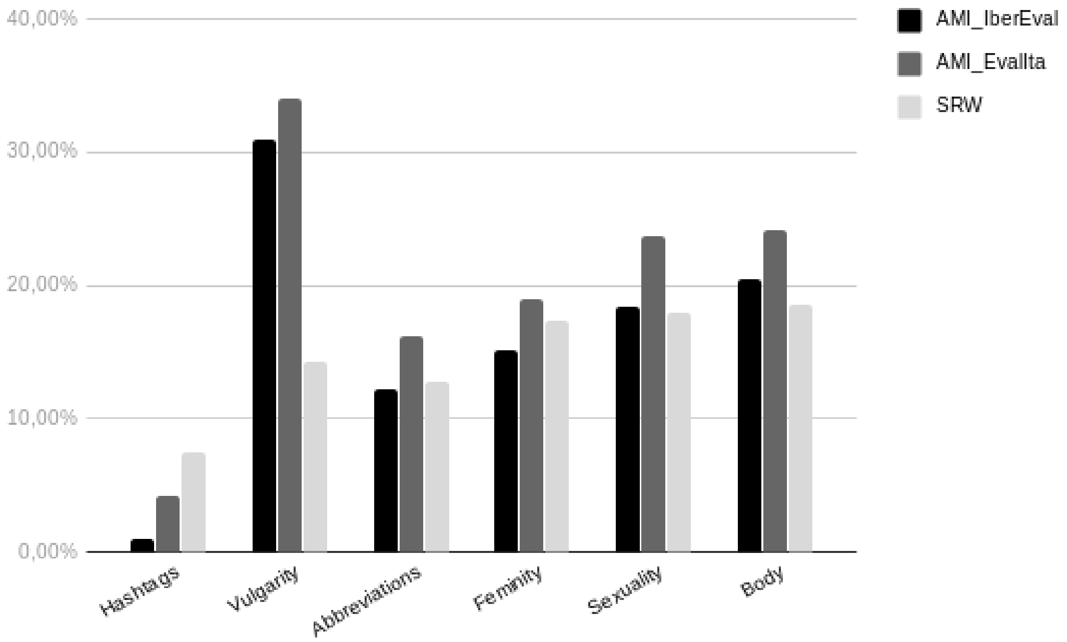


Figure 3.1 – IG of Lexica in each Dataset.

With regard to the used lexica, we observed the IG of each lexicon in the datasets. Looking at Figure 3.1 we can notice that the differences among the corpora are minor suggesting that these lexica could be important for both classification tasks. In particular, Sexuality and Human Body lexica play an important role, showing that the discrimination even in sexist tweets is mainly towards women. Also, words relative to Femininity have similar relevant distribution in all corpora, as well as the use of abbreviations. We report some examples in Table 3.18.

Observing Tables 3.15 and 3.16 and Figure 3.1 we notice that the list of profanities and offensiveness has a higher impact especially on misogynistic corpora, suggesting that the tweets of these corpora aim to offend more than the sexist tweets. However, in general, these results do not hide the fact that sexist tweets are usually addressed to women and that sexist discrimination, as the various reported examples show, could contribute to offend women. To catch linguistic patterns in spite of the orthographic errors, typical of spontaneous writing in tweets, we adopted a range of characters from 1 to 7. Examining some character n-grams extracted from each dataset, the tweets seem to present similar constructions typical of the informal speech (“gonn”, “I m” , “yo”). In addition, offensive words such as “hoe” and “bitc” are located. Table 3.19 reports some BoC with the highest Information Gain values extracted from the datasets.

We noticed a similar good performance using unigrams, bigrams and trigrams (UBT). Specifically, bigrams and trigrams help the system to recognize syntactical and lexical

corpora	**text**
AMI_IberEval	Me trying to flirt- You have really nice eyebrows... I'd like to cum on them to see if they wash off RT @KGJump12: Sometimes I want a *girlfriend*, but then I quickly remember how i hate *women*
AMI_EVALITA	Love a b***h for what ? all she good for is sucking dick! They've made it almost impossible for Men to be dominant, in the *matriarchal* western society. fuckfeminism
SRW	RT @of_The_Guild It really pisses me off when anime girls don't have big boobs #NotSexist - A Misogynist @parody_guy A *woman* wants *her* man to treat *her* like a *princess* to the world and f*** *her* like a whore. - Someone

Table 3.18 – Misogynistic and Sexist Tweets Containing Words from Sexuality (underlined) and Femininity (in italics) Lexica.

patterns of abusive language which are difficult to be captured taking into account only the bag of words (BoW), preserving the order of the words. Some examples of significant bigrams and trigrams are reported in Table 3.20. This table shows an interesting lexical similarity between the corpora: purpose of hate and women as target.

AMI_IberEval	**AMI_EVALITA**	**SRW**
‘ the ’	‘ thi’	‘ thei’
‘#male’	‘a bi’	‘thr’
‘#ye’	‘a bit’	‘ thes’
‘our a’	‘out o’	‘swe’
‘#yes’	‘to r’	‘n an’
‘tim ’	‘a c’	‘s i ’
‘ ther’	‘ thou’	‘ #m’
‘ a b’	‘end ’	‘hel’
‘hea’	‘ a bi’	‘ #mk’
‘eon ’	‘hen i’	‘ #mkr’

Table 3.19 – The Most Relevant Bag of Characters in each Dataset.

bigrams			**trigrams**		
AMI_IberEval	AMI_EVALITA	SRW	AMI_IberEval	AMI_EVALITA	SRW
(‘a’, ‘b***h’)	(‘a’, ‘woman’)	(‘but’, ‘women’)	(‘f***’, ‘off’, ‘you’)	(‘Shut’, ‘f***’, ‘up’)	(‘I’, ‘am’, ‘sexist’)
(‘a’, ‘girl’)	(‘a’, ‘hoe’)	(‘sexist’, ‘but’)	(‘What’, ‘the’, ‘difference’)	(‘WomenSuck’, ‘don’, ‘t’)	(‘I’, ‘m’, ‘sexist’)
(‘a’, ‘whore’)	(‘a’, ‘b***h’)	(‘girls’, ‘are’)	(‘a’, ‘ass’, ‘b***h’)	(‘a’, ‘stupid’, ‘b***h’)	(‘I’, ‘not’, ‘sexist’)
(‘b***h’, ‘I’)	(‘women’, ‘are’)	(‘women’, ‘are’)	(‘a’, ‘good’, ‘girl’)	(‘a’, ‘whore’, ‘you’)	(‘Not’, ‘sexist’, ‘but’)
(‘your’, ‘ass’)	(‘she’, ‘s’)	(‘but’, ‘girls’)	(‘a’, ‘hoe’, ‘I’)	(‘don’, ‘t’, ‘get’)	(‘but’, ‘I’, ‘hate’)
(‘a’, ‘woman’)	(‘re’, ‘a’)	(‘women’, ‘should’)	(‘a’, ‘hoe’, ‘b***h’)	(‘don’, ‘t’, ‘have’)	(‘but’, ‘women’, ‘are’)
(‘a’, ‘hoe’)	(‘a’, ‘whore’)	(‘girls’, ‘should’)	(‘a’, ‘stupid’, ‘b***h’)	(‘don’, ‘t’, ‘know’)	(‘call’, ‘me’, ‘sexist’)
(‘stupid’, ‘b***h’)	(‘stupid’, ‘b***h’)	(‘no’, ‘sexist’)	(‘a’, ‘whore’, ‘you’)	(‘f***’, ‘up’, ‘you’)	(‘sexist’, ‘I’, ‘hate’)
(‘she’, ‘s’)	(‘a’, ‘girl’)	(‘but’, ‘female’)	(‘b***h’, ‘suck’, ‘my’)	(‘is’, ‘a’, ‘b***h’)	(‘sexist’, ‘I’, ‘just’)
(‘b***h’, ‘you’)	(‘Women’, ‘are’)	(‘when’, ‘women’)	(‘on’, ‘my’, ‘dick’)	(‘is’, ‘a’, ‘cunt’)	(‘sexist’, ‘but’, ‘female’)
(‘my’, ‘dick’)	(‘I’, ‘hate’)	(‘women’, ‘can’)	(‘re’, ‘a’, ‘b***h’)	(‘is’, ‘a’, ‘whore’)	(‘sexist’, ‘but’, ‘girls’)
(‘ass’, ‘b***h’)	(‘to’, ‘rape’)	(‘promo’, ‘girls’)	(‘re’, ‘a’, ‘whore’)	(‘women’, ‘are’, ‘stupid’)	(‘sexist’, ‘but’, ‘hate’)
(‘you’, ‘stupid’)	(‘f***’, ‘you’)	(‘all’, ‘female’)	(‘the’, ‘difference’, ‘between’)	(‘you’, ‘stupid’, ‘b***h’)	(‘sexist’, ‘but’, ‘women’)

Table 3.20 – The Most Relevant Bigrams and Trigrams of Words in each Dataset.

The preprocessing of data aims at deleting all symbols, emoticons and urls from tweets' text. To perform a correct match between the words in the texts and each manually modeled lexicon, we applied the lemmatizer provided by NLTK for English. Thus, each tweet is represented as a vector composed of: the weights of n-grams of characters and words calculated with TF-IDF; the weights of lexical features calculated using IG. In order to take into account also words that are relevant for the classification, but not in the lexica, we added in the vector their weights calculated by means of IG.

The evaluation is performed using 10-fold cross-validation to inspect deeply the performance of the used approach, and considering the balanced nature of the datasets as measure of evaluation we employed the *Accuracy*. The obtained results have been compared with a baseline obtained using unigrams weighted by means of TF-IDF. Table 3.21[21] reports also the singular contribution of each feature.

	AMI_IberEval	**AMI_EVALITA**	**SRW**
Baseline	*0.751*	*0.783*	*0.884*
UBT	0.745	0.786	0.893
BoC	0.754	0.788	0.871
Lexica	0.699	0.735	0.735
All Features	**0.760**	**0.795**	**0.894**

Table 3.21 – Results Based on Accuracy for each Dataset.

Stylistic features in general seem to perform well in spite of the limited increase with respect to the baseline. Whereas, the combination of all features reached better results, confirming that a system improves its knowledge if informed with specific lexica that encourage the attention on specific topics, stereotypes, and insults towards women. Moreover, we can notice a difference between misogyny and sexism detection in these corpora: in the first task character n-grams appear as the best performing feature, whereas in the second the best one is the combination of bag and sequences of words.

Looking at Table 3.19 and Table 3.20, we can hypothesize that character n-grams are important in a context where the vocabulary is composed of a great amount of insults because help the system to perceive differences less evident between negative and positive classes. Whereas in the sexist corpus, where insults are less frequent, the co-occurrence of words could represent an important pattern for the classification. Finally, in order to better understand how our approach works, we reported some examples correctly predicted using the combination of all the features in Table 3.22 and an error analysis on all corpora.

[21] The most contributing features are underlined.

corpora	**text**
AMI_IberEval	@EbersohlMyra @Lolomonet @mandee_boo @_morganmariaa *B***h* you look like a dog Control, I own your ass now *b***h* This *stupid* Bitch Lied. https://t.co/KOeOdaibX7
AMI_EVALITA	Too many *women* confuse their hearts with their *vaginas* The more *WOMEN* in the workplace means LESS *WOMEN* IN THE KITCHEN. Like and retweet if you see this as a big problem #ilovefood #homecookedm @JWMofficial: You're wearing yoga pants to show off your ass. Don't give me this *bull* shit 'they're comfy!' we all know the truth *ladies* *#ASS #ASS #ASS*
SRW	@mannythemenace Call me sexist, but I hate *female* receptionists. They always give me attitude like I'm the reason their *bf* broke up with 'em RT @RobDurbinn I'm not *sexist* but *women* rappers are a bigg*er* joke than the WNBA @terrencewoods I'm not *sexist* but *girls* our age that drive is super scary

Table 3.22 – Correctly Predicted Tweets with the Combination of Features: N-Grams (underlined) and Words from Lexica (in italics).

3.1.2.3 Error Analysis

Despite the natural differences among misogynistic and sexist corpora such as the schema of annotation or the keywords used to collect the tweets, the misclassified instances are similar. For instance, the lack of world knowledge is an evident obstacle for the system. One of these cases can be observed in the sexist corpus. Some tweets talk about the Australian competitive cooking game show, My Kitchen Rules (MKR), and they refer to participants or events in the show that are difficult to understand without the context:

(35) *"@USER: Not enough Lemon in their lemon tart #MKR woops". Plenty of tart though*

Waseem and Hovy [2016] specified that they used the hashtags relative to MKR to collect the data, because often the tweets containing #mkr prompt sexist issues directed at the women participating at the show. Among the wrong predictions in sexism detection, few tweets have not been predicted as sexist because their targets are men:

(36) *@USER @USER I made up nothing. A stoner said that and you are too ignorant of world events to know he's wrong. #sorryitsaboy*

This kind of error was foreseeable, considering the nature of the corpus. However, for the aim of our analysis, this 'error' suggests that the system works well. Another problem, arose from the lack of context, concerns the presence of URLs in tweets. Indeed, the interpretation of the text sometimes relies on external information conveyed by the links:

(37) *@USER Can you explain why this is wrong? http://t.co/pTkwk45P9P*

In particular, this URL links to another tweet which expresses the common idea that when women are angry, it is supposedly due to the lack of sex in their lives. For this reason, considering the information in the link, the tweet could be annotated as `sexist`. Finally, there are some tweets that describe a sexist or misogynistic situation/event and, despite their label is misogynistic/sexist, they are not predicted as hostile towards women:

(38) *@USER It is Muslim Jihad culture to rape yagidi in Iraq,Christian & Hindu in Pak.purchase poor Muslim girl for sex slave*

For example, in Example (38), the attack is against some cultural group and not against women. Therefore, the system classified this kind of tweets correctly for our proposal. Finally, as seen in previous experiments on misogyny detection some misclassifications are due to the presence of ironic devices:

(39) *@USER Knickers is a tangle, aye? Don't like women much do you? Hysterical things aren't they? Only good a few things, aye?*

Generally, the collections of data extracted from the web are a little representation of all the texts that are published daily on the web by millions of people, and thus a little representation of their opinions. Focusing on the corpora here analyzed, it is possible to notice that the texts that are annotated as `sexist` and that theoretically discriminate male as well as women, actually they tend to discriminate and offend mainly women. This observation is confirmed by the corpora analysis and the analysis of n-grams of characters and words extracted from the texts.

Moreover, the fact that the implemented system oriented to identify the hostility against women works well in the prediction of sexist tweets and of misogynistic tweets reveals that these kinds of collections contain texts that have similar characteristics. Therefore, it seems that the sexist discrimination of women could suggest a hateful attitude against women.Could the sexist 'common way of thinking' be considered innocent? These computational experiments seem to indicate that in general sexist opinions hide a sentiment of hate, making more subtle the hostility, and, in this particular case, the misogynistic attitude.In this line, it could be interesting to investigate the intensity of hostility against women in sexist tweets, little explored until now [Pamungkas et al., 2020a].

3.1.3 Stance or Insults?

As Manne [2017] mentioned, misogyny involves even *third-personal indignation* that is the type of hostility shown towards women who choose the abortion. In addition, as seen in previous experiments, sexist ideology tends to support misogynistic attacks, hiding sometimes, even a sentiment of hate. To investigate these types of misogyny in a multi-genre framework, we analyzed the stance expressed in some data from Twitter and newspapers online about two of the most active and debated issues that involve women: legalization of abortion and the raising of feminist movements. Taking into account the definition of stance as the "intellectual or emotional attitude" towards an issue (Merriam-Webster Dictionary), our investigation focuses on two hypotheses:

H1 such issues, especially on Twitter, where the communication is spontaneous, give rise to polarized debates that could involve aggressive tones and stereotyped ideas which express misogynistic and sexist offenses;

H2 even newspapers (Section 2.1.1) tend to express a stance, in most of the cases implicitly on these hot issues, despite the deontological code.

Expressing a stance, in fact, tends to be a justification and, thus, a mask for hostile attitudes. And it is common that debates on delicate issues, that involve a specific group of people, tend to be discussed vigorously and with utterances that incite to violence and hate:

(40) *One day I'm gonna set an abortion clinic on fire. Anyone wanna join? #prolife*[22]

(41) *MARRIAGE for a man is MURDERAGE, That's right MURDER'RAGE! Women have ruined the trust of men, and destabilized their own future. #feminism*[23]

(42) headline: *Kansas Supreme Court To Rule On Dismemberment Ban In Abortion Case*
text: *Should a majority of the Kansas Supreme Court follow this predictable course and invent a state constitutional right to abortion, it would decimate the will of the people, who have consistently elected one of the most pro-life and pro-active state legislatures in this country. Sadly, in this case, it would not just be democracy that dies in the dark.*[24]

Observing Examples (40) and (41), it is clear that they do not express only disapproval towards the legalization of the abortion and feminist thought, but instigate to violence and express sexist and misogynistic attitudes. In (42), the headline and the text extracted from an article of an American newspaper, the journalist expresses his/her stance against abortion using a strong title and an implicit accusation of homicide against women who choose the abortion. This expression of unfavorable stance implicitly transmits a very negative image of women who have an abortion that could encourage the hostility against them.

To prove our hypotheses, we approached a task of stance detection in tweets and newspapers, investigating particularly the presence of misogynistic and sexist attacks. In

[22] This tweet is extracted from StanceDataset.
[23] This tweet is extracted from StanceDataset.
[24] Article extracted from GDELT-FM.

particular, we created a set of features specific to capture explicit and implicit hostility towards women, and we experimented their contribution from a multi-genre perspective. This kind of approach allows also to bring to light if the detection of especially unfavorable stance, could benefit in general from the information about hostility against women. In particular, we conceive stance detection as a task of prediction of the highest probability of a text to belong to the `against`, `favor` or `neutral` classes. The set of the engineered features focuses on extracting: the stylistic characteristics by means of BoC (with a range from 3 to 5 grams); relevant expressions measuring the TFIDF of unigrams and bigrams; and lexical and semantic information concerning specifically misogynistic and sexist speech. To this purpose, we used the offensive lexica and the lexica described in Table 3.4, and we computed the similarity of the text with misogynistic and sexist context using a word embedding built exploiting the English datasets of AMI competitions [Fersini et al., 2018b,a] and SRW [Waseem and Hovy, 2016]. These features inform linguistically a system based on an SVM classifier with an RBF kernel, and $C = 5$ and $\gamma = 0.1$.

3.1.3.1 Misogyny in Stance on Twitter

To investigate H1, we used the StanceDataset released by the organizers of the Stance Detection shared task at SemEval 2016[25] [Mohammad et al., 2016] and in particular the sections of the corpus about the feminist movement (called here Feminism) and the legalization of abortion (called here Abortion). The distribution of labels is reported in Table 3.23. The use of a benchmark corpus allows us to compare the performance of our models with the others participating at the competition.

dataset	training set			test set			total
	favor	against	none	favor	against	none	
Feminism	210	328	126	58	183	44	949
Abortion	121	355	177	46	189	45	933

Table 3.23 – Distribution of Labels in Feminism and Abortion.

Firstly, we analyzed the lexicon of these corpora and, in particular, of the favorable and unfavorable tweets of the training set (Table 3.24), calculating: the size of vocabulary, the lexical richness by means of the Type-Token Ratio (TTR) and the offensiveness of the data looking at the words belonging to lexica in Table 3.4, NoSwearing lexicon[26] and the English version of Hurtlex. To obtain these values we preprocessed the texts cleaning off every symbol, punctuation and urls; whereas hashtags[27] and mentions (@USER) are taken into account as tokens. All words are lemmatized with WordNet lemmatizer provided by NLTK.

[25] http://alt.qcri.org/semeval2016/task6/

[26] https://www.noswearing.com/dictionary

[27] The query hashtags used to extract tweets from Twitter are deleted by the organizers during the creation of the dataset to exclude obvious cues for the classification [Mohammad et al., 2016].

	Abortion		Feminism	
	`favor`	`against`	`favor`	`against`
Number of tokens	1,961	6,052	3,948	6,027
Vocabulary	678	1,739	1,313	1,881
Type-token ratio	34.57%	28.73%	33.25%	31.21%
Sexuality	113	291	202	308
Human body	20	68	26	28
Femininity	186	380	410	684
Offenses	9	26	66	120
NoSwearing lexicon	55	127	116	201
Hurtlex conservative	221	540	432	773
Hurtlex inclusive	524	1,604	1,029	1,696

Table 3.24 – Lexicon Analysis of the Training Set in Feminism and Abortion.

Table 3.24 shows that unfavorable tweets contain the majority of offensive words. We need to consider that the number of unfavorable tweets in both datasets is higher than favorable ones, however, the amount of some offensive words (for example from Hurtlex) is very high. Overall, the tweets labeled as `against` from Feminism are more offensive than the ones collected towards Abortion.

Secondly, performing various experiments with 10-fold cross-validation and weighting features with Mutual Information, we selected specific features for stance detection towards the legalization of abortion (SDA) and feminist movements (SDF) as shown in Table 3.25.

feature	**SDA**	**SDF**
Sexuality		v
Profanity		
Femininity		v
Human body		
NoSwearing		v
Hurtlex		v
Similarity		v
Unigrams		v
Bigrams	v	
BoC	v	v

Table 3.25 – Selected Features for the SDA and SDF Tasks.

From our experiments lexical information does not seem to support stance detection towards the legalization of abortion suggesting that these tweets are not very offensive against women. However, analyzing the most relevant bigrams we noticed some connotative expressions that transmit the idea that women that have or want to have an abortion are criminal (such as *killing baby*, *let live*, *use someone*) or are insensitive and deserve to die (*death penalty*, *black life*, *dead woman*). Similar expressions appear extracting the most informative unigrams for stance detection towards feminism (*spankafeminist*, *feminazi*). For this second task we selected the lexica with the highest values in Table 3.24 (as showed in Table 3.25); and differently from the first task, the computation of similarity with misogynistic and sexist contexts prove to be a useful feature. In particular, we calculated the cosine similarity between the tweets of the FEMINISM dataset and misogynistic and sexist tweets. The used word embedding was built from AMI_IBEREVAL, AMI_EVALITA and SRW datasets taking into account a window of 5 words and a vector with a length of 100 items.

Experiments and Results These tasks present some particularities respect to the previous experiments on misogynistic and sexist contents: the collections of data are imbalanced (Table 3.23) and the classification problem is not binary but multiclass. For this, we used the function to balance the weights of the classes provided by the *scikit-learn* library, and we predicted the probability of a tweet to belong to the `against`, `favor`, and `none` classes and then, we chose the class with the highest probability.

We compared the performance of our models with the results obtained by the first ranked system at the Stance Detection shared task and the baselines provided by the organizers [Mohammad et al., 2016]. Therefore, we used the same measures of evaluation used in the competition. In particular, Mohammad et al. [2016] considered the $f1$-scores for the `favor`, and `against` classes calculated on precision and recall values, and for the ranking used their average ($f1$-avg). The `none` class is considered as negative class of the favorable and unfavorable tweets. Table 3.26 reports these results.

The provided baselines include 4 approaches: *Majority class*, *SVM-unigrams* using unigrams of words, *SVM-ngrams* exploiting word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) as features, *SVM-ngrams-comb* that combines all the sets of training data for the training using words and character n-grams. The teams that performed the best performance employed classifiers based on deep leaning approaches, without addressing the peculiarities of each target. Their principal novelty lies in the techniques of transfer learning [Zarrella and Marsh, 2016] and augmentation of data [Vijayaraghavan et al., 2016] used to face the scarcity of training data. Moreover, whereas Vijayaraghavan et al. [2016] proved the advantages of a character-level model in an augmented set of data, our results show that in a context of scarce opinions towards issues that involve a specific group of people, the lexical information could help the systems to detect stance correctly. In particular, our models achieved the highest result in the classification of the stance towards the legalization of abortion, and overcame the challenging baselines for the detection of stance towards feminist movement. In this last case,

[28] Vijayaraghavan et al. [2016]

[29] Zarrella and Marsh [2016]

approach	**SDA**	**SDF**
	$f1$-avg	$f1$-avg
Baseline		
Majority class	0.403	0.391
SVM-unigrams	0.601	0.556
SVM-ngrams	*0.664*	*0.575*
SVM-ngrams-comb	0.637	0.528
First Ranked System	0.633[28]	**0.621**[29]
Our Approach		
char-ngrams	0.663	0.526
+unigrams-words	–	0.538
+bigrams-words	**0.685**	–
+lexica	–	0.581
+similarity	–	*0.603*

Table 3.26 – Results obtained for both Tasks in Tweets.

the technique of transferring features from other systems trained on large and unlabeled datasets adopted by Zarrella and Marsh [2016] allows the system to acquire more general knowledge than our model.

In order to understand the benefits and disadvantages of our approaches, we analyzed the values of precision and recall obtained by each feature. On the one hand, *precision* evaluates how well the systems classify only the relevant documents; on the other hand, *recall* reports the sensitivity of the model estimating how well the systems identify all relevant documents [Blair and Maron, 1985]. The scores are reported in Tables 3.27 and 3.28, and, to have a comparison with the baseline, we reproduced the performance of the most challenging baseline: *SVM-ngrams*[30].

For this kind of investigation, we reported in Table 3.27 even the performance obtained employing lexical features related to abusive and sexist language (*Model 1*) and semantic features (*Model 2*). As we can observe, the use of lexical features related to abusive language against women achieved a higher value of recall in the `against` class than the best performing model (*Our Model*) in SDA. This means that in a real context where the social platforms or ITC companies try to retrieve all the possible offensive opinions towards a specific issue, lexica prove to be a helpful feature. However, the specialists should go through the false positives (i.e., the tweets predicted as `against` but actually favorable). In our case, the purpose is to find a balanced model, tuned on precision and recall, that is able to predict correctly both classes (`against` and `favor`).

[30] Although we used the same approach described in Mohammad et al. [2016], the f-avg scores are slightly different from the ones reported in the overview of the task (64.07 for SDA and 55.52 for SDF).

	precision	**recall**	**$f1$-score**
Baseline			
SVM-ngrams			
Favor	0.466	0.739	0.571
Against	0.805	0.635	0.710
Our model			
char-ngrams			
Favor	0.500	0.739	0.596
Against	0.821	0.656	0.729
+bigrams-words			
Favor	0.507	**0.761**	0.609
Against	**0.835**	0.698	0.761
Model 1			
char-ngrams+bigrams-words+lexica			
Favor	0.481	0.283	0.356
Against	0.724	**0.804**	0.762
Model 2			
char-ngrams+bigrams-words+similarity			
Favor	**0.519**	0.587	0.551
Against	0.794	0.757	0.775

Table 3.27 – Precision and Recall Scores for SDA in Tweets.

Similarly to *Model 1*, *Model 2*, using the cosine similarity between the analyzed tweets and misogynistic and sexist contexts, obtained a higher recall in the `against` class than *Our Model*. Although *Model 2* seems to be more balanced than the previous one, the low recall in the `favor` class suggests that is not really able to retrieve favorable opinions. As showed by the f-score values, *Our Model* seems to perform a more balanced prediction of both classes, even compared to baseline scores.

From Table 3.28 we can notice that our approach seems to make the system more sensitive to retrieve unfavorable opinions than the baseline model. However, *Our Model* achieves more balanced values of recall and precision for both classes than the baseline model. In addition, compared to the initial scores obtained with character n-grams and unigrams of words, we can observe that lexical and semantic features related to misogynistic and sexist speeches guided the system to retrieve favorable and unfavorable tweets correctly. Such finding suggests that unfavorable opinions are expressed with hostile tones, corroborating H1.

Error Analysis Finally, we carried out an error analysis to better understand where our systems fail and where match the correct classification of the tweets. In general, we noticed that the systems have difficulties to predict correctly tweets that do not contain contextual information such as a hashtag. Indeed, as affirmed in Mohammad et al. [2016],

	precision	**recall**	$f1$**-score**
Baseline			
SVM-ngrams			
Favor	0.339	**0.690**	0.454
Against	**0.797**	0.557	0.656
Our model			
char n-grams			
Favor	0.323	0.534	0.403
Against	0.741	0.579	0.650
+unigrams-words			
Favor	0.330	0.534	0.408
Against	0.729	0.617	0.669
+lexica			
Favor	0.380	0.603	0.467
Against	0.768	0.634	0.695
+similarity			
Favor	**0.416**	0.638	0.503
Against	0.763	**0.650**	0.702

Table 3.28 – Precision and Recall Scores for SDF in Tweets.

some hashtags have been used as queries to extract tweets and replaced with '#SemST' to exclude obvious cues for the classification. Some of these tweets are:

(43) *Some men do not deserve to be called gentlemen #SemST*[31]

(44) *A much needed 3 days with these guys @rory3burke @Im_ Brady missed @JimmahTwittah but what a weekend #SemST*[32]

(45) *In civilian clothes and someone laughs at me thinking its a joke that I'm apart of the U.S. Navy. #SemST*[33]

Others are hard to understand because refer to specific events or contexts, such as episodes or tv shows:

(46) *As I rewatced Charmed episodes! LOVING IT EVEN MORE! #SemST*[34]

(47) *..Can I also add that I really enjoyed looking at @TahirRajBhasin in #Mardaani :P Tahir, you were a dashing baddie! #Bollywood #SemST*[35]

[31] The original tweet is: *Some men do not deserve to be called gentlemen #WomenAgainstFeminism.*

[32] The original tweet is missing.

[33] The original tweet is missing.

[34] The original tweet is:*As I rewatced Charmed episodes! LOVING IT EVEN MORE! #feminism.*

[35] The original tweet is:*..Can I also add that I really enjoyed looking at @TahirRajBhasin in #Mardaani :P Tahir, you were a dashing baddie! #Bollywood #Feminism.*

or, because contain figurative language such as irony:

(48) *Equality is the police burying a domestic violence accusation against a female sports star, too #wedidit #usa #SemST*[36]

(49) *@LifeSite Right, where are the pre-born women's rights? #allLivesMatter #equalRights #SemST*[37]

The analysis of the counterpart confirms the optimal results obtained, particularly, in the `against` class. The used features, aimed at putting attention on lexical information, help the system to recognize unfavorable stance, especially when it could be dangerous for a target:

(50) *I am about to deck these 2 b***hes in the f***ing mouth. #1A #2A #NRA #COS #CCOT #TGDN #PJNET #WAKEUPAMERICA #SemST*

(51) *Meanwhile, @JustinTrudeau wants to waste your money to kill innocent children in the womb. #dangerous #hypocrite #noChoice #SemST*

(52) *Women are taught to put their values into their hymens, rather than their intelligence, accomplishments, goals or character #feminism #SemST*

(53) *You should start using Google translate @baedontcare, it is sooooo easy even retarded feminists like you can use it. #SemST*

Expressing stance with hostile tones surely does not guarantee constructive debates or the *democratic virtual space* expected by the first enthusiasts of Internet [Barlow, 2001]. On the contrary, it incites misbehavior that could have violent effects also in real life.

3.1.3.2 Misogyny in Stance on Newspapers

As Examples (21) and (42) show, these hostile tones crowd into newspapers, even if they appear less explicit and expressed in a very formal manner. To investigate better the presence of implicit misogyny even in the stance expressed in newspapers, we created GDELT-FM, a collection of news from GDELT[38].

GDELT-FM Dataset Considering the results of our experiments on Twitter that show that debates on feminist movements report a very offensive and misogynistic language against women (see Tables 3.24 and 3.25), we collected news about feminist movements. In particular, we considered the news published by newspapers online from the 1st of October to 31st of December in 2017 in Europe, Japan and USA, and linked to

[36] The original tweet is missing.

[37] The original tweet is: *@LifeSite Right, where are the pre-born women's rights? #prolife #allLivesMatter #equalRights*

[38] https://www.gdeltproject.org/

the spread of the hashtag #metoo in occasion of the legal case of sexual assault and harassment in the workplace risen against the movie producer Harvey Weinstein.

The choice to use the database provided by the GDELT Project, supported by Google Jigsaw, relies on the amount of broadcast, print, and web news that they collect every day from every country identifying also automatically people, locations, organizations, themes, sources, emotions, counts, quotes, images, and events. The project provides a free open platform that allows to monitor events on the entire world from the 1st of January 1979. In particular, we downloaded news from the GDELT Global Knowledge Graph (GKG) dataset that is an ensemble of news articles complete with related information (such as persons, organizations, locations, themes, events, sources, and sentiment) extracted from newspapers published every day in all the globe from April 2013. The metadata have been exploited to collect our dataset of news about the Weinstein's scandal. Moreover, we used the automatically generated *tone* (i.e., polarity) of the article to annotate automatically the stance of the newspapers/news-websites towards feminist movements.

Following the technique described by Yoshioka et al. [2018], we calculated the Polarity-based Stances (PS) of each news-website. Their formula exploits the *tone* of the article that resumes in terms of average its attitude, computed by the difference between the percentage of positive and negative terms in the text of the article. Let d be the article published by the news-website w, and t the tone of d. Looking at t we annotated the stance of d s_d as `favor` (1), `none`/`neutral` (0) and `against` (−1). To determine the stance we used a threshold σ as follows:

$$s_d = \begin{cases} 1 \; t > \sigma \\ 0 \;\; -\sigma < t < \sigma \\ -1 \; t < \sigma \end{cases} \tag{3.1}$$

Considering the s_d we can calculate the $\overrightarrow{PS}_w$ using the equation

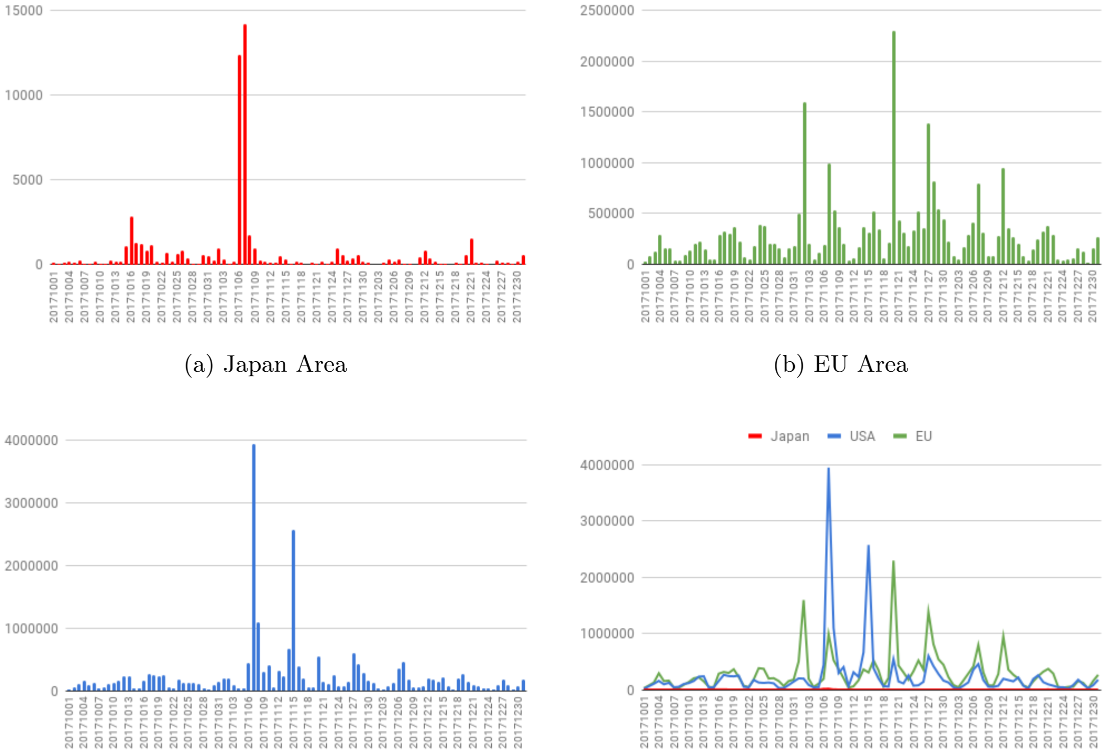
$$\overrightarrow{PS}_w(\tau) = \left(\frac{\sum_{d \in w_\tau} (\mathbf{1}\,[s_d = 1])}{|w|}, \frac{\sum_{d \in w_\tau} (\mathbf{1}\,[s_d = -1])}{|w|} \right) \tag{3.2}$$

where τ is the theme (i.e., feminist movements) and w_τ is the set of articles regarding the selected τ. In order to obtain a consistent value of $\overrightarrow{PS}_w$ for each news-website, we applied this formula only to news-websites containing more than 10 articles about feminist movements, that is 184 news-websites containing a total of 3,673 news. Table 3.29 shows the distribution of labels in this dataset.

Statistical Analysis Before carrying out the computational experiments, we analyzed the content of these news per geographical area. In particular, we show in Figure 3.2 how many events are related to Europe, USA and Japan areas per day in the news. In this way, we can analyze the most important events about feminist movements in each area between October and December 2017.

label	training set	
	news-websites	news
favor	122	2,468
against	52	1,048
none	10	157
total	184	3,673

Table 3.29 – Distribution of Labels in GDELT-FM.



(a) Japan Area

(b) EU Area

(c) USA Area

(d) Overall

Figure 3.2 – Number of Events per Area between October and December 2017 in GDELT-FM.

The peaks of the curves in the graphs of Figure 3.2 increase especially during the month of November (d), when the group of victims released a list of the instances of sexual abuse (rape and harassment) committed by the movie producer Harvey Weinstein. In particular in EU and USA areas refer to (b) and (c), this scandal had a large media coverage, whereas in Japan (a) the attention of media was focused especially on the discourse of Ivanka Trump at the World Assembly for Women in Tokyo and the relative 'womenomics' initiative of Japanese government. Some news are also related to the safety of Tokyo for women compared to other big towns. In Europe, the majority of events in the peaks of the curves in (b) regards to the spread of the novel hashtag #metoo, and other related to hashtags such as #BalanceTonPorc, new reports about gender gap released by World Economic Forum and various protests and parades organized for the International day for the Elimination of Violence Against Women. In the USA (c), the scandal of media and movie industry is the main issue treated by newspapers/news-websites along with a list of protests organized for women's rights, reports about domestic violences and new political cases that accused parliamentarians of sexual abuses and offenses against women.

Experiments and Results As said before, we employed the same algorithm of classification used for tweets applying the same function to balance the weights of the classes and the same strategy to predict the class of each document (i.e., the news) in a multi-class environment. Although the set of experiments is similar, here we applied only the 10-fold cross-validation considered our main proposition of valuation of the contribution of features aimed at capturing implicit and explicit hostility against women. Table 3.30 reports the results obtained with the same set of features used for SDF on Twitter (see Table 3.25). The scores represent the average of the f-avg scores.

approach	**SDF**
	$f1$-avg
Baseline	
SVM-ngrams	*0.702*
Our Approach	
char-ngrams	0.732
+unigrams-words	0.774
+4 lexica	0.794
+all lexica	0.834
+similarity	**0.874**

Table 3.30 – Results obtained for SDF in News-Websites.

As baseline for our experiments, we proposed the same model of the best performing baseline in Mohammad et al. [2016] for SDF: *SVM-ngrams* (see Section 3.1.3.1). Differently from SDF on Twitter, our experiments on news suggest that to obtain a good performance in stance detection in news our system needs to be informed also with lexica about profanities and human body characteristics. To investigate their usefulness, we analyzed some news that contain the offensive words and words related to body. We noticed that some of them, even if they are favorable to feminist movement and encourage women to declare the suffered abuses, use rude tones that could be moderated in a non-offensive perspective, like in the following extract:

(54) headline: *Victim-blaming, misogyny, and conspiracy theories: A Rebuttal to Blogger JJ Gross*
text: *One can, of course, ask what young lady with half a brain would go up to a mogul's hotel room unchaperoned and expect to find anything other than a dirty old man bartering stardom for sexual favors. What exactly where (sic) these girls thinking? And who forced them to comply, or even remain in Weinstein's room? Surely he did not hold them at gunpoint. Had Gross read up on the case he would have been supplied with an answer: Weinstein was aided and abetted by an entire industry, as in this sad piece. And that because he wielded so much power, women were terrified of coming forward. He was holding their careers and their reputations at virtual gunpoint. Not to mention the accounts of him physically blocking doorways.*

(55) headline: *Sarah Vine on the 'hysterical Westminster witch hunt'*
text: *Ms Leadsom should be careful, then. She doesn't want to end up being the McCarthy de nos jours. Because make no mistake: this so-called sex scandal has all the hallmarks of a moral panic. [...] What started as a WhatsApp group of parliamentary employees swapping notes on their bosses has turned into a mob of aggrieved 'victims' claiming a million sexual micro-aggressions against a number of unnamed individuals who, it seems, are not even allowed to know where they are supposed to have overstepped the mark. Words like 'handsy' and 'inappropriate' seem to make up the bulk of the accusations — terms that can mean almost anything but, in reality, prove nothing. If someone is upset and an MP puts a reassuring arm around her shoulder, is that inappropriate? If they make a clumsy joke, is that an 'unwanted advance'? [...] In other words, it's the revenge of the millennials, many of whom will have had their senses of humour surgically removed at university. Theirs is a generation that seems permanently aggrieved, in a perpetual state of disgust at anyone over the age of 30. [...] Anne Robinson put her finger on the button when she pointed out that in the Seventies, pioneering young feminists such as herself had a more robust attitude to men behaving badly than the 'fragile' women of today. [...] But the problem with the current generation of young women is that they have somehow got it into their heads that they don't have to stick up for themselves, or take responsibility for their own safety. Feminism has taught them that they are entitled to equality and respect, even if they have done nothing to earn it. Common sense and the intelligent rules of human behaviour have been replaced by a childish desire to push boundaries and a touchy, uppity tendency to take offence at the slightest thing. Thus you have women waving their breasts around in public in so-called 'free the nipple' protests — and then complaining when men are caught ogling them.*

The lexical information proves its usefulness in our experiments, especially in combination with semantic context given by the cosine similarity calculated on the basis of a word-embedding created from a misogynistic and sexist collection of tweets. This last feature, indeed, allows the system to interpret the lexica in specific contexts, such as the hostility against women. Finally, the important contribution of lexical and semantic features in news is due to the style and lexical richness of an article in a newspaper, a genre very different from the informal and short text of a tweet. However, these results show that, like in tweets, the employment of features aiming at capturing linguistic abuses against women tends to increase the performance of stance detection, confirming H2.

Although the interesting results obtained on news, we recognize some limitations, especially about the technique adopted to annotate the stance based exclusively on the polarity of the documents. Even though its interpretation could appear generic, we think that this technique could help humans to have a first sight on the stance of newspapers containing big amount of news that hardly could be annotated manually, such as in our case. Our study is a first proposal to approach the unsolved problem connected to the spread and feeding of sexist ideology and enforcement of misogyny even in newspapers.

3.1.4 Discussion

Looking at our purposes described at the beginning of this Section 3.1, we can now resume some main observations emerged from the experiments. In particular, we focused on:

1) the automatic detection of misogyny in multilingual texts (English, Spanish, and Italian tweets) looking at the different types of misogynistic attacks more common in each language;

2) the differences and analogies between sexist and misogynistic tweets from the automatic language processing perspective observing principally the social and conventional biases;

3) the expression of misogyny as third-personal indignation or as result of sexist attacks, and its coexistence with stance on Twitter and newspapers.

First, we were surprised that, despite the distribution of labels in the multilingual AMI datasets (Tables 3.1 and 3.2) highlights some differences among languages, actually the selection of features proves that all the considered lexica can help the system to perform better across the language English, Italian, and Spanish **(1)**.

Moreover, these first experiments in misogyny detection confirm the fact that from a computational perspective it is difficult to mark clear boundaries between sexism against women and misogyny; somehow, the former tends to reinforce the stereotypes that feed the hostility and attacks against women. And this is evident even in the categorization adopted by the organizers of the different editions of the AMI shared task: sexist stereotypes offend women, although that appear superficially less ‘dangerous’ than the threats of violence (see examples in Tables 3.3 and 3.17).

As highlighted above, the speeches with the purpose of discriminating a group perceived as outgroup, tend to amplify, justify and motivate the violence of one group against another. For this reason, sexist speeches against women, apparently less dangerous than misogynistic ones, actually affect the perception of women making ‘normal’ some behaviors (institutional or violent) of domination (from men) or submission (from women). As mentioned by Manne [2017], sexism is the ideology that rationalizes the patriarchal social relation, and it could affect or be accepted by men as well as by women. Some psychological studies [Bearman et al., 2009] analyzed how *internalized sexism* takes place in everyday conversations through some dialogic practices: assertion of incompetence, which expresses an internalized sense of powerlessness; competition between women; the construction of women as objects; and the invalidation or derogation of women. Therefore, sexist ideology, as well as the stereotypes that support it, operates implicitly and deeply in the process of *normalization* of discrimination that flows into misogynistic attacks, tending to put order in social and domestic relations **(2)**.

This process of normalization is supported also by sexist humor that contributes to increase the tolerance of sexist events [Ford et al., 2001]. It is a very usual form of communication that, making the reader laugh, masks the offensive intention and the real negativity of the message. Like for humans, deciding whether an ironic message is hateful is an open challenge in abusive language detection. For instance, in our experiments the classifiers struggle to predict correctly ironic and sarcastic messages, even when we tried to inform the system with a feature that captures the sentiment imbalance in the tweets. This suggests the necessity to approach irony and sarcasm with dedicated techniques that make the system aware of this type of figurative language.

Finally, looking at the results obtained with the performed experiments, we can start answering also the first research question of our thesis (**RQ1**). Firstly, we brought to light the necessity to consider the manifestation of hate speech, even in messages that seem to ‘only’ express a stance towards a subject. This emerged from the addition, in a system of stance detection, of semantic references to hurtful messages against women (i.e., by means of features capturing similarities with misogyny content). The system of stance detection towards delicate issues such as feminism movement in tweets and newspapers, improves its performance, discovering an implicit hostile attitude towards women. Secondly, the proposed approaches prove to lead the systems of misogyny, sexism and stance detection to perceive indirect interpretations of the messages in tweets and newspapers. In particular, the obtained results, show the importance of using lexical resources, as showed also by García-Díaz et al. [2021], to take into account specific words or sequence of words that allow inferring negative stereotypes, prejudices and offenses expressed even implicitly in different textual genres **(3)**.

3.2 Hate Speech, Stereotypes and Aggressiveness

Taking into account the taxonomy proposed by Waseem et al. [2017], the boundaries among the different types of abusive language could be defined looking at the target and the degree of connotation of the hurtful expressions. Precisely in this section, we analyze hate speech and its relative dimensions of aggressiveness and stereotypes. Hate speech is a type of utterance that aims at spreading, justifying, inciting violence and hate against a target that could be an individual or a group [Sanguinetti et al., 2018d]. The illocutory force of these utterances could be expressed employing aggressive statements or stereotypes that reinforce the intention of the speaker/user. However, stereotypes and aggressiveness in our intuition should be considered as orthogonal dimension of hate speech; thus they could be expressed even in non-hateful messages.

Adopting this perspective, we wondered mainly about:

1) how these dimensions interact between them and what is the benefit that the systems of hate speech and stereotypes detection could gain from mutual information (Sections 3.2.1 and 3.2.2);

2) if this possible advantage has a distinct impact on different textual genres (Section 3.2.2);

3) and, if the employment of some specific features could help the system to infer implicit biases, double meanings or specific emotions like in the previous experiments on misogyny detection (Sections 3.2.2 and 3.2.3).

3.2.1 The HaSpeeDe 2020: Shared Task

One of our main efforts to address these points was the creation of a dataset of tweets and news headlines annotated considering the presence of hate speech and stereotype. This dataset was released in occasion of the second edition of HaSpeeDe[39] proposed at EVALITA 2020[40].

In HaSpeeDe 2020 [Sanguinetti et al., 2020], we proposed three tasks to participants:

- **Task A - Hate Speech Detection:** binary classification task aimed at determining the presence or the absence of hateful content in the text towards a given target (among immigrants, Muslims and Roma);

- **Task B - Stereotypes Detection:** binary classification task aimed at determining the presence or the absence of a stereotype towards the same targets as the task A;

- **Task C - Identification of Nominal Utterances:** sequence labeling task aimed at recognizing Nominal Utterances (NUs) in data previously labeled as hateful.

[39] http://www.di.unito.it/~tutreeb/haspeede-evalita20/index.html

[40] https://www.evalita.it/2020/tasks

3.2.1.1 HaSpeeDe2020 Dataset

As we can see, the principal novelty of this edition is about the possibility to analyze at computational level the relation between hate speech and stereotype. Indeed, in Francesconi et al. [2019] the error analysis of the main systems on the dataset released in HaSpeeDe 2018, here called HaSpeeDe2018, showed that the occurrence of stereotypes constitutes a source of error in hate speech identification. The other novelties rely on the possibility to test the performance of systems in a multi-genre test set from new and traditional media, and on the occasion of recognizing the NUs as textual reference of hateful attitudes. This last task allows understanding the syntactic strategies used in the expression of hate speech in formal (news headlines) and informal (tweets) textual genres.

HaSpeeDe2020 contains tweets and news headlines collected as follows.

Twitter Corpus The training set (Train_TW) of HaSpeeDe2020 consists of the Twitter portion of the data from HaSpeeDe2018 (4,000 tweets posted from October 2016 to April 2017), and a new subset of the tweets gathered for the Italian hate speech monitoring project "Contro l'Odio" [Capozzi et al., 2019]. This part was retrieved using the Twitter Stream API and filtered using the set of keywords described in Poletto et al. [2017]. The newly annotated tweets were posted between September 2018 and May 2019 and were annotated by Figure Eight (now Appen[41]) contributors for hate speech (`hs`/`non-hs`) and by the task organizers for the stereotype (`stereo`/`non-stereo`) category, employing the same guidelines of annotation used for HSC [Sanguinetti et al., 2018d]. In particular, only data posted between September and December 2018 were included in the training set; whereas the test set (Test_TW) contains the tweets posted between January and May 2019. Differently from this new subset, the part coming from the Twitter portion of HaSpeeDe2018 was already annotated for stereotype, since it was part of the Italian Hate Speech corpus described in Sanguinetti et al. [2018b].

About the task C, the annotation of NUs was carried out by the organizers and only on hateful tweets following an updated version of the guidelines elaborated in Comandini and Patti [2019a][42].

News Corpus The corpus containing news headlines about immigrants related events was used only in the test set (Test_NW). These data were retrieved between October 2017 and February 2018 from online newspapers such as *La Stampa*, *La Repubblica*, *Il Giornale*, *Liberoquotidiano*; and annotated for hate speech and stereotype categories within the context of a Master degree thesis discussed in 2018 at the Department of Foreign Languages at the University of Turin [Speranza, 2018]. The annotation was performed according to the same guidelines used for developing the Twitter corpus. Finally,

[41] https://appen.com/

[42] https://github.com/msang/haspeede

similarly to the Twitter part of HaSpeeDe2020, the organizers individuated the presence of NUs in Test_NW. Table 3.31 shows the data distribution for each task, and Table 3.32 reports some examples for the tasks A and B. The data for the task C reports the typical IOB (Inside Outside Beginning) format[43] as follows:

#Text=maledetti terroristi dell'inferno ...[44]

7331-1	maledetti	B-NU-CGA
7331-2	terroristi	I-NU-CGA
7331-3	dell'inferno	I-NU-CGA
7331-4	.	I-NU-CGA
7331-5	.	I-NU-CGA
7331-6	.	I-NU-CGA

	Task A		**Task B**		**Task C**		
set	hs	non-hs	stereo	non-stereo	w/ NUs	w/o NUs	**total**
Train_TW	2,766	4,073	3,042	3,797	1,565	1,201	6,839
Test_TW	622	641	569	694	379	243	1,263
Test_NW	181	319	175	325	152	29	500

Table 3.31 – Distribution of Labels in HaSpeeDe2020.

3.2.1.2 Approaches, Results, and Error Analysis

The shared task started on 29th May 2020 with the release of the development data and finished on 25th September of the same year with the evaluation of participating systems on the test set. For each task, participants were allowed to submit up to 2 runs, evaluated according to specific metrics. In particular, we used the standard measures of Precision, Recall, and $f1$-score (showed below) for the tasks A and B: the scores were computed for each class separately, and finally we used the average of $f1$-macro to get the overall results for the rankings.

$$Precision_{class} = \frac{\#correct\_class}{\#assigned\_class} \tag{3.3}$$

$$Recall_{class} = \frac{\#correct\_class}{\#total\_class} \tag{3.4}$$

$$F_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}} \tag{3.5}$$

[43] The Course-Grained Annotation (CGA) is necessary in order to bypass the substandard syntactic constructions of Twitter's Italian writings, which are often without a reliable punctuation and rely heavily on parataxis.

[44] *cursed terrorists of hell ...*

hs	stereo	text
1	1	*per avere i soldi.... e poi uno non deve essere razzista ... ma andate a fare in culo immigrati di merda!!!!* → to get the money.... and then one must not be racist... but f*** immigrants of s**t!!!!
1	0	*#Asselborn vuole migranti? Lo facciamo contento. Noi ne abbiamo almeno 700.000 di troppo. Se ci fa la cortesia di lasciarci il suo indirizzo di casa, penso che Salvini almeno un 500.000 glieli mandi volentieri a casa sua...* → Does #Asselborn want migrants? We make him happy. We have at least 700,000 too many. If you do us the courtesy to leave us his home address, I think that Salvini will gladly send at least 500,000 to his home...
0	1	*Iniziano le scuole, immigrati già pronti a spacciare URL* → Schools begin, immigrants ready to deal drugs URL
0	0	*#promesse Se poi non se ne farà nulla, sarà colpa degli#immigrati, dell'#Europa, del #PD e della #Costituzione. Molti nemici, molte scuse URL* → #promises If nothing is done about it, it will be the fault of #immigrants, #Europe, #PD and #Costitution. Many enemies, many excuses URL

Table 3.32 – Examples Extracted from the Training Set of HaSpeeDe2020.

For the task C, the token-wise scores were computed, and a NU was considered correct only in case of exact match, i.e., if all tokens that compose it were correctly identified. For this task, Precision, Recall, and $f1$-score are thus computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$

$$F_{class} = 2\frac{Precision * Recall}{Precision + Recall} \tag{3.8}$$

where True Positive (TP) is the number of tokens that are part of a NU in both NU-true and NU-predicted, False Positive (FP) is the number of tokens that are part of a NU in the NU-predicted but not in the NU-true, and False Negative (FN) is the number of tokens that are part of a NU in the NU-true but not in the NU-predicted. As baselines scores, we used the performance of different systems:

- for the tasks A and B, a typical classifier based on the most frequent class (Baseline_MFC) and a Linear SVM with TF-IDF of unigrams and 2–5 char-grams were used (Baseline_SVC);

- for the task C, a memory-based approach was used, which identifies as correct in the test, the NUs that appear in the training set inspired by the one presented for the COSMIANU corpus [Comandini et al., 2018].

A total amount of 14 teams participated in the task A, and among them 6 teams also submitted their results for the task B. Unfortunately, we did not receive any submission for the task C on NUs identification. A quick look at the three best scored systems in both tasks is provided in Table 3.33. A complete overview about the participating systems and the employed language resources is provided in Sanguinetti et al. [2020].

team		**description**
TheNorth	Lavergne et al. [2020]	TheNorth team experimented the fine-tuning of various versions of BERT (such as AlBERTo [Polignano et al., 2019], UmBERTo[45] and so on) for both runs. In particular, they used a linear layer with a softmax on top of the CLS token, applying a novel technique of layer-wise learning rate. TheNorth is the only team that takes into account the possible correlation between texts containing hate speech and texts expressing stereotyped ideas about targets. Indeed, they tested the performance of a multitasking approach for both tasks (second run) against a fine-tuned UmBERTo model (first run).
CHILab	Gambino and Pirrone [2020]	The CHILab team experimented transformer encoders in the first run creating specifically two transformer/convolution blocks for each input (texts and Part-of-Speech or PoS tags) averaged through max pooling and processed finally by a dropout and dense layer to obtain the predictions; and a depth-wise Separable Convolution techniques in the second one. Even in this second system the input consists of two embeddings of texts and PoS tags. To adapt the PoS-tagging model provided by the Python's *spaCy* library to social media language, they added emoticons, emojis, hashtags and URLs to vocabulary. In addition, in order to preprocess the texts, they used a sentiment lexicon for replacing emoticons with appropriate labels about the expressed sentiment. CHILab used also additional tweets taken from TWITA by means of some keywords extracted from the provided training set to extend the embedding layer of their model.
UO	Rodriguez Cisnero and Ortega Bueno [2020]	The UO team employed a Bi-LSTM with the addition of semantic and lexical features in the first run, and the pre-trained DBMDZ learning model[46] in the second one. As features, they used: lexicons such as HurtLex and SenticNet[47] to feature words with hateful categories and sentiment information; WordNet to catch lexical ambiguity, syntactic patterns and similarity among words; and, finally, they calculated information gain to capture the most relevant words. Moreover, they used as additional data to improve the knowledge of systems SENTIPOLC2016 dataset [Barbieri et al., 2016].
Montanti	Bisconti and Montagnani [2020]	Differently from the majority of the participating teams, Montanti explored classical machine learning. In particular, they tested the performance of LR, SVM with kernel and RF emploing different feature vectors. The submitted runs were obtained using the best performing model obtained with LR with a TF-IDF vector for the first run and concatenation of TF-IDF and DistilBert vectors for the second one.

Table 3.33 – Best Performing Systems at HaSpeeDe 2020.

		Task A	
team	**id**	$f1$_**Tw**	$f1$_**Nw**
TheNorth	2	**0.809**	
TheNorth	1	0.790	
CHILab	1	0.789	**0.774**
UO	2		0.731
Montanti	1		0.726
Baseline_SVC		*0.721*	*0.621*
Baseline_MFC		*0.337*	*0.389*

Table 3.34 – The Best Results for the Task A in Tweets and News Headlines in HaSpeeDe2020.

		Task B	
team	**id**	$f1$_**Tw**	$f1$_**Nw**
TheNorth	1	**0.772**	
TheNorth	2	0.768	
CHILab	1	0.761	**0.720**
CHILab	2		0.7184
Montanti	1		0.7166
Baseline_SVC		*0.715*	*0.669*
Baseline_MFC		*0.355*	*0.394*

Table 3.35 – The Best Results for the Task B in Tweets and News Headlines in HaSpeeDe2020.

Tables 3.34 and 3.35 report the results of the best performing systems on both test sets for each task, along with the baselines scores[48]. In general, participating systems obtained a good performance in both tasks, despite no teams has not designed a dedicated system for stereotypes recognition. In our view, stereotypes and hate speech are meant as orthogonal dimensions of abusive language, which do not necessarily coexist (see Table 3.32). However, the teams focused on developing a hate speech detection model, adapting the same to stereotypes recognition, considering actually stereotypes as a characteristic of hate speech. Observing the different scores, we supposed that this could be one of the reasons that led the systems to not generalize well when applied to the task B, especially in texts that are not hateful but contain stereotypes.

To investigate it, we analyzed the percentages of false and true positives and negatives in the task B, taking into account the set of common incorrect predictions of the three

[45] `https://github.com/musixmatchresearch/umberto`

[46] `https://huggingface.co/dbmdz/bert-base-italian-uncased`

[47] `https://www.sentic.net/`

[48] For a complete overview of the results, see Sanguinetti et al. [2020].

best runs (see Table 3.35), and calculating them in relation to the actual distribution of `hs` and `stereo` in the test sets. In particular, we noticed a higher percentage of `FN` (21% in tweets and 35% in news headlines) of the stereotype class in non-hateful texts, respect to `FN` (5% in tweets and 28% in news headlines) in hateful ones. A similar increase was observed also in `FP` in hateful texts. These values suggest that stereotypes appear as a subtle phenomenon that could not give rise to hurtful message. Analyzing the predictions of the three best runs even in the task A (Table 3.34), a similar influence of stereotypes is observed in false negative and positive, but to a minor extent. These results are in line with the observations emerged from the error analysis of HaSpeeDe 2018 [Francesconi et al., 2019].

3.2.2 Hate Speech and Stereotypes Detection

Taking into account the findings coming from the results obtained in the second edition of HaSpeeDe, we proposed a statistical and computational analysis that could clarify the relation between hate speech and stereotypes and the benefit of the knowledge of biased and generalized beliefs in abusive language detection.

3.2.2.1 Statistical Analysis

To our knowledge, only few scholars experimented on the contribution of less explicit information, such as metaphors [Lemmens et al., 2021] and stereotypes [Lavergne et al., 2020], for abusive language detection. This motivated us to investigate more about how different dimensions of hate interact among them. To this purpose, we applied a statistical analysis to study the association between stereotypes and hate speech, interpreted as nominal variables of a population. To have a longer sample, we extended the training set of HaSpeeDe2020 with the section of tweets of IronITA2018 [Cignarella et al., 2018b] coming from HSC. This extended version allowed also to train the systems on more evidences with the same annotations. We call here this extended version as HaSpeeDe20_ext. The final composition of the used dataset is showed in Table 3.36.

	Task A		**Task B**		
set	`hs`	`non-hs`	`stereo`	`non-stereo`	**total**
Train_TW_ext	3,035	5,226	3,554	4,707	8,261
Test_TW	622	641	569	694	1,263
Test_NW	181	319	175	325	500

Table 3.36 – Distribution of Labels in HaSpeeDe20_ext.

On this dataset, we computed:

- χ^2 test of independence that, by means of the interpretation of p-value, gives information on the existence or not of significant relations between nominal variables;

- Yule's Q to indicate if the association between two binary variables is positive (values close to 1), negative (values close to -1), or null (values close to 0).

To reject the null hypothesis (hypothesis that the variables are independent) of the χ^2 test of independence, the p-value should be minor than the significance level set by convention to 0.05; and to calculate the p-value, we considered a degree of freedom based on the number of observations. Considering that, the results reported in Table 3.37 show that the variable `hs` is strongly associated with `stereotype` especially in news headlines.

	Tweets	**News**
	`stereotype`	`stereotype`
`hs`	0.00/**0.80**	0.00/**0.96**

Table 3.37 – p-Values/Yule's Q Values in HaSpeeDe2020.

The results of this first analysis confirm the fact that hate speech tends to be characterized, especially in implicit contexts such as news headlines, by cognitive biases such as stereotypes.

3.2.2.2 Computational Approaches

Taking into account this strong association and the benefits, in terms of performance, showed in misogyny detection by linguistic features, the second analysis aims mainly at analyzing:

a) the contribution of specific features, i.e., lexical, semantic, syntactic and stylistic, that help the system to infer implicit contents such as emotions or negative connotations;

b) the advantage of having data informed also about the presence of stereotypes for hate speech detection.

Moreover, considering the good performance of BERT related-language models observed in HaSpeeDe 2020, we propose to investigate:

c) the benefit obtained by the knowledge coming from language models pre-trained on different textual genres in Italian.

Systems Description We designed three different set of systems.

FT_model (Fine-Tuned based model). Inspired by Tamburini [2020], we selected three language models, trained on different genres of texts in Italian and available on the Hugging Face platform[49]:

[49] https://huggingface.co/

- AlBERTo [Polignano et al., 2019] is a transformer developed on the example of the classical BERT model [Devlin et al., 2019a], except for the fact that the authors implemented only the masked learning strategy. Excluding the step based on the next sentence prediction, AlBERTo is made suitable for types of short texts such as tweets or headlines (and not dialog), and, thus, for the tasks of classification and prediction. Moreover, it is trained on TWITA, a large dataset composed of Italian tweets from February 2012.

- UmBERTo is designed looking at the RoBERTa base model architecture [Liu et al., 2019] employing, in particular, two innovative approaches: SentencePiece[50] [Kudo and Richardson, 2018] and Whole Word Masking[51]. The model employed here is trained on Commoncrawl ITA exploiting OSCAR (Open Super-large Crawled ALMAnaCH coRpus) Italian large corpus [Ortiz Suárez et al., 2020].

- GilBERTo[52] architecture is based on the RoBERTa model and CamemBERT text tokenization approach. Like UmBERTo, it is trained on Italian texts coming from OSCAR but, differently from the former, using the subword masking technique exploiting the SentencePiece tokenizer.

In this first set, we simply fine-tuned these language models on hate speech and stereotypes classification tasks, taking into account only the CLS token of the BERT-based model. Indeed, in accordance with Devlin et al. [2019b], the purpose of this token is to contain the information useful for the classification task at the end of the forwarding process. Then a simple classifier can just take this CLS token as input to classify the whole text. Moreover, we added a dropout layer and a final linear layer to get the class-related probability employing a Sigmoid function.

FT+Feat_model (Fine-Tuned and Features based model). The main idea beside this model is to converge the awareness coming from a pre-trained language model with the specific knowledge derived from dedicated linguistic features. On the one hand, the learning transferred by a language model trained on different Italian texts should help the classifier to generalize better ranging from informal and more formal writings and make the system able to 'understand' better the unseen tweets and news headlines. On the other hand, engineered features lead the system to pay attention to specific elements, expressed or unexpressed, in the text. At this purpose, we add at the previous network (FT_model) a new input layer consisting in the batch normalization of the features vector[53], combined with CLS token of BERT-based pre-trained model.

MTL_model (Multi-task Learning based model). The choice of employing a multi-task learning (MTL) based model is motivated by the strong correlation observed in

[50] This language-independent subword tokenizer and detokenizer creates sub-word units specifically to the size of the chosen vocabulary and the language of the corpus.

[51] This technique applies a mask to an entire word.

[52] `https://github.com/idb-ita/GilBERTo`

[53] The batch normalization technique helps to standardize the layer and stabilize the learning process.

Table 3.37 between the expression of stereotypes, that mark the believed negative traits of the target, and the possibility to hurt her/him (Table 3.32), seen also in TheNorth's results. Moreover, at computational level, the advantages derived from the use of MTL techniques such as the *hard parameter sharing* are various. Firstly, this technique gives systems more evidences to evaluate whether a feature is relevant or not, focusing strictly on the most relevant ones for each task. Then, the hard parameter sharing allows a better generalization for each task: learning simultaneously more tasks means to find a representation that is appropriate for learning all the tasks, reducing consequently the overfitting on the original task [Baxter, 1997]. To understand the real contribution of the simultaneous learning of correlated tasks, we employed a network similar to FT_model. The only difference is due to double linear output layers, one for each task.

Linguistic Features With respect to the creation of the features vector representation for FT+Feat_model, a data preprocessing phase is performed in accordance with the information that we wanted to extract from the tweets. For the majority of the features, we took into account a dictionary of words weighted with TF-IDF. To create this dictionary and the word embedding model used to extract semantic information, we preprocessed the tweets as follows: deleting URLs and symbols like @ and # to maintain the lexical information of hashtags and users' names; tokenizing and lemmatizing words using the TreeTagger tool[54] [Schmid, 1994] implemented for Python in the *treetaggerwrapper* library[55]; and removing stopwords[56] to retain lexical significant words. Moreover, to extract PoS tags and syntactic dependencies from texts, we used the *spacy-udpipe* library with the TWITTIRÒ model for the Italian language in Twitter[57] [Cignarella et al., 2019]. Finally, the majority of the features have been standardized using *MinMaxScaler* of *scikit-learn*[58] with default range of scaling.

Inspired by our previous works on hostility detection against women, and with the purpose of helping the system to infer connotative meanings or figurative language, we developed a set of features that consists of stylistic, syntactic, and semantic information. The overview of all features is reported on Table 3.38.

Stylistic Features Especially in short and informal texts such as tweets, punctuation helps authors to express better their intention (i.e., quotation marks to underline the opposite of the literal meaning: *"risorsa"*). Like punctuation, negation patterns show to play an important role in the process of comprehension of figurative language [Giora et al., 2015b, 2018, Karoui et al., 2015, 2017]. Therefore, these patterns and their relevance are caught by the system, providing as vectorized inputs the sum of TF-IDF weights of punctuation characters (`punct`) and negation elements (`negation`) in the text.

[54] Using this tool the numbers are replaced by `@card@` tag.

[55] https://treetaggerwrapper.readthedocs.io/en/latest/

[56] For the list of stopwords see: http://di.unito.it/stopwordsit

[57] http://di.unito.it/twittirotreebank

[58] https://scikit-learn.org/stable/index.html

Syntactic Features As shown in other works [Cignarella et al., 2020], syntactic features are proven to be useful to detect ironic language in social media. In particular, we helped the system to capture some syntactic dependencies that could reveal pragmatic information, such as: adverbial locutions (`adv_loc`), intensifiers (`intens`), discourse connections (`disc_conn`), mentions (`mention`) and nominal phrases (and the number of nominal phrases in the tweet) (`nom_phrase` and `num_nom_phrase`).

Semantic Features We used lexical resources (Sentix[59], HurtLex and EmoLex[60]) to extract emotional and affective information from the texts following the example of Corazza et al. [2020], and an ensemble of features aimed at helping the system to understand the semantic incongruities and similarities revealed by words and pairs of words used in implicit and figurative messages.

Sentiment Lexicon In Sentix [Basile and Nissim, 2013] each entry consists of an Italian lemma followed by information as PoS tag, WordNet synset ID, a positive and a negative score from SentiWordNet, a polarity and an intensity score. Using this information, we calculated the average of positive and negative scores of words in the tweet (`avg_positive` and `avg_negative`), the standard deviation (σ) of polarity inside the tweet and the intensity score average to indicate whether the tweet expresses an objective or subjective message (`avg_intensity`).

Hurtful Words HurtLex is a multilingual lexicon of hateful words created from the Italian lexicon “Le Parole per Ferire” by Tullio de Mauro. The entries in the lexicon are categorized in 17 types of offenses (see Table 3.9) enclosed in two macro-categories: *conservative* (words with literally offensive sense) and *inclusive* (words with not literally offensive sense, but that could be used with negative connotation). To extract features from tweets relative to the 17 categories, we used a specific *featurizer*[61] created specifically for this lexicon. As weight for each category, we computed the sum of TF-IDF of words in the tweet belonging to each category without omitting the macro-category of reference.

Emotional Lexicon EmoLex [Mohammad and Turney, 2013b] is a multilingual lexicon containing sentiment and affective information for each word. For our purposes, we principally used the annotation relative to the 8 principal emotions of Plutchik [Plutchik and Kellerman, 1980]. Inspired by Plutchik [2001], we exploited the wheel of emotions (Figure 3.3) to capture in the message the principal emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), the primary dyads or feelings (aggressiveness, optimism, love, submission, awe, disapproval, remorse, contempt), and the variability of opposite emotions and contrary feelings by means of σ. The weight of emotions and feelings are computed summing the TF-IDF of words belonging to the specific categories.

[59] http://valeriobasile.github.io/twita/sentix.html

[60] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[61] https://github.com/valeriobasile/hurtlex

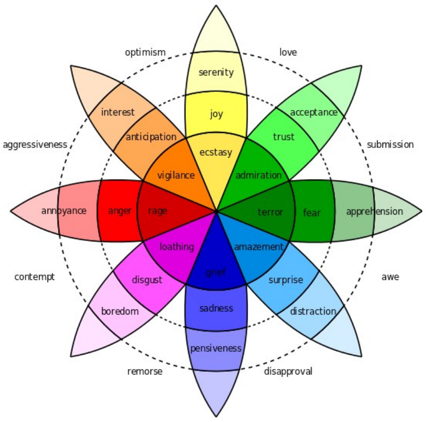


Figure 3.3 – Plutchik's Wheel of Emotions.

Incongruity/Similarity Features In this set of features, we calculated: the variability of the TF-IDF weights of the words inside the tweet by means of σ and the coefficient of variation (`cv`), the average of weights (`avg`), and the maximum (`max`), minimum (`min`) and median (`med`) values of list of TF-IDF weights of words (`W`) and of bigrams of words (`B`) of a text to take into account the most significant tokens (such as interjections and hashtags). The values related to bigrams are computed using the weights' normalization on maximum and minimum scores (`C1`) and on standard deviation and average (`C2`). Additionally, we created a word embedding model starting from a pre-trained model on TWITA. Firstly, using the *Gensim* library[62], we updated the vocabulary and the word embeddings of the TWITA model with SENTIPOLC2016. Secondly, we extended the updated word embedding model with 'out of vocabulary words' predicting their most probable embedding vectors considering their context. The prediction is based on a language model built on HASPEEDE20_EXT using bidirectional Long-Short Term Memory (bi-LSTM) cell[63]. The final word embedding model is used to calculate the similarity ($\cos(\theta)$) between pairs of words (vector of bigram of words) and the sentence context (corresponding to sentence vector) (cos(θ)_BS), and between the bigrams of words within the sentence (cos(θ)_BB). To create the feature vector for our system we computed σ, the coefficient of variation, the average, and maximum, minimum and median scores of lists of cosine similarity values.

[62] https://radimrehurek.com/gensim/

[63] This methodology is inspired by a work presented in http://di.unito.it/oov by Shabeel Kandi in 2018.

type	group	feature		
Stylistic		punct	negation	
Syntactic		num_nom_phrase	disc_conn	nom_phrase
		adv_loc	intens	mention
Semantic	Sentiment Lexicon	avg_positive	σ_pos_neg	avg_negative
		avg_intensity		
	Hurtful Words	inclusive_an	conservative_an	inclusive_asf
		conservative_asf	inclusive_asm	conservative_asm
		inclusive_cds	conservative_cds	inclusive_ddf
		conservative_ddf	inclusive_ddp	conservative_ddp
		inclusive_dmc	conservative_dmc	inclusive_is
		conservative_is	inclusive_om	conservative_om
		inclusive_or	conservative_or	inclusive_pa
		conservative_pa	inclusive_pr	conservative_pr
		inclusive_ps	conservative_ps	inclusive_qas
		conservative_qas	inclusive_rci	conservative_rci
		inclusive_re	conservative_re	inclusive_svp
		conservative_svp		
	Emotional Lexicon	anger	aggressiveness	anticipation
		contempt	disgust	remorse
		fear	disapproval	joy
		awe	sadness	submission
		surp	love	trust
		optimism	σ_joy_sad	σ_agg_awe
		σ_trust_disg	σ_cont_sub	σ_fear_ang
		σ_rem_love	σ_surp_ant	σ_dis_opt
	Incongruity/Similarity	W_max	B_C1_max	W_med
		B_C1_med	W_min	B_C1_min
		W_avg	B_C1_avg	W_σ
		B_C1_σ	W_cv	B_C1_cv
		B_max	cos(θ)_BB_max	B_med
		cos(θ)_BB_med	B_min	cos(θ)_BB_min
		B_avg	cos(θ)_BB_avg	B_σ
		cos(θ)_BB_σ	B_cv	cos(θ)_BB_cv
		B_C2_max	cos(θ)_BS_max	B_C2_med
		cos(θ)_BS_med	B_C2_min	cos(θ)_BS_min
		B_C2_avg	cos(θ)_BS_avg	B_C2_σ
		cos(θ)_BS_σ	B_C2_cv	cos(θ)_BS_cv

Table 3.38 – List of Features.

Features Relevance Figure 3.4 shows the most relevant features for each task of classification.

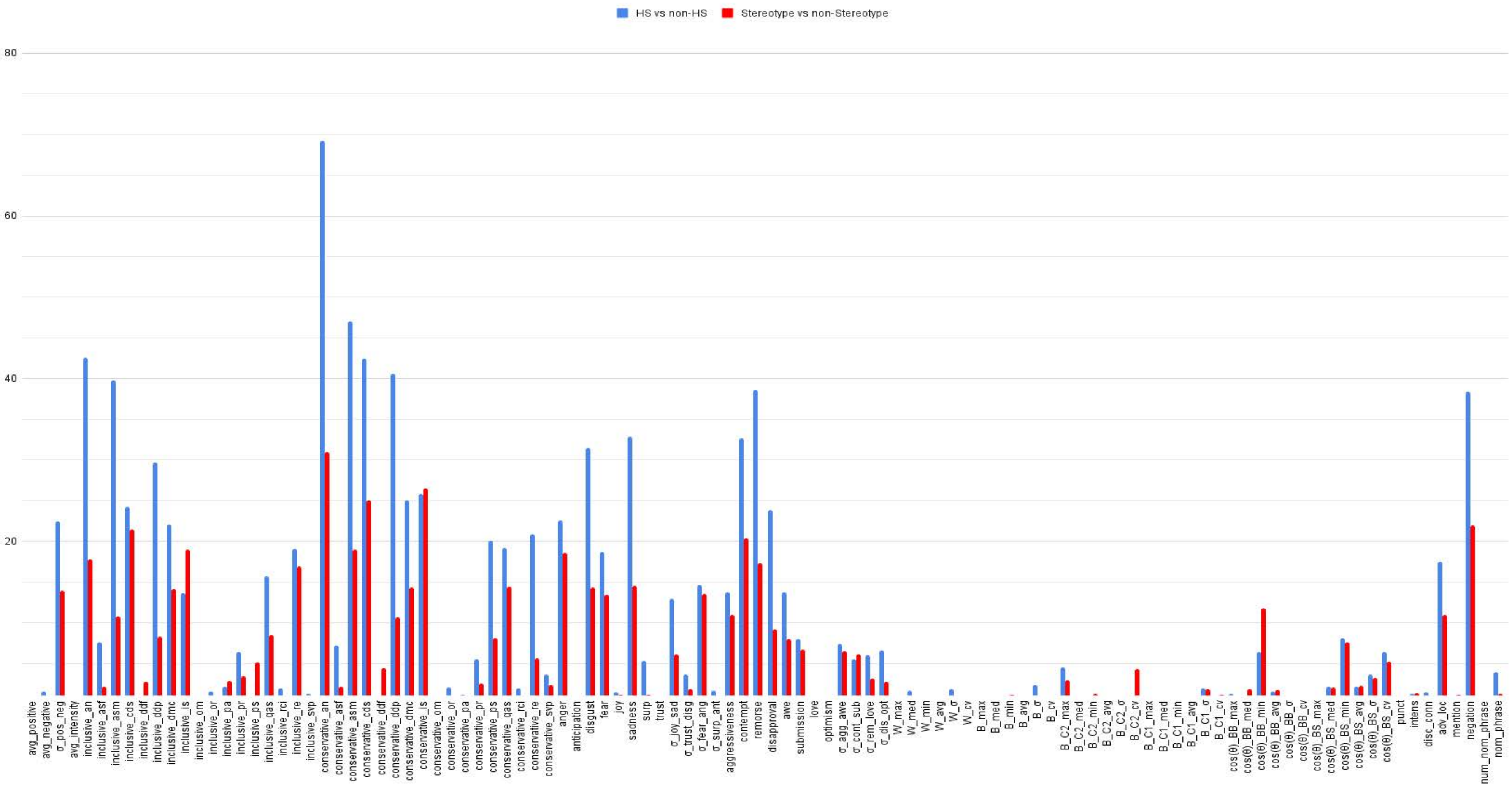


Figure 3.4 – Relevance of Features in TRAIN_TW_EXT for the Tasks A and B.

Looking at Figure 3.4 we can notice that Hate Speech and Stereotypes are characterized by very similar features. In general, both are featured by negative emotions and feelings (anger, awe, disgust, aggressiveness, fear) and by offensive words with conservative and inclusive interpretation. Some categories of offenses related especially to animals, physical disabilities or diversity, behaviors/morality and general swear words are more relevant in hate speech; whereas in stereotyped messages the offenses more significant are linked in particular to economic and social issues, cognitive and ethnic sphere, even if in a more indirect way. At semantic level, the minimum score of similarity between the bigrams of words within the sentence (`cos`(θ)`_BB`) appears to be relevant in stereotypes recognition. This specific feature brings out the semantic incongruity in the text: a common technique used to express an ironic utterance [Riloff et al., 2013, Joshi et al., 2015, Pan et al., 2020]. Finally, both phenomena appear characterized by specific syntactic patterns, such as negation and adverbial locutions. From a manual examination, we found that the former are used especially to mark some characteristics of outgroup, juxtaposing it sometimes with the ingroup. While the latter tend to increase the intensity of some beliefs or make the sentences mainly nominal. Some of these markers are reported in Table 3.39.

`hs`	`stereo`	`text`
1	1	*Ora ankio andrò ad emigrare dato che qui sarà tutto occupato da stranieri.. tanto di noi italiani non gliene frega nessuno.... vergognaaa* → Now me too I will go to emigrate since everything here will be occupied by foreigners.. *so much of* us Italians do *not* give a damn.... shame on
1	1	*I nostri migranti non erano assolutamente come questa gentaglia qui! Continuare a fare questo paragone è un'offesa per tutti quelli che si sono rotti mani e schiena nelle miniere e nelle fabbriche e vivevano nascosti il resto del tempo.* → Our migrants were absolutely *not* like this scum here! Continuing to make this comparison is an offense to all those who broke their hands and backs in mines and factories and lived in hiding the rest of the time.
1	1	*Odiano i nemici più di quanto amino i figli, o forse una figlia femmina vale poco o nulla, islam se lo conosci lo eviti* → They hate enemies *more than* they love their children, or maybe a daughter is worth *little or nothing*, Islam if you know it you avoid it
1	1	*e lui parla dei migranti, pensa quanto gli frega degli italiani* → and he talks about migrants, thinks *how much* he cares about Italians

Table 3.39 – Tweets Extracted from TRAIN_TW_EXT.

Experiments and Results Our experimental setting could be divided in two principal steps. Firstly, we employed a 5-fold cross-validation on the entire TRAIN_TW_EXT to understand the capacity of generalization of the proposed approaches and to adjust the parameters of the neural network. Secondly, we used the 20% of the training set as validation set to tune the systems and evaluate them on TEST_TW and TEST_NW in order to compare the obtained results with the ones of the competition and to evaluate the proposed approaches in a cross-genre framework. Table 3.40 reports the information about the parameters of the neural network used in both tasks and the additional functions applied to improve its performance.

parameter/function	**value/description**
max sequence length	100
dropout rate	0.1
learning rate	2e-5
batch size	64
maximum epochs	15
optimizer	AdamW
scheduler	we employed the *Constant schedule* with warmup[64] to define dynamic learning rates during the training phase.
early stopping	we applied a custom early stopping function to avoid the overtraining of the neural network, looking at the values of the loss (or common loss in MTL) obtained on the validation set with a patience of 3 epochs.
seed	we applied to the environment a seed function from the *random* library to make the results reproducible.
loss	to minimize the loss function during the training, we used the BCEWithLogitsLoss function that combines a Sigmoid layer and the BCELoss in one single class. It is provided by *pytorch*.

Table 3.40 – Parameters' Values and Functions.

In regard to the second set of models, the experiments include the evaluation of systems informed firstly with all the designed features (Feat), and secondly, with a set of selected features (SelectedFeat) for each task. To select the best features, we considered their χ^2 value (greater than 10) for a total of 27 features for hate speech and 25 for stereotypes detection.

In both phases of evaluation, we used the same metrics proposed by the organizers in HaSpeeDe 2020: $f1$-macro as average score of $f1$ value of each class. Especially for the second phase, we compared the obtained results with the scores provided as

[64] https://huggingface.co/transformers/main_classes/optimizer_schedules.html

baselines (*Baseline_ MFC* and *Baseline_ SVC*) and reached by the first ranked systems. Tables 3.41 and 3.42 report the results obtained in the 5-fold cross-validation setting. In particular, they show the average of the $f1$-macro scores (*avg_folds*) per each proposed approach and per each adopted language model (LM). Moreover, to understand their general performance in the tasks A and B, we calculated the average of the *avg_folds* (Average).

approach	FT	FT+All_Feat	FT+27_Feat	MTL (`stereo`)
	avg-folds	*avg-folds*	*avg-folds*	*avg-folds*
AlBERTo	0.749	0.740	**0.752**	0.727
UmBERTo	0.735	0.747	0.738	0.726
GilBERTo	0.737	0.733	0.741	0.733
Average	0.740	0.740	***0.743***	0.729

Table 3.41 – Results obtained in 5-Fold Cross-Validation Setting in the Task A.

approach	FT	FT+All_Feat	FT+25_Feat	MTL (`hs`)
	avg-folds	*avg-folds*	*avg-folds*	*avg-folds*
AlBERTo	0.691	0.677	**0.695**	0.678
UmBERTo	0.682	0.680	0.690	0.643
GilBERTo	0.662	0.683	0.655	0.667
Average	0.678	***0.680***	***0.680***	0.662

Table 3.42 – Results obtained in 5-Fold Cross-Validation Setting in the Task B.

For this first step of evaluation, looking at the Average scores, we can notice that the best approach, in both tasks, involves the presence of linguistic features, confirming once again that despite the incredible contribution of transfer learning, the injection of linguistic information helps to infer additional or implicit meanings.

Observing, instead, the performance of the different LMs adopted for this study, we can notice that, in general, they perform similarly. The standard deviation calculated for each proposed approach shows that LMs' performance varies very slightly from the Average score (from 0.00 to 0.01 in the task A and from 0.00 to 0.02 in the task B). However, in the majority of the cases, the best avg_folds scores are obtained employing the AlBERTo pre-trained model. We hypothesize that the reason lies in the characteristics of this transformer, which make it very suitable to the classification of short texts.

Nevertheless, proposing these experiments in the second step of the evaluation on TEST_TW and TEST_NW in the tasks A and B (Tables 3.43, 3.44, 3.45, 3.46) shows that the knowledge inherited from AlBERTo is useful only for the detection of hate speech in news headlines and, especially, in MTL architecture. Differently from the task A, in Table 3.46 GilBERTo reports the highest scores in the majority of the proposed approaches for detecting stereotypes in news headlines. We think that the size and the type of corpus used to create the pre-trained model would have helped to generalize better. Also, UmBERTo performs very well in tweets in both tasks.

approach	FT	FT+All_Feat	FT+27_Feat	MTL (`stereo`)
AlBERTo	0.741	0.759	0.773	0.757
UmBERTo	0.790	0.797	0.797	0.775
GilBERTo	0.762	0.784	0.777	0.775
Avg_LMs	0.800	**0.810**	0.807	0.789

Table 3.43 – Results obtained on TEST_TW in the Task A.

approach	FT	FT+All_Feat	FT+27_Feat	MTL (`stereo`)
AlBERTo	0.630	0.652	0.613	**0.677**
UmBERTo	0.606	0.664	0.658	0.635
GilBERTo	0.602	0.561	0.617	0.675
Avg_LMs	0.612	0.628	0.630	0.668

Table 3.44 – Results obtained on TEST_NW in the Task A.

approach	FT	FT+All_Feat	FT+25_Feat	MTL (`hs`)
AlBERTo	0.718	0.740	0.727	0.732
UmBERTo	0.767	0.774	0.765	0.785
GilBERTo	0.746	0.756	0.768	0.755
Avg_LMs	0.784	0.785	**0.793**	0.787

Table 3.45 – Results obtained on TEST_TW in the Task B.

This optimal performance of UmBERTo on tweets appears also in the ranking of the tasks A and B in the HaSpeeDe 2020 shared task. As seen in Table 3.33, TheNorth employed the UmBERTo pre-trained model in both submitted runs, juxtaposing its fine-tuning (run 1) with the additional learning of the stereotype detection task in run 2. This second run achieved the best score (0.809) in the task A (Table 3.34). Comparing these results with the performance of our approach based on MTL (0.775 in Table 3.43), we can suppose that, in spite of the fact that the scores obtained with the FT of UmBERTo are similar in TheNorth run 1 and in our experiments (0.790 Table 3.34 and Table 3.43), the technique of layer-wise learning rate introduced by TheNorth team, that aims at limiting the update in the first layers, could help the system to take advantage of the information in such layers during the fine-tuning in both tasks.

Looking specifically at the difference of genre, we can notice that the awareness of stereotypes does not report very high scores for the detection of hate speech in news headlines.

approach	FT	FT+All_Feat	FT+25_Feat	MTL (`hs`)
AlBERTo	0.731	0.728	0.678	0.731
UmBERTo	0.681	0.684	0.717	0.695
GilBERTo	0.733	**0.781**	0.727	0.725
Avg_LMs	0.712	0.724	0.722	0.724

Table 3.46 – Results obtained on TEST_NW in the Task B.

In Table 3.44, MTL of stereotypes and hate speech for the task A appears the best approach for news headlines. But comparing this score with the ranking (Table 3.34) we notice that it is lower. And even a lower performance is obtained by TheNorth with their multi-task approach in the second run (0.660)[65]. Indeed, despite the very high association in Table 3.37, the majority of the headlines marked as hate speech actually report intolerant slogans typical of political communication, very different from the hate speech expressed in tweets (Table 3.47).

In this second setting of evaluation, we added another model, called Avg_LMs, that takes into account the mean of the probabilities obtained employing every LM for each text. In this way, the system considers the decisions of all the models for each approach. It proves its usefulness in both tasks only in the TEST_TW set. Finally, as observed in the first set of experiments, the addition of linguistic information improves the performance of the systems in both textual genres. The complete rankings that include also our best approaches and an additional baseline based on the average of the probabilities obtained by fine-tuning LMs (*Baseline_Avg_LMs*) are provided in Tables 3.48 and 3.49.

In these final tables, we can notice that the increase of our approach in the task A is modest compared to result obtained by TheNorth (run 2). Therefore, it is clear that, in general, both linguistic features and MTL (`stereo`) allow the system to be more sensitive to recognize hate speech especially in tweets. However, as further investigation, we would like to test the application of the layer-wise learning rate to understand the contribution of the features even in a similar model to TheNorth one. Indeed, as different layers tend to capture different types of information from the texts [Yosinski et al., 2014], it could be useful to adopt a discriminative fine-tuning, adjusting each layer with different learning rates [Howard and Ruder, 2018].

About the task B, the increase produced by our systems is substantial with respect to the best scores in the ranking of the HaSpeeDe 2020 shared task: more than 2% in tweets and 6% in news headlines. One important difference between our two models is about the different set of features in relation to textual genre. As we can see in Table 3.46, the employment of all the set of designed features produces in general better results than the selection of specific features in the TEST_NW set. We can better understand this difference, looking at the style of the news headlines in Table 3.50.

[65] The complete ranking for both tasks is reported in Sanguinetti et al. [2020].

hs	stereo	genre	text
1	0	nw	*Marco Bussetti: “A scuola tuteliamo gli studenti immigrati ma prima i nostri figli”* → Marco Bussetti: “At school we protect immigrant students but first our children”
1	1	nw	*Sanremo 2019, don Salvatore Picca il prete pro-Salvini massacra il festival: “Vinci se sei musulmano e drogato”* → Sanremo 2019, Don Salvatore Picca the pro-Salvini priest massacres the festival: “Win if you are Muslim and drug addict”
1	0	tw	*@user @user @user @user Rimandate tutti i migranti a casa loro.* → @user @user @user @user Send all migrants back to their home.
1	1	tw	*@user Profughi denunciano??? E dove trovano i soldi se sono scappati dalla fame???? Perché non mandano i soldi che qualche ignorante gli dà per questa cazzata ai loro parenti bisognosi??? Sono tutti dei buffoni. Avanti così Salvini! Grazie* → @user Refugees denounce??? And where do they find the money if they ran away from hunger???? Why don't they send the money that some ignoramus gives them for this bullshit to their needy relatives??? They are all fools. Go on like this Salvini! Thanks

Table 3.47 – Examples Extracted from HaSpeeDe2020.

In spite of the types of stereotypes against immigrants, Romas and Muslims (parasites, privileged people, invaders) are similar in the different genres, it is possible to notice a different tone, that in tweets tends to express louder anger and disgust about the perceived *outgroup* than in news headlines. Finally, observing the comparison with Baseline_Avg_LMs we can notice that the performance obtained simply fine-tuning LMs is higher in tweets than in news headlines in both tasks. This is explicable for the type of training set used, containing only tweets. Although the use of a pre-trained LM leads the system to generalize better in various NLP tasks, fine-tuning on another genre evidently does not help. News headlines, as we already noticed, have some characteristics that make the expression of hate speech or stereotypes different from tweets.

3.2.2.3 Error Analysis

Hate Speech in News Headlines Considering the lower scores obtained on headlines in the task A respect to tweets, we wonder what kind of hate speech is expressed in headlines. Indeed, as seen in Table 3.37, the association between hate speech and stereo-

		Task A	
approach	**id**	**$f1$_Tw**	**$f1$_Nw**
Avg_LMs (FT+All_Feat)		**0.810**	
TheNorth	2	0.809	
TheNorth	1	0.790	
CHILab	1	0.789	**0.774**
UO	2		0.731
Montanti	1		0.726
AlBERTo (MTL `stereo`)			0.677
Baseline_Avg_LMs		*0.800*	*0.612*
Baseline_SVC		*0.721*	*0.621*
Baseline_MFC		*0.337*	*0.389*

Table 3.48 – Results in the Task A in the Ranking of HaSpeeDe 2020.

		Task B	
approach	**id**	**$f1$_Tw**	**$f1$_Nw**
Avg_LMs (FT+25_Feat)		**0.793**	
GilBERTo (FT+All_Feat)			**0.781**
TheNorth	1	0.772	
TheNorth	2	0.768	
CHILab	1	0.761	0.720
CHILab	2		0.718
Montanti	1		0.717
Baseline_Avg_LMs		*0.784*	*0.712*
Baseline_SVC		*0.715*	*0.669*
Baseline_MFC		*0.355*	*0.394*

Table 3.49 – Results in the Task B in the Ranking of HaSpeeDe 2020.

types, especially in news headlines, is very high. This means that, in the majority of the headlines marked as hate speech, we find also negative stereotypes against minorities. Therefore, why MTL on stereotypes detection does not work? Observing the difference of hate speech expressed in news headlines and tweets, we noticed some stylistic differences. For example, in Table 3.47, the headlines labeled as hate speech tend to contain quotes of public figures who support a politics less tolerant. From this perspective, we looked at the composition of these texts exploiting the NUs contained in the headlines labeled with `hs` = 1.

In particular, we extracted all the coarse-grained NUs from the news headlines labeled as containing hate speech, and then we took from them the most frequent trigrams. These are: ‘basta balle ecco’, ‘via i migranti’, and ‘immigrati la verità’, involved in specific syntactic contexts such as the ones in Table 3.51.

hs	stereo	genre	text
0	1	nw	*Falsi permessi di soggiorno: tutti assolti gli immigrati coinvolti* → False residence permits: all immigrants involved were absolved
1	1	nw	*Gli immigrati fanno perdere all'Italia 5 miliardi l'anno* → Immigrants make Italy lose 5 billion a year
1	1	nw	*Ecco il reddito di cittadinanza Pacchia per immigrati e rom* → Here is the citizenship income Gravy train for immigrants and roma
0	1	tw	*"Mi piacciono gli stranieri in Italia perché sono già in ciabatte" Cit Federica* → "I like foreigners in Italy because they are already in slippers" Federica
1	1	tw	*Anche ieri 6000 arrivi. Quando avranno la maggioranza gli islamici ci sottometteranno con la forza.* → Even yesterday 6000 arrivals. When they have a majority, Muslims will subdue us by force.
1	1	tw	*Purtroppo alcuni immigrati creano ancora forti problemi a dire il poco 'animaleschi'.... Occorre mettere ordine, dobbiamo tutelare le nostre famiglie...* → Unfortunately, some immigrants still create serious problems to say the least 'animal-like'.... We must put order, we must protect our families...

Table 3.50 – Examples Extracted from HaSpeeDe2020.

From these sentences, the stance of journalists, clearly, shows a certain degree of intolerance against immigrants, even with sharp tones. In particular, in our sample, we noticed that this strong position is transmitted with simple NUs that remember a political rhetoric feeding the juxtaposition between the ingroup and the outgroup. Comandini and Patti [2019a] analyzed the presence of NUs in the hateful tweets of HSC. A NU is conceived as an utterance whose main clause does not have a verb in a finite form. Among all the NUs, the authors found a specific type of NUs, called *Slogan-like NUs*, that conveys populist slogans and a sharp adherence to a point of view. This type of NUs tends to encourage actions for expelling immigrants from Italy, killing or imprisoning them, or generally inciting authoritative attitude typical of the populist ingroup's thinking. From this perspective, the detected trigrams actually could be defined as slogans.

In general, NUs are typical of news headlines and their higher distribution compared to tweets (Table 3.31) is not surprising. However, it appears interesting how slogan-like NUs are recurrently employed to convey hate speech within news headlines. This observation, moreover, justifies the performance obtained by the CHILab team (see Table 3.33) who

hs	stereo	text
1	1	*Immigrati, ammiraglio brutale: ora basta balle. "Ecco chi trama contro l'Italia, serve una guerra"* → Immigrants, brutal admiral: no more lies now. "Here is who is plotting against Italy, we need a war"
1	1	*C'è la scuola, via i migranti: "Siamo contrari all'apartheid ma ora serve più sicurezza"* → There is the school, out the migrants: "We are against apartheid but now we need more security"
1	1	*Immigrati, la verità nei numeri: "Crimini 6 volte più di italiani"* → Immigrants, the truth in the numbers: "Crimes 6 times more than Italians"

Table 3.51 – News Headlines Extracted from TEST_NW.

innovately employed a PoS tag embedding jointly with the classical token representation for tranformers. Evidently, the nominal structure, present also in the training set of tweets, was caugth by the systems informed with syntactical knowledge and associated with the presence of hateful discourses. On the basis of this, we can suppose that in the genre of tweet, the expression of hate conveys various elements that go beyond the stylistic and syntactic aspects. To this purpose, as further work, we want to investigate specific syntactic and pragmatic features in hateful news headlines exploiting also GDELT-FM.

The Impact of Linguistic Knowledge In order to understand the contribution of linguistic features in a LM-based neural network for the detection of hate speech and stereotypes, we looked at the percentage of TP, TN[66], FP and FN produced by our best models and by the Baseline_Avg_LMs. Tables 3.52 and 3.54 show the values of the confusion matrices respectively for hate speech detection in tweets (TW) and for stereotypes detection in tweets and news headlines (NW).

approach	**FP** (%)	**FN** (%)	**TP** (%)	**TN** (%)
Avg_LMs (FT+All_feat)	24	**14**	**86**	76
Baseline_Avg_LMs	**20**	21	79	**80**

Table 3.52 – Values of Confusion Matrix for the Task A in Tweets.

In Table 3.52, we can notice that the employment of linguistic features helps the system to recognize correctly the tweets containing hate speech, ameliorating its sensibility toward the positive class. In Table 3.53 we report some examples extracted from the FN cases of

[66] True Negative

Baseline_Avg_LMs that are correctly predicted as TP with Avg_LMs (FT+All_feat).

FN	TP	
Baseline_Avg_LMs	Avg_LMs (FT+All_feat)	
`hs`	`hs`	`text`
0	1	*setta di odio oppressione soprusi violenza morte la sua illegalità in Occidente* → sect of hatred oppression abuses violence death its illegality in the West
0	1	*Reddito di cittadinanza, boom di richieste dagli stranieri, ma ad aprile le domande totali rallentano? STRANIERI???? E che cazzo succede? Gli italiani lavorano per mantenere loro?* → Citizenship income, boom in requests from foreigners, but in April the total applications slow down? FOREIGNERS???? What the fuck is going on? Do the Italians work to support them?
0	1	*Siamo italiani per noi non ci sono strutture, servizi sociali, bandi milionari, hotel e residenze di lusso. I nostri connazionali lasciati a marcire e morire di freddo per strada, le passerelle solo per i migranti.*
0	1	→ We are Italian, for us there are no structures, social services, millionaire tenders, hotels and luxury residences. Our compatriots left to rot and freeze to death on the street, the catwalks only for migrants.

Table 3.53 – Tweets Correctly Recognized as Hate Speech.

These tweets show a strong emotional component. They express fear, anger, aggressiveness, underlining also the differences between Italians and foreigners. These characteristics have been caught principally thanks to lexical and syntactic features that put the attention on some elements that maybe have not been considered relevant in the simple fine-tuning.

	approach	**FP** (%)	**FN** (%)	**TP** (%)	**TN** (%)
TW					
	Avg_LMs (FT+25_feat)	**25**	16	84	**75**
	Baseline_Avg_LMs	27	**15**	**85**	73
NW					
	GilBERTo (FT+All_feat)	8	**39**	**61**	92
	Baseline_Avg_LMs	**6**	55	45	**94**

Table 3.54 – Values of Confusion Matrix for the Task B in Tweets and News Headlines.

A similar behavior is observed in Table 3.54 for the detection of stereotypes in news headlines, where the best performance is obtained exploiting the entire set of features. Also, here, we notice an interesting increase of TP that shows especially negative feelings such as disgust and frustration, semantic incongruities within headlines, and sentiment or negative connotation of some specific words (Table 3.55). As seen already in Table 3.50, the different style between tweets and headlines relies on the louder tones used in tweets compared to headlines. But the expressed feelings and the impact on society of this kind of messages are similar: headlines like tweets spread negative beliefs against minorities. Thanks to the designed features, we are able to allow the system to go beyond the style in order to look at the pragmatical and semantic aspects of text.

FN	TP	
Baseline_Avg_LMs	GilBERTo (FT+All_feat)	
stereo	stereo	text
0	1	*Italiani in strada a -20°, loro qui: la villa del Cinquento agli immigrati. Giardini, lusso e sfarzo: ecco le foto* → Italians in the street at -20 °, they are here: the villa of the Cinquento for immigrants. Gardens, luxury and pomp: here are the photos
0	1	*Dopo aver tagliato ai disabili, Renzi regala un miliardo ai richiedenti asilo* → After cutting the disabled, Renzi gives a billion to asylum seekers
0	1	*Dà la casa ai profughi. E gli italiani?* → It gives the house to the refugees. And the Italians?
0	1	*Ecco il reddito di cittadinanza Pacchia per immigrati e rom* → Here is the citizenship income Gravy train for immigrants and roma

Table 3.55 – News Headlines Containing Stereotypes Correctly Detected.

Looking instead at the values of the prediction for the detection of stereotypes in tweets, we can notice a slight increase of the precision on the positive class compared to the baseline. As we see in Table 3.56, with the set of the selected features, the system is able to distinguish with more attention the text containing actually stereotypes. This set of features contains especially hurtful words related to defects, morality, crimes and social advantages, variation of polarity and emotions, negations and negative feelings such as contempt, anger, and sadness.

Finally, observing the FN and FP cases remained unsolved (Tables 3.57 and 3.58), we noticed that a good percentage of tweets (about 30%) are ironic. As seen also in previous experiments on misogyny, sometimes users, especially in the social network contexts, tend to use various figures of speech even to express hate. Humor in social media could be thought as an engagement tool used to create consensuses, amusement or avoid possible censorship in case of delicate subjects.

Differently from tweets, misclassified news headlines in the task B only in few cases contain humor (like in "*Milano , Sala coccola gli immigrati e loro gli occupano lo stabile*"[67]). Indeed, they are commonly characterized by shortness; and interpreting the presence of stereotypes in very few words could be hardly even for humans, like in "*Rogo tossico dal*

[67] *Milan, Sala pampers immigrants, and they occupy the building*

FP/FN	TN/TP	
Baseline_Avg_LMs	Avg_LMs (FT+25_feat)	
stereo	stereo	text
1	0	*I rom nell'accogliente quartiere dei Parioli subito!* → The Roma in the welcoming Parioli district immediately!
1	0	*I campi rom devono essere fatti su Venere... una bella navicella spaziale... e via... tutti al calduccio?* → Roma camps must be done on Venus... a beautiful spaceship... and go... all warm?
1	0	*Tagliamogli le palle vedrete che non lo faranno più che siano italiani o stranieri* → Let's cut their balls and you will see that they will no longer do it whether they are Italians or foreigners
0	1	*I terroristi sono islamici: solo quando le cose si chiamano con il loro nome si possono sconfiggere fuoridalcoro* → Terrorists are Islamic: only when things are called by their name can they be defeated out of the way
0	1	*Non conoscono la benché minima nozione di IGIENE i migranti!!!!!!* → Migrants do not know the slightest notion of HYGIENE!!!!!!
0	1	*Solamente un cretino puó rispondere alla domanda i migranti vengono da noi perchè anche noi siamo andati in Belgio o negli Usa. A parte che eran richiesti, chi c'era nel Belgio e negli Usa? Africani? Islamici? No, EUROPEI, e stessa cultura millenaria. Solo cosí funziona.* → Only an idiot can answer the question migrants come to us because we too went to Belgium or the USA. Apart from that they were required, who was there in Belgium and the USA? Africans? Islamic? No, EUROPEANS, and the same millennial culture. This is the only way it works.

Table 3.56 – Tweets Correctly Classified.

campo nomadi di Castel Romano , a Pomezia residenti barricati in casa"[68] and "*Eraclea, profughi protestano per il cibo e lo buttano in strada*"[69].

3.2.3 Aggressiveness Detection

As said in Section 2.3.4, a message is considered aggressive if its purpose is to humiliate or discredit a target using a rude language [Carmona et al., 2018]. Scholars like Waseem et al. [2017] and Sanguinetti et al. [2018d] consider aggressiveness as an additional orthogonal aspect of abusive language (Figure 1.1), that could reinforce with a pejorative language the aim to spread violence towards the target.

A first shared task that put the attention on this aspect of abuse online was proposed by

[68] *Toxic fire from the nomad camp of Castel Romano, residents barricated in their homes in Pomezia*

[69] *Eraclea, refugees protest for food and throw it on the street*

Task A		
Baseline_Avg_LMs	Avg_LMs (FT+All_feat)	
hs	hs	text
1	1	*Lui è Gianni il pomodoro. Per colpa di Salveenee che non vuole far sbarcare i migranti della SeaEye in Italia non verrà mai raccolto. Non possiamo permettere una simile barbarie, VERGOGNA! facciamorete restiamoumani portiaperti Amburgo* → He is Gianni the tomato. Due to Salveenee who does not want to disembark the migrants of the SeaEye in Italy, it will never be collected. We cannot allow a barbaric simile, SHAME! let's stay human open-door Hamburg
0	0	*TorreMaura Orfini PD: "I neo - fascisti ci fanno schifo". Ma che strano. Non si sono mai espressi in questi termini contro terroristi islamici, killer di ragazzine e sgozzatori di ragazzi felici. (StefanoLeo) Chissà perché...* → TorreMaura Orfini PD: "The neo-fascists make us sick". How strange. They have never spoken in these terms against Islamic terrorists, killers of little girls and slaughterers of happy boys. (StefanoLeo) Who knows why...

Table 3.57 – Ironic Tweets Misclassified in the Task A.

Carmona et al. [2018] at IberEval 2018 in the forum of MEX-A3T[70]. To our knowledge, for the first time they fostered the investigation on Mexican, a variation of Spanish that differs lexically from Castilian standard. Indeed, the linguistic context (such as distinctive lexica, syntactic characteristics and specific conventional meanings) plays an important role in the recognition of abusive language because it leads to different interpretations (see examples 27 and 28).

In the first and second edition of MEX-A3T, the boundaries between what is considered offensive and aggressive are fuzzy, and this leads to very low scores (see Section 2.3.4). In its third edition, Díaz-Torres et al. [2020] clarified that a message is aggressive if a target is involved and if its aim is hurting or inciting to violence towards the target; whereas the presence of derogatory words make it also offensive. Considering this premise, the corpus MEX-A3T2018 collecting Mexican Spanish tweets was annotated by two annotators looking at the intention of the user and at the rudeness of the language. From this perspective, tweets reporting quotes, pornographic content, self-attacks or attacks against objects are not considered aggressive.

The tweets in MEX-A3T2018 have been collected between August and November 2017 considering their geolocalization in Mexico City. The organizers used specific controversial hashtags (about politics, homophobia, and discrimination in general) and a fixed vocabulary about Mexican vulgarities and insults to extract these tweets. As seen in the previous experiments, the difficulty to process automatically such spontaneous texts, such as tweets, is due to the variety of language and expressions used by users with

[70] https://sites.google.com/view/mex-a3t2018/home?authuser=0

Task B		
Baseline_Avg_LMs	Avg_LMs (FT+25_feat)	
stereo	stereo	text
1	1	*Questa gente razzista non merita i rom. Trasferite il campo ai parioli... lì si troveranno tutti bene.* → These racist people do not deserve the Roma. Transfer the camp to Parioli. . . everyone will be happy there.
0	0	*L'Inghilterra discrimina i cristiani perseguitati a favore dei rifugiati musulmani e accoglie i jihadisti che hanno cacciato i cristiani. In due anni accolti solo 21 cristiani, mentre tornavano 400 terroristi andati a combattere con l'Isis. Lo chiamano multiculturalismo* → England discriminates against persecuted Christians in favor of Muslim refugees and welcomes jihadists who have expelled Christians. In two years only 21 Christians were welcomed, while 400 terrorists returned to fight with Isis. They call it multiculturalism

Table 3.58 – Ironic Tweets Misclassified in the Task B.

different backgrounds, and to the shortness of the text containing often orthographic or grammar mistakes as well as ironic devices. Therefore, they could contain jokes or humorous expressions, derogatory adjectives, profanities and also nicknames that underline, for instance, physical defects of the target. Some examples are reported in Table 3.60.

The entire dataset has been split into training (70%) and test set (30%) as showed in Table 3.59.

training set		**test set**
agg	non-agg	
2,727	4,973	3,156

Table 3.59 – Distribution of Labels in MEX-A3T2018.

As we can see from this extract of the dataset (Table 3.60), in general, people tend to use colloquially swear words with non-offensive purposes, making the task more challenging. To cope with this task and the difficulties that it brings with it, we carried out some experiments with a computational approach that combines a CNN with a specific set of designed features. For instance, to distinguish non-offensive from offensive expressions, we manually modeled a list of direct insulting expressions, such as *chinga a tu madre* or *vete a la verga*, and we considered as offensive cases the presence of the target.

3.2.3.1 System Description

One of the main difficulties related to the treatment of the data in Mexican is the lack of adequate linguistic resources. In fact, Mexican is a variation of Spanish and the

agg	text
0	*@USER Borracha jajajja me haces el día hdp jajajja también te amo y te extraño ♡♡* → @USER You're drunk hahaha you make my day mf hahaha Me too I love you and I miss you ♡♡
0	*@USER El arte estaba bien hermoso y creativo. Qué triste que ahora hasta en eso balgan berga.* → @USER The art was very beautiful and creative. How sad that now even in that they screw it up.
1	*@USER chinga tu madre ratero de mierda. Eres un vulgar populista retrógrado* → @USER f*** your mother f***ing thief. You are a vulgar retrograde populist
1	*@USER Que puto problema tienes con mis gustos musicales puta* → @USER What a f***ing problem you have with my musical tastes, b***h

Table 3.60 – Examples Extracted from MEX-A3T2018.

vocabulary often is not the same or has other meanings[71]. Moreover, these linguistic differences are more evident in the informal register, daily used also in social platforms. Therefore, considering these complexities, we experimented a combination of features with a neural network, trying to inform the system as much as possible. Indeed, a supervised system based on a deep learning architecture usually derives the features from the data without the necessity of feature engineering efforts. However, some more implicit and complex aspects of the language (as seen in the previous sections) are difficult to perceive.

We employed a CNN architecture [Banerjee et al., 2018], combining the features as showed in Figure 3.5. The idea, as according to the objectives reported at the beginning of this Section 3.2, is to understand the impact of linguistic feature on a simple deep learning (DL) framework; and to this purpose, two models have been implemented:

- simple CNN architecture (**DL**),
- CNN architecture with a set of linguistic features (**DL+FE**) (Figure 3.5).

The DL+FE model takes in input a vector of the extracted features (Features-layer) and an embedding representation of the tweets; whereas the DL model takes as input only the embedding representation. The neural network is articulated as follows:

[71] About that, the *Academia Mexicana de la Lengua* proposes a dictionary of Mexican language: https://www.academia.org.mx/obras/obras-de-consulta-en-linea/diccionario-de-mexicanismos

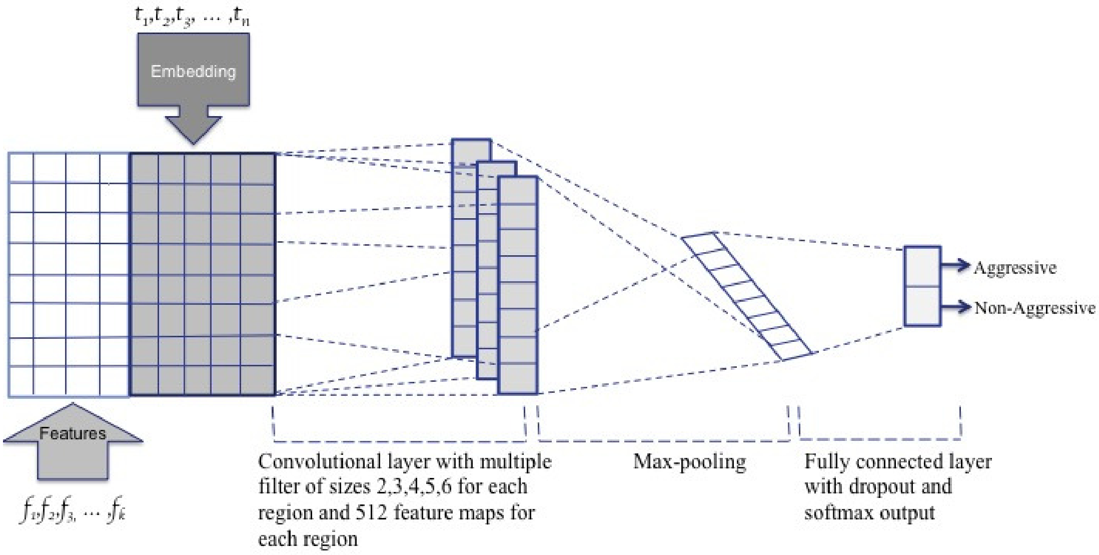

$t_1, t_2, t_3, \ldots, t_n$
Embedding
Features
$f_1, f_2, f_3, \ldots, f_k$
Aggressive
Non-Aggressive
Convolutional layer with multiple filter of sizes 2,3,4,5,6 for each region and 512 feature maps for each region
Max-pooling
Fully connected layer with dropout and softmax output

Figure 3.5 – Architecture based on CNN and Features.

Embedding Layer For the embedding, a vocabulary table is prepared by compiling the training data. Therefore, the embedding layer acts as a lookup table which maps vocabulary word indices into low-dimensional vector representations. The length of all the aggressive tweets is not the same. Therefore, the zero-padding (i.e., the missing part replaced by zeros) has been employed to maintain the input vector to a fixed size.

Convolutional Layer Let $W_j \in \mathbb{R}^p$ be the p-dimensional vector corresponding to the j-th word in the tweet. Here, a tweet is represented as $W_{1:m} = W_1 \oplus W_2 \oplus ... \oplus W_m$, where, $W_1, W_2, \ldots, W_m$ are the words in the tweet and $\oplus$ is the concatenation operator. Also, let $F_{1:n} = F_1 \oplus F_2 \oplus ... \oplus F_n$ be the feature set for the tweet $W_{1:m}$. The resulting vector is $r1 : n + m = F_{1:n} \oplus W_{1:m}$ after combining the feature set $F_{1:n}$ with the vector representation of the tweet $W_{1:m}$. Therefore, $r_{1:n+m} = r_1 \oplus r_2 \oplus ... \oplus r_{n+m}$, where either $r_i \in F_{1:n}$ or $r_i \in W_{1:m}$.

Let $r_{j:j+i}$ refer to the concatenation of $r_j, r_{j+1}, \ldots, r_{j+i}$. In the convolution operation, the filter $t \in \mathbb{R}^{hp}$ is applied to a window of h words to produce new features such as feature A_j is generated from a window of words $r_{j:j+h-1}$ by $a_i = f(t.r_{j:j+h-1} + b)$, where, $b \in \mathbb{R}$ is a bias term and f is a non-linear function. A feature map $O = [O_1, O_2, \ldots, O_{m-h+1}]$ (where, $O \in \mathbb{R}^{m-h+1}$) is produced by applying the aforesaid filter to each possible window of h words (i.e., $\{r_{1:h}, r_{2:h+1}, \ldots, r_{m-h+1:m}\}$) in the tweet. After applying the max-pooling operation to the feature map O, the maximum value $O' = \max\{O\}$ is obtained for the particular filter. The prime goal of the max pooling operation is to capture the most important feature with the highest value for each feature map. The proposed framework uses multiple filters with varying window sizes to obtain multiple features. The extracted features are provided as input to the fully-connected layer.

Fully-connected Layer In the fully-connected layer (sometimes called as the dense layer), the best features which are selected by the max-pooling operation from the convolutional kernel are combined. The output of this layer is passed to the next layer, i.e., the output layer.

Output Layer This layer is the final layer of this proposed architecture. This layer is made of 2 neurons. Each neuron is for a target class, i.e., one neuron for the aggressive (`agg`) class and another for the non-aggressive (`non-agg`) class. The softmax is used as the nonlinear activation function.

3.2.3.2 Linguistic Features

About the Feature-layer, we created a vector representing the features' representation of the texts. As underlined before, the set of features comes from an accurate analysis of the dataset, and involves stylistic, emotional and semantic aspects of Mexican tweets annotated like aggressive. To extract these features, the dataset was preprocessed deleting symbols and urls, that can hinder the process of extraction of features, and PoS-tagged the texts using FreeLing.

Stylistic Features Aspects like affect and personality could be captured by stylistic information [Argamon et al., 2007]. For this reason, specific traits of users writing have been taken into account, like: the presence of Mexican abbreviations more used in tweet context, such as *hdp*, *alv*, the use of punctuation elements (question *¿?*, exclamation marks *¡!* and sequences of dots ...) and uppercase characters. In particular, the system inspects if the user writes all the words in uppercase, or only the first letter, or uses capital letters inside the words. Moreover, quantitative features, such as the number of characters and words per sentence and the average word length, are considered. Among the stylistic features, the emoticons play an important role in the digital writing. Thus, the use of emoticons[72] annotated with polarity (positive/negative/neutral) is taken into account. Actually, emoticons are used as representations of facial expressions giving contextual information to readers.

Syntactic Features Considering the fact that the purpose of an aggressive tweet is to insult and offend someone, identifying the presence of a target is important for this task. Thus, one of the methods employed is to locate the mention of the target by means of specific syntactic patterns: the mention of the proper name or *@USER* followed or preceded by words or expressions from the lists of impolite adjectives or vulgar expressions.

[72] The annotated list of used emoticons for this work is provided by the Unicode Consortium: `http://www.unicode.org/`

Affective and Lexical Features Aggressive texts aim to offend and attack psychologically the victims, addressing their dignity with insults, humiliating adjectives or vulgar expressions. Therefore, two collections of profanities and derogatory adjectives have been created in order to help the system to detect aggressive texts like:

(56) *Te vas a chingar a tu madre pinche estupido pendejo!!!*[73]

Among lexical features, we considered also trigrams of words. In particular, the 100 most relevant sequences of trigrams of words have been chosen and weighted with TF-IDF. In order to understand the importance of these features for this task, Table 3.61 reports some of the most frequent trigrams in the analyzed dataset in comparison with unigrams and bigrams of words.

unigrams	**bigrams**	**trigrams**
('verga')	('la', 'verga')	('a', 'la', 'verga')
('madre')	('a', 'la')	('hasta', 'la', 'madre')
('putas')	('de', 'la')	('tu', 'puta', 'madre')
('putos')	('que', 'me')	('me', 'vale', 'verga')
('loca')	('que', 'no')	('a', 'chingar', 'a')
('pinche')	('los', 'putos')	('su', 'puta', 'madre')
('puta')	('la', 'madre')	('chingar', 'a', 'su')
('todo')	('en', 'la')	('sus', 'putas', 'madres')
('joto')	('en', 'el')	('todos', 'los', 'putos')
('ser')	('que', 'se')	('a', 'la', 'chingada')
('q')	('su', 'madre')	('chingas', 'a', 'tu')
('vida')	('las', 'putas')	('hijo', 'de', 'tu')
('vale')	('lo', 'que')	('mandar', 'a', 'la')
('marica')	('a', 'su')	('hijos', 'de', 'su')
('ver')	('y', 'no')	('la', 'puta', 'madre')
('luchona')	('que', 'te')	('me', 'vale', 'madre')
('mierda')	('puta', 'madre')	('chinga', 'tu', 'madre')
('solo')	('voy', 'a')	('estoy', 'hasta', 'la')

Table 3.61 – The Most Frequent Words N-Grams.

Actually, the different possible n-grams have been tested, and the trigrams obtained the best performance. Indeed, as Table 3.61 shows, the trigrams of words appear as the most significant respect to unigrams and bigrams, because they capture the typical multiword expressions used in Mexican language as vulgar or semantically altered expressions. Moreover, the data analysis reveals that the majority of aggressive expressions in Mexican language are long combinations of insults:

(57) *@USER Adios, hijo te toda tu perra celestial puta madre.*[74]

[73] *F*** you, stupid as***le!!!*

[74] *@USER Bye, son of a fu***ng b***h.*

Semantic Features In this study, one of the purposes is to examine the emotions linked to aggressive language. Therefore, the system tends to capture the emotions that are expressed in the aggressive texts by the use of the SEL. Each word in this lexicon is associated with the six principal Ekman emotions (Joy, Anger, Surprise, Disgust, Sadness and Fear) in accordance with the *Probability Factor of Affective Use* in Spanish. We considered, specifically, the words with a higher probability factor for each emotion. Moreover, the lexicon is extended by synonyms and slang forms usually used in social networks [Posadas-Durán et al., 2015].

Features Relevance Finally, by means of the Information Gain analysis, the impact of emotions and the relevance of the delineated features on the recognition of aggressive tweets has been analyzed. Figure 3.6 shows the results.

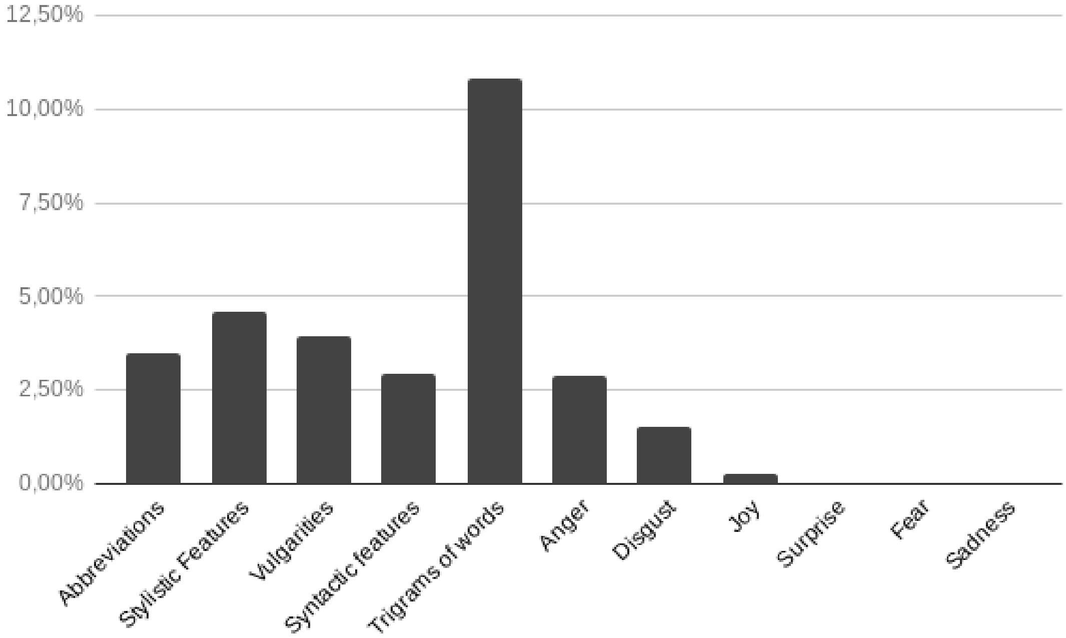


Figure 3.6 – Relevance of Features in the Training Set for Aggressiveness Detection.

From this chart, we can notice that anger and disgust are the principal emotions involved in aggressive language. Indeed, anger is the main emotion that drives the hateful message toward someone. Moreover, capturing trigrams of words appears to be relevant for this kind of task, confirming the fact that insults in Mexican language are, in most of the cases, longer composition of slurs.

3.2.3.3 Experiments and Results

In order to assess the contribution of features, in a first preliminary phase, we created a validation set splitting the training data provided by the organizers of the Aggressiveness

Detection shared task. Taking into account the unbalanced distribution of aggressive and non-aggressive tweets (see Table 3.59), the set of 7700 tweets is split into 7000 samples for training and 700 for validation, perfectly separated in 350 positive and 350 negative samples. While for the final evaluation, we used the test set provided by the organizers, comparing the performance of DL and DL+FT with those of the participants. For the tuning phase, the same measure used in the competition is employed. Considering the unbalanced corpus, the organizers preferred to use the $f1$-score for the positive class (i.e., aggressive tweets): *$f1$-score (1).*

Table 3.62 shows the results obtained on the validation set and those obtained on the test set; showing also the ranks that this approach could have obtained in the framework of the competition: 3rd for DL and 10th for DL+FE. In particular, we report only the scores of the best performing system and the baselines provided by the organizers of the shared task. These baselines are obtained training an SVM classifier with linear kernel and C=1 with trigrams of characters (MEX-A3T-Baseline_1) and with BoW (MEX-A3T-Baseline_2) [Carmona et al., 2018].

approach	**validation Set**	**test set**			
	$f1$-score (1)	precision	recall	$f1$-score (1)	rank
Graff et al. [2018]		–	–	*0.49*	*1*
DL	0.82	0.37	0.53	**0.44**	3
MEX-A3T-Baseline_1		–	–	*0.43*	–
DL+FE	**0.83**	0.38	0.42	0.40	10
MEX-A3T-Baseline_2		–	–	*0.37*	–

Table 3.62 – $f1$-score obtained for Positive Class on Aggressiveness Detection.

Looking at this table. we can see that using the validation set, the linguistic features seem to help the classification task. However, in the evaluation with the test set, the values for DL+FE decrease compared to DL. Moreover, comparing the obtained results with the baselines, it is evident that the results of DL+FE overcomes only the BoW baseline and not the trigrams of characters baseline. A possible justification of the high result obtained by the trigrams of characters is that with n-grams of characters, the system could detect also the typographical mistakes or variations often found in informal texts like tweets.

Considering these low results, an error analysis has been carried out. As seen in our previous analyses, we noticed that humor is prevalent in misclassified tweets. The users tend to disguise aggressive comments as humorous, involving, principally, sarcasm or irony in their utterances:

(58) *@USER @USER El señor tiene el superpoder de hablar mierda, cagar la madre y cambiar su color de piel a color naranja*[75]

[75] *@USER @USER This man has the superpower of producing shit with his mouth, f***ing and changing the colour of his skin to orange.*

(59) *@USER #LOS40MeetAndGreet 9. Por q es una mamá luchona que cuida a su bendiciòn*[76]

(60) *Gracias Facebook, pero no son personas que "quizá conozca", son personas que conozco pero que me valen verga y no las quiero agregar.*[77]

(61) *Aunque me cagues.... brilla, pero.. Algún día construiré mi súper misil para mandarte a la Merga #namaste #mamaste*[78].

We observed also a notable presence of certain elements such as the *laughter* that generally implies that the text is not aggressive. However, in some cases, the laughter seems to emphasize the offensive mockery expressed in the tweet. For instance, the following tweets are not classified as aggressive probably because of these misleading elements:

(62) *TRUMP, ESTADO UNIDOS Y SY PUTO MURO SE FUERO A CHINGAR A SU MEDRE SE QUEDARON SI MUNDIAL POR PUTOS JEJEJE*[79]

(63) *LOS PUTOS SIEMPRE QUIEREN TODO DE A GRATIS jajaja no mamen :D*[80]

In misclassified tweets we noticed also that users tend to use hyperbola, like superlative adjectives, to emphasize the anger or the disgust toward someone, such as:

(64) *Mándame una de 1000 por que te voy a mandar a chingar a su reputisima madre por putos y ratas*[81]

(65) *@USER Otra rata más!!, por igual que lo consiguen éstos imbéciles corruptos HDP no tienen madre! !*[82]

As typical of digital writing, users tend to abbreviate the words, especially the functional words such as pronouns or relative connectors. These abbreviations, not taken into account in our approach, seem to hinder the correct classification:

(66) *@USER @USER @USER A ti t da pena mostrar tu foto, por tu cara de estupido y Maricon que tienes ve con el america a chingar a su madre*[83]

[76] *@USER #LOS40MeetAndGreet 9. Because she is a fighter mother who takes care of her kid.*

[77] *Thanks Facebook, but they are not persons who 'maybe' I know, they are persons that I know but I don't care about them and I don't want to join them.*

[78] *Even I don't like you...shine, but.. one day I will build my super missile for sending you to Marshit #namaste #yousuck*

[79] *Trump, USA and its f***ing wall go to hell, they are out of the World Cup because they are motherf***er jejeje.*

[80] *The motherf***ers always want all free jajaja f*** off*

[81] *Send me one of the 1000 because of you're b***h, f*** you.*

[82] *@USER Other rat more!!, however they reach it, these f***ing corrupt son of b***h have no shame*

[83] *@USER @USER @USER You have shame to show your photo because of your face of jerk and gay, go to hell with America.*

(67) *@USER HDP! Citen 5 cosas q pasan en Venezuela y q temen los fanáticos y empinados a EPNdejo! D las q digan, mencion...*[84]

Although we informed the system with direct insults and not common swear words used also without an offensive aim, few misclassified tweets contain vulgar expressions that do not insult someone:

(68) *No hay mejor sensación que darte cuenta que algo te vale chingos de verga*[85]

Indeed, this kind of tweets aims to emphasize a subjective opinion, and does not address a target. Like those, the vulgar expressions are also used as exclamations, for instance:

(69) *Puta madre quiero dormirrrrrer*[86].

These reported examples reveal the real difficulty to detect automatically aggressiveness, especially in a very spontaneous context such as social platforms. The typical misspellings or grammar mistakes with our approach are difficulty treated. Moreover, the informal language is complex and overflowing with semantic exceptions that mislead the decision of the system.

3.2.4 Discussion

The described experiments allow us to investigate the objectives defined at the beginning of this Section 3.2:

1) how these dimensions interact between them and what is the benefit that the systems of hate speech and stereotypes detection could gain from mutual information;

2) if this possible advantage has a distinct impact on different textual genres;

3) and, if the employment of some specific features could help the system to infer implicit biases, double meanings or specific emotions like in the previous experiments on misogyny detection.

Firstly, from our analysis on the corpus of HaSpeeDe2020, the correlation between hate speech and stereotypes is clear in both textual genres considered here: tweets and news headlines. Nevertheless, taking advantage of this correlation, in terms of multitasking learning, injecting stereotypes knowledge to detect hate speech reports optimal performance only in tweets **(1)**.

A deep analysis showed that news headlines tend to involve specific nominal forms that like political slogans incite intolerant measures, especially, against immigrants. This

[84] *@USER Son_of_b***h! Cite 5 events in Venezuela and of whom the fanatic people and raised to jerk are afraid! Of those I mention...*

[85] *There is nothing like realizing that you don't care about something.*

[86] *F*** I want to sleeeeeeep.*

finding, supported also by the results obtained by CHILab, suggests the idea that in this kind of texts a syntax-based approach could help the system to recognize the cases of hate speech **(2)**.

About the role of linguistic knowledge, the experiments on hate speech and stereotypes detection show that no matter the contribution of transfer learning, the classifier needs to be informed linguistically to infer additional or implicit meanings. The entire set of the designed features, indeed, helps our systems to retrieve the indirect expressions of emotions, feelings, and also pragmatic and semantic aspects that could appear veiled in texts such as news headlines **(3)**.

A similar trend is visible in the experiments on the validation set created using the training set provided by the organizers of the MEX-A3T shared task [Carmona et al., 2018], but not confirmed on the blind data of the test set. To this purpose, in future work we plan to improve the set of features trying to cover all the issues emerged from error analysis, and experiment also with recurrent neural network that proved their efficacy in different NLP related tasks [Minaee et al., 2021].

3.3 Conclusions

In this chapter, we presented various experiments aimed at answering, mainly, our first research question:

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?

Firstly, we unveiled the characteristics of abusive language, and, secondly, we investigated the best techniques that help the system to infer implicit meanings. In particular, we observed from data that hate speech against immigrants and women involves very deep social biases, pervasive even in discussions related to the targets. For example, in newspapers online (articles or simple headlines) and in tweets talking about controversial issues such as abortion or feminist movements, the presence of these forms of abuse is persistent.

Therefore, to make the system aware of specific knowledge about stereotypes, prejudices and implicit meanings, we noticed that the use of lexica and related features is very useful, even in transformers-based models. In fact, lexica such as the ones created specifically to capture misogynistic characteristics, help the system to recognize when an indirect hateful message is actually addressed to the targeted victim. Whereas, not target-based lexica could lead the system to detect general aggressiveness, less useful in a real-world context [Fersini et al., 2020]. Moreover, considering the statistical correlation between orthogonal dimensions of abusive language, such as hate speech and stereotypes, we exploited also the technique of multi-task learning. We noticed that, in general, the injection of stereotypes knowledge reports an interesting performance for the detection of hate speech in tweets, but not in news headlines. Analyzing how hate speech is expressed in headlines, we noticed that the nominal structure of slogans is mainly employed, and for this reason, approaches that infer syntactic information could be preferred for this type of texts. Nevertheless, we observed that, also in a racist context, lexical information about conservative and inclusive interpretation of offensive words proves to be useful in hate speech and stereotypes detection revealing even some topic biases.

A constant in the error analysis of our experiments is the presence also of humorous and ironic expressions. The ironic language, adding a contrary or a new meaning to the message, makes the detection of verbal aggression and microaggressions hard. As Pexman and Olineck [2002] pointed out, the speakers/users tend to lower the cost of criticizing someone using ironic language. Unfortunately, the victims of these utterances do not perceive them as humorous having the same effect, or worse, as verbal attacks [Bowes and Katz, 2011].

Cognitively, the implicitness in language forces humans to process the message in more *steps of comprehension*, employing linguistic and contextual knowledge that helps them to interpret correctly its meaning. And this implicitness, that makes the true meaning veiled or masks the intention of the speakers/users, is realized using also figurative devices such as irony and sarcasm. To understand the impact of ironic language on verbal attacks and, consequentially, to make the system of abusive language detection aware also of these complex forms of expression, in the next chapters, we propose new analyses and new approaches.

Part II

Irony and Sarcasm Detection

Chapter 4

Traits of Irony and Sarcasm

Ironic language gives humans the possibility to express lightly their opinions without exposing themselves and obtaining a general consent and sharing, even for the utterances perceived unacceptable if said directly. Although some scholars stress only on this capacity of irony to mute the negative meaning [Dews and Winner, 1995], others insist on the amplification of the negative meaning that this linguistic device provokes, especially in its sarcastic form [Pexman and Olineck, 2002, Bowes and Katz, 2011]. Looking at the delicate contexts analyzed in previous chapters, where the sensitivity of a group or individual could be offended, irony and sarcasm prove to play a very important role, sometimes of reinforcement and sometimes of lessening of the attacks.

Considering this premise, in this part of our thesis, we focus mainly on the characteristics of ironic language, modeling the problem of irony and sarcasm detection. The computational approach helps us to bring to light the multilingual traits of irony even in a multi-genre framework, and the peculiar characteristics of sarcasm as a special form of irony (Chapter 5).

4.1 Theoretical Background

Over the centuries, irony as communicative and cognitive strategy, was studied by multidisciplinary scholars: philosophers attributed negative (falsehood by Plato) or positive values (medium to reach new knowledge by Socrates) to irony. Rhetoricians identified in irony the ability of embellishing the discourse, predisposing the listener/reader to receive the content of the speaker/author. From Cicerone in De Oratore (II, 269-270), the definition of irony is not substantially changed. Irony is saying something that is the contrary or different of what is literally said.

However, only in the last century, communication and language experts like John Austin, Paul Grice and John Searle, emphasized the concepts of *intention*, *efficacy* and *consequences* of the communicative act. Irony, in this context, was considered as an *indirect* illocutory act where the meaning of what is said is different of the one intended by the speaker/author [Marchetti et al., 2007]. In particular, Grice [1975] defined irony as a conversational implicature that breaks the maximum of quality used to communicate a meaning different from the conventional one established by semantics. Indeed, for Grice the meaning of words is not fixed, and depends on the communicative intention of the one who decided to use them.

Looking at how it works, other authors reevaluated irony in a new non-antiphrasic perspective. Sperber and Wilson [1981], for instance, highlighted the *echoic* function of irony in which the speaker/author evokes a 'meaning' of an expression with a critical attitude (i.e., 'Today it's a beautiful day!' when it is raining). According to Clark and Gerrig [1984], the Echoic Mention Theory of Sperber and Wilson [1981] could not explain other instances of irony such as the Jonathan Swift's essay "A Modest Proposal" (1729), where the author encourages serving Irish children as food to rich people. Clark and Gerrig [1984] stated that this essay could not be considered as a *mention*, because it is rather impossible that it could be considered a cultural practice or that someone would have proposed it. This essay is perfectly explainable if the *pretense* function of irony is taken into account. The Pretence Theory of Irony and the Echoic Mention Theory, actually, provide important concepts that Williams [1984] involved in a more generic frame: the Display Theory of Verbal Irony. For Williams [1984], some other verbal situations could not be explained only in terms of *echo* and *pretense*, and for this reason she introduced the idea of the *contradiction*: irony is realized in the juxtaposition of two incompatible elements.

To cover all the situations in which an utterance could be perceived as ironic, Utsumi [2000] proposed the idea of a prototype of irony, and the closer the sentence is to this prototype, the more ironic it is considered. An utterance, indeed, implicitly displays all the conditions for a perfect *ironic environment* when it:

1. alludes to the speaker's expectation E;
2. includes pragmatic insincerity by intentionally violating one of the pragmatic principles;
3. implies the speaker's emotional attitude toward the failure of E. [Utsumi, 1996]

The idea of the pragmatic insincerity is reconsidered also by Attardo [2007]. In his theory of Relevant Inappropriateness, the ironic utterance is both inappropriate and relevant to its context. In the communicative process, there are two factors that lead to the inference of the ironic value of the utterance: the maximum relevance, that is, the pertinence with the context and the intention; and the antiphrasic assumption of irony. This theory, extending the Grice's perspective, appears exhaustive in the explication of how the ironic meaning is elaborated and expressed [Marchetti et al., 2007].

Moreover, even cognitive studies have given their contribution to understand how listeners/readers interpret an utterance as ironic. In particular, Giora [1995] considered irony as indirect negation, that is, a type of negation that does not make use of an explicit negation marker and has not a scalar interpretation as results (i.e., ‘He is not silly’ does not evoke its opposite ‘He is clever’ but other similarly meanings such as ‘He is naive/boring’). This is the case when a positive/affirmative utterance, such as ‘What a lovely party!’, is used to express implicitly that a real state of affairs is different or far from our expected state of affairs, and that is clearly expressed in the utterance (‘lovely party’). In this, the comprehension of irony involves both the explicit message and the implicated negation in order to solve their dissimilarity. On this view, its interpretation requires more efforts than non-ironic utterance [Reyes and Rosso, 2014].

Later, focusing especially on sarcasm, Giora et al. [2015b] advanced the Defaultness Hypothesis with the aim to analyze the default and non-default meanings activated by specific stimuli in order to decide whether a sentence or text is sarcastic or not. A response generated by default should be novel, free of internal ambiguities, and free of explicit contextual marks that invite non-literal interpretation (such as #sarcasm, #irony). In accordance with this hypothesis, Giora et al. [2015b] discovered, considering a *sarcasm scale*[1] and the reading time, that the sarcastic interpretation is activated by default in negative sentences (‘He is not the most organized student’) and that the literal interpretation is activated by default in positive counterpart (‘He is the most organized student’). Moreover, looking specifically at sarcastic cases of negative sentences, Giora et al. [2015a] displayed that the function of negation used to mitigate/attenuate highly positive concepts triggers a default sarcastic interpretation, as well as the function of strongly attenuation of highly positive concept (‘He is not exceptionally bright’) [Giora et al., 2018]. This last work proves that also the rhetorical questions (‘Do you really believe he is sophisticated?’) could play the role of mitigators lessening the implicit assertions and conveying a sarcastic meaning, especially in stronger mitigation cases.

This brief *excursus* of the theories that delineate the concept, the functions, the characteristics, and even the comprehension of irony demonstrates the difficulty of defining irony under a unique perspective. Also for this reason, the literature of computational approaches designed to recognize automatically ironic language is wide and various.

4.1.1 Irony and Sarcasm Characteristics

The works that investigated the specific characteristics of irony and sarcasm are not that many. From a more linguistic and cognitive perspective, sarcasm could be distinguished from other forms of irony for involving negative evaluation against the victim [Alba-Juez and Attardo, 2014]. The negativity of sarcasm covered by apparent positivity is found out in qualitative and quantitative analyses carried out on English self-tagged tweets by Wang [2013]. The aim of this work was to answer precisely four questions:

1) Is sarcasm more aggressive than irony?
2) Is there a specific target attacked in sarcasm, but not in irony?

[1] The participants in the experiments had to decide at what degree the sentence could be sarcastic.

3) Is the tweeter aware of his/her sarcastic or ironic tweets?

4) Are there any overlapped features between sarcasm and irony? [Wang, 2013]

Firstly, the quantitative analysis on ironic and sarcastic tweets confirms that users use more positive words to express more aggressive meaning. Secondly, the qualitative analysis reveals that: also ironic tweets could address a specific target; users are aware to be ironic/sarcastic; and, finally, users tend to use the hashtag #sarcasm to express more subjective utterances and #irony to identify an event as ironic. In some cases these hashtags could be used in interchangeable way explaining why tweets containing #irony could attack a victim. This last point underlines the fact that speakers commonly perceive irony and sarcasm as similar phenomena.

Similar findings are reported by Sulis et al. [2016]. The authors examined qualitatively and quantitatively the dataset released by the organizers of SemEval2015-Task11 [Ghosh et al., 2015] containing English self-annotated tweets that include specific hashtags (i.e., #not, #sarcasm and #irony). In particular, they investigated the impact of sentiment, emotions, various affective lexica, tweets length and punctuation in this dataset, revealing some important differences especially between tweets containing #irony and #sarcasm, such as:

- in general, shorter messages are mostly negative, and sarcastic tweets result to be shorter than ironic ones;

- colons are specially used in ironic texts, while exclamation marks in those containing #sarcasm and #not;

- tweets with #irony are especially related to negative sentiment and emotions (anger, disgust, fear, and sadness), differently from those with #sarcasm that contain words expressing mainly joy, anticipation, trust, surprise and positive sentiment (in line with Wang [2013]);

- polarity reversal [Bosco et al., 2013] is more relevant in tweets with #sarcasm, showing a particular shift from literal positive to real negative polarity;

- tweets with #irony prove to be more creative and implicit than the ones with #sarcasm.

These observations are supported also at computational level. For instance, in various English datasets of tweets, Hernández-Farías et al. [2016] demonstrated the discriminating power of negative sentiment in irony detection, and of positive sentiment (and words expressing 'love') in sarcasm detection and the relevance of features such as the presence of mentions and the length of tweets, especially in sarcasm detection.

4.2 Computational Approaches to Irony Detection

The detection of irony and sarcasm is gaining more and more interest in scientific communities and companies. In fact, it proves to be relevant in Sentiment Analysis for recognizing correctly the opinion or orientation of users about a specific subject (product, service, topic, issue, person, organization, or event) [Reyes and Rosso, 2012, Ghosh et al., 2015] as well as on Hate Speech detection [Nobata et al., 2016, Cambria et al., 2017]. Considering the change of type of communication in favor of the computer-mediated communication, users tend to express their opinions on social platforms such as Twitter and Facebook, or directly in commentary sections of vendors such as Amazon. In this scenario, researchers are encouraged to explore figurative language principally in these texts online that appear spontaneous and in some cases brief; and these efforts of analysis are evident in the amount of shared tasks that recently have been proposed.

Among them, let us mention: SENTIPOLC (SENTIment POLarity Classification) 2014 and 2016 subtask on *Irony detection in Italian tweets* [Basile et al., 2014, Barbieri et al., 2016], DEFT2017-Task2 on *Figurative language detection* in French tweets [Benamara et al., 2017], SemEval2018-Task3 on *Irony detection in English tweets* [Van Hee et al., 2018b] that asked participants to distinguish also among four classes (irony by clash, situational irony, other verbal irony and non-irony), IroSvA2019 on *Irony Detection in Spanish Variants* [Bueno et al., 2019] where also the topics were provided to understand to what ironic comments referred to, IDAT 2019 on *Irony Detection in Arabic Tweets* [Ghanem et al., 2019] that is the first shared task that approaches a such difficult language to process, ALTA2019 shared task on *Sarcastic Target Identification* [Molla and Joshi, 2019], focused on the presence of a victim in sarcastic messages. This last task is very related to sentiment target identification [Liu, 2012]. More recently, FigLang2020-Task2 on *Sarcasm Detection* [Ghosh et al., 2020] focused on sarcastic texts identification in English conversations on Twitter and Reddit. These shared tasks prove the efforts, at international level, in the modeling of figurative language understanding, in order to let the machine understands perfectly natural language.

4.2.1 Irony and Sarcasm Detection

Considering the fuzzy boundaries among different types of irony, the majority of computational techniques address irony and sarcasm similarly and considering them as synonyms. As in many NLP tasks, and as confirmed in the majority of the above-mentioned shared tasks, deep learning-based approaches obtain very competitive results also in irony and sarcasm detection. Especially transformers models, such as BERT and its variants [Potamias et al., 2020], have been largely employed in the last competition in FigLang2020-Task2, confirming the importance for an automatic system of having extended language knowledge.

Along with language models, other techniques have been explored on irony and sarcasm detection, using various automatic approaches (rule-based systems [Sentamilselvan et al., 2021], classical machine learning [Reyes et al., 2012] and deep learning [Zhang et al., 2019]), such as data augmentation [Lee et al., 2020], multi-task learning [Cimino et al., 2018, Wu et al., 2018], neural attention mechanisms [Tay et al., 2018, Xiong et al., 2019], multi-modal analyses [Cai et al., 2019, Castro et al., 2019] and feature engineering. Taking advantage of the explainability of models based on classical machine learning algorithms, scholars examined the specific impact of stylistic features [Buschmeier et al., 2014], pragmatic symbols (such as hashtags, mentions and emojis in tweets) [Kunneman et al., 2015, González-Ibáñez et al., 2011], syntactic patterns [Hao and Veale, 2010, Riloff et al., 2013], sentiment and emotional lexica [Hernández-Farías et al., 2016], semantic context and users information [Bamman and Smith, 2015, Joshi et al., 2015][2]. Therefore, discriminating characteristics of ironic and sarcastic texts that could help the classifier were investigated. In particular, scholars studied aspects such as a potential incongruity of information within ironic or sarcastic messages, as well as language ambiguities [Reyes et al., 2012, Barbieri et al., 2015b, Naseem et al., 2020], semantic contrast [Pan et al., 2020], sentiment discordance [Zhang et al., 2019], emotional shift [Agrawal et al., 2020], dissonance between positive sentiment and negative situations [Riloff et al., 2013] and contrast between the orientation of a specific community (e.g., forum) and the published message [Wallace et al., 2015, Joshi et al., 2015].

Another aspect previously investigated in irony and sarcasm detection is the contribution of emotion and sentiment in various languages and in different contexts. In particular, Hernández-Farías et al. [2016] and Babanejad et al. [2020] showed, in various English corpora of tweets, the robustness of models based on affective features; in a Facebook context, Raghavan et al. [2017] proposed a prototype where identifying emotions helps sarcasm detection; and Chauhan et al. [2020] underlined the effectiveness of using sentiment and emotion detection tasks in a multi-task learning framework to recognize sarcasm in a multi-modal conversational scenario.

With respect to hate speech, the intuition about the use of sarcasm to disguise hateful and offensive utterances was preliminary investigated in Justo et al. [2014] and Nobata et al. [2016]. Justo et al. [2014] delineated differences and analogies in sarcasm and nastiness detection. In particular, they observed that the length and linguistic information are relevant especially for sarcasm detection, whereas semantic information improves results for both tasks. However, specific lexical cues seem to work really well for nastiness detection, demonstrating that nasty opinions tend to be expressed by users overtly and without ambiguities. Nobata et al. [2016] showed instead how abusive contents sometimes are disguised by sarcasm, making hate speech more subtle and, thus, more difficult to be recognized. Nevertheless, the intuitive correlation between sarcasm and abusive language is poorly discussed and experimented [Cimino et al., 2018].

Looking in particular at the languages that we will analyze in this second part of the thesis, Spanish and Italian, we briefly describe the existing approaches.

[2] Exhaustive overviews are presented in the following surveys: Wallace [2015], Joshi et al. [2017], Sarsam et al. [2020].

4.2.1.1 Irony Detection in Spanish

The literature about ironic language detection in Spanish is actually limited and cover various forms of irony, such as humor and satire. Castro et al. [2018] and Chiruzzo et al. [2019, 2021] organized the HAHA shared task in the context of IberEval 2018 and IberLEF 2019 and 2021, about identification of humor in Spanish tweets annotated by various users of "Clasifica Humor" platform[3]. About satire detection, del Pilar Salas-Zárate et al. [2017] proposed a psychological based approach exploiting news satirical sources on Twitter for Mexican and Castilian variants of Spanish, while Barbieri et al. [2015a] employed linguistic and semantic features (such as ambiguity and synonyms), sentiment analysis and slang words in a similar collection of tweets from Spain.

The studies focused especially on irony and sarcasm are really few. To the best of our knowledge, the study proposed by Jasso and Meza-Ruíz [2016] is the first to explore the irony in Spanish tweets, considering the sarcasm as a subclass of irony. In particular, they explored word and character level of the texts employing n-grams of words and characters and word embedding, using classical classifiers such as the SVM and the RF. Recently, Blanco et al. [2018] explored deeply the function of sarcasm in Spanish dialogues online, creating a corpus annotated taking into account the presence of sarcasm and the tone of nastiness on a three-grade scale without previous definitions. Indeed, the annotators, mainly from Spain, used their own perception and provided contextual information to annotate the data. The dialogs were extracted considering some specific topics and a statistical analysis of the agreement for each issue reveals that terrorism, abortion, and gay marriage report the highest percentage of nasty and sarcastic language.

Only in 2019, Bueno et al. [2019] organized, for the first time, a shared task at IberLEF 2019[4] of irony detection on three variants of Spanish (IroSvA)[5]: Castilian, Mexican and Cuban. The corpus provided for each linguistic variant is a collection of short texts annotated as ironic and non-ironic. Differently from Mexico and Spain, the access to social media in Cuba is still limited and, for this reason, the organizers preferred selecting news comments for this variant and tweets for Mexican and Castilian Spanish. These data were collected taking into account specific topics that are controversial and generate major discussion, such as: digital television, tourism, sports scandals, internet, transport in Cuba; the divorce of ex Mexican president, the fuel shortage, the funding cuts, CONACYT problems, the foreign politics in Mexico; and the book written by the Spanish Prime Minister, the exhumation process of dictator Franco, reality show, the tendency of freethinkers in favor of the flat Earth, the mediator/*relator* for Catalan issues in Spain. The labels relative to these issues are reported in each corpus. Like in Blanco et al. [2018], the data have been annotated by native speakers of each variant, considering the predefined context and without a specific definition of irony but basing on their own perception of irony. In this frame, all the forms of irony, such as sarcasm, have

[3] https://clasificahumor.com/

[4] https://sites.google.com/view/iberlef-2019/

[5] https://www.autoritas.net/IroSvA2019/

been covered by the `ironic` label. The particularity of this task relies on the presence of three different variants of Spanish language and on two textual genres (along with different topics); thus, it gives the possibility to observe the behavior of participating systems even in cross-variant and cross-genre perspective.

For each variant, Bueno et al. [2019] proposed a binary task on irony detection for a total of 3 subtasks. In general, the best performing systems employed transformers and classical approaches with different kinds of representation of texts (i.e., word embeddings, LSTM derived representation, character and word n-grams) and feature engineering (such as morphological and dependency-based features). The highest results in binary classifications have been obtained on the Castilian variant (with a macro F-score 0.717), followed by the Mexican (0.680) and the Cuban (0.653) one. Moreover, observing the scores obtained by the participants in each topic for each variant, the organizers noticed that the systems tend to perform better on economic and internal political subjects (funding cuts, political mediator and tourism) and worse in sport, reality show, and foreign politics. Looking at the cross-variant/genre performance, Bueno et al. [2019] noticed that the best results on Mexican and Cuban variants are obtained training the models, respectively, on Cuban and Mexican texts. While, when the models are trained on the Castilian variant, the scores of the testing on Mexican and Cuban data are worse and similar. This confirms the expectation about the linguistic proximity of American variants. Testing, instead, the models on Castilian variant, the best performance is obtained with Cuban data. This is possibly related to the fact that the news commentary are less informal and more similar to the Spanish standard.

4.2.1.2 Irony Detection in Italian

About Italian, one of the first works on irony detection online is described in Barbieri et al. [2014]. The authors created a corpus of tweets labeled as ironic and non-ironic. In particular, they collected all tweets coming from Twitter accounts “spinozait” and “LiveSpinoza” in the set of ironic ones; and all the tweets coming from daily Italian newspapers in non-ironic set. Spinoza is an Italian collective blog that includes ironic posts on politics on the style of news. This corpus is very unbalanced (12.5% of the corpus is ironic and 87.5% non-ironic), but represents a possible real scenario. To detect irony, they used a Decision Trees based classifier informed with different features. They aimed at capturing: the rarest words and frequency gap (exploiting the retrieved frequencies from the CoLFIS corpus[6] [Laudanna et al., 1995]), the synonyms, the ambiguity looking at the number of WordNet synsets associated to a word, the style by means of PoS tags, the sentiments, and the reverse of sentiments using Sentix lexica. The experiments reveal a very high performance of this approach on the test set (F-score of 0.76) and in 10-fold cross-validation (0.71) compared to the baseline model that uses only BoW (0.65 in both sets of experiments).

[6] https://www.istc.cnr.it/en/grouppage/colfis

The SENTIPOLC shared task at EVALITA 2014 is the first effort of Italian NLP community to provide a benchmark for irony detection on social media. It appears as a subtask of a more general task focused on subjectivity and polarity detection in tweets. Irony tends to reverse the polarity of a message, making sentiment analysis hard. In this context, Basile et al. [2014] proposed a corpus of tweets annotated with the presence of subjectivity (binary classification), of what type of polarity (negative, positive, both or neutral sentiments), and of irony (binary classification). Most of the participats used supervised approaches, based especially on the SVM classifier. This gave the possibility to experiment with a lot of different linguistic features based on emoticons, punctuation, links, usernames, hashtags and specific vector space to identify words that are out of vocabulary. Even some rule-based systems were employed [Delmonte, 2014]. The results for irony detection task were lower than other subtasks: 0.576 F-score compared to 0.714 in subjectivity detection and 0.677 in polarity identification. A lower score (0.541) was reported by systems participating in the second edition of SENTIPOLC at EVALITA 2016 [Barbieri et al., 2016]. A possible reason relies on the inconsistency of the topic in the provided training and test sets. In general, the approaches to irony detection are similar to the first competition, although some teams represented texts also as word embeddings and employed deep learning approaches.

Building on the experience of SENTIPOLC and SemEval2018-Task3, we proposed a new shared task at EVALITA 2018, IronITA[7], that, taking into account the different forms of ironic language, aims at analyzing the presence of irony and, if the tweet is ironic, of sarcasm. In Cignarella et al. [2018b], on the basis of theoretical literature, we defined sarcasm as a type of more aggressive irony, with a cutting tone and addressed to a victim[8]. Sarcasm, until now, is a type of irony not accurately explored. Indeed, as said before, the majority of works focused on figurative language processing, treated irony and sarcasm as synonyms.

4.3 Conclusions

Taking into account the existing literature on irony and sarcasm from a computational perspective, we can notice that the attention on languages different from English is still scarce. From a rhetorical perspective, figures such as irony modify the logic value of utterances suggesting other interpretations that could be inferred from language-based knowledge, such as idiomatic expressions and proverbs [Basile et al., 2018]. For this reason, it is important to extend the analysis of the phenomenon in other languages, in order to capture: on the one hand, the specific ironic patterns typical of a language/culture; and, on the other hand, the universal linguistic and pragmatic criteria connected to the understanding of ironic language. To this purpose, in the next chapter we aim at investigating ironic language in Spanish and Italian, poorly explored until now, and at

[7] See Section 5.2.

[8] Later, in ALTA 2019 Molla and Joshi [2019] presented a shared task asking participants to detect the target of sarcastic utterances in English.

delineating multilingual common traits.

Moreover, although higher performance is achieved by systems that employ the pre-trained language models and complex neural networks, Ghosh et al. [2020] noticed that subtle humor is missed by these best scoring systems. A type of subtle humor is done by implicit or world knowledge-based incongruities difficult to be perceived by the system without a contextual or extra information, such as the hashtag in 'Took 6 hours to reach work today. #yay' [Joshi et al., 2017]. Other cases, such as news comments (71), miss explicit elements that suggest the opposite/secondary meaning, such as:

(70) *#8aprile giornata dei #rom allora facciamo anche la giornata delle piattole, scarafaggi, topi e altri parassiti*[9]

(71) *no se preocupe tanto que el dia que eso llegue, va a ser el mismo día que tendremos internest de banda ancha disponible para todos*[10]

As human, we can infer the sentiment incongruity in (70) between the first part of the sentence ('#8aprile giornata dei #rom') and the rest, and in (71) between the positive expectation of 'no se preocupe tanto' and the frustration that 'that day' will never come. But this is not so obvious for the system. Moreover, we can perceive the negativity of emotions (frustration, anger, and disgust) especially in the second part of the utterances in contrast with the neutrality of the first one in (70) and the positivity (anticipation, hope) in (71). And, it is clear, for example, that in (70), although in an ironic form, the user is expressing hate towards the Roma community excluding the possibility to celebrate a day dedicated to them. However, for a system based exclusively on a pre-trained language model or a complex neural network, detecting the affective aspects expressed in the sentence appears difficult, especially if they contribute to make the message ironic. About it, we hypothesize that the collaboration between the general perspective of the pre-trained language models and the linguistic knowledge inferred by specific designed features, could help the system to infer complex mechanisms behind irony interpretation.

Finally, taking into account the previous observations on sarcasm and its more offensive intention [Bowes and Katz, 2011, Lee and Katz, 1998], we would contribute to the discussion about its similarities and differences with irony, often used as synonym. In particular, we aim at especially identifying the impact of hurtful and emotional language on its expression in controversial social issues, such as the integration of minorities, where sarcasm could be marked by a clearer hostile behavior.

[9] *#8april day of the #roma community then we also do the day of lices, cockroaches, mice and other parasites.* Tweet extracted from IronITA2018.

[10] *do not worry so much that the day when it will come, it will be the same day that we will have broadband internet available for everyone.* News comment extracted from the IroSvA-Cuba corpus.

Chapter 5

Irony and Sarcasm: The Case of Abusive Context

Taking into account the previous works on irony and sarcasm, in this chapter we present, firstly, some computational experiments that give us the possibility to bring to light important characteristics of ironic language in a multilingual/monolingual and multigenre perspective. Secondly, we focus on the abusive context, where ironic language is used to both lessen the tone and reinforce the negativity of the toxic message. To this purpose, we investigate:

1) what are the common elements in various languages and genres that could trigger the interpretation of irony and its detection online (Sections 5.1 and 5.2);

2) the role played by affective aspects in ironic language, especially in the abusive context (Sections 5.1 and 5.2);

3) which characteristics of ironic language are peculiar of sarcasm (Section 5.2).

And, considering the previous results obtained on abusive language with combined approaches (Section 3.2.2), we wonder also here:

4) if transformer-based architectures aimed to identify ironic language could benefit from the injection of linguistic features (Section 5.3).

Let us highlight that the first three objectives, inspired by the literature in ironic language, touch multidisciplinary fields and, especially, cognitive mechanisms that are far from the scope of this thesis. Therefore, our approach of investigation is limited and centered on giving a contribution to the discussion with the analysis of specific features involved in the automatic detection of this linguistic phenomenon in Spanish and Italian, even in a neural network context. Investigating the multilingual features of ironic language helps us to better define the characteristics of ironic language and in particular of sarcasm, as well as its role in the abusive context.

Thanks to these experiments, we are able to answer also the second research question of this thesis:

RQ2 What is the role played by sarcasm in hateful messages online?

5.1 Irony Detection in Spanish

As seen in the previous chapter (Section 4.2), the majority of works on irony detection focused on the English language. IroSvA represents the first effort as shared task centered on irony detection in Spanish. This shared task is organized in three subtasks:

- Subtask A: Irony detection in Spanish tweets from Spain
- Subtask B: Irony detection in Spanish tweets from Mexico
- Subtask C: Irony detection in Spanish news comments from Cuba

For each subtask the organizers provided a collection of short texts about specific social and politic issues, discussed on Twitter or in newspapers such as Cubadebate (`http://www.cubadebate.cu/`), Granma (`http://www.granma.cu/`) and OnCubaNews (`https://oncubanews.com/`). Each text is annotated as ironic (`iro`) and non-ironic (`non-iro`) and contains the label of the referred topic. Table 5.1 describes the distribution of labels in the released datasets, and Table 5.2 reports some examples extracted from each dataset (called here IroSvA-Spain, IroSvA-Mex and IroSvA-Cuba).

subtask	**name**	**training set**		**test set**		**n_topics**	**genre**
		`iro`	`non-iro`	`iro`	`non-iro`		
A	IroSvA-Spain	800	1,600	200	400	10	Tweets
B	IroSvA-Mex	800	1,600	200	400	10	Tweets
C	IroSvA-Cuba	800	1,600	200	400	9	Comments
Total		7,200		1,800			

Table 5.1 – Distribution of Labels, Number of Topics (N_topics) and Genre for each Dataset.

As we can notice in Table 5.1, the distribution of labels in the datasets is consistent among all the variants but is unbalanced between the two considered classes. Nevertheless, the organizers preferred to evaluate the performance considering this real distribution of irony online, and, thus, using the $f1$-macro, implying that both classes have equal weight in the final score. The ranking, instead, is based on the average of $f1$-score of all the subtasks. Analyzing the confusion matrices, Bueno et al. [2019] noticed that the major confusion appears from ironic to non-ironic text in all the three subtasks. This suggests that the systems, regardless the type of Spanish variant and the genre, tend to perform similarly. Moreover, this type of error is expected, considering the difficulty to recognize some of these texts as ironic even by humans. Looking at the Cuban comments in Table 5.2, for instance, it is hard without the commented news or the real user's experience to decide if they are ironic or not. Indeed, in both comments, we found very positive words ('muy', 'eficiente', 'bello', 'excelente', 'orgullo') that could suggest that these hyperboles are used as indirect negations [Giora, 1995].

`iro`	`language`	`topic`	`text`
0	Castilian	Book Sánchez	*Pedro Sánchez publica en febrero un libro que ha terminado de escribir en La Moncloa URL vía @USER* → Pedro Sánchez publishes in February a book that he has finished writing in La Moncloa URL by @USER
0	Mexican	Foreign Politics	*Váyanse a luchar a Venezuela. A México no le incumbe* → Go fight in Venezuela. Mexico is not concerned
0	Cuban	Tourism	*Es un bello hotel, con un excelente colectivo, no tengo dudas de que será orgullo para nuestro turismo.* → It is a beautiful hotel, with an excellent group, I have no doubt that it will be pride for our tourism.
1	Castilian	Mediator	*@USER También que le expliquen que un relator" no es un anticonceptivo* → @USER Also that they explain to him that a relator" is not a contraceptive
1	Mexican	CONACYT	*¿De cuándo acá tan preocupados por la ciencia y la investigación?* → Since when this concern here about science and research?
1	Cuban	Digital TV	*LA TV DIGITAL ESTA SIENDO MUY EFICIENTE.* → DIGITAL TV IS BEING VERY EFFICIENT.

Table 5.2 – Examples Extracted from IroSvA-Spain, IroSvA-Mex and IroSvA-Cuba.

Considering that, we tried to get as much advantage as possible from the topic information provided in the datasets. Therefore, contextual and semantic features, with lexical, stylistic and affective ones, have been analyzed in a simple classifier based on an SVM classifier with an RBF kernel and the parameters $C = 5$ and $\gamma = 0.01$. The kernel and the parameters of the SVM classifier have been set on the basis of various experiments. Considering the imbalanced collection of data (Table 5.1), we used the function to balance the weights of the classes provided by the *scikit-learn* library for Python. In the frame of our participation in this shared task, we called our system SCoMoDI (Spanish Computational Models to Detect Irony).

5.1.1 Linguistic Features

The set of the designed features can be organized as follows.

Stylistic Features Taking into account the corpora-based analyses carried out in Karoui et al. [2017] for English, French and Italian, we examined the impact of features such as hyperbole expressed by exclamation marks (!, ¡), ellipsis expressed by dots (...), questions denoted by question marks (?, ¿) and quotes expressed by inverted commas (“”, ‘’). Considering the fact that some ironic texts could be characterized by a sarcastic tone against someone, we took into account also the typical symbol of mention in Twitter (@). In the features vector, these features are represented by a simple count of the number of times each item appears in the text.

Lexical Features As lexical features, we used unigrams of words weighted with TF-IDF. To extract the unigrams, we pre-processed the texts, deleting all symbols and numerical characters and selecting words using a tokenizer able to take into account the compound nouns. Finally, in order to weight the words without considering their inflectional morphology, we used the SnowballStemmer for the Spanish language provided by NLTK.

Semantic Features In this group we gather features such as: semantic contexts and topic distribution. *Semantic contexts* of each text are computed by calculating the cosine similarity between the vocabulary of the text and the vocabularies extracted from each group of ironic texts labeled with the same topic. The cosine similarity is calculated on the basis of pre-trained word embedding of the Spanish Billion Words Corpus[1] provided by Cardellino [2016]. To lead the classifier to capture similarities between texts belonging to the same topic, we extracted the *topic distribution* of the text, considering the number of topics of each subtask (Table 5.1). To this purpose, we created a Latent Dirichlet Allocation model on the provided training sets using the *Gensim* library [Řehůřek and Sojka, 2010] for Python, taking into account also bigrams and trigrams of words. The idea is to cluster the texts that talk about the same topic similarly in the same class.

Affective Features Inspired by previous work on computational models for irony detection [Hernández-Farías et al., 2016, Sulis et al., 2016, Pamungkas and Patti, 2018], we explored the role of features related to the affective information present in the tweets and the psychological response stimulated by the message.

[1] Available at http://crscardellino.github.io/SBWCE/

Emotional Categories To identify the emotions involved in each text, we counted the number of words that belong to emotional lexica, such as the multilingual EmoLex provided by Mohammad and Turney [2013a] and the SEL. For each variant, we considered only the emotions that are relevant for the classification. It is surprising that, for all the variants, the most relevant emotions are negative, such as anger, fear, disgust, and sadness.

Dimensional Models of Emotions In order to understand the mental responses to stimuli of ironic texts, we investigated the impact of psychological dimensions such as imagery, activation, and pleasantness. Inspired by Reyes et al. [2013] and Hernández-Farías et al. [2016], we used an automatic translated Spanish version of the Dictionary of Affect in Language (DAL) [Whissell, 1989]. From our analysis, the dimensions that turned out useful for the classification in all the three variants are pleasantness and imagery.

Abusive Language Inspired by Blanco et al. [2018] and considered the relevance of negative emotions, we analyzed also the impact of abusive language, counting the words included in the Spanish lexica of derogatory expressions and profanities described in Section 3.2.3.2. These lists of words prove to be significant for the classification, especially, in Mexican and Castilian tweets.

5.1.2 Experimental Setting, Results and Observations

To evaluate these features and the parameters for the classifier, we performed a 5-fold cross-validation on the training sets, tuning the systems on the metric used for the competition: the average of $f1$-scores of the classes. Moreover, we carried out also the ablation feature test and, on the basis of these analyses, we created the models for each variant. The highest $f1$-scores values obtained with the relevant features are reported in Table 5.3.

The organizers provided four baselines calculated considering different representations of the text: n-grams of words (*Word nGrams*), word embeddings (*W2V*) and low dimensionality statistical embedding (*LDSE*) [Rangel-Pardo et al., 2016]. They used also the majority voting (*Majority*) technique as additional baseline. Table 5.4 reports the results obtained in the competition compared with these baselines. In this table, we can see that only the model built for the Cuban variant (Subtask C) overcomes slightly all the provided baselines, while the other models overcome only the *Majority* baseline. This difference could be due to the textual genre of news comments which does not contain Twitter mentions (@USER), hashtags or emojis. Another influential factor could be the use of lexical features, such as unigrams, that in general help the text classification.

However, although the different textual genres, analyzing the misclassified texts, we noticed that in all the three variants of Spanish irony is expressed similarly. Actually, we individuated various figures of speech involved in the expression of irony. With the proposed models, we aimed at capturing some of them by exploiting textual marks,

	Subtask	**Subtask**	**Subtask**	**Subtask**	**Subtask**	**Subtask**
	A	**B**	**C**	**A**	**B**	**C**
Lexical Features	v	v	v			v
Stylistic Features						
hyperbole	v	v	v			v
ellipsis	v	v	v			v
question	v	v	v		v	
quotation	v	v	v			v
mention	v	v			v	
Semantic Features						
semantic context	v	v	v	v	v	v
topic distribution	v	v	v		v	v
Affective Features						
Emotional Categories						
anger	v	v	v		v	
fear	v	v	v	v		
disgust	v	v	v	v		
sadness	v	v	v			v
Dimensional Models of Emotions						
imagery	v	v	v	v	v	
activation	v	v	v			
pleasantness	v	v	v		v	v
Abusive Language						
derogatory expressions	v	v	v	v		
profanities	v	v	v		v	
***f*1-scores**	0.456	0.477	0.501	**0.549**	**0.553**	**0.523**

Table 5.3 – Experimental Results on the Training Sets.

although the error analysis highlights that it was not sufficient. In particular, we found that hyperbole (72) and ellipsis (73) are expressed also at a semantic level. See, for instance, the following examples from the IroSvA test set where irony was not recognized:

(72) *Felicidades director muy buena tarifa así se hace.*[2]

(73) *Cuando yo sea grande quiero ser como los inventores del paquete.*[3]

As defined in Lanham [1996], hyperbole is expressed by "exaggerated or extravagant terms used for emphasis and not intended to be understood literally". In fact, Example (72) from IroSvA-Cuba is a clear example of hyperbole that aims at exaggerating positively the actions of someone who is doing bad in his job. The importance of hyperbole for irony and sarcasm detection has been already underlined by Kunneman et al. [2015]. We found this same phenomenon in the misclassified tweets (74) and (75) respectively from IroSvA-Spain and IroSvA-Mex:

(74) *@USER @USER Te falta Pisarello, Echenique y nuestro gran concejal de tráfico el*

[2] *Congratulations director it is a very good rate, this is how it should be done.*

[3] *When I grow up, I want to be like the inventors of this offer.*

	Subtask A	**Subtask B**	**Subtask C**	$f1$-avg
Baselines				
LDSE	0.679	**0.661**	0.633	0.658
W2V	**0.682**	0.627	0.603	0.638
Word nGrams	0.670	0.620	0.568	0.619
Majority	0.400	0.400	0.400	0.400
Our approach				
SCoMoDI	0.665	0.557	**0.634**	0.619

Table 5.4 – Results obtained in the IroSvA Competition.

señor Grezzi.[4]

(75) *El pueblo sabio y bueno salio a expresar su voz @USER*[5]

In Example (73) from IroSvA-Cuba, irony is expressed by ellipsis. In Lanham [1996], ellipsis is defined as "omission of a word easily supplied". In this news comment, the author wanted to subtract, on purpose, some words containing information that could complete the meaning of the sentence. This subtraction is possible because of the presence of context that give us some intuition about the real meaning of the message. This same phenomenon is found especially in IroSvA-Mex (76):

(76) *Será que le da clases particulares el @USER*[6]

Unfortunately, the simple syntactic features that we used especially for Subtask C are not sufficient to capture these more complex puns based on semantic incongruity.

Another common linguistic phenomenon found during the error analysis in Mexican and Cuban variants is the use of apostrophe to stimulate the ironic interpretation of the message. In Lanham [1996], the apostrophe is defined as the action of "breaking off a discourse to address some person or personified thing either present or absent", as we can see in Examples (77) and (78) extracted from the misclassified texts in IroSvA-Mex and IroSvA-Cuba:

(77) *Con todo respeto señor presidente, le solicito atentamente que haga una auditoría al @USER cuyos miembros se rayan y donde la mafia de Octavio Paz se ha instalado.*[7]

(78) *Y ahora es que usted se entera que la honestidad pasó de moda?*[8]

[4] *@USER @USER You miss Pisarello, Echenique and our fantastic city councillor of the traffic Mister Grezzi.*

[5] *The wise and good people came out to express their voice @USER*

[6] *It is possible that @USER teaches him private lessons.*

[7] *With all due respect, Mr. President, I kindly ask you to do an audit at the @USER whose members benefit and where the mafia of Octavio Paz has settled.*

[8] *And only now you realize that honesty went out of fashion?*

Moreover, the rhetorical question seems to be one of the most used devices to express irony in all the variants of Spanish. In Lanham [1996], the rhetorical question is defined as the question "which implies an answer but does not give or lead us to expect one". We noticed that although for Subtask B we took into account the presence of question marks, this expedient is not enough to classify correctly irony. Observe the following texts:

(79) *Cuando pedirán perdón Alemania e Italia a los Valencianos , por mandarnos a la Oltra y Grezzi ? URL*[9]

(80) *Disculpa, sabes si para trabajar en el @USER ¿Debo llevar mi curriculum impreso o depilado?*[10]

(81) *Otra interrogante, por qué nadie fuera de Cuba ha denunciado que los cubanos violamos abiertamente los derechos de los productores de esas programaciones?? O será que el paquete ha venido a ser el primer 'embajador' en el restablecimiento de las relaciones??*[11]

In Examples (79) from IroSvA-Spain, (80) from IroSvA-Mex and (81) from IroSvA-Cuba, we can see that rhetorical questions involve also other figures of speech such as apostrophe in (79) and (80), and metaphor in (80) and (81) which fill the messages with various allusions, making its interpretation more difficult. These observations suggest that there are some similar ways on how Spanish speakers prefer to express irony. Moreover, some of these ways have been already explored also in English, French, and Italian ironic tweets in Karoui et al. [2017], such as hyperbole and rhetorical questions. Therefore, it seems that some kinds of puns tend to characterize the ironic expression, independently of the language and the genre of the text.

5.2 Irony and Sarcasm Detection in Italian

Differently from the mentioned shared tasks, the IronITA shared task at EVALITA 2018 proposed a deeper analysis of ironic text asking participants to recognize, firstly, whether a tweet is ironic or not, and, secondly, to discriminate sarcastic tweets from non-sarcastic ones in Italian. Its purpose was to investigate the possibility to approach these two different linguistic phenomena and analyze their characteristics in hateful and general context.

5.2.1 The IronITA 2018: Shared Task

In continuity with the previous shared tasks at EVALITA on irony detection in Italian, we organized the IronITA shared task dedicated to identify the presence of irony and, if

[9] *When Germany and Italy will apologize to Valencians, for sending us Oltra and Grezzi? URL*

[10] *Excuse me, do you know if to work at the @USER Should I bring my curriculum printed or shaved?*

[11] *Another question, why anyone outside of Cuba have not denounced that Cubans openly violated the rights of the producers of these programs? Or has the package become the first 'ambassador' in the restoration of relations ??*

the text is ironic, also specifically of sarcasm in tweets. The attention on sarcastic form of irony is the main novelty of IronITA. Already Van Hee et al. [2018b] in SemEVal2018-Task3 proposed as additional task to identify the type of irony individuated in the first subtask, distinguishing among *verbal irony by means of a polarity contrast*, *other verbal irony* (when no polarity contrast between the literal and the intended meaning is expressed) and *situational irony* (situations that fail to meet some expectations)[12]. Sarcasm, intended as a specific type of irony, is less studied at computational level. As Attardo [2007], Bowes and Katz [2011] and others noticed, sarcasm is more offensive than other forms of irony, with the intent to convey scorn or mock a clear victim without excluding the possibility of having fun. For its peculiarity, sarcasm appears adequate to express hurtful opinions. To understand this possible role of irony and sarcasm, a part of the provided dataset, is extracted from a corpus of hate speech against minorities such as the Roma community, immigrants, and Muslims.

The IronITA shared task consisted in automatically classifying messages from Twitter for irony and sarcasm. It was organized in a main task (task A) centered on irony, and a second task (task B) centered on sarcasm, whose results were separately evaluated:

- **Task A - Irony Detection:** binary classification where systems have to predict whether a tweet is ironic (`iro`) or not (`non-iro`);
- **Task B - Different types of irony with special focus on sarcasm identification:** since sarcasm is defined as a specific type of irony, this task consists in a multi-class classification where systems have to predict one out of the three following labels: **i) sarcasm** (`sarc`), **ii) irony not categorized as sarcasm** (`iro non-sarc`) such as other kinds of verbal irony or descriptions of situational irony which do not show the characteristics of sarcasm, and **iii) non-irony** (`non-iro`).

The participants are allowed to submit up to 4 predictions to both the tasks or only to the task A, and they could be obtained with 'constrained' or 'unconstrained' runs (or both). The constrained runs are mandatory and have to be produced by systems whose only training data is the dataset provided by the task organizers. On the other hand, the participant teams are encouraged to train their systems on additional annotated data and submit the resulting unconstrained runs.

5.2.1.1 IronITA2018 Dataset

Taking into account the properties of sarcasm, the tweets could be classified as sarcastic (`sarc` = 1) only if irony is present (`iro` = 1). We can see some example of this cascade annotation in Table 5.5.

IronITA2018 is a collection of tweets coming from different sources: HSC and TWITTIRÒ [Cignarella et al., 2018a], composed of tweets from LaBuonaScuola (TW-BS) [Stranisci

[12] To explain this type of irony, we can imagine a situation where firefighters who have a fire in their kitchen while they are out to answer a fire alarm [Shelley, 2001].

iro	sarc	text
0	0	*Le critiche al governo monti da parte di chi ci ha portato sull'orlo del fallimento sono intollerabili.* → The criticisms towards Monti's government by those who have brought us to the verge of bankruptcy are just intolerable.
1	0	*@USER le risorse della scuola pubblica alle private... Questa è la buona scuola!* → @USER resources of public schools to private ones... This is the good school!
1	1	*Gli islamici non sopportano manco Peppa Pig. Manco io,ma per altri motivi. Gli islamici, francamente, avrebbero anche rotto i coglioni* → Muslims can't even stand Peppa Pig. Me neither, but for other reasons. Muslims, frankly, would also have pissed off

Table 5.5 – Examples Extracted from IronITA2018.

et al., 2016], SENTIPOLC2016, Spinoza (TW-SPINO) [Barbieri et al., 2016]. Only in the test set, some tweets have been also included from the TWITA collection [Barbieri et al., 2016]. The distribution of tweets according to the various source datasets is shown in Table 5.6.

	training set				test set				
	iro	non-iro	sarc	iro non-sarc	iro	non-iro	sarc	iro non-sarc	total
TW-BS	467	646	173	294	111	161	51	60	3,109
TW-SPINO	342	0	126	216	73	0	32	41	
SENTIPOLC2016	461	625	143	318	0	0	0	0	
TWITA	0	0	0	0	67	156	28	39	
HSC	753	683	471	282	184	120	105	79	1,740
TOTAL	**3,977**				**872**				**4,849**

Table 5.6 – Distribution of the Annotated Tweets According to the Source.

As described in Cignarella et al. [2018b], the annotation was organized in two steps. Firstly, the dataset was split in two halves and two couples of Italian native speakers (working in figurative language) annotated sarcasm in each half. Secondly, to solve the disagreement, the couple previously involved in the annotation of the first half of the dataset produced a new annotation for the tweets in disagreement of the second portion of the dataset and vice versa. Then, the cases where the disagreement persisted (131 tweets) have been discarded as too ambiguous to be classified. The final Inter-Annotator Agreement calculated with Fleiss' kappa is $\kappa = 0.56$ for the tweets belonging to the TWITTIRÒ corpus and $\kappa = 0.52$ for the data from the HSC corpus, and it is considered moderate[13] and satisfying for the purpose of the shared task.

[13] According to the parameters proposed by Fleiss [1971].

In this process, the annotators relied on a specific definition of 'sarcasm', and followed detailed guidelines[14]. In particular, we defined **sarcasm** as *a kind of sharp, explicit and sometimes aggressive irony, aimed at hitting a specific target to hurt or criticize without excluding the possibility of having fun* [Du Marsais et al., 1981, Gibbs, 2000]. The factors we have taken into account for the annotation are, the presence of:

1) a clear **target**,

2) an obvious **intention** to hurt or criticize,

3) **negativity** (weak or strong).

A single training set has been provided for both tasks A and B, which includes 3,977 tweets. Following, a single test set has been distributed for both tasks A and B, which includes 872 tweets, hence creating a $82\% - 18\%$ balance between training and test data. Moreover, IronITA2018 overlaps with HaSpeeDe2018 [Bosco et al., 2018a]. In the training set we count 781 overlapping tweets, while in the test set we count an overlap of just 96 tweets.

The data were released in the following format[15]:

```
idtwitter text irony sarcasm topic
```

where `idtwitter` is the Twitter ID of the message, `text` is the content of the message, `irony` is 1 or 0 (respectively for ironic and not ironic tweets), `sarcasm` is 1 or 0 (respectively for sarcastic and not sarcastic tweets), and `topic` refers to the source corpus from where the tweet has been extracted.

5.2.1.2 Approaches and Results

The shared task started on 30th May 2018 with the release of the development data and finished on 23rd September of the same year with the evaluation of the participating systems on the test set. They have been evaluated according to specific metrics for each task. In particular, we used $f1$-score calculated respectively for binary and multi-class evaluation. For the task A, the systems have been evaluated against the gold standard test set on their assignment of a 0 or 1 value to the `irony` field. We measured the precision, recall and $f1$-score of the prediction for both the classes:

$$precision_{class} = \frac{\#correct\_class}{\#assigned\_class}$$

$$recall_{class} = \frac{\#correct\_class}{\#total\_class}$$

[14] For more details on this regard, please refer to the guidelines: https://github.com/AleT-Cig/IronITA-2018/blob/master/Definition%20of%20Sarcasm.pdf

[15] Link to the datasets: http://www.di.unito.it/~tutreeb/ironita-evalita18/data.html

$$f1_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}}$$

The overall score is the average of the $f1$-scores for the `iro` and `non-iro` classes (i.e., $f1$-macro).

For the task B, the systems have been evaluated against the gold standard test set on their assignment of a 0 or 1 value to the `sarcasm` field, assuming that the value for `irony` has been provided. We have measured the precision, recall and $f1$-score for each of the three classes:

- non-ironic
  `irony` = 0, `sarcasm` = 0
- ironic-non-sarcastic
  `irony` = 1, `sarcasm` = 0
- sarcastic
  `irony` = 1, `sarcasm` = 1

The evaluation metric is the $f1$-macro computed over the three classes. Note that for the purpose of the evaluation of the task B, the following combination is always considered wrong:

- `irony` = 0, `sarcasm` = 1

Our scheme imposes that a tweet can be annotated as sarcastic only if it is also annotated as ironic, which corresponds to interpreting sarcasm as a specific type of irony, as reported in the examples in Table 5.5.

For the ranking, we implemented two straightforward baseline systems:

- *Baseline_MFC* (Most Frequent Class) assigns to each instance the majority class of the respective task, namely `non-iro` for the task A and `non-sarc` for the task B.
- *Baseline_Random* assigns uniformly random values to the instances. Note that for the task A, a class is assigned randomly to every instance, while for the task B, the classes are assigned randomly only to eligible tweets which are marked as ironic.

A total amount of 7 teams participated in the task A, and in particular we received 17 runs for the task A and 7 runs (from 4 teams) for the task B. A short description of the three best scored systems in both tasks is provided in Table 5.7. A complete overview about the participating systems is provided in Cignarella et al. [2018b].

team		description
ItaliaNLP	[Cimino et al., 2018]	The ItaliaNLP team developed a multi-task learning approach based on bi-LSTM networks exploiting the correlation between irony and sarcasm (ItaliaNLP-MTL)[16], and among various related sentiment analysis tasks, specifically between irony/sarcasm and polarity identification in run 1, and among irony/sarcasm, polarity and hate speech detection in run 2. To this purpose, they used additional tweets from SENTIPOLC2016 and HaSpeeDe2018, in addition to automatically generated and translated sentiment polarity lexica, semantic (word embeddings) and morpho-syntactic features.
UNIBA	[Basile and Semeraro, 2018]	The UNIBA team employed an SVM classifier taking advantage of sentiment information [Basile and Novielli, 2014], unigrams, bigrams, trigrams, microblogging features and word embedding vectors from TWITA as semantic representation of tweets and to capture the usage of words in Twitter context.
X2Check	[Di Rosa and Durante, 2018]	Principally exploiting n-grams word representation, X2Check built a system based on a Multinomial Naïve Bayes classifier trained on additional tweets annotated as ironic from SENTIPOLC2016.
UNITOR	[Santilli et al., 2018]	The UNITOR team created a cascade of kernel-based SVM classifiers: the first classifier discriminated between *ironic* and *non-ironic* tweets, while the second one distinguished *sarcastic* and *non-sarcastic* tweets. To generalize lexical information of training texts, they created a word embedding using about 10 millions of tweets downloaded in July 2016, and, on the basis of this word space representation, they computed the cosine similarity between words and sentences to capture the unconventional use of a word and PoS tag, in addition to the respective mean and variance value for tweet. Finally, they used various sizes of character n-grams, synthetic features (number of punctuation, symbols, uppercase letters and so on), sentiment information for words and PoS tags extracted by a distributional polarity lexicon built in [Castellucci et al., 2016]. Only for the unconstrained run that reaches the first rank in the task B classification, the team built a specific ironic dataset collecting 6,000 tweets assuming to be ironic on specific hashtags (#irony or #ironia) to get, also, specific words or patterns of ironic texts.
Aspie96	[Giudice, 2018]	Aspie96 used a Gated Recurrent Units exploiting the advantages of character level representation.

Table 5.7 – Best Performing Systems at IronITA 2018.

Tables 5.8 and 5.9[17] report the results of the three best performing systems for each task along with the baselines scores and the model ItaliaNLP-MTL not officially ranked but tested on the released test set by Cimino et al. [2018]. As we can notice, no matter the challenging task and the lower amount of linguistic resources available for the Italian language, the systems obtained high results in the task A. The complete ranking for both tasks is published in Cignarella et al. [2018b].

Looking at Table 5.8, the official first ranked system reported the trend to identify

[16] This run was not submitted by the team during the competition, but they reported its performance on the test set of IronITA2018 in Cimino et al. [2018].

[17] Unconstrained runs are in gray background.

team	run	rank	*f*1-score		
			`non-iro`	`iro`	*macro*
ItaliaNLP-MTL	–	–	–	–	***0.736***
ItaliaNLP	1	1	0.707	**0.754**	**0.731**
UNIBA	1	3	0.689	0.730	0.710
X2Check	1	5	**0.708**	0.700	0.704
Baseline_Random	–	–	*0.503*	*0.506*	*0.505*
Baseline_MFC	–	–	*0.668*	*0.000*	*0.334*

Table 5.8 – The Best Results for the Task A at IronITA 2018.

team	run	rank	*f*1-score			
			`non-iro`	`iro`	`sarc`	*macro*
ItaliaNLP-MTL	–	–	–	–	–	***0.530***
UNITOR	2	1	0.668	**0.447**	**0.446**	**0.520**
ItaliaNLP	1	3	**0.707**	0.432	0.409	0.516
Aspie96	1	5	0.668	0.438	0.289	0.465
Baseline_Random	–	–	*0.503*	*0.266*	*0.242*	*0.337*
Baseline_MFC	–	–	*0.668*	*0.000*	*0.000*	*0.223*

Table 5.9 – The Best Results for the Task B at IronITA 2018.

correctly ironic messages more than non-ironic ones, and obtained a macro *f*1-score of 0.731, revealing a performance in line with the results in SemEval2018-Task3 (0.705) about irony detection in English tweets [Van Hee et al., 2018a]. About the task B, we can notice lower *f*1-scores in Table 5.9 due probably to the difficulty to distinguish sarcasm from other types of irony, even in a multi-task learning context, and to the scarce amount of sarcastic data respect to the rest (see Table 5.6).

This difficulty of detecting sarcasm encouraged us to carry out an error analysis in order to understand whether the systems did not detect sarcastic tweets because they confused sarcasm with other types of irony, or because of its peculiar characteristics. To these purposes, we exploited the multi-source composition of IronITA2018 to recover the original labels for each instance and perform a deeper qualitative and quantitative analysis, looking at the dimensions of hate from HSC and the rhetorical and pragmatic elements from TWITTIRÒ.

5.2.1.3 Error Analysis

At the beginning of this section, we describe the extension of annotation performed on IronITA2018 retrieving the original labels of the HSC and TWITTIRÒ corpora, and extending the annotation for the instances that missed some labels.

HSC annotation, as described in Sanguinetti et al. [2018c], consists of various labels referring to *dimensions of hate*, such as aggressiveness (`agg`), offensiveness (`off`)[18], stereotype (`stereo`) and hate speech (`hs`);

TWITTIRÒ schema has three levels of annotation, as described in Cignarella et al. [2018a]. In particular, we applied two levels of annotations related to *linguistic characteristics*:

1 **Contradiction Type**[19]: If the tweet is ironic, one can individuate the type of contradiction that activates irony [Giora et al., 2015b]. Actually, irony is often expressed through a contradiction that could occur between two lexicalized clues (such as opposite terms or propositions) within the sentence (`explicit`), or between an internal lexicalized cue and an external pragmatic context echoed in the sentence (`implicit`). For example:

(82) *Vedo che c'è molta disinformazione sul referendum del 17 maggio. [@USER]*[20] (`iro` and `non-sarc`)

(83) *Trovato l'ispiratore delle ricette del governo Monti: Bisogna prendere il denaro dove si trova. Presso i poveri.... URL*[21] (`iro` and `sarc`)

2 **Linguistic Categories**: If the tweet is ironic and a type of contradiction has been individuated, the final level of annotation specifies the linguistic elements creating the contradiction, and, therefore, the ironic expression. The figures of speech and pragmatic clues relative to implicit and explicit contradiction are listed in Table 5.10.

The preexisting annotations of the source corpora HSC and TWITTIRÒ in the tweets of IronITA2018 covered only the data of the training set. In order to perform the analysis on the whole IronITA dataset, we applied these two fine-grained annotations, also, to the tweets of the test set, following the respective guidelines[22]. The annotation process involved four Italian native speakers working in irony and hate speech. In accordance with the source of the tweets, the test set was split in two halves and annotated by different couple of annotators. To solve the disagreement between annotators on each half, the couple previously involved in the annotation of dimensions of hate produced a new annotation according to the TWITTIRÒ schema, while the other couple did the

[18] Although the original annotation established a range of strength (no, weak and strong) for aggressiveness and offensiveness, in our work we took into account only the presence of these phenomena.

[19] In accordance with the TWITTIRÒ schema of annotation, the labels of levels 2 and 3 are applied only to ironic tweets (see Table 5.12).

[20] The referendum was indeed on April 17th, 2016: "*I see there's a lot of disinformation on the referendum of May 17th. [@USER]*"

[21] *Found the inspirer of Monti's government's recipes: One must take the money where it lies. From poor people.... URL*

[22] For the tweets coming from HSC, the schema of annotation applied is available at `http://di.unito.it/hsc.`; and for the data coming from the other sources related to political or more general topics (TW-BS, TW-SPINO, SENTIPOLC2016 and TWITA) the schema of annotation of TWITTIRÒ applied is available at `http://di.unito.it/twittiro`.

categories	**label**	**definition**
Analogy$^{\text{Both}}$	`an`	Analogy covers figures of speech, such as metaphor, analogy, simile and similarity, used to compare different ontological concepts or domains.
Hyperbole$^{\text{Both}}$	`hyp`	Hyperbole is used to emphasize or exaggerate something.
Euphemism$^{\text{Both}}$	`euph`	Euphemism allows reducing the duress of an idea or a fact to soften the reality.
Rhetorical Question$^{\text{Both}}$	`r_q`	Rhetorical question is used to make a point about an issue rather than to elicit an answer.
Context Shift$^{\text{Expl}}$	`c_s`	Context shift involves a sudden change of topic or frame, such as the use of exaggerated politeness in an inappropriate situation.
False Assertion$^{\text{Impl}}$	`f_a`	False assertion assumes the assertion of a unreal fact or declaration.
Oxymoron/Paradox$^{\text{Expl}}$	`o/p`	Oxymoron and Paradox concern an explicit lexical (antonyms) and pragmatic contradiction.
Other$^{\text{Both}}$	`other`	"Other" category covers humor and situational irony, where the contradiction involves events and not the use of words.

Table 5.10 – Linguistic Categories in TWITTIRÒ.

same with the HSC annotation. At the end of the process of extension, the tweets of IronITA2018 are labeled with IronITA, HSC and TWITTIRÒ schema of annotation, as shown in Tables 5.11 and 5.12.

Exploiting the extension of annotation of IronITA2018, we tried to reveal the difficulties of existing approaches on irony and sarcasm detection in Italian tweets carrying out a deep error analysis. In particular, we studied the set of the common predictions (correct and incorrect) of the three best runs for each task, applying two main types of analyses. Firstly, a qualitative analysis on the common misclassified ironic and sarcastic tweets; secondly, we deepened the qualitative observations with a quantitative analysis exploiting the multi-label annotation of IronITA2018, and the morphosyntactic information extracted by PoS-tagging and parsing the misclassified ironic/sarcastic tweets with the *UDPipe* pipeline [Straka and Straková, 2017].

Since the differences between runs of the same systems are not significant, we considered the predictions of the best run submitted by the teams that obtained the best scores. This choice allowed us to take into account the predictions that were obtained with different approaches (see Table 5.7). Therefore, we considered:

- for the task A: the first runs of the teams ItaliaNLP, UNIBA and X2Check (un-

iro	sarc	hs	agg	off	stereo	text
0	0	yes	yes	no	yes	*@USER tutto tempo danaro e sacrificio umano sprecato senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile* → @USER all the time money and human sacrifice wasted without purge of talibans and muslim radicals it's all useless
1	0	no	yes	yes	yes	*Gentili proprietari dei resort alle #maldive... accogliete il profugo dall'Italia per dieci giorni. #profughi #esiamonoi #notengodinero* → Respectable owners of the resorts at the #maldives... welcome the refugee from Italy for ten days. #refugees #andit'sus #notengodinero
1	1	yes	yes	no	no	*Dai ragazzi, è Natale! Portiamo un po' di calore al campo nomadi. Io penso alla benzina, voi portate i fiammiferi?* → Come on guys, it's Christmas! Let's bring some warmth to the nomads camp. I'll take care of the gasoline, you'll bring the matches?

Table 5.11 – Examples from HSC Source in IronITA2018.

constrained)

- for the task B: the second run of the team UNITOR (unconstrained) and the first runs of the teams ItaliaNLP and Aspie96.

The majority of them used the same system to detect irony and sarcasm, except UNITOR that employed a cascade architecture of classifiers that selected automatically the most distinctive information for each task among a consistent set of features.

Collecting the predictions of the best performing systems in the IronITA shared task, we selected the set of *hard cases* (HC henceforth) composed of the common misclassified tweets, and the set of *simple cases* (SC henceforth) composed of the common tweets correctly classified. Tables 5.13 and 5.14 show the sizes of HC and SC sets for each task and their percentage calculated on the total of tweets in the test set for each class. Considering our interest in the comprehension of the role played by affective aspects, such as hate, in irony and sarcasm, in Tables 5.13 and 5.14 we divided the sets of tweets in two principal domains: HSC and NO-HSC. The latter collects tweets coming from TW-BS,

iro	sarc	level 2	level 3	text
0	0	0	0	*Come fare in modo che gli studenti sperimentino l'entusiasmo della scoperta scientifica? #AmgenTeach URL #labuonascuola* → How to make students experiment the enthusiasm of scientific discovery? #AmgenTeach URL #labuonascuola
1	0	explicit	an	*Crolla la borsa di Shanghai. Ora bisogna risollevarla senza muovere le altre. [@USER]* → Shanghai's stock market crashes. Now we should raise it again, but without moving the others. [@USER]
1	1	implicit	im:f_a	*E comunque @USER alla lezione di sillabazione de #labuonascuola era assente URL* → Anyway @USER was absent at the lesson of the #labuonascuola on hyphenation URL

Table 5.12 – Examples from TWITTIRÒ Source in IRONITA2018.

TW-SPINO, SENTIPOLC2016 and TWITA and covering general issues not necessarily related to abusive context.

	HARD CASES		SIMPLE CASES	
	iro	non-iro	iro	non-iro
NOHSC	18	39	125	153
HSC	10	23	112	48
TOTAL CLASS	28 (6%)	62 (14%)	237 (54%)	201 (46%)
TOTAL CASES	90		438	

Table 5.13 – Hard and Simple Cases in the Task A.

Comparing the distribution of HC and SC in the tasks A and B, we can notice that: ironic tweets are in general correctly identified, whereas sarcastic ones result more difficult to detect; and, looking at the difference between the sets of HSC and NO-HSC in Table 5.14, sarcastic tweets tend to be identified correctly in hateful contexts. Moreover, to measure the impact of the low inter-annotator agreement in the results obtained in the competition in the task B, we observed if the common misclassified tweets by the three best systems in the competition (88 HC in Table 5.14) caused also disagreement during the annotation. Among these 88 HC, only 4 tweets were considered hard to interpret even by the annotators. However, during the second phase of the annotation, the disagreement was solved. Considering this low percentage (4.5% of HC), we can state that the low inter-annotator agreement did not affect the results in the competition.

	Hard Cases			Simple Cases		
	`sarc`	`iro non-sarc`	`non-iro`	`sarc`	`iro non-sarc`	`non-iro`
NOHSC	66	0	1	0	91	258
HSC	16	4	1	19	31	83
TOTAL CLASS	82 (38%)	4 (2%)	2 (0.5%)	19 (9%)	122 (56%)	341 (78%)
TOTAL CASES	88			482		

Table 5.14 – Hard and Simple Cases in the Task B.

Looking especially at HC, we performed two main error analyses: qualitative and quantitative analysis.

Qualitative Analysis Our first step is to examine qualitatively HC carrying out a manual error analysis with the purpose to find stylistic, syntactic and semantic markers that made irony and, especially, sarcasm difficult to identify. Secondly, we deepened these findings with a quantitative analysis. It is important to underline that our attention in the task B is focused on understanding if unidentified sarcasm is confused with other types of irony, or is not recognized for its peculiarities. Considering that, our analysis in the task B will concern only *sarcastic* and *ironic non-sarcastic* tweets.

Stylistic Markers refer to those patterns related to the writing style in a social media like Twitter, such as discursive and informal elements. In particular, in ironic/sarcastic HC we noticed a great number of quotation marks, ellipsis, and intensifiers ('sempre più', '150k', 'solo'). Especially sarcastic HC contain also negation markers ('non', 'nemmeno', 'né'), interjections ('boh', 'GRAZIE', 'ah beh'), and informal language (such as swear words, dialectal and colloquial expressions).

Syntactic Markers involve phrase types and syntactic coarse-grained classes. In particular, in ironic/sarcastic HC, we noticed a high frequency of: noun phrases that work sometimes as slogan ('Stop profughi', 'città sotto assedio', 'buona scuola o buona propaganda') [Comandini and Patti, 2019b]; adverbial locutions ('altro che', 'bene', 'di certo') and, especially, discourse connectors with function adversative ('invece', 'ma'), causal ('perché') or sequential ('prima', 'ora').

Semantic Markers cover elements that could be caught analyzing the meaning of the message. Ironic/sarcastic HC tend to have a surprise effect caused by a contrast between phrases or sentences within the message (84) or by an unexpected answer or solution (85):

(84) *@USER "ti aggiorneremo sull'avvio della consultazione" Sto ancora aspettando #labuonascuola*[23]

(85) *@USER Anche noi abbiamo la nostra via x i rom: quella dei forni della Italsider.*[24]

[23] *"@USER "we will let you know regarding the start of consultation" I'm still waiting #labuonascuola"*

[24] *@USER We too have our own way for romas: the ovens of Italsider.*

Another common semantic element is the assertion of false events, such as:

(86) *Wojtyla era pronto alle dimissioni. Ma non riusciva a firmarle. [fedgross]*[25]

Sarcastic HC, moreover, involve echoic mentions (87) and context shifting (88):

(87) *La moglie di Bobo Craxi scippata ad Hammamet. In un commosso ricordo del suocero. [fdecollibus]*[26]

(88) *Frattini pubblica sul sito del ministero le foto delle sue vacanze. La mia preferita è quella dove sta alla scrivania. [stenit]*[27]

All these elements are far from the textual markers and require an extended knowledge of the language, as well as of the world, to be captured. This makes irony and sarcasm detection a real challenging task.

Quantitative Analysis At a deeper level, we carried out a more quantitative analysis aimed at identifying specific elements of irony and sarcasm that could make their detection hard. Firstly, we focus on stylistic and syntactic markers, examining morphosyntactic information extracted by PoS-tagging and parsing the misclassified ironic and sarcastic tweets. Secondly, we exploit the multi-label annotation of IronITA2018 to analyze, at a semantic level, the impact of the dimensions of hate, inherited from HSC, as well as of rhetorical and pragmatic elements from TWITTIRÒ on irony and sarcasm detection.

Morphosyntactic Analysis We conducted an error analysis investigating the morphosyntactic characteristic of the language used in misclassified tweets, taking advantage of the fact that a portion of IronITA2018 has been annotated accordingly to the format of *Universal Dependencies*[28] (henceforth UD) [Cignarella et al., 2019]. By training the *UDPipe* pipeline on other available Italian treebanks ISDT [Simi et al., 2014], PoSTWITA [Sanguinetti et al., 2018a], and TWITTIRÒ-UD [Cignarella et al., 2019] we easily tokenized, lemmatized, PoS-tagged and parsed the remaining tweets that were not released as part of a gold standard in the official UD repository[29] obtaining a full morphosyntactic annotation for the test set of IronITA2018. We proceeded in two steps: firstly we observed the distribution of PoS tags in the entire test set and compared it with the PoS tags distribution in HC of both tasks, and later we focused only on *ironic* tweets that were wrongly classified as *non-ironic* (28 tweets for the task A) and on *sarcastic* tweets that were wrongly classified as *ironic non-sarcastic* (82 tweets for the task B) (see Table

[25] *Wojtila was ready to write his resignation. But he wasn't able to sign it. [fedgross]*

[26] *The wife of Bobo Craxi mugged in Hammamet. In a moved memory of her father-in-law. [fdecollibus]*

[27] *Frattini posts photos of his vacations on the ministry website. My favorite one is that where he's behind his work-desk. [stenit]*

[28] https://universaldependencies.org/.

[29] http://di.unito.it/uditaliantwittiro.

	Test Set	All HC				Ironic or Sarcastic HC			
PoS tags	**(782 tweets)**	**HC Task A (90 tweet)**	**freq (%)**	**HC Task B (88 tweet)**	**freq (%)**	**HC Task A (28 tweets)**	**freq (%)**	**HC Task B (82 tweets)**	**frequ (%)**
ADJ	816	73	8.95	86	10.54	26	3.19	82	10.05
ADP	1,964	207	10.54	218	11.10	52	2.65	197	10.03
ADV	870	103	11.84	91	10.46	15	1.72	81	9.31
AUX	579	79	13.64	59	10.19	16	2.76	56	9.67
CCONJ	338	41	12.13	34	10.06	6	1.78	28	8.28
DET	1,999	203	10.16	237	11.86	52	2.60	213	10.66
INTJ	100	7	7.00	15	15.00	2	2.00	14	14.00
NOUN	2,583	288	11.15	275	10.65	80	**3.10**	249	9.64
NUM	172	18	10.47	18	10.47	3	1.74	18	10.47
PRON	900	111	12.33	94	10.44	26	2.89	84	9.33
PROPN	879	56	6.37	92	10.47	17	1.93	81	9.22
PUNCT	2,247	186	8.28	272	12.11	47	2.09	208	9.26
SCONJ	200	19	9.50	22	11.00	2	1.00	17	8.50
SYM	1,557	157	10.08	144	9.25	68	4.37	134	8.61
VERB	1,572	185	11.77	166	10.56	48	3.05	148	9.41
X	168	10	**5.95**	12	**7.14**	4	2.38	12	**7.14**
Total	16,944	1,743	10.29	1,835	10.83	464	2.74	1,622	9.57

Table 5.15 – Distribution of PoS Tags in HC.

5.15). In a following step, we applied the same procedure accordingly to the distribution of dependency relations (see Table 5.16).

Observing Table 5.15, we are able to see how PoS tags are distributed across the test set and examine whether the PoS tags in HC report any significant difference in their distribution. For instance, the high number of NOUN PoS tag (3.10% [in red]) in ironic HC suggests that these tweets could contain noun phrases or slogans with ironic meaning not recognized by the systems. On the other hand, it seems that the presence of the SYM PoS tag (8.61% [in green]) and of the X PoS tag (5.95% and 7.14% [in magenta]) is lower especially in sarcastic HC, suggesting that the tokens with these PoS tags (e.g., foreign words, emojis, hashtags, mentions and URLs) might be good indicators for the detection of sarcasm. Moreover, we can notice a high frequency of DET PoS tag (10.66% [in orange]) in sarcastic HC. Accordingly to the UD tagset[30], DET PoS tag includes quantifiers and various determiners (indefinite, exclamatory, demonstrative and so on). All these elements could be used as intensifiers. Another interesting value is the frequency of INTJ PoS tag (14.00% [in cyan]), that as seen before seems to play an important role in sarcasm detection.

[30] https://universaldependencies.org/u/pos/

	Test Set	All HC				Ironic or Sarcastic HC			
Deprels	**(782 tweets)**	**HC Task A (90 tweet)**	**freq (%)**	**HC Task B (88 tweet)**	**freq (%)**	**HC Task A (28 tweets)**	**freq (%)**	**HC Task B (82 tweets)**	**freq (%)**
`acl`	128	10	7.81	11	8.59	3	2.34	10	7.81
`acl:relcl`	149	23	**15.44**	14	9.40	5	3.36	13	8.72
`advcl`	191	24	12.57	16	8.38	4	2.09	13	6.81
`advmod`	842	96	11.40	87	10.33	14	1.66	77	9.14
`amod`	682	61	8.94	72	10.56	25	3.67	69	10.12
`appos`	55	2	3.64	1	1.82	–	–	1	1.82
`aux`	293	45	15.36	24	8.19	8	2.73	23	7.85
`aux:pass`	42	6	**14.29**	7	16.67	1	2.38	7	16.67
`case`	1,760	188	10.68	202	11.48	49	2.78	184	10.45
`cc`	338	39	11.54	34	10.06	6	1.78	28	8.28
`ccomp`	114	13	11.40	15	13.16	4	3.51	9	7.89
`compound`	54	5	9.26	2	3.70	–	–	1	1.85
`conj`	391	37	9.46	36	9.21	6	1.53	32	8.18
`cop`	244	28	11.48	28	11.48	7	2.87	26	10.66
`csubj`	19	2	10.53	1	5.26	–	–	1	5.26
`dep`	473	34	7.19	19	4.02	21	4.44	18	3.81
`det`	1,901	194	10.21	224	11.78	51	2.68	201	10.57
`det:poss`	73	6	8.22	12	16.44	1	1.37	11	15.07
`det:predet`	22	4	18.18	2	9.09	–	–	1	4.55
`discourse`	97	9	9.28	14	14.43	2	2.06	13	13.40
`discourse:emo`	48	8	16.67	8	16.67	2	4.17	9	18.75
`dislocated`	2	–	–	–	–	–	–	–	–
`expl`	161	11	6.83	23	14.29	3	1.86	18	11.18
`expl:impers`	17	7	**41.18**	1	5.88	1	5.88	1	5.88
`expl:pass`	6	1	**16.67**	–	–	–	–	–	–
`fixed`	38	3	7.89	2	5.26	–	–	2	5.26
`flat`	18	2	11.11	2	11.11	–	–	2	11.11
`flat:foreign`	40	1	2.50	1	2.50	1	2.50	1	2.50
`flat:name`	157	8	5.10	13	8.28	2	1.27	12	7.64
`iobj`	110	11	10.00	9	8.18	2	1.82	10	9.09
`mark`	398	38	9.55	38	9.55	5	1.26	30	7.54
`nmod`	1,081	99	9.16	102	9.44	34	3.15	96	8.88
`nsubj`	791	91	11.50	87	11.00	22	2.78	79	9.99
`nsubj:pass`	48	3	6.25	4	8.33	–	–	4	8.33
`nummod`	146	16	10.96	14	9.59	3	2.05	14	9.59
`obj`	791	105	13.27	91	11.50	30	3.79	81	10.24
`obl`	749	93	12.42	100	13.35	23	3.07	87	11.62
`obl:agent`	19	–	–	1	5.26	–	–	1	5.26
`parataxis`	435	49	11.26	74	17.01	17	3.91	66	15.17
`parataxis:appos`	1	–	–	–	–	–	–	–	–
`parataxis:hashtag`	228	26	11.40	22	9.65	15	6.58	20	8.77
`punct`	**2,245**	186	8.29	271	12.07	47	2.09	208	9.27
`root`	872	90	10.32	88	10.09	28	3.21	82	9.40
`vocative`	17	2	11.76	1	5.88	–	–	–	–
`vocative:mention`	487	51	10.47	52	10.68	16	3.29	49	10.06
`xcomp`	171	16	9.36	10	5.85	6	3.51	12	7.02
Total	16,944	1,743	10.29	1,835	10.83	464	2.74	1,622	9.57

Table 5.16 – Distribution of Dependency Relations in HC.

In the same way, we then calculated the distribution of *dependency relations* (deprels). In Table 5.16 we illustrate a list of all the dependency relations and their frequency in the three different subsets[31]. Considering the style of the user-generated contents, it is not surprising to see that the most frequent deprel is `punct` (used 2,245 times [**in bold**], being 13.25% of the total) [Bazzanella, 2011, Sanguinetti et al., 2017].

For what concerns the distribution of other deprels in the subset of misclassified tweets of the task A, we noticed a distribution that deviates from the standard of the follow-

[31] With the hyphen '–' we indicate that a dependency relation is not present in a subset. As reference of the UD deprels see `https://universaldependencies.org/u/dep/`.

ing relations: `acl:relcl` (relative clauses), `aux:pass` (auxiliary verbs in a passive voice construction), `expl:impers` and `expl:pass` (expletive particles), indicating that tweets with these syntactic features tend to be misclassified [in blue]. On the other hand, tweets containing the following deprels, seem to be correctly classified the majority of the times: `appos` (appositional modifiers), `flat:foreign` (foreign words) and `flat:name` (multi-word expressions) [in green]. The deprel `discourse:emo` seems to have an unbalanced distribution in the task B, suggesting that it might be creating noise and making more difficult the detection of sarcasm (18.75%, $\Delta = 9.18$ deviation from the average distribution) [in red]. Moreover, the `parataxis` dependency relation has a greater distribution in the misclassified tweets of the task B in both scenarios (all HC: 17.01%, and sarcastic HC: 15.17%), deviating $\Delta = 6.18$ in the first case and $\Delta = 5.6$ in the second [in orange], but presents an average distribution in the two scenarios of the task A. Similarly, the deprel `parataxis:hashtag` presents a $\Delta = 3.84$ with regard to the average distribution in the misclassified tweets of the task A, in the scenario where we look at all the misclassified tweets (6.58%), but then its distribution is around average values in all the other cases [in magenta]. Finally, `xcomp` seems to be less present in the misclassified tweets of the task A (5.85%) [in cyan], presenting a deviation of $\Delta = 4.98$.

Semantic and Pragmatic Analysis To enrich the qualitative semantic markers identified before, we examined the percentages of FP and FN, and, equally, TP and TN in presence of the dimensions of hate and linguistic characteristics. The percentages are calculated considering the absolute frequency of each dimension of hate/linguistic characteristic in HC and SC and its distribution in the test set. Taking into account the low values of HC and SC in both tasks, below we report the most relevant observations.

To analyze the impact of hurtful language, we considered the presence of hate speech, aggressiveness, offensiveness, and stereotype in *ironic/non-ironic* and *sarcastic/ironic non-sarcastic* tweets (as shown in Table 5.17).

In the task A, high percentages of TPs in presence of hate speech (70.27%), aggressiveness (62.71%), offensiveness (65.57%) and stereotypes (61.04%) and of TNs in non-hateful contexts (respectively 44.90%, 45.84%, 44.45% and 46.43%) suggest that systems tend to correctly classify tweets as ironic when the texts contain a more hurtful language. Indeed, observing the highest values of FN cases in both tasks (7.32% in the task A and 19.40% in the task B), we can hypothesize that the lack of offenses could conduct to predict ironic/sarcastic tweet as non-ironic/non-sarcastic, but, conversely, the presence of derogatory speech could increase the FPs, as shown in the task A (29.17% and 22.27%) and in the task B (9.09% and 8.70%). Therefore, it appears necessary to balance the information about hateful language given to the system. In NO-HSC, the highest percentages of false predictions are related to FP cases (12.30%). Analyzing these tweets that the systems tend to predict as ironic, we noticed that are principally characterized by negative emotions, such as rage or frustration. It is clear that negative emotions and a more hurtful language have an impact on the detection of irony and sarcasm.

				Task A							Task B		
	Test set (304 HSC tweets)						Test set (184 HSC tweets)					
Dimensions of Hate	**IRONIC (184 tweets)**	**NON-IRONIC (120 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**	**SARC (105 tweets)**	**IRO NON-SARC (79 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**
`hs yes`	37	22	27.27	5.40	**70.27**	18.18	26	11	**9.09**	19.23	7.69	36.36
`hs no`	147	98	17.35	5.44	58.50	44.90	79	68	4.41	13.92	**21.52**	39.70
`agg yes`	59	24	**29.17**	6.78	62.71	16.67	44	15	6.67	18.18	15.91	13.34
`agg no`	125	96	16.67	4.80	60.00	45.84	61	64	4.69	13.11	19.67	45.31
`off yes`	61	21	19.05	1.64	65.57	19.04	38	23	8.70	7.89	13.16	26.09
`off no`	123	99	19.19	**7.32**	58.54	44.45	67	56	3.57	**19.40**	20.90	44.64
`stereotype yes`	77	36	19.45	5.19	61.04	25.00	48	29	6.89	10.42	14.58	27.59
`stereotype no`	107	84	19.05	5.61	60.75	**46.43**	57	50	4.00	19.30	21.05	**46.00**

Table 5.17 – Distribution of Dimensions of Hate.

Since the annotation schema of TWITTIRÒ focuses only on the ironic texts, Table 5.18 does not report FP and TN values calculated on the negative class for the task A. Taking into account the percentages of TP, we can delineate some important linguistic markers in ironic texts that could help irony detection: context shift (60.47%), oxymoron (55.77%) and hyperbole (53.85%). Other more subtle linguistic categories, such as euphemism (89) and rhetorical question that could be confused as simple question (90), tend to increase the FN values (respectively 26.09% in explicit contradictions and 11.11% in implicit ones):

(89) *Altro che 'merito', #labuonascuola ha anche profumo di incostituzionalità URL #sapevatelo @USER @USER*[32]

(90) *Si può fare "buona scuola" senza Geografia? | Orizzonte Scuola URL via @USER*[33]

With respect to the task B, since HC are *sarcastic* and in SC are only *ironic non-sarcastic*, Table 5.18 does not report FP and TP percentages computed respectively on the negative and positive classes. Moreover, in the task B, the TNs represent the *ironic non-sarcastic* texts.

We can observe that the percentage of FNs is higher than in the task A, probably for the complexity of the task. Examining the FN cases, sarcasm tends to be predicted as *non-sarcastic irony* especially when it contains rhetorical questions (that make the correct identification difficult also in the task A), hyperbole (more related to irony) and situational irony. The `other` category, normally observed in *ironic non-sarcastic* texts for its references to specific funny situations, as explained in Wang [2013] could involve also sarcastic situations, even if in a more subtle manner than in ironic ones:

(91) *Quando mi dicono: "stai zitta che bevi ancora il latte" io rispondo: "si ma con il cioccolato perché io sono già grande" ahahhahaha*[34] (`iro non-sarc`)

(92) *@USER @USER @USER La buona scuola in cui tutti parleranno solo inglese.Come Renzi.Che pena.*[35] (`sarc`)

[32] *What 'merit'? #labuonascuola also smells as unconstitutional URL #sapevatelo @USER @USER*

[33] *Is it possible to have a "good school" without Geography? | Orizzonte Scuola URL via @USER*

[34] *When they tell me: "shut up since you're still drinking milk" I reply "yes, but with cocoa since I'm already grown up" ahahhahaha*

[35] *@USER @USER @USER The good school in which everyone will speak English.As Renzi.What a shame.*

	Task A						Task B					
	Test set (568 NO-HSC tweets)						Test set (251 NO-HSC tweets)					
Linguistic Categories	IRONIC (251 tweets)	NON-IRONIC (317 tweets)	FP (%)	FN (%)	TP (%)	TN (%)	SARC (111 tweets)	IRO NON-SARC (140 tweets)	FP (%)	FN (%)	TP (%)	TN (%)
Explicit												
`an`	14	–	–	7.14	50.00	–	8	6	–	62.50	–	50.00
`euph`	23	–	–	**26.09**	34.78	–	14	9	–	64.29	–	**77.78**
`ex: c_s`	43	–	–	2.33	**60.47**	–	15	28	–	46.67	–	50.00
`ex: o/p`	52	–	–	7.69	55.77	–	30	22	–	53.33	–	72.73
`hyp`	13	–	–	7.69	53.85	–	4	9	–	75.00	–	66.67
`other`	26	–	–	3.85	38.46	–	5	21	–	**80.00**	–	76.19
`r_q`	20	–	–	10.00	45.00	–	13	7	–	76.92	–	71.43
Implicit												
`euph`	3	–	–	–	33.33	–	1	2	–	**100.00**	–	**100.00**
`hyp`	2	–	–	–	**100.00**	–	–	2	–	–	–	**100.00**
`im: f_a`	25	–	–	4.00	68.00	–	11	14	–	45.45	–	35.71
`other`	21	–	–	–	19.05	–	9	12	–	66.67	–	66.67
`r_q`	9	–	–	**11.11**	55.56	–	1	8	–	–	–	87.50

Table 5.18 – Distribution of Linguistic Categories.

5.3 The Unbearable Hurtfulness of Sarcasm

The IronITA contest provides a framework suitable for investigating irony and sarcasm in linguistic and computational terms. Indeed, the multi-source composition of IronITA2018 allowed us to understand the difficulties of state-of-the-art systems to detect irony and sarcasm in Italian tweets, and revealed that systems based on supervised approaches recognize a certain connection between offensive language and irony and sarcasm. To avoid the increase of FNs and FPs in both tasks, it appears necessary to inform the system with more specific information related to abusive language and emotions.

On the basis of these findings, we performed two analyses in order to disclose specific characteristics of sarcastic irony looking at the data and at the results of some computational experiments. Firstly, we exploited the composition of IronITA2018 to carry out statistical analyses able to disclose specific characteristics of sarcasm related especially to hostility that moves sarcastic expressions, and to rhetorical and pragmatic elements that distinguish sarcasm from other types of irony. Secondly, we experimented computationally the contribution of specific linguistic features for irony and sarcasm recognition, exploiting in particular the transfer learning approach by means of the pre-trained language models.

5.3.1 Statistical Analysis

This statistical analysis lets us study the association between irony/sarcasm and the dimensions of hate/linguistic characteristics interpreted as nominal variables of a population (i.e., data). In particular, we computed the χ^2 test of independence and the Yule's Q (see Section 3.2.2.1).

Dimensions of Hate Table 5.19 shows the p-values for the χ^2 test of independence and the Yule's Q values of the possible associations between *irony*[36]/*non-sarcastic irony*/*sarcasm* and each dimension of hate considered in the HSC. We remember that to reject the null hypothesis (variables are independent) of the χ^2 test of independence, the p-value should be minor than 0.05. Looking at this table, we noticed that: *sarcasm* is related to some degree to all the dimensions of hate and, especially, to aggressiveness, whereas *non-sarcastic irony* and, in general, *irony* are strongly associated with offensiveness, showing that, in presence of specific targets in the discussed issues, irony could be also offensive:

(93) *@USER Mi hanno insegnato che non tutti i musulmani sono terroristi ma il 99% dei terroristi nel mondo sono musulmani.*[37]

These results confirm our initial intuitions: sarcasm appears more aggressive than other types of irony and, considering the high values for hate speech, could perfectly fit to disguise negative messages.

[36] This label includes all types of irony.

[37] *@USER They have taught me that not all Muslims are terrorists, but 99 percent of the world's terrorists are Muslims.*

		`hs`	`agg`	`off`	`stereotype`
Task A	irony	0.00/0.22	0.00/0.35	**0.00/0.45**	0.00/0.37
Task B	sarcasm	0.00/0.37	**0.00/0.59**	0.01/0.23	0.02/0.19
	non-sarcastic irony	0.65/-0.05	0.28/-0.11	**0.00/0.32**	0.00/0.26

Table 5.19 – p-Values/Yule's Q Values for Dimensions of Hate.

Linguistic Characteristics Since the TWITTIRÒ schema of annotation is only focused on ironic texts, the set of observations is composed of *sarcastic* and *ironic non-sarcastic* tweets only. In this context, we could calculate statistical values for *sarcasm* and infer possible association for *non-sarcastic irony* by the sign of the Yule's Q values. Therefore, in Table 5.20, positive Q values refer to associations with *sarcasm* (maximum value in bold) and negative Q values to associations with *non-sarcastic irony* (minimum value in italic); while p-values indicate in general the existence or not of a dependence.

Table 5.20 reports significant signals of association, on the one side, between *non-sarcastic irony* and the `other` category (containing, indeed, other types of irony, such as situational irony) in the explicit class, and with hyperbole (`hyp`) in the implicit one; and, on the other side, between *sarcasm* and euphemism (`euph`) (maybe used to hide the negativity of messages) in the explicit class, and with false assertion (`f_a`) in the implicit one. Moreover, looking at the distribution of the *sarcastic/ironic non-sarcastic* tweets with respect to the explicit/implicit type of contradiction, we noted that sarcastic tweets tend to be slightly more explicit (83%) than non-sarcastic ones (79%) (see Examples 82 and 83), even if in general the implicit category is less represented in both classes[38]. A similar trend was observed also in English by Sulis et al. [2016]. In general, although the lower distribution of sarcastic texts in IRONITA2018 (see Table 5.6), the statistical measures helped to delineate some typical features of irony and sarcasm.

	`an`	`euph`	`ex:c_s`	`ex:o/p`	`im:f_a`	`hyp`	`other`	`r_q`
Explicit sarcasm	0.28/0.08	**0.02/0.25**	0.00/-0.28	0.01/0.17	–	0.28/-0.14	*0.00/-0.30*	0.24/0.09
Implicit sarcasm	0.18/-0.23	0.92/0.03	–	–	**0.01/0.31**	*0.23/-0.54*	0.47/-0.11	0.31/-0.24

Table 5.20 – p-Values/Yule's Q Values for Linguistic Characteristics.

From this first analysis it is clear that especially sarcasm is statistically related to aggressiveness, confirming previous theoretical works [Bowes and Katz, 2011], and, in general, to the false assertion and euphemism ironic devices. This, jointly with error analysis, helped us to design a computational approach of irony and sarcasm detection applicable also in abusive context.

[38] Indeed, we have a total of 294 implicit ironic tweets against 1227 explicit ones.

5.3.2 Computational Approach

The error analysis carried out on the results of the best performing systems at the IronITA shared task brings to light a clear difficulty to recognize sarcasm, showing that the predictions of the systems are affected by the hurtful language and the implicitness of expedients such as euphemism or a rhetorical question. Looking at the good performance obtained in NLP tasks, nowadays, with transformer-based approaches (like in HaSpeeDe 2020 shared task) also with small training sets, we proposed a computational analysis based on a hybrid system that exploits the general knowledge coming from the pre-trained language models and specific information that leads the system to infer hidden meanings from the text.

Moreover, the IronITA shared task suggests a novel computational interpretation of the sarcasm detection task as a sub-task of irony detection: if a tweet is ironic it could be sarcastic or not. Therefore, to detect sarcasm, we need to recognize before the presence of irony in the text. From this perspective, we adopted a *cascade architecture* where tweets that were predicted as ironic in the task A are classified as sarcastic and non-sarcastic in the task B. Although we used the same neural network for both tasks, the selected features in each classification task are different. Indeed, computing the χ^2 value for each feature, we are able to observe which feature is more significant for irony and which for sarcasm detection.

Our main idea is to converge in a unique system the awareness coming from the learning of a pre-trained language model with the linguistic knowledge derived from dedicated features. On the one side, the learning transferred by a language model trained on Italian tweets should help the classifier to be more sensitive to style and semantics of a more informal writing and make the system able to 'understand' better unseen cases. On the other side, engineered features lead the system to pay attention to specific elements, expressed or unexpressed in the text, that characterize irony and sarcasm.

As a pre-trained language model specific for Italian on social media texts, we used AlBERTo, the model for Twitter Italian language understanding created by Polignano et al. [2019] (see Section 3.2.2.2). This language model was trained on TWITA, a large dataset collecting Italian tweets from February 2012. The model that we used in this experiment was trained on 200M tweets published from 2012 to 2015 using 12 hidden layers with size of 768 neurons[39]. From now on, we refer to this hybrid system as AlBERToIS: AlBERTo based model for Irony and Sarcasm detection.

5.3.2.1 System Description

AlBERToIS takes into account two principal sets of inputs: AlBERTo's inputs and the features vector representation. In accordance with standard BERT input representation [Devlin et al., 2019b], the text is represented for AlBERTo as tokens, segments

[39] https://github.com/marcopoli/AlBERTo-it

and masked input. In order to load the trainable model of AlBERTo and tokenize the texts, we used the *keras-bert* implementation for BERT[40]. Moreover, we used *keras*[41] and *tensorflow*[42] as principal libraries to build our system. To create the features vector representation, we used the same process described in Section 3.2.2.2, extracting the same features listed in Table 3.38. Before combining these features with AlBERTo, we applied the batch normalization technique to the input-layer of features to standardize it and stabilize the learning process. In the end, the combination is attained concatenating the final-layer of AlBERTo with the input-layer of the features vector representation.

Taking into account the considerable size of AlBERTo transformer, after the concatenation step, we used a dropout layer to prevent the overfitting. And, at the end of our neural network, we added a dense-layer with standard ReLU activation with an input of 256 neurons and an output-dense-layer with a sigmoid function for binary classification in the task A (*ironic* and *non-ironic* classes) and in the task B (*sarcastic* and *non-sarcastic* classes).

As said before, we employed in our input the most relevant features for each task. The relevance of features is computed by means of the χ^2 value, and, in spite of the difference of the distribution of ironic and sarcastic tweets in the training set, looking at Figure 5.1, we can observe an important lexical trend in ironic and sarcastic tweets. Users tend to use hurtful words especially to express sarcasm, and words related to emotions to express irony.

With respect to other features, we can notice that: the variability of sentiment polarity in the message is characteristic of ironic and sarcastic statements, the variation of weights of words and pairs of words in a tweet appears more significant in sarcastic expressions, whereas ironic messages imply semantic similarities and incongruities disclosed by means of the computation of cosine similarity. About the syntactic features, Figure 5.1 shows that, in general, the punctuation plays an important role in the expression of irony in short texts. However, also the other syntactic features investigated here show to be involved mainly in ironic utterances.

[40] https://github.com/CyberZHG/keras-bert

[41] https://keras.io/

[42] https://www.tensorflow.org/

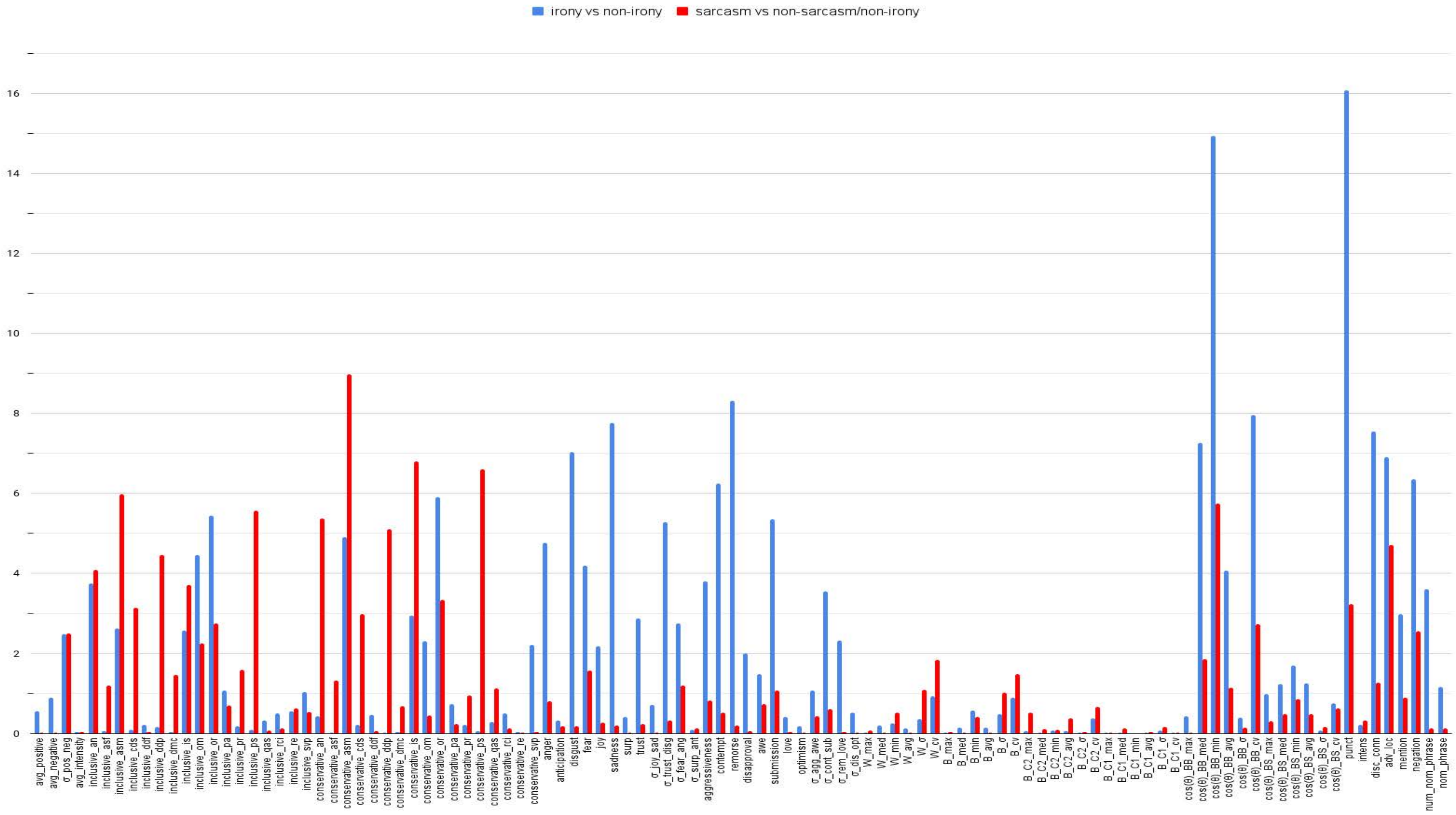


Figure 5.1 – Relevance of Features in the Training Set for the Tasks A and B.

5.3.2.2 Experiments and Results

The experimental phase was performed using the 20% of the training set as validation set; and focused mainly on searching for the best parameters and functions for our neural network, and on discovering the contribution of the features for irony and sarcasm detection. The former are resumed in Table 5.21.

parameter/function	**value/description**
max sequence length	80
dropout rate	0.3
learning rate	1e-5
batch size	8
maximum epochs	10
optimizer	Adam
early stopping	we applied the early stopping function provided by *keras* to avoid the overtraining of the neural network, looking at the values of the loss obtained on the validation set with a patience of 3 epochs.
seed	we applied a seed function from the *tensorflow* library to make the results reproducible.
initial bias	specifically for the task B, we adopted a technique to care about the initial bias calculated taking into account the imbalance between sarcastic and non-sarcastic classes[43].
learning rate finder	we found the learning rate of 0.00001 by means of a specific callback function[44].
loss	to minimize the loss function during the training, we used the binary cross-entropy function for binary classification provided by *keras*.

Table 5.21 – Parameters' Values and Functions.

For the latter, we carried out various experiments taking into account the χ^2 value of the features, and the binary accuracy scores obtained on the validation set. Indeed, binary accuracy metric is typically used for calculating how often predictions match binary labels. In particular, for the task A, the best binary accuracy score (0.817) is obtained with the set of 24 features with a χ^2 greater than 3. As shown in Figure 5.1, the most contributing set of features for this task includes: hurtful words, most of the statistical values calculated considering the cosine similarity between bigrams vectors and the sentence/vector context, stylistic features, adverbial locutions, discourse connections,

[43] We train our model on sarcastic and non-sarcastic tweets, including ironic/non-ironic ones, to ensure that the system could recognize specific characteristics of sarcasm.

[44] http://di.unito.it/lrfinder

number of nominal phrases among the syntactic features, and, finally, all the negative emotions (such as anger, disgust, fear, sadness, as well as the variability of trust and disgust) and negative feelings (such as aggressiveness, contempt, remorse and submission, as well as the variability of contempt and submission). Differently from the task A, the best model selected for the task B, with a binary accuracy score of 0.772, involves all the extracted features.

The selected best models for the task A and the task B are evaluated on the test set used in the IronITA shared task. To this purpose, we used the same evaluation metrics proposed by the organizers: $f1$-score for each class and $f1$-macro as average score. Specifically for the task B, we adopted a cascade architecture. Therefore, the predictions are obtained only for the tweets that were predicted as ironic in the task A. As baselines, we used the models provided in the competition (*Baseline_MFC* and *Baseline_Random*, see Section 5.2.1.2) and the AlBERTo-based model without linguistic features.

approach	**run**	**$f1$-score**		
		`non-iro`	`iro`	*macro*
AlBERToIS	–	**0.739**	**0.768**	**0.754**
AlBERTo	–	*0.722*	*0.747*	*0.735*
ItaliaNLP-MTL	–	–	–	0.736
ItaliaNLP	1	0.707	0.754	0.731
Baseline_Random	–	*0.503*	*0.506*	*0.505*
Baseline_MFC	–	*0.668*	*0.000*	*0.334*

Table 5.22 – Comparison of the Results for the Task A.

approach	**run**	**$f1$-score**			
		`non-iro`	`non-sarc iro`	*sarc*	**macro**
AlBERToIS	–	**0.739**	**0.471**	0.518	**0.576**
AlBERTo	–	***0.739***	*0.416*	***0.527***	*0.561*
ItaliaNLP-MTL	–	–	–	–	0.530
UNITOR	2	0.668	0.447	0.446	0.520
Baseline_Random	–	*0.503*	*0.266*	*0.242*	*0.337*
Baseline_MFC	–	*0.668*	*0.000*	*0.000*	*0.223*

Table 5.23 – Comparison of the Results for the Task B.

Tables 5.22 and 5.23 report the results obtained respectively in the task A and the task B. As we can notice, in both tasks AlBERToIS performs better in both classes, overcoming the best systems and the provided baselines. In spite of the $f1$-score achieved in the sarcastic class with a simple AlBERTo-based system is slightly higher than the one obtained with AlBERToIS, the proposed model reveals to be more balanced and solid to discriminate between sarcasm and non-sarcastic irony.

In line with previous works in various languages and genres (see Section 4.2.1), our results confirm the relevance of affective features for irony detection also in Italian. In English, Sulis et al. [2016] showed that negative sentiment and emotions are peculiar of ironic tweets, and, with our previous experiments on the corpora released at IroSvA (Section 5.1), we confirmed this finding also in Spanish. In particular, we can observe that the most discriminating emotions for irony detection are all negative (anger, disgust, fear, and sadness). In addition, negative feelings (aggressiveness, contempt, remorse, and submission) and the variability of contempt and submission prove to be significant. A different trend is visible in the task B. Indeed, as shown in Figure 5.1 the emotional features, in general, report a really low score except for fear, submission, and the variability of fear and anger.

Moreover, in our experiments in Spanish, we noticed that aggressive language is present in ironic texts. Looking at Figure 5.1, especially for discriminating *sarcastic* from *non-sarcastic* tweets, hurtful language seems to play an important role. Therefore, we carried out an additional experiment in sarcasm detection using in AlBERToIS the features with a χ^2 greater than 3 like in the task A. This set of 15 features includes the minimum value of cosine similarity calculated between pairs of words and the sentence context, the weight of the punctuation in the tweet, adverbial locutions and various hurtful words with a conservative and inclusive negative connotation. These words are mainly related to animals, male genitalia, physical disabilities/diversity, social and economic advantages, ethnicity, plants and general insults, such as:

(101) *Ma di quella senza fissa dimora "rom" ke ha danneggiato beni al Pantheon nn se ne parla? Dalla serie facciamoli tutti entrare, cani, porci..*[45]

The $f1$-macro obtained on the test set with this model is really competitive (0.573) showing that the contribution of features linked to the hurtful intention of sarcasm is notable. With respect to irony detection, the best selected model uses as features some categories of hurtful words related to plants, animals, male genitalia and homosexuality. These words are especially inclusive.

Finally, in order to understand the contribution of each feature in AlBERToIS, we carried out an ablation test. Observing its results in Table 5.24, we notice that in general the system tends to perform worse when the information about hurtful words is subtracted in both tasks. Moreover, it is interesting to note that knowledge about sentiment, and in particular about the variation of polarity in the message (see Figure 5.1), proves to be essential for sarcasm detection just as the features used to extract semantic incongruities and similarities are for irony detection.

5.3.2.3 Error Analysis

Looking at the values of the confusion matrices in Table 5.25, AlBERToIS shows an increase of sensibility in the predictions compared to the best performing systems in the

[45] *But, is there no mention of that homeless "Roma" who damaged the assets of the Pantheon? From the series let's get them in, dogs, pigs..*

	Task A	**Task B**
	$f1$-macro	$f1$-macro
AlBERToIS	**0.754**	**0.576**
Stylistic Features	0.749 (↓0.5%)	0.551 (↓2.5%)
Syntactic Features	0.738 (↓1.6%)	0.556 (↓2%)
Semantic Features		
- *Sentiment Lexicon*	–	0.532 (↓**4.4%**)
- *Hurtful Words*	0.725 (↓**2.9%**)	0.534 (↓**4.2%**)
- *Emotional Lexicon*	0.737 (↓1.7%)	0.551 (↓2.5%)
- *Incongruities and Similarities*	0.727 (↓**2.7%**)	0.545 (↓3.1%)

Table 5.24 – Ablation Test in AlBERToIS for the Tasks A and B.

IronITA shared task. In particular, we notice a reduction of 5% of FPs in the task A, and a notable increment of TPs of 11% in the task B. The error analysis in Section 5.2.1.3 revealed, mainly, how the lack of offenses on the one hand, and the presence of derogatory speech on the other hand, tend to improve, respectively, FN and FP in both tasks. Using the selected categories of hurtful words and specific affective features may have allowed AlBERToIS to improve the detection of ironic tweets when they contain or not offensive language. However, in the task B, the confusion matrix reports an increase of FPs of 8%. Analyzing the set of ironic tweets misclassified as sarcastic, we noted that most of the tweets containing especially stereotypes and offensive expressions, such as:

(94) *I rom saranno pure l'etnia più meschina, ladra, bugiarda del globo, ma NON GIUSTIFICA QUESTO. Manco allo zoo dai, a me viene il vomito #lidl*[46]

Nevertheless, looking at the TP and FN cases of AlBERToIS, we notice that in presence of aggressive language, sarcasm is correctly detected like in (95):

(95) *Ma pensa te! I ladri rampicanti sono rom quelli che portano cultura!! #Roma URL*[47]

In addition, we can see in Table 5.25 a similar trend observed in Section 5.2.1.3: the percentage of FPs is higher than FNs in irony detection. The tweets misclassified as ironic by AlBERToIS contain, especially, questions: rhetorical (96) and simple (97):

(96) *@USER bel programma #labuonascuola ma come è possibile per noi giovani andare a scuola senza avere i soldi per il pane?*[48]

(97) *@USER alla fine t'han messa dentro o no?*[49]

[46] *The Roma will also be the meanest, thief, liar ethnic group in the world, but IT DOES NOT JUSTIFY THIS. Not even at the zoo, come on, I feel like vomiting #lidl*

[47] *Can you believe it? The climbing thieves are Roma who bring culture!! #Roma URL*

[48] *@USER nice program #labuonascuola but how is it possible for us to go to school without having money for bread?*

[49] *@USER in the end did they put you in jail or not?*

	approach	id	**FP** (%)	**FN** (%)	**TP** (%)	**TN** (%)
Task A						
	ItaliaNLP	1	**36**	18	82	64
	AlBERToIS	–	31	18	82	**69**
Task B						
	UNITOR	2	22	**59**	41	**78**
	AlBERToIS	–	**30**	48	**52**	70

Table 5.25 – Values of Confusion Matrices for the Tasks A and B.

We suppose that the questions need to be addressed more specifically at a syntactic level, as well as exclamations:

(98) *@USER tra mezz'ora?! Ok... mi tocca aspettare ancora... ce la posso fare!*[50]

Another typical aspect of irony that makes its detection hard, also with AlBERToIS, is the use of euphemisms like in:

(99) *Messico, uccisa reginetta di bellezza. È quel piccolo difetto che la valorizza. [mukenin]*[51]

However, differently from values in Table 5.18, AlBERToIS could classify correctly the majority of situational ironic/sarcastic tweets. Examining the TP and FP cases, we noticed that the semantic features helped our model to detect correctly sarcastic tweets containing false assertions and oxymoron, whereas texts involving a context shift tend to be misclassified as sarcastic:

(100) *Mattarella batte le mani al ritmo di Bella ciao. Batterie non incluse. [@USER]*[52]

Actually, for sarcasm detection, AlBERToIS takes into account all the engineered features that could, as in this last case, capture some patterns that are more related to irony as shown in Table 5.20.

5.4 Discussion and Conclusions

In the experiments and analysis of this second part of our thesis, we tried, on the one hand, to contribute to the multidisciplinary discussion on which are the (linguistic and pragmatic) elements involved in ironic interpretations, and on the other hand, to understand at computational level how these elements contribute to make an automatic system aware of ironic language. In particular, our purpose was to investigate:

[50] *@USER in half an hour?! Ok... I have still to wait... I can make it!*

[51] *Mexico, beauty queen killed. It is that small flaw that valorizes her. [mukenin]*

[52] *Mattarella claps his hands to the rhythm of Bella ciao. Batteries not included. [@USER]*

1) if there are some common elements in various languages and genres that could trigger the interpretation of irony and its detection online;

2) the role played by affective aspects in ironic language, especially in the abusive context;

3) which characteristics of ironic language are peculiar of sarcasm.

4) if transformer-based architectures aimed to identify ironic language could benefit from the addition of linguistic features.

Looking at the two languages investigated in this chapter (Spanish and Italian) and at the existing literature on other languages (Section 4.2), we are able to draw some common and multilingual traits of irony. First, observing the characteristics of multilingual ironic utterances online (in tweets and comments), we noticed that there are some figures of speech often used by users. Karoui et al. [2017] analyzed tweets (self tagged or manually annotated as ironic) in different languages (Italian, English, and French), looking for specific linguistic devices in explicit and implicit expression of irony: analogy, metaphor, hyperbole, euphemism, rhetorical question, oxymoron, paradox and other elements such as false assertion, context shift, situational irony or specific markers (capital letters, quotation marks, and so on). In particular, they noticed that some categories have a similar high distribution in the three languages, such as oxymoron, false assertion and situational irony. Others such as analogy, context shift and euphemism are more common in Italian, while rhetorical questions and hyperbole reported a higher distribution especially in Italian and French tweets.

The characteristics of the ironic Italian tweets are confirmed in the deep error analysis carried out on the predictions in the test set of IronITA2018. In particular, observing these predictions made by supervised systems, we noticed that some of these linguistic devices trigger actually the ironic interpretation of the system, such as context shift, oxymoron, and hyperbole[53]. Whereas, other traits such as euphemism and rhetorical questions that need more contextual or external information tend to hinder the correct identification of ironic tweets, even in our well-informed AlBERToIS model. Moreover, with our participation at the IroSva shared task, we were able to investigate also Spanish ironic texts, and we noticed that if some figures such as hyperbole and rhetorical question are common even in Spanish, others such as ellipsis and apostrophe stand out[54] **(1)**.

Another common trait at the multilingual level is the role played by emotions and, in general, by affective information in irony understanding. For example, Hernández-Farías et al. [2016], looking at various affective information, individuated how some of them

[53] These linguistic devices, jointly with the `other` category, showed to be in a dependent relation with irony in Table 5.20.

[54] Unfortunately, a comparable multilingual observation could not be drawn also for the sarcastic form of irony. Indeed, to the best of our knowledge, sarcasm as a specific form of irony is represented only in IronITA2018.

have a relevant discriminative power in the classification of ironic and non-ironic tweets in English. Among them, we have pleasantness, imagery, and activation (see the DAL dictionary) and negative sentiment. In Sulis et al. [2016] similar findings came out from a statistical analysis of English #ironic self-labeled tweets that involve very negative emotions, compared to the #sarcastic ones.

These affective dimensions proved to be useful in irony detection even in the IroSvA datasets. In our experiments, we employed the automatic translated version of the DAL dictionary and the SEL emotional lexicon, and, independently of the genre of text, we found that only when the system is informed with very negative emotions, imagery, and pleasantness dimensions, it is able to detect better irony. About imagery, it is a dimension that captures the difficulty (from hard to easy) of imaging the reference of the word. Sulis et al. [2016] individuated that this specific dimension is higher in English #ironic tweets, suggesting the idea that they are more creative than the #sarcastic ones. As emerged from our experiments in Spanish, this dimension, differently from pleasantness, is useful especially to detect irony in tweets and not in comments, confirming the expectation of the more spontaneous and informal the context is, the more creative users tend to be.

Unfortunately, we have not yet an Italian version of the DAL dictionary and these findings could not be demonstrated also in IronITA2018. However, we were able to investigate the role of emotions in ironic Italian tweets by means of the qualitative error analysis, features analysis and computational experiments. All these analyses confirm the usefulness of negative emotions (anger, disgust, fear, and sadness) and negative feelings (aggressiveness, contempt, remorse, and submission) to detect irony, suggesting that these affections are involved in the expression of irony regardless of language, and in some way, they are perceived by the reader and thus, also by the system **(2)**.

In general, when scholars investigate affective information involved in ironic and sarcastic texts, focus especially on the role of dimensions that are behind the meaning of the words, such as emotions, sentiments and psychological aspects. And considering the fact that the majority of the existing literature agrees on the fact that irony tends to involve negative affects, in our experiments, we examined also the hostility expressed in the ironic texts. In this sense, in Spanish, we exploited manual modeled lexica of derogatory expressions and profanities that prove to have a discriminative power, especially in the detection of ironic tweets. In Italian, the multi-source composition of IronITA2018 allowed us to investigate deeper this aspect in the abusive context, observing the role of specific dimensions of hate such as hate speech, aggressiveness, offensiveness, and stereotype in ironic and sarcastic tweets.

Carrying out the error analysis on the common predictions of the best ranked systems in the competition, we found that supervised systems recognize a certain relation between hurtful language and the ironic one. This finding was confirmed by the statistical analysis of the dataset that revealed strong associations between irony and the dimension of offensiveness and, especially, between sarcasm and aggressiveness. In the same line, feature analysis, computational experiments and the final ablation test showed that aggressive and offensive language captured by using the HurtLex lexicon, proves to be discriminating mainly for sarcasm detection, and, only when it is less explicit also for irony detection. These findings confirm our intuition about the use of sarcasm for expressing hateful messages towards immigrants, helping thus to answer the second research question of the thesis:

RQ2 What is the role played by sarcasm in hateful messages online?

Jointly with the hurtful language, sarcasm seems to be characterized at linguistic level by markers such as the adverbial locutions and punctuation, devices such as euphemism and false assertion, semantic and sentiment shifts internal at the text. This latter is in line with the analysis reported in Sulis et al. [2016], where polarity reversal appears as a peculiarity of #sarcastic English tweets **(3)**.

In our set of experiments, all these features proved their contribution to irony and sarcasm detection in Italian, even in the transformer-based system. Indeed, linguistic information made AlBERToIS more sensible to ironic texts (Table 5.22). About the sarcasm detection, we got a general improvement, but not specifically on sarcastic tweets. We suppose that the scarcity of sarcastic samples in IronITA2018 could have impacted such outcome. This suggests, from an extended perspective, that automatic detection systems that have to process natural language hardly could obtain better performance if aware of linguistic knowledge **(4)**.

In this line, in the next and last part of this thesis, we want to face the detection of abusive language taking into account the important role played by ironic language in delicate issues such as immigration and social integration.

Part III

Abusive and Ironic Language

Chapter 6

Creativity in Abusive Language Online

In the previous chapter, we have seen that sarcasm, in abusive contexts, shows to be characterized by a very aggressive language aimed at mocking and hurting the victims, compared to other forms of irony that appear indirectly offensive and are characterized by a background of negative emotions and feelings. Moreover, from the overviews of various shared tasks on abusive language detection [Basile et al., 2019, Carmona et al., 2018] and earlier works [Nobata et al., 2016], the presence of ironic language proves to hinder the correct prediction of abuses online.

Similar findings in sentiment analysis have highlighted the importance to recognize creative language to reveal the real meaning of messages online. For example, Hernández-Farías and Rosso [2017] underlined a significant gap between the performance of sentiment analysis systems on non-figurative content and the performance reached on sarcastic content. In order to investigate the impact of linguistic creativity even in abuses detection, in this last part of our thesis, we focus mainly on the third research question:

RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

To give the system an overall perception of the linguistic phenomena that co-occur with abusive language, we can employ some approaches that tend to make it more sensitive even to indirect abuses. In this sense, some studies propose to train the models in different related tasks, as ItaliaNLP and TheNorth teams approached respectively IronITA and HaSpeeDe shared tasks. Multi-task learning, indeed, gives systems more evidences to evaluate whether a feature is relevant or not, especially, in the cases where the data have various labels.

In Section 3.2.2.2, we already experimented in hate speech and stereotypes detection as complementary tasks, although the performance of the model is lower than the one that exploits linguistic features. These, evidently, help the system to infer elements such as emotions or meanings far from the literal sense of the message (as seen in Table 3.53). However, this approach does not allow the system to identify perfectly those cases where ironic language is employed (Tables 3.57 and 3.58). To this purpose, in this chapter, we present: firstly, a statistical analysis on a new version of HaSpeeDe20_ext annotated also with the presence of irony and sarcasm, extending the findings of the previous chapter; secondly, new computational experiments to test the efficacy of making the system of abusive language detection sensible to ironic language; and finally, a deep analysis of the obtained results confirming our initial hypothesis.

6.1 Irony in Hateful Contents

The performed error analysis on our previous experiments on the detection of different abuses online (misogyny, aggressiveness, hate speech), already, underlined the need to make the system aware of ironic language. To deepen understanding related to this aspect, we performed a statistical analysis aimed to reveal the presence of irony in abusive context. To this purpose, we extended the levels of annotation of HaSpeeDe20_ext with the dimensions of irony (`iro` and `non-iro`) and sarcasm (`sarc` and `iro non-sarc`). For some instances of the training set (Train_TW_ext in Table 3.36) we recovered these dimensions from IronITA2018, for the rest of the dataset, we applied the same schema of annotation used in IronITA2018 (see Section 5.2.1.1). This new process of annotation involved three pairs of expert annotators due to the large size of this dataset (with a total of 8,602 instances). Therefore, firstly, the dataset was split in three parts and each couple annotated one part; and then, only one annotator of each couple performed a new annotation to solve the cases of disagreement. The distribution of the labels in this new versions of HaSpeeDe20_ext and some examples extracted from it are reported respectively in Tables 6.1 and 6.2.

set	hs	non-hs	stereo	non-stereo	iro	non-iro	sarc	iro non-sarc	**total**
Train_TW_ext	3,035	5,226	3,554	4,707	1,806	6,455	1,111	695	8,261
Test_TW	622	641	569	694	361	902	239	122	1,263
Test_NW	181	319	175	325	40	460	21	19	500

Table 6.1 – Distribution of the Existing and New Labels in HaSpeeDe20_ext.

A first look at the distribution of labels suggests that, as expected, the sarcastic form of irony is more used in abusive context than other forms of irony (as can be seen from the examples in Table 6.2); and that, in general, creative language is not used often in news headline. However, even if the instances of sarcastic headlines are few, these tend to express hate and stereotypes with a very cutting tone:

(102) *Italia, prima gli immigrati: ecco dove gli regaleranno un ristorante e un lavoro*[1]

[1] *Italy, immigrants first: this is where they will give him a restaurant and a job*

(103) *Ecco come rovinano l'Italia: immigrati, scippi e rapine? Le cifre-verità sulle 'risorse'*[2]

hs	stereo	iro	sarc	text
1	0	1	0	*"Anziché far venire gli immigrati diamo il Reddito di Cittadinanza*[3] *e gli italiani incominceranno a trombare come ricci. . . " (Massimo Baroni, deputato M5S)* → "Instead of letting immigrants come, we give the Citizenship Income and the Italians will begin to f**k like hedgehogs..." (Massimo Baroni, M5S deputy)
1	0	1	1	*Per l'ONU la capotreno sarebbe colpevole di di 'razzismo' e 'intolleranza' verso un'immigrata.. Come si è permessa di chiedere il biglietto ad un nigeriana?? Insomma, noi italiani non sappiamo proprio... URL* → According to the UN, the conductor would be guilty of 'racism' and 'intolerance' towards an immigrant.. How did she dare to ask a Nigerian for a ticket? In short, we Italians just don't know ... URL
1	1	1	1	*Oggi domenica delle palme! Qui in Italia è festa... f***ulo all'islam, per loro tutto il nostro calore URL* → Palm Sunday today! Here in Italy it's holyday... f**k Islam, for them all our warmth URL
0	1	1	0	*Mentre la Sinistra voleva imporre lo Ius Soli agli italiani, "come fanno nei paesi civili come gli USA", Trump vuole abolire lo Ius Soli in USA per impedire che immigrati e clandestini vadano a partorire in USA per prendere la cittadinanza. URL* → While the Left wanted to impose the Ius Soli to Italians, "as they do in civilized countries like the USA", Trump wants to abolish the Ius Soli in the US to prevent immigrants and illegal immigrants from giving birth in the US to take citizenship. URL
0	1	1	1	*@USER @USER @USER Accettiamo scommesse sul tipo di 'lavoro' che sta andando a fare il rom in. . . URL* → @USER @USER @USER We accept bets on the type of 'work' the roma guy is going to do in. . . URL

Table 6.2 – Examples Extracted from HaSpeeDe20_ext.

[2] *Here's how they ruin Italy: immigrants, muggings and robberies? The truth-figures on 'resources'*

[3] The Reddito di Cittadinanza is an economic support to combat poverty, inequality and social exclusion.

6.1.1 Statistical Analysis

Using the same statistical approach applied to IronITA2018 and HaSpeeDe2020, we computed the dependence and intensity of the relation between the dimensions of hate (hate speech and stereotype) and of ironic language (irony and sarcasm).

	Tweets		News	
	hs	stereotype	hs	stereotype
irony	0.00/-0.17	0.00/0.10	**0.00/0.72**	0.00/0.70
sarcasm	**0.00/0.24**	0.00/0.15	0.85/0.07	0.58/0.19
non-sarcastic irony	0.00/-0.27	0.75/0.01	**0.00/0.70**	0.00/0.65

Table 6.3 – p-Values/Yule's Q Values in HaSpeeDe20_ext.

Considering this analysis as an extension of the statistics already presented in Section 5.3.1, we compare the results in Table 6.3 with those reported in Table 5.17. We can notice that, in spite of the expanded sample of tweets, the association between sarcasm and abusive language maintains high scores, especially in cases of hate speech. Differently from sarcasm, the values related to the relation between irony and stereotypes are lower than the ones computed on IronITA2018, and in the case of non-sarcastic irony the relation is even absent. About the genre of news headlines, the association between hateful and ironic language appears stronger than in tweets, especially, in the cases where irony is not sarcastic. Although the values related, particularly, to news headlines are based on very few data (Table 6.1), this analysis gives us a first look of the possible characteristics of indirect language in messages containing hate speech and stereotypes in different textual genres.

6.2 Computational Experiments

To test these associations also at computational level, we employed the same multi-task learning approach (MTL_model) described in Section 3.2.2.2[4]. Thus, we experimented with a simple language model-based system that learns how to solve two tasks at the same time. In this set of experiments, we do not employ external linguistic features, because we count on the fact that the system during the learning process could be able to evaluate which features are relevant for each task and be aware of both tasks. Following the idea that we want to understand if the system aware of ironic language is able to improve its performance in abusive language detection, to evaluate these last experiments, we used the 20% of Train_ext as validation set for training and Test_TW and Test_NW for testing, employing the same evaluation measures used in HaSpeeDe 2020: $f1$-macro as average of the $f1$ of each class.

Like in Section 3.2.2.2, we tested the single language models for Italian (AlBERTo, UmBERTo and GilBERTo) and the average of their predicted probabilities for each instance

[4] Also the parameters are the same reported in Table 3.40.

(Avg_LMs). The following tables (Tables 6.4 and 6.5 for the task A, Tables 6.6 and 6.7 for the task B) report the $f1$-macro score obtained with these models compared to the value obtained with a simple fine-tuned system (the FT_model).

approach	FT	MTL (`iro`)	MTL (`sarc`)
AlBERTo	0.741	0.753	0.765
UmBERTo	0.790	0.780	**0.816**
GilBERTo	0.762	0.756	0.778
Avg_LMs	0.800	0.792	0.795

Table 6.4 – Results obtained on TEST_TW in the Task A.

approach	FT	MTL (`iro`)	MTL (`sarc`)
AlBERTo	**0.630**	0.600	0.611
UmBERTo	0.606	0.621	0.569
GilBERTo	0.602	0.582	0.605
Avg_LMs	0.612	0.597	0.584

Table 6.5 – Results obtained on TEST_NW in the Task A.

approach	FT	MTL (`iro`)	MTL (`sarc`)
AlBERTo	0.718	0.735	0.732
UmBERTo	0.767	0.760	0.774
GilBERTo	0.746	0.753	0.758
Avg_LMs	0.784	0.781	**0.789**

Table 6.6 – Results obtained on TEST_TW in the Task B.

approach	FT	MTL (`iro`)	MTL (`sarc`)
AlBERTo	0.731	0.670	0.708
UmBERTo	0.681	0.716	0.650
GilBERTo	**0.733**	0.694	0.698
Avg_LMs	0.712	0.700	0.691

Table 6.7 – Results obtained on TEST_NW in the Task B.

Looking at the performance of each language model in genres, we notice that UmBERTo performs very well especially in tweets as already noted in Tables 3.43 and 3.45. Moreover, making the system aware of ironic language seems to work well only in social media contents and not in news headlines, showing very low results especially in hate speech detection. Nevertheless, observing the scores obtained in headlines by models trained also on irony detection, the results obtained with UmBERTo (0.621 in the task A and 0.716 in the task B) are slightly higher than the ones achieved training sarcasm detection (0.611 in the task A and 0.698 in the task B). On the contrary, the system appears to be robust in tweets, especially, when it is trained also on sarcasm detection. These findings reflect the relevance of irony and sarcasm respectively for informal and formal texts in abusive context, already seen in statistical analysis (Table 6.3).

Finally, we compared, in Tables 6.8 and 6.9, these new results with the best ranked systems and baselines models of the HaSpeeDe 2020 shared task, and the results obtained adding linguistic features and training on stereotypes and hate speech detection for respectively the task A and the task B (see Section 3.2.2.2).

		Task A	
approach	**id**	$f1$_**Tw**	$f1$_**Nw**
UmBERTo (MTL `sarc`)		**0.816**	
Avg_LMs (FT+All_Feat)		0.810	
TheNorth	2	0.809	
TheNorth	1	0.790	
CHILab	1	0.789	**0.774**
UO	2		0.731
Montanti	1		0.726
AlBERTo (MTL `stereo`)			0.677
UmBERTo (MTL `iro`)			0.621
Baseline_Avg_LMs		*0.800*	*0.612*
Baseline_SVC		*0.721*	*0.621*
Baseline_MFC		*0.337*	*0.389*

Table 6.8 – Results in the Task A in the Ranking of HaSpeeDe 2020.

Observing these tables, it is clear that the linguistic information enriched with external features is very important to make the system able to understand the meaning. However, the awareness about sarcasm proves to be another important element to take into account in the detection of abusive messages. Indeed, if in the task A, the model trained also on sarcasm detection achieves the best score, in the task B it performs similar to the one informed with selected features on stereotypes detection. Considering this, in the future, we want to experiment with a more complex system that could combine the general knowledge coming from pre-trained language models with linguistic information and figurative language awareness.

Among all our experiments, we found that the most challenging task is the detection of hate speech in news headlines. In fact, our models did not achieve competitive results in it. As seen before, CHILab's approach is relevant for this task, because of the use of syntactic representation of the text jointly with common embedding textual representation.

		Task B	
approach	**id**	$f1$_**Tw**	$f1$_**Nw**
Avg_LMs (FT+25_Feat)		**0.793**	
Avg_LMs (MTL `sarc`)		0.789	
GilBERTo (FT+All_Feat)			**0.781**
TheNorth	1	0.772	
TheNorth	2	0.768	
CHILab	1	0.761	0.720
CHILab	2		0.718
Montanti	1		0.717
UmBERTo (MTL `iro`)			0.716
Baseline_ Avg_ LMs		*0.784*	*0.712*
Baseline_ SVC		*0.715*	*0.669*
Baseline_ MFC		*0.355*	*0.394*

Table 6.9 – Results in the Task B in the Ranking of HaSpeeDe 2020.

These two representations of texts are then processed with two different transformers, whose max pooling's results are averaged with a softmax function. The particular attention to the PoS tag representation of the text has evidently allowed the system to capture the nominal patterns that are not caught by our set of features. In this line, as further investigation, we want to analyze the contribution and discover, in particular, the kind of syntactic information that is helpful for abusive language detection in such formal message.

6.3 Analysis of Results

Sarcasm Understanding for Hate Speech Detection In order to understand the advantage to inject sarcasm knowledge in the system of hate speech detection, we carried out the analysis of TPs, TNs, FPs and FNs obtained with MTL comparing it with the baseline model based on LM fine-tuning and the best model informed with linguistic features. Considering the low performance obtained in news headlines, we performed this analysis only on tweets (Table 6.10).

approach	**FP** (%)	**FN** (%)	**TP** (%)	**TN** (%)
UmBERTo (MTL `sarc`)	24	**12**	**88**	76
Avg_LMs (FT+All_feat)	24	14	86	76
Baseline_Avg_LMs	**20**	21	79	**80**

Table 6.10 – Values of Confusion Matrix for the Task A in Tweets.

As we can notice, the system based on MTL of sarcasm shows an improvement of its accuracy in detecting hateful messages. For instance, the tweets in Tables 3.57 and 3.58 that have been misclassified by *Avg_ LMs (FT+All_feat)* are actually detected correctly by *UmBERTo (MTL `sarc`)*. Another interesting finding that emerged from a manual analysis is that making aware the system of sarcasm, even other tweets containing figures of speech, such as rhetorical questions and hyperboles, have been correctly classified (Table 6.11).

FN	TP		
Baseline_Avg_LMs	Avg_LMs (FT+All_feat)	UmBERTo (MTL `sarc`)	
`hs`	`hs`	`text`	
0	0	1	*Sri Lanka: colti e ricchi i terroristi ISIS. la madre si è fatta saltare uccidendo 3 figli e dei poliziotti... MA CON CHI DOVREMMO INTEGRARCI?* → Sri Lanka: ISIS terrorists educated and rich. the mother blew herself up, killing 3 children and some policemen... BUT WHO SHOULD WE INTEGRATE WITH?
0	0	1	*Quindi se un italiano muore in ospedale in mezzo alle formiche è 'episodio' mentre se un nigeriano muore per una circoncisione si richiede la sanità gratuita per gli immigrati. Roba da guerra civile e di sommosse fino ai bastioni di Orione.* → So if an Italian dies in the hospital in the midst of ants it is an 'episode' while if a Nigerian dies of a circumcision, free healthcare is required for immigrants. Stuff from civil war and riots up to the ramparts of Orion.
1	1	0	*Ma con tutti i problemi che hanno gli italiani bisogna pensare agli stranieri?* → But with all the problems that Italians have, should we think about foreigners?
1	1	0	*In una Tv pubblica con canone fisso in bolletta quello che vogliono imporci è parlare di "politica e migranti"? Se ami l'Italia* IT *boicottasamremo* → On a public TV with a fixed fee in the bill, what they want to impose on us is to talk about "politics and migrants"? If you love Italy IT boycottsamremo

Table 6.11 – Tweets Correctly Classified in the Task A.

Moreover, we noticed that the additional information about the figurative language that we gave to the system, actually, helped it to recognize the absence of hate speech in those tweets where users talk about the issue but without a hateful intention:

(104) *Si, così si deve fare, i criminali stranieri devono scontare la pena al loro paese. In Romania poi io so che sono molto più duri di noi. Questa è la strada giusta.*[5]

(105) *I rom sono tutti cittadini europei e una buona parte italiani. Non li puoi espellere. Sono perfettamente a conoscenza dei delitti che dici ma la situazione non si risolve con l'odio*[6]

[5] Yes, that's the way it has to be done, foreign criminals have to serve their sentences in their own country. In Romania, I know that they are much tougher than us. This is the right way.

[6] *Roma are all European citizens and a good part of them Italians. You cannot expel them. I am perfectly aware of the crimes you mention, but the situation cannot be resolved with hatred*

A similar attention towards negative class coming from the analysis of the predictions obtained with *Avg_LMs (MTL `sarc`)* in the task B on tweets in comparison with the baseline and *Avg_LMs (FT+25_Feat)*. *Avg_LMs (MTL `sarc`)* tends to improve indeed the recognition of TN cases, showing to be able to distinguish better the tweets with a quiet tone not expressing stereotypes:

(106) *Questa generazione di adolescenti è fantastica. Loro ci salveranno dal disastro e ci insegnano a difendere ciò che conta Rom razzismo TorreMaura*[7]

(107) *Non sono sicuramente i rom a rendere un inferno la tua ed altrui esistenza, ma il sistema in cui vivi e che subisci senza neppure il coraggio di immaginare che avresti il diritto di vivere in un mondo migliore.*[8]

Significance of Results Looking at the final tables of results (Tables 6.8 and 6.9) as well as at the error analysis in Section 3.2.2.3, we can summarize saying that:

i. only very spontaneous texts, such as tweets, tend to be characterized by a sharp ironic language;

ii. the awareness of sarcasm certainly helps the system of hate speech detection to retrieve positive examples, and in a real world-context, it could be convenient;

iii. a broad spectrum of linguistic features makes the system sensible especially to positive classes in hate speech and stereotypes detection respectively in tweets and news headlines;

iv. differently from news headlines, a selected group of linguistic features, related for example to typical topics of stereotypes against minorities (such as defects, morality, crimes and social advantages), makes the system more precise;

v. hate speech tends to be expressed in news headlines with *slogans-like NUs* and approaches such as the injection of stereotypes knowledge or the use of features that proves to be useful in tweets, are not effective in this genre.

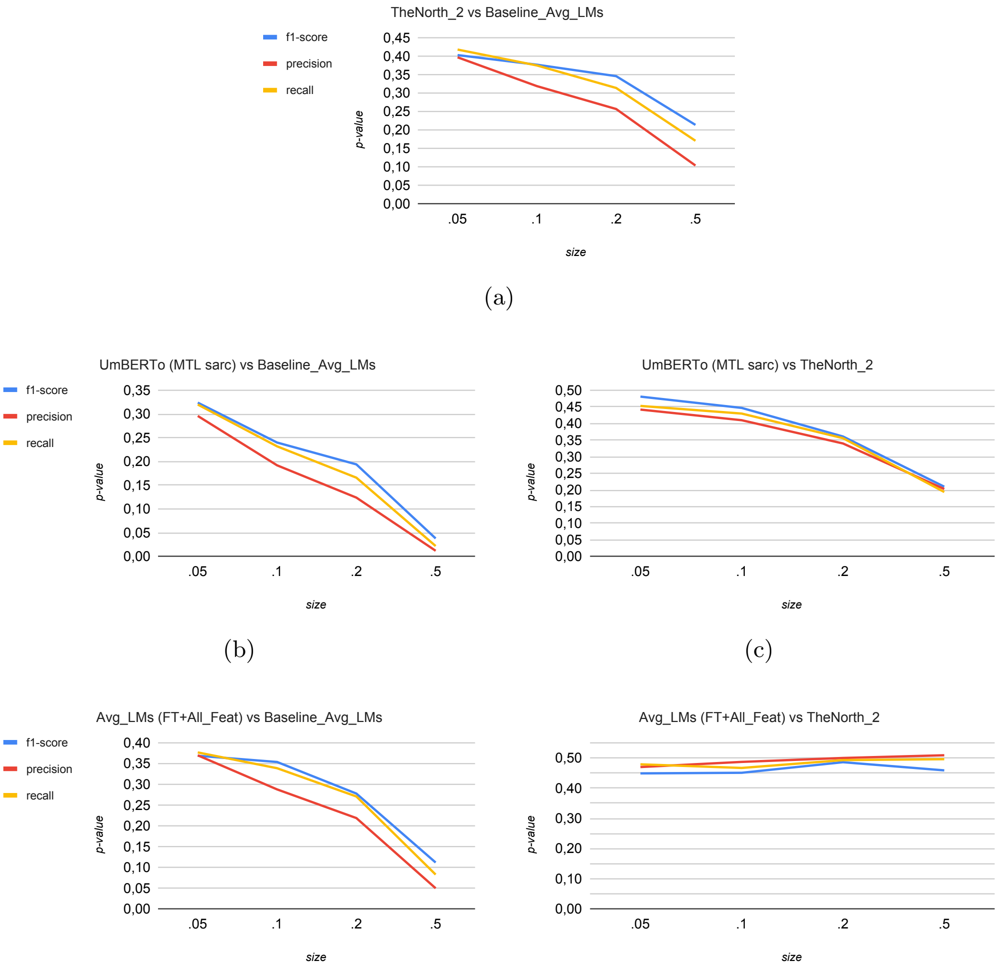
Considered these results, we carried out also a statistical experiment to understand how significant our approaches are. Taking into account the fact that we cannot meet the assumptions of perfect metrics and unbiased dataset, we followed the suggestions of Søgaard et al. [2014], reporting the significance results across the used dataset with regard to all available metrics (i.e., precision, recall and $f1$-score). As thresholds for the p-value, we adopted the typical cut-off at 0.05 and the Søgaard et al. [2014]'s one at 0.0025 (considered most reliable for NLP tasks). In particular, for this experiment

[7] *This generation of teenagers is fantastic. They will save us from disaster and teach us to defend what matters Roma racism TorreMaura*

[8] *It is certainly not Roma who make your existence and others' lives hell, but the system in which you live and which you suffer without even the courage to imagine that you would have the right to live in a better world.*

we used the library *boostsa*[9] that gives us the possibility to run a bootstrap sampling significance test, observing how fair are the obtained significance values with respect to different sizes of the unbalanced dataset.

Taking into account the results summarized above, we propose to investigate the significance comparing the baseline *Baseline_Avg_LMs* and best scored models in the HaSpeeDe 2020 shared task, with the systems that are aware of sarcasm or informed with linguistic knowledge used for hate speech and stereotypes detection. As parameters, we used a sample size that ranges from 0.05 to 0.5 of the total size of the test set, and 1000 iterations for computing the bootstrap sampling.



(a)

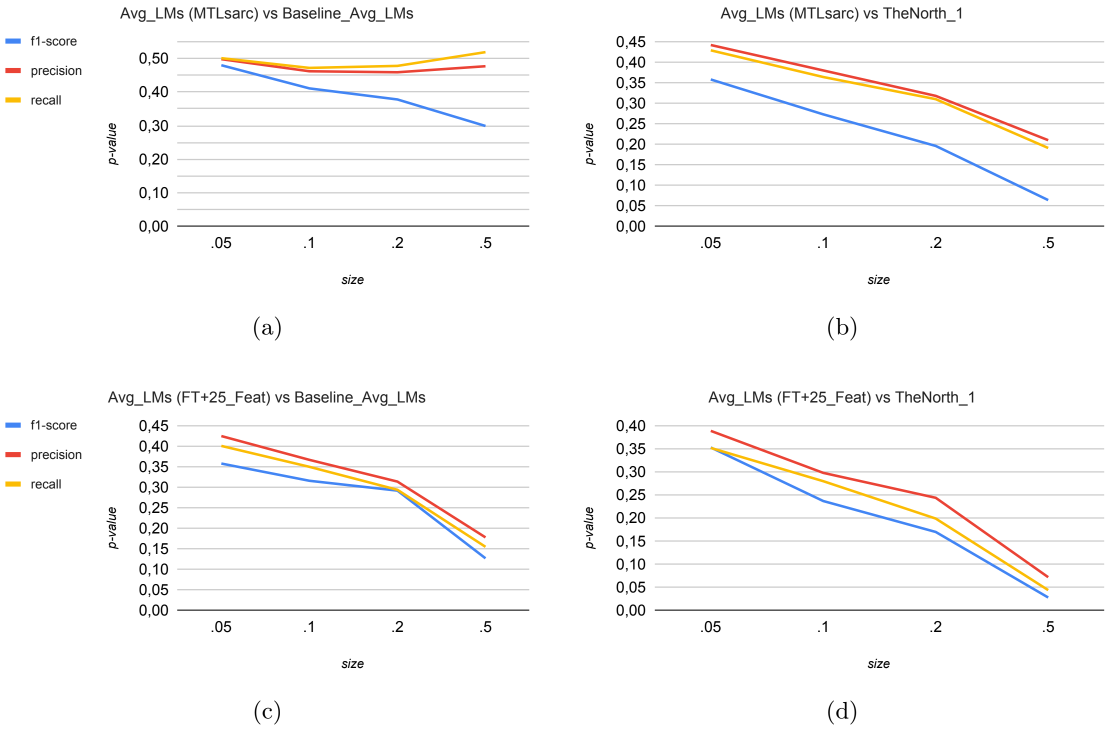
(b) (c)

(d) (e)

Figure 6.1 – Levels of Significance of Models for Hate Speech Detection in Tweets.

[9] https://github.com/fornaciari/boostsa#readme

Observing the curves of levels of significance of the three considered models compared with *Baseline_Avg_LMs* in Figure 6.1 for the task A, the system aware of sarcasm (b) reports significant p-value lower than typical 0.05 (0,038 for $f1$-score, 0.012 for precision and 0.022 for recall). Moreover, increasing the size of samples we can notice that the p-values tend to decrease proving that with a bigger test set, despite unbalanced, the model could perform optimally. Although the scores obtained with linguistic knowledge (d) are not minor than 0.05, they are still lower than the ones obtained with the MTL (`stereo`) approach used by TheNorth (*TheNorth_2*) in (a). In addition, we reported also the comparison of our models with *TheNorth_2*, and only the predictions of *UmBERTo (MTL* `sarc`*)* in (c) show a relevant trend.



(a)

(b)

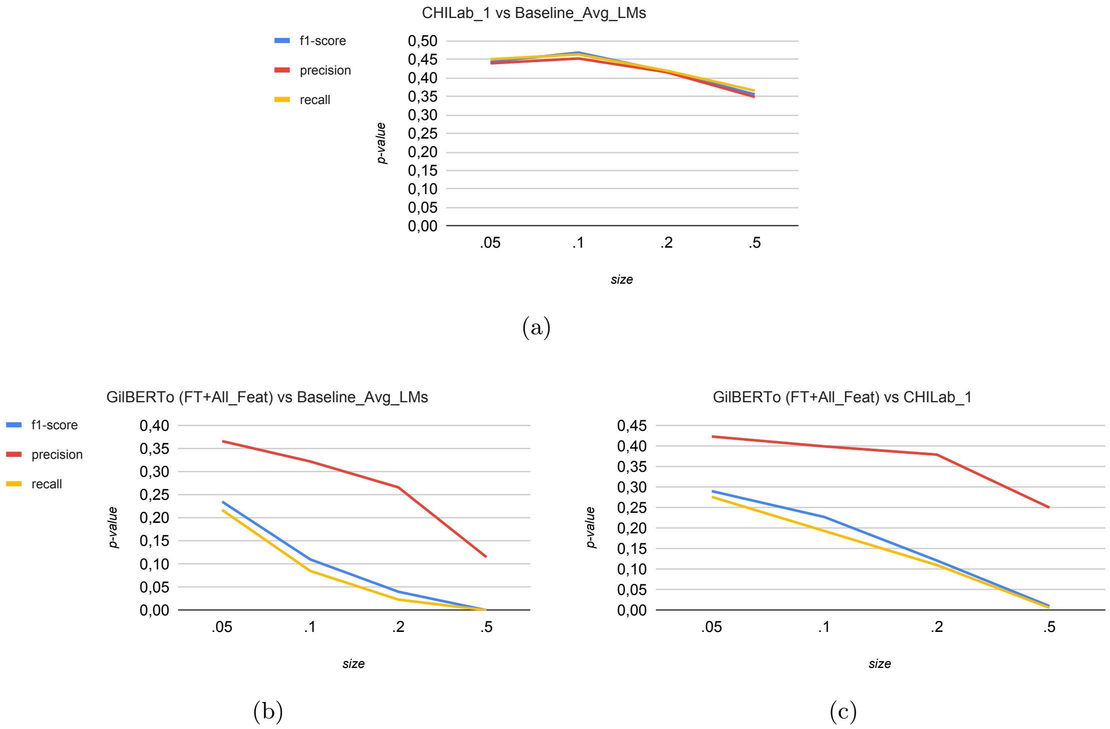
(c)

(d)

Figure 6.2 – Levels of Significance of Models for Stereotypes Detection in Tweets.

Differently from hate speech, in stereotypes detection in tweets (Figure 6.2) the approaches based on MTL with sarcasm (a) and linguistic features (c) do not show significant results compared to the challenging baseline model based on the fine-tuned LMs. However, their performance in (b) and (d) shows to be significant if compared with *TheNorth_1* system based only on UmBERTo fine-tuning. And in particular when the system is informed with linguistic features (p-values for $f1$-score of 0.028 and for recall of 0.04) in (c).

Finally, in detection of stereotypes in news headlines (Figure 6.3), increasing the size



(a)

(b)

(c)

Figure 6.3 – Levels of Significance of Models for Stereotypes Detection in News Headlines.

of samples, the recall and $f1$-score curves of our model obtain very low values in both comparisons. In particular, in (b) these values are very significant even with few examples reaching a p-value of 0.00 in recall and $f1$-score when the sample is bigger. The precision curves in (b) and (c) show a slow decrease, suggesting that for this specific task we need to improve the ability of the system to distinguish more finely the cases with stereotypes. Indeed, looking at the headlines that are misclassified by *GilBERTo (FT+All_feat)*, they tend not to contain verbs or in general predicative structures, demonstrating the need to approach this textual genre with specific syntactic information. As seen in Section 3.2.2.3, also hate speech tends to be expressed with *Slogan-like* NUs typical of political communication. And, despite linguistic features aimed at capturing emotions, offensiveness (related to specific topics), or semantic incongruities, helped in the detection of stereotypes, there are some texts that contain only nominal structures that are confused as `stereo` $= 0$, such as:

(108) *Immigrazione nel caos Gli irregolari in libertà i centri rimpatrio vuoti*[10]

(109) *Come rimanere clandestini a norma di legge*[11]

[10] *Immigration in chaos The illegal immigrants are free and the repatriation centers are empty*

[11] *How to stay clandestine according to the law*

(110) *L'auto dei rom di Casal Bruciato? Targa falsa e senza assicurazione*[12]

6.4 Discussion and Conclusions

In this last part of our thesis, we investigated at statistical and computational level the efficiency of making the systems of abusive language detection aware of ironic language. From our early studies, indeed, we noticed that systems aimed at detecting abusive language in various domains (misogynistic, racial or simply aggressive one), showed some difficulties when the message is not explicitly abusive, but it contains indirect references to prejudices or ironic expressions addressed to victims. About the former, we noticed in the experiments of the first part of the thesis, that linguistic features help the system to infer implicit or secondary meanings. While, about the latter, we investigated in literature and, then, at computational level, what is the type of irony that could reinforce the negative message and, at the same time, lessen the hurtful tones, hindering the detection of abusive text.

Linguistic and pragmatic studies suggest sarcasm as the kind of irony that is perceived as aggressive, and, in this line, we investigated linguistically its features, especially in tweets about delicate issues such as the integration of other cultures in Italy. Taking into account the statistical findings, in this last chapter, we aimed at answering our third research question:

RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

Approaching the problem as a learning question, we trained the system of abusive language detection also on irony and sarcasm identification. The reached performance shows that a system, even when it is able to generalize better because of the use of pre-trained language model, needs to be fed with linguistic knowledge (especially relative to the task), and needs to be aware mainly of sarcasm to understand the sarcastic messages that are abusive towards minorities. These observations are generalized to both dimensions of hate analyzed here: hate speech and stereotypes.

The only very challenge still open is the detection of hate speech in formal texts such as news headlines. In this textual genre, indeed, the message inciting to hate is not expressed like in tweets, and even if the simultaneous training in stereotypes identification seems to help slightly the detection, the realized performance is still low. In our experiments presented in Sections 3.2.2.2 and 6.2, we used an extended version of the original training set of HaSpeeDe2020. We think that, even if in stereotypes detection the longer sample of tweets helps also to detect stereotypes in news headlines, we cannot say the same for the detection of hate speech.

Considering that, as future work we want to investigate the possibility of joining linguistic knowledge, coming from a features' representation of text, with figurative language

[12] *The Roma car of Casal Bruciato? False and without insurance license plate*

awareness in a pre-trained language model-based architecture. In this case, we hypothesize that the system could be completely informed about the various expediences that are used to express hate. Taking into account the low results in hate speech detection in news headlines, inspired by the approach used by the CHILab team in the HaSpeeDe 2020 shared task, we want to capture the nominal form of hateful expressions in abusive headlines, analyzing the 'superficial' (PoS tags) and deep (deprels) structure of the sentences and exploiting the annotation provided for task C in the HaSpeeDe 2020 shared task (Section 3.2.1). Finally, in order to make the knowledge of the system independent of data, we want to experiment also with unsupervised approaches based especially on specific representations of the text able to cover the different ways of inciting intolerance and reinforcing beliefs towards the perceived outgroups.

Chapter 7

Conclusions and Future Work

Various scholars in linguistics highlighted on the mutual relation between language and society, that is composed of the speakers of that language, affecting each other and changing together. Therefore, with *words* we are not just speaking, but we do things, things that could help people or things that could marginalize or hurt people [Bianchi, 2021]. New technologies give us the possibility to stay constantly connected with other people, communicating with them and exchanging opinions. Unfortunately, despite their positive and important impact in society, they have also amplified negative behaviors or ideologies that could worsen the situation of some categories of people already excluded or considered inadequate for the society because of their sexual identity, physical abilities and origins.

Our investigation aimed at contributing to the comprehension of how abuses, such as misogyny and racism, are expressed directly and indirectly, and how they could be recognized by machines. To this purpose, we tried to answer the following questions:

RQ1 How to make abusive language detection systems sensitive to implicit manifestations of hate?

RQ2 What is the role played by sarcasm in hateful messages online?

RQ3 Could the awareness of the presence of sarcasm increase the performance of abusive language detection systems?

In Chapter 3, the corpora-based analysis, the statistical tests and computational experiments on various benchmark datasets showed that abusive language towards women and immigrants involves: on the one side, very offensive words; and on the other side, deep social biases that appear to be pervasive even in discussions that involve these targets, and figurative devices that create a secondary meaning different or opposite to the literal one. To make the systems of abusive language detection aware of stereotypes or prejudices, we experimented various approaches, discovering that especially lexica-based features are very useful even in the systems with neural architectures. Indeed, systems that tend to generalize better because they take advantage from pre-trained language models, show to perform better when linguistic knowledge is provided. In this way, they are able to infer the possible interpretations that are implicit in the text, especially when these refer to conventional and not explicitly expressed ideas such as the patriarchal social order or incivility of immigrants (**RQ1**).

Approaching abusive language detection as a classification problem, we noticed that one of the points that remained unsolved was related to the presence of ironic devices. Irony, in fact, is used to mask the purpose of haters to insult specific vulnerable targets. Ironic texts, analyzed in Chapter 5, have been found to be aggressive, above all when the sarcastic form of irony is employed; proving, therefore, some arguments in favor of linguistic and pragmatic theories [Bowes and Katz, 2011] (**RQ2**).

Considering that, we designed a new approach of detection in Chapter 6, exploiting the presence of irony in manual annotated texts (the benchmark dataset released in occasion of HaSpeeDe 2020). In particular, we designed a system that fine-tune Italian language models simultaneously on the tasks of hateful and ironic language recognition in a multi-task framework. We compared its results with the one obtained with the previous approach that combines general knowledge, coming from language models, and linguistic information, provided by means of specific linguistic features. We discovered that the awareness of sarcasm helps the system to retrieve correctly hate speech in social media texts, such as tweets; and that linguistic features make the system sensible to stereotypes in both tweets and news headlines (**RQ3**).

Despite the success of these approaches that overcame the current state of the art in Italian, they proved to be less effective in detecting hate speech in news headlines. News headlines have, indeed, a singular structure that we propose to address in future works exploiting especially syntactic information.

7.1 Research Contributions

To deal with the detection of abusive language in creative and implicit messages, we contributed to research designing different methodologies and approaches, and creating linguistic resources such as lexica and benchmark datasets already shared with the NLP community in the context of the shared tasks that we organized.

Firstly, we carried out various qualitative and quantitative corpora-based analyses that allow us to delineate explicit and implicit characteristics of abusive language against, specifically, women and cultural minorities, even in a multi-language and multi-genre context. The most important elements that make the recognition of the abuses online difficult, are, specifically:

- social biases, such as stereotypes, that especially in misogynistic contexts tend to motivate the hateful attacks against women;

- stance expression towards delicate issues that involve a specific target such as the legalization of abortion and feminist manifestations;

- and ironic language that especially in its sarcastic form shows some characteristics, such as aggressive language and sharper tone, that make it suitable to hurt the victim.

About irony, we performed various analyses that allowed us to contribute also to the more theoretical and linguistic discussion on: 1) the peculiarities of sarcasm compared to other forms of irony, and 2) mono and multilingual characteristics of irony. Sarcasm, defined in literature as a sharp form of irony with the intent of scorning a victim without excluding the possibility to amuse, proved to be characterized by: aggressiveness and hurtful language, explicit contradictions marked with adverbial locutions, semantic and polarity shifts, and false assertions and euphemistic forms. The computational experiments carried out especially on irony detection revealed, instead, that negative emotions are involved in the expression of irony, regardless the language, the context, and the genre. As linguistic traits, we noticed some commonalities among English, Spanish, Italian and French, where irony tends to be expressed by hyperboles, rhetorical questions and oxymoron. Moreover, we observed that in Spanish it involves specifically ellipsis and apostrophe, whereas in Italian context shift and euphemism.

Secondly, to approach indirect abusive language detection, we explored various computational techniques ranging from classical to recent ones. The employment of specific set of features, the comparison of basic and complex approaches, as well as the ablation tests supported the interpretability of the models and the comprehension of the analyzed phenomenon. Trying to design, into technical solutions, the linguistic and cognitive mechanisms to make the system able to capture the intentional meaning of the message, we examined the contribution of: lexica-based features, distributional semantics models, transfer learning techniques based on pre-trained language models, and multi-task learning approaches. Among them, we emphasize the performance of the approaches based on the combination of pre-trained language models and linguistic features (like AlBERToIS) and on the simultaneous learning of ironic language (like MTL_model) that realized the best scores respectively in irony and sarcasm detection and hate speech detection in Italian. These scores have been compared with the results obtained by participating teams on the benchmark datasets proposed at the two shared tasks that we organized in the context of EVALITA 2018 and 2020: IronITA about irony and sarcasm detection in tweets, and the second edition of HaSpeeDe about hate speech and stereotypes detection in tweets and news headlines.

In order to set proper experimental settings to answer our research questions, we created various resources, such as lexica and corpora, resumed briefly below.

Corpora

- **IronITA2018**[1]: it is a benchmark corpus [Cignarella et al., 2018b] released for the IronITA shared task proposed at EVALITA 2018. This dataset is a collection of Italian tweets from two different contexts (political and abusive) and it is annotated with the presence of irony and sarcasm, considered as a form of irony. Taking into account the multisource composition of this dataset, we extended the levels of annotation as reported in Section 5.2.1.3, and made this version available.

- **HaSpeeDe2020**[2]: it is a benchmark corpus [Sanguinetti et al., 2020] released in the second edition of HaSpeeDe shared task at EVALITA 2020. It collects Italian tweets and newspapers headlines against minorities such as Muslims, Romas and migrants. The provided labels during the competition are *hate speech* and *stereotype*, but the extended version with the annotation of the presence of irony and sarcasm is available.

- **GDELT-FM**: this corpus was created extracting news articles from the GDELT platform. The collected news are about feminist movements related to events happened from the 1st of October to 31st of December in 2017 in Europe, Japan, and USA. This corpus was annotated automatically, looking at the polarity-based stance of the journal.

Lexica

- We created misogynistic multilingual lexica including words and expressions related to stereotypes in abusive language towards women in Spanish, Italian and English. If the core of these lexica is manually treated, we extended them automatically, exploiting GloVe and TWITA words embeddings.

- We manually created for the Mexican variations of Spanish lexica that collect implicit and explicit offensive and derogatory words and expressions coming from informal Mexican speech.

7.2 Relevant Publications

Below, we outline the works published during the Ph.D. by organizing them into five groups.

Journals

- Frenda, S., Cignarella, A. T., Basile, V., Bosco, C., Patti, V., and Rosso, P. (2022). **The Unbearable Hurtfulness of Sarcasm**. In Expert Systems with Applications (ESWA) 193, 116398, Elsevier.

[1] https://live.european-language-grid.eu/catalogue/corpus/7372

[2] https://live.european-language-grid.eu/catalogue/corpus/7498

- Frenda, S., Patti, V. and Rosso, P. (2021). **Killing Me Softly: Creative and Cognitive Aspects of Implicitness in Abusive Language Online**. In Natural Language Engineering (NLE). Accepted after minor revision.

- Frenda, S., Banerjee S., Rosso P. and Patti V. (2020). **Do linguistic features help deep learning? The case of aggressiveness in Mexican tweets**. In Computación y Sistemas, 24(2).

- Frenda, S., Ghanem, B., Montes-y-Gòmez, M. and Rosso, P. (2019). **Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter**. In Journal of Intelligent & Fuzzy Systems (JIFS), vol. 36.

Conferences

- Frenda, S., Noriko, K., Patti, V., and Rosso, P. (2019). **Stance or insults?**. In Proceedings of Ninth International Workshop on Evaluating Information Access (EVIA2019), a Satellite Workshop of the NTCIR-14 Conference (pp. 15-22).

- Frenda, S. (2018). **The role of sarcasm in hate speech. A multilingual perspective**. In Proceedings of Doctoral Symposium of the 33rd Conference of the Spanish Society for Natural Language Processing (SEPLN 2018).

Overviews of Shared Tasks

- Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M. A., Bosco, C., Caselli, T., Patti, V. and Russo, I., (2020). **HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task**. In Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020).

- Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., and Rosso, P. (2018). **Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA)**. In Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), Vol. 2263.

Reports of Participation at Shared Tasks

- Frenda, S. and Patti, V. (2019). **Computational Models for Irony Detection in three Spanish variants**. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019).

- Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes-y-Gòmez, M. and Villasenor-Pineda L. (2018). **Automatic Lexicons Expansion for Multilingual Misogyny Detection**. In Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), Vol. 2263.

- Frenda, S., Ghanem, B. and Montes-y-Gòmez, M. (2018). **Exploration of misogyny in Spanish and English tweets**. In Notebook Papers of 3rd SEPLN Workshop IBEREVAL 2018.

- Frenda, S. and Banerjee, S. (2018). **Deep analysis in aggressive Mexican tweets**. In Notebook Papers of 3rd SEPLN Workshop IBEREVAL 2018.

Other Related Publications

- Stranisci, M. A., Frenda, S., Ceccaldi, E., Basile, V., Damiano, R., Patti, V. (2022). **AppReddit: a Corpus of Reddit Posts Annotated for Appraisal**. Language Resources and Evaluation Conference 2022 (LREC 2022). Accepted after minor revision.

- Stranisci, M. A., Cignarella, A. T., Frenda, S., Lai, M., Bosco, C., Patti, V. (2021).**Hate Speech e Dangerous Speech in Twitter**. In Rassegna Italiana di Linguistica Applicata. Bulzoni Editore. In press.

- Frenda, S., Cignarella, A. T., Stranisci, M. A., Lai, M., Bosco, C. and Patti, V. (2021). **Recognizing Hate with NLP: The Teaching Experience of the #DeactivHate Lab in Italian High Schools**. In Proceedings of Eighth Italian Conference on Computational Linguistics (CLiC-it 2021).

- Frenda, S. (2017). **Ironic Gestures and Tones on Twitter**. In Proceedings of Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Vol. 2006.

7.3 Future Work

Looking at all the approaches that we implemented for this investigation, some other challenges in abusive language detection remain unaddressed. For instance, in some misclassified cases, we noticed that swear words used with non-abusive intent, created ambiguity for the system. Especially in informal and spontaneous contexts such as social platforms, the texts frequently contain swear words that cover different functions (surprise, insult, friendly nicknames). Already, some colleagues like Pamungkas et al. [2020b] coped with this issue and, in further work, we want to improve our approach taking into account the positive and negative function of swearing.

The other open challenge, when we process the texts only at message level, is surely the absence of contextual information. Reading the text in isolation actually oversimplifies how hate speech happens in reality. It is true that some texts are themselves clearly abusive, but in the other cases, context could help to give a more informed perspective to interpret them as abuses or not, making the approach fairer for a real-world application. To understand how our system, aware of ironic and implicit expressions, performs informed with context, we plan to experiment it on datasets created for this purpose, as for instance, the dataset built by Menini et al. [2021] for English language.

Very recently, some scholars started to approach abuses detection also in images [Menini et al., 2021]. Among them, memes are a new form of immediate communication, very common online, that plays with the relation between image and text to create the funny meaning. Traditionally, this communicative strategy is used for marketing or political campaigns in billboards. But recently it is used also in daily interactions, and like other forms of expressions also memes show to be abusive. To the best of our knowledge, the detection of abuses in memes is poorly explored until now [Zhou et al., 2021]. Only in this last year, a shared task on multimedia automatic misogyny identification (MAMI) has been organized at SemEval 2022[3]. Taking into account our findings on ironic and abusive language, it would be very interesting to experiment also with multi-modality approaches to face these new forms of toxicity online.

Finally, our approaches proved to be less effective in the detection of hate speech in news headlines. The provided analyses in Chapter 3 showed that the journalists tend to express a stance against the presence of immigrants in Italy using very simple NUs that remember the political rhetoric that feeds the juxtaposition between the ingroup and the outgroup. In this line, we would deepen the analysis of the role of these nominal structures in Italian news headlines exploiting the annotation provided for task C in the HaSpeede 2020 shared task, and in English examining the NUs of the news headlines automatically annotated as unfavorable towards feminist movements in GDELT-FM. Inspired by the architecture of the system proposed by the CHILab team, it would be interesting to employ a neural architecture able to take also into account a deeper syntactic representation of text.

[3] https://competitions.codalab.org/competitions/34175

Bibliography

A. Agrawal, A. An, and M. Papagelis. Leveraging transitions of emotions for sarcasm detection. In J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1505–1508. ACM, 2020. URL https://doi.org/10.1145/3397271.3401183.

R. Ahluwalia, H. Soni, E. Callow, A. C. A. Nascimento, and M. D. Cock. Detecting hate speech against women in English tweets. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2263/paper032.pdf.

S. Akhtar, V. Basile, and V. Patti. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896, 2021. URL https://arxiv.org/abs/2106.15896.

L. Alba-Juez and S. Attardo. The evaluative palette of verbal irony. *Evaluation in context*, 242:93—-116, 2014. URL https://doi.org/10.1075/pbns.242.05alb.

M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, volume 10859 of *Lecture Notes in Computer Science*, pages 57–64. Springer, 2018. URL https://doi.org/10.1007/978-3-319-91947-8_6.

M. E. Aragón, M. Á. Á. Carmona, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma. Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, and A. Rosá, editors, *Proceedings of the Iberian Languages*

Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, volume 2421 of *CEUR Workshop Proceedings*, pages 478–494. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2421/MEX-A3T_overview.pdf.

M. E. Aragón, H. J. Jarquín-Vásquez, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, and G. Bel-Enguix. Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. In M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, S. M. J. Zafra, J. A. O. Zambrano, A. Miranda, J. P. Zamorano, Y. Gutiérrez, A. Rosá, M. Montes-y-Gómez, and M. G. Vega, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020*, volume 2664 of *CEUR Workshop Proceedings*, pages 222–235. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2664/mex-a3t_overview.pdf.

S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007. URL https://doi.org/10.1002/asi.20553.

G. Attanasio and E. Pastor. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identification in Italian tweets. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper142.pdf.

S. Attardo. Irony as relevant inappropriateness. *Irony in language and thought: A cognitive science reader*, pages 135–170, 2007. URL https://psycnet.apa.org/record/2007-10000-006.

J. L. Austin. *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955.* Oxford university press, 1975. URL https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198245537.001.0001/acprof-9780198245537.

N. Babanejad, H. Davoudi, A. An, and M. Papagelis. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.20.

S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, volume 10, pages 2200–2204, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

A. Bakarov. Vector space models for automatic misogyny identification. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 211–213. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2263/paper035.pdf.

D. Bamman and N. A. Smith. Contextualized sarcasm detection on Twitter. In M. Cha, C. Mascolo, and C. Sandvig, editors, *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 574–577. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10538.

S. Banerjee, S. Naskar, P. Rosso, and S. Bandyopadhyay. Code mixed cross script factoid question classification - A deep learning approach. *Journal of Intelligent & Fuzzy Systems*, 34(5):2959–2969, 2018. URL https://doi.org/10.3233/JIFS-169481.

F. Barbieri and H. Saggion. Modelling irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics. URL https://aclanthology.org/E14-3007.

F. Barbieri, F. Ronzano, and H. Saggion. Italian irony detection in twitter: a first approach. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 : 9-11 December 2014, Pisa*, pages 28–32. Pisa University Press, 2014. URL http://digital.casalini.it/3043505.

F. Barbieri, F. Ronzano, and H. Saggion. Is this tweet satirical? A computational approach for satire detection in Spanish. *Procesamiento del Lenguaje Natural*, 55:135–142, 2015a. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5225.

F. Barbieri, F. Ronzano, and H. Saggion. UPF-taln: SemEval 2015 tasks 10 and 11. sentiment analysis of literal and figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 704–708, Denver, Colorado, June 2015b. Association for Computational Linguistics. URL https://aclanthology.org/S15-2119.

F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti. Overview of the EVALITA 2016 SENTIment POLarity Classification task. In *Proceedings of 3rd Italian*

Conference on Computational Linguistics (CLiC-it 2016) & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA'16), Naples, Italy, 2016. CEUR.org. URL https://hal.archives-ouvertes.fr/hal-01414731.

J. P. Barlow. A Declaration of the Independence of Cyberspace. In *Crypto Anarchy, Cyberstates, and Pirate Utopias*. The MIT Press, 05 2001. URL https://doi.org/10.7551/mitpress/2229.003.0006.

R. Barthes and A. Lavers. *Mythologies*. Vintage classics. Vintage, 1993. URL https://books.google.es/books?id=wsGDVdYoRA4C.

A. Basile and C. Rubagotti. Crotonemilano for AMI at EVALITA 2018. A performant, cross-lingual misogyny detection system. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2263/paper034.pdf.

P. Basile and N. Novielli. Uniba at EVALITA 2014-SENTIPOLC task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014 : 9-11 December 2014, Pisa*, pages 58–63. Pisa University Press, 2014. URL http://digital.casalini.it/3044388.

P. Basile and G. Semeraro. UNIBA - Integrating distributional semantics features in a supervised approach for detecting irony in Italian tweets. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, 2018. CEUR-WS. URL http://ceur-ws.org/Vol-2263/paper024.pdf.

V. Basile. I computer e il linguaggio naturale. *Ithaca: Viaggio nella Scienza*, 2020 (16):151–166, 2020a. URL http://siba-ese.unisalento.it/index.php/ithaca/article/view/23009.

V. Basile. It's the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. In G. Vizzari, M. Palmonari, and A. Orlandini, editors, *Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020), Anywhere, November 27th, 2020*, volume 2776 of *CEUR Workshop Proceedings*, pages 31–40. CEUR-WS, 2020b. URL http://ceur-ws.org/Vol-2776/paper-4.pdf.

V. Basile and M. Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-1614.

V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. Overview of the EVALITA 2014 SENTIment POLarity Classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy, 2014. Pisa University Press. URL https://hal.archives-ouvertes.fr/hal-01228925.

V. Basile, N. Novielli, D. Croce, F. Barbieri, M. Nissim, and V. Patti. Sentiment polarity classification at EVALITA: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*, 12(2):466–478, 2018. URL https://doi.org/10.1109/TAFFC.2018.2884015.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel-Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/S19-2007.

E. Bassignana, V. Basile, and V. Patti. Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, and F. Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2253/paper49.pdf.

J. Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997. URL https://doi.org/10.1023/A:1007327622663.

C. Bazzanella. Oscillazioni di informalità e formalità: scritto, parlato e rete. *Formale e informale. La variazione di registro nella comunicazione elettronica*, pages 68–83, 2011. URL http://www.carocci.it/index.php?option=com_carocci&task=schedalibro&Itemid=72&isbn=9788843061310.

S. Bearman, N. Korobov, and A. Thorne. The fabric of internalized sexism. *Journal of Integrated Social Sciences*, 1(1):10–47, 2009. URL https://jiss.org/documents/volume_1/issue_1/JISS_2009_1-1_10-47_Fabric_of_Internalized_Sexism.pdf.

F. Benamara, C. Grouin, J. Karoui, V. Moriceau, and I. Robba. Analyse d'opinion et langage figuratif dans des tweets : Présentation et résultats du Défi Fouille de Textes DEFT2017. In *Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017)*, pages pp. 1–12, Orléans, France, June 2017. URL https://hal.archives-ouvertes.fr/hal-01912785.

E. Bender. The #BenderRule: On naming the languages we study and why it matters. *The Gradient*, 2019. URL https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

S. Benesch. Dangerous speech: A proposal to prevent group violence. *Voices That Poison: Dangerous Speech Project proposal paper*, 2012. URL http://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf.

C. Beukeboom and C. Burgers. How stereotypes are shared through language: A review and introduction of the Social Categories and Stereotypes Communication (SCSC) framework. *Review of Communication Research*, 7:1–37, 2019. URL https://research.vu.nl/en/publications/how-stereotypes-are-shared-through-language-a-review-and-introduc.

C. Bianchi. *Hate speech: Il lato oscuro del linguaggio*. Editori Laterza, 2021. URL https://books.google.it/books?id=arYYEAAAQBAJ.

E. Bisconti and M. Montagnani. Montanti @ HaSpeeDe2 EVALITA 2020: Hate speech detection in online contents. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper177.pdf.

D. C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3):289–299, 1985. URL https://doi.org/10.1145/3166.3197.

R. J. Blanco, J. M. Alcaide, M. I. Torres, and M. A. Walker. Detection of sarcasm and nastiness: New resources for Spanish language. *Cognitive Computation*, 10(6): 1135–1151, 2018. URL https://doi.org/10.1007/s12559-018-9578-5.

T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, volume 29, pages 4349–4357, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE intelligent systems*, 28(2):55–63, mar 2013. URL https://www.computer.org/csdl/magazine/ex/2013/02/mex2013020055/13rRUxAAT3i.

C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, and

P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, 2018a. CEUR-WS. URL http://ceur-ws.org/Vol-2263/paper010.pdf.

C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018b. URL http://ceur-ws.org/Vol-2263/paper010.pdf.

A. Bowes and A. Katz. When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal*, 48(4):215–236, 2011. URL https://doi.org/10.1080/0163853X.2010.532757.

C. F. Boxer and T. E. Ford. Sexist humor in the workplace: A case of subtle harassment. In J. Greenberg, editor, *Insidious Workplace Behavior*, pages 175–205. Routledge/Taylor & Francis Group, 2010. URL https://psycnet.apa.org/record/2010-11517-006.

D. Bray and V. Cerf. The unfinished work of the internet. *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*, page 403, 2019. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198843498.001.0001/oso-9780198843498-chapter-25.

L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674. Association for Computational Linguistics, Nov 2019. URL https://aclanthology.org/D19-1176.

A. Brown. What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3):297–326, 2018. URL https://doi.org/10.1177/1468796817709846.

R. O. Bueno, F. M. Rangel-Pardo, D. I. H. Farías, P. Rosso, M. Montes-y-Gómez, and J. Medina-Pagola. Overview of the task on irony detection in Spanish variants. In M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, and A. Rosá, editors, *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages

229–256. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2421/IroSvA_overview.pdf.

P. Burnap and M. L. Williams. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242, 2015. URL https://doi.org/10.1002/poi3.85.

K. Buschmeier, P. Cimiano, and R. Klinger. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL https://aclanthology.org/W14-2608.

Y. Cai, H. Cai, and X. Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1239.

R. Calvin, K. Winters, S. B. Wyatt, D. R. Williams, F. C. Henderson, and E. R. Walker. Racism and cardiovascular disease in African Americans. *The American journal of the medical sciences*, 325(6):315–331, 2003. URL https://pubmed.ncbi.nlm.nih.gov/12811228/.

E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80, 2017. URL https://doi.org/10.1109/MIS.2017.4531228.

A. T. E. Capozzi, M. Lai, V. Basile, C. Musto, M. Polignano, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, G. Semeraro, and M. Stranisci. Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2481/paper14.pdf.

D. Card, J. Gross, A. Boydstun, and N. A. Smith. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1148.

C. Cardellino. Spanish Billion Words Corpus and Embeddings, August 2016. URL https://crscardellino.github.io/SBWCE/.

M. Á. Á. Carmona, E. Guzmán-Falcón, M. Montes-y-Gómez, H. J. Escalante, L. V. Pineda, V. Reyes-Meza, and A. R. Sulayes. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings* *of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf.

X. Carreras, I. Chao, L. Padró, and M. Padró. FreeLing: An open-source suite of language analyzers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 239–242, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/271.pdf.

T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, and M. Granitzer. I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.760.

G. Castellucci, D. Croce, and R. Basili. A language independent method for generating large scale polarity lexicons. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, pages 38–45. European Language Resources Association (ELRA), 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/summaries/449.html.

S. Castro, L. Chiruzzo, and A. Rosá. Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 187–194. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2150/overview-HAHA.pdf.

S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4619–4629. Association for Computational Linguistics, 2019. URL https://doi.org/10.18653/v1/p19-1455.

D. S. Chauhan, D. S R, A. Ekbal, and P. Bhattacharyya. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 4351–4360, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.401.

L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. J. Prada, and A. Rosá. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, pages 132–144. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2421/HAHA_overview.pdf.

L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. Meaney, and R. Mihalcea. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67:257–268, 2021. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/viewFile/6394/3814.

F. Chiusaroli. "Scritture brevi" nel diasistema delle scritture digitali. *CLUB Working Papers in Linguistics*, pages 5–18, 2017. URL http://corpora.ficlit.unibo.it/CLUB/index.php?slab=club-wpl.

A. T. Cignarella, C. Bosco, and V. Patti. TWITTIRÒ: A social media corpus with a multi-layered annotation for irony. In R. Basili, M. Nissim, and G. Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS, 2017. URL http://ceur-ws.org/Vol-2006/paper070.pdf.

A. T. Cignarella, C. Bosco, V. Patti, and M. Lai. Application and analysis of a multi-layered scheme for irony on the Italian Twitter corpus TWITTIRÒ. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018a. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1664.

A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, and P. Rosso. Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA). In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS, 2018b. URL http://ceur-ws.org/Vol-2263/paper005.pdf.

A. T. Cignarella, C. Bosco, and P. Rosso. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, Aug. 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-7723.

A. T. Cignarella, V. Basile, M. Sanguinetti, C. Bosco, P. Rosso, and F. Benamara. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.116.

A. Cimino, L. De Mattei, and F. Dell'Orletta. Multi-task learning in deep neural networks at EVALITA 2018. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018. URL https://pdfs.semanticscholar.org/5d6c/8a971091a9f3348157066884104de6c5f7ed.pdf.

A. Clark, C. Fox, and S. Lappin. *The Handbook of Computational Linguistics and Natural Language Processing*. Blackwell Handbooks in Linguistics. Wiley, 2013. URL https://books.google.es/books?id=zBmom42eWPcC.

H. H. Clark and R. J. Gerrig. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121—-126, 1984. URL https://doi.org/10.1037/0096-3445.113.1.121.

I. Clarke and J. Grieve. Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3001.

H. L. Colston. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse processes*, 23(1):25–45, 1997. URL https://doi.org/10.1080/01638539709544980.

G. Comandini and V. Patti. An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, Florence, Italy, Aug. 2019a. Association for Computational Linguistics. URL https://aclanthology.org/W19-3518.

G. Comandini and V. Patti. An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, Florence, Italy, Aug. 2019b. Association for Computational Linguistics. URL https://aclanthology.org/W19-3518.

G. Comandini, M. Speranza, and B. Magnini. Effective communication without verbs? Sure! Identification of nominal utterances in Italian social media texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2253/paper31.pdf.

M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22, 2020. URL https://doi.org/10.1145/3377323.

K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Feminist Legal Theory: Readings in Law and Gender*, pages 57–80, 1991. URL https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1052&context=uclf.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, volume 11, pages 512–515. AAAI Press, 2017. URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665.

C. Davies. *Jokes and targets*. Indiana University Press, 2011. URL https://id.erudit.org/iderudit/1026156ar.

M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, and G. Alor-Hernández. Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33, 2017. URL https://doi.org/10.1016/j.knosys.2017.04.009.

R. Delgado. First Amendment formalism is giving way to First Amendment legal realism. *Harvard Civil Rights–Civil Liberties Law Review*, 29:169–174, 1994. URL https://heinonline.org/HOL/LandingPage?handle=hein.journals/hcrcl29&div=12&id=&page=.

R. Delmonte. ITGETARUNS a linguistic rule-based system for pragmatic text processing. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 : 9-11 December 2014, Pisa*, pages 64–69. Pisa University Press, 2014. URL http://digital.casalini.it/3044389.

S. Dev, T. Li, J. M. Phillips, and V. Srikumar. On measuring and mitigating biased inferences of word embeddings. *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 34(05):7659–7666, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6267.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019a. URL https://doi.org/10.18653/v1/n19-1423.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019b. URL https://doi.org/10.18653/v1/n19-1423.

S. Dews and E. Winner. Muting the meaning a social function of irony. *Metaphor and Symbolic Activity*, 10(1):3–19, 1995. URL https://doi.org/10.1207/s15327868ms1001_2.

E. Di Rosa and A. Durante. Irony detection in tweets: X2Check at IronITA 2018. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, 2018. CEUR-WS. URL http://ceur-ws.org/Vol-2263/paper025.pdf.

M. J. Díaz-Torres, P. A. Morán-Méndez, L. V. Pineda, M. Montes-y-Gómez, J. Aguilera, and L. Meneses-Lerín. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 132–136. European Language Resources Association (ELRA), 2020. URL https://aclanthology.org/2020.trac-1.21/.

K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Workshops*. AAAI, 2011. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841.

K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. W. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3):1–30, 2012. URL https://doi.org/10.1145/2362394.2362400.

T. L. Dixon and C. L. Williams. The changing misrepresentation of race and crime on network and cable news. *Journal of Communication*, 65(1):24–39, 12 2015. URL https://doi.org/10.1111/jcom.12133.

S. Douglass, S. Mirpuri, D. English, and T. Yip. "they were just making jokes": Ethnic/racial teasing and discrimination among adolescents. *Cultural Diversity and Ethnic Minority Psychology*, 22(1):69–82, 2016. URL https://doi.org/10.1037/cdp0000041.

J. F. Dovidio, M. Hewstone, P. Glick, and V. M. Esses. Chapter 1: Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, pages 3–28. SAGE Publications, 2010. URL http://dx.doi.org/10.4135/9781446200919.n1.

C. C. Du Marsais, J. Paulhan, and C. Mouchard. *Traité des tropes*. Le Nouveau Commerce, 1981. URL https://books.google.it/books?id=LgqFnAEACAAJ.

J. Dubois, F. Edeline, J.-M. Klinkenberg, P. Minguet, F. Pire, and H. Trinon. *Rhétorique générale*, volume 7 of *Langue et langage*. Larousse, 1970. URL https://books.google.it/books?id=F1yBQgAACAAJ.

E. Fersini, D. Nozza, and P. Rosso. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018a. URL http://ceur-ws.org/Vol-2263/paper009.pdf.

E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at IberEval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS, 2018b. URL http://ceur-ws.org/Vol-2150/overview-AMI.pdf.

E. Fersini, D. Nozza, and P. Rosso. AMI @ EVALITA2020: Automatic misogyny identification. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper161.pdf.

S. T. Fiske. Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, 2(4):357–411, 1998. URL https://psycnet.apa.org/record/1998-07091-025.

S. T. Fiske, A. J. 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. URL https://doi.org/10.1037/0022-3514.82.6.878.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. URL https://doi.org/10.1037/h0031619.

A. Fokkens, N. Ruigrok, C. Beukeboom, G. Sarah, and W. van Atteveldt. Studying Muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1590.

T. E. Ford, E. R. Wentzel, and J. Lorion. Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. *European Journal of Social Psychology*, 31(6): 677–691, 2001. URL https://doi.org/10.1002/ejsp.56.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30, 2018. URL https://doi.org/10.1145/3232676.

J. Fox and W. Y. Tang. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Computers in Human Behavior*, 33:314–320, 2014. URL https://doi.org/10.1016/j.chb.2013.07.014.

J. Fox, C. Cruz, and J. Y. Lee. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, 52:436–442, 2015. URL https://doi.org/10.1016/j.chb.2015.06.024.

A. Foxman and C. Wolf. *Viral Hate: Containing Its Spread on the Internet.* St. Martin's Publishing Group, 2013. URL https://books.google.es/books?id=yVBeWQ61I94C.

C. Francesconi, C. Bosco, F. Poletto, and M. Sanguinetti. Error analysis in a hate speech detection task: The case of HaSpeeDe-TW at EVALITA 2018. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481, 2019. URL https://core.ac.uk/download/pdf/302359305.pdf.

S. Frenda, A. T. Cignarella, M. A. Stranisci, M. Lai, C. Bosco, and V. Patti. Recognizing hate with NLP: The teaching experience of the #DeactivHate Lab in Italian high schools. In E. Fersini, M. Passarotti, and V. Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS, 2021. URL http://ceur-ws.org/Vol-3033/paper35.pdf.

R. Fulper, G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe. Misogynistic language on Twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, 2014. URL https://figshare.com/articles/journal_contribution/Misogynistic_Language_on_Twitter_and_Sexual_Violence/1291081.

B. Gambäck and U. K. Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3013.

G. Gambino and R. Pirrone. CHILab @ HaSpeeDe 2: Enhancing hate speech detection with Part-of-Speech tagging. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper175.pdf.

W. A. Gamson and A. Modigliani. The changing culture of affirmative action. *Equal employment opportunity: labor market discrimination and public policy*, pages 373–394, 1994. URL https://books.google.it/books?id=Es0ecSl0H4YC&printsec=frontcover&hl=it#v=onepage&q&f=false.

B. M. Garavelli. *Manuale di retorica*. Saggi tascabili. Bompiani, 1997. URL https://books.google.it/books?id=FUyIQQAACAAJ.

J. A. García-Díaz, M. Cánovas-García, R. C. Palacios, and R. Valencia-García. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518, 2021. URL https://doi.org/10.1016/j.future.2020.08.032.

J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518, 2021. URL https://www.sciencedirect.com/science/article/pii/S0167739X20301928.

B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, and P. Rosso. IDAT at FIRE2019: overview of the track on irony detection in arabic tweets. In P. Majumder, M. Mitra, S. Gangopadhyay, and P. Mehta, editors, *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE '19), Kolkata, India, December, 2019*, pages 10–13. ACM, 2019. URL https://doi.org/10.1145/3368567.3368585.

A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. A. Barnden, and A. Reyes. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In D. M. Cer, D. Jurgens, P. Nakov, and T. Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 470–478. The Association for Computer Linguistics, 2015. URL https://doi.org/10.18653/v1/s15-2080.

D. Ghosh, A. Vajpayee, and S. Muresan. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.figlang-1.1.

R. W. Gibbs. Irony in talk among friends. *Metaphor and Symbol*, 15(1-2):5–27, 2000. URL https://doi.org/10.1080/10926488.2000.9678862.

R. Giora. On irony and negation. *Discourse Processes*, 19(2):239–264, 1995. URL https://doi.org/10.1080/01638539509544916.

R. Giora, A. Drucker, O. Fein, and I. Mendelson. Default sarcastic interpretations: On the priority of nonsalient interpretations. *Discourse Processes*, 52(3):173–200, 2015a. URL https://doi.org/10.1080/0163853X.2014.954951.

R. Giora, S. Givoni, and O. Fein. Defaultness reigns: The case of sarcasm. *Metaphor and Symbol*, 30(4):290–313, 2015b. URL https://doi.org/10.1080/10926488.2015.1074804.

R. Giora, I. Jaffe, I. Becker, and O. Fein. Strongly attenuating highly positive concepts. The case of default sarcastic interpretations. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, 16(1):19–47, 2018. URL https://doi.org/10.1075/rcl.00002.gio.

N. D. Gitari, Z. Zuping, H. Damien, and J. Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10 (4):215–230, 2015. URL https://gvpress.com/journals/IJMUE/vol10_no4/21.pdf.

V. Giudice. Aspie96 at IronITA (EVALITA 2018): Irony detection in Italian tweets with character-level convolutional RNN. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2263/paper026.pdf.

R. González-Ibáñez, S. Muresan, and N. Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-2102.

M. Graff, S. Miranda-Jiménez, E. S. Tellez, D. Moctezuma, V. Salgado, J. Ortiz-Bejar, and C. N. Sánchez. INGEOTEC at MEX-A3T: author profiling and aggressiveness analysis in twitter using μtc and evomsa. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 128–133. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2150/MEX-A3T_paper6.pdf.

H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Speech acts*, Syntax and Semantics 3, pages 41–58. Academic Press, 1975. URL https://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf.

A.-M. Hancock. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on Politics*, 5(1):63–79, 2007. URL https://www.scinapse.io/papers/2145382344.

Y. Hao and T. Veale. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650, 2010. URL https://doi.org/10.1007/s11023-010-9211-1.

I. Hare and J. Weinstein. *Extreme speech and democracy.* Oxford University Press, 2009. URL https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199548781.001.0001/acprof-9780199548781.

D. I. Hernández-Farías and P. Rosso. Chapter 7 - irony, sarcasm, and sentiment analysis. In F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, editors, *Sentiment Analysis in Social Networks*, pages 113 – 128. Morgan Kaufmann, Boston, 2017. URL http://www.sciencedirect.com/science/article/pii/B9780128044124000073.

D. I. Hernández-Farías, V. Patti, and P. Rosso. Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24, 2016. URL https://doi.org/10.1145/2930663.

S. Hewitt, T. Tiropanis, and C. Bokhove. The problem of identifying misogynist language on Twitter (and other online social spaces). In W. Nejdl, W. Hall, P. Parigi, and S. Staab, editors, *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 333–335. ACM, 2016. URL https://doi.org/10.1145/2908131.2908183.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *ACL*. Association for Computational Linguistics, 2018. URL http://arxiv.org/abs/1801.06146.

W.-C. Hwang and S. Goto. The impact of perceived racial discrimination on the mental health of Asian American and Latino college students. *Cultural Diversity and Ethnic Minority Psychology*, 14(4):326–335, 2008. URL https://pubmed.ncbi.nlm.nih.gov/18954168/.

E. A. Jane. 'back to the kitchen, cunt': Speaking the unspeakable about online misogyny. *Continuum*, 28(4):558–570, 2014. URL https://doi.org/10.1080/10304312.2014.924479.

G. Jasso and I. V. Meza-Ruíz. Character and word baselines for irony detection in Spanish short texts. *Procesamiento del Lenguaje Natural*, 56:41–48, 2016. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285.

A. Joshi, V. Sharma, and P. Bhattacharyya. Harnessing context incongruity for sarcasm detection. 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, pages 757–762, Beijing, China, July 2015. Association for Computational Linguistics. URL https://aclanthology.org/P15-2124.

A. Joshi, P. Bhattacharyya, and M. J. Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017. URL https://doi.org/10.1145/3124420.

R. Justo, T. C. Corcoran, S. M. Lukin, M. A. Walker, and M. I. Torres. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133, 2014. URL https://doi.org/10.1016/j.knosys.2014.05.021.

J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, and L. H. Belguith. Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 644–650. Association for Computational Linguistics (ACL), 2015. URL https://doi.org/10.3115/v1/p15-2106.

J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 262–272, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1025.

E. Kim and R. Klinger. A survey on sentiment and emotion analysis for computational literary studies. *CoRR*, abs/1808.03137, 2018. URL http://arxiv.org/abs/1808.03137.

D. R. Kinder. Opinion and action in the realm of politics. *The handbook of social psychology*, 2(4):778—-867, 1998. URL https://static1.squarespace.com/static/5d6d2f03677cfc00014c9153/t/5f453c650c0cea7d3c19501a/1598372974785/Kinder+1998.pdf.

D. Küçük and F. Can. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37, 2020. URL https://doi.org/10.1145/3369026.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-2012.

G. Kuipers and B. Van der Ent. The seriousness of ethnic jokes: Ethnic humor and social change in the Netherlands, 1995–2012. *Humor*, 29(4):605–633, 2016. URL https://doi.org/10.1515/humor-2016-0013.

F. Kunneman, C. Liebrecht, M. Van Mulken, and A. Van den Bosch. Signaling sarcasm: From hyperbole to hashtag. *Information Processing & Management*, 51(4):500–509, 2015. URL https://doi.org/10.1016/j.ipm.2014.07.006.

M. Z. Kurdi. *Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax*. Number v. 1 in Cognitive Science. Wiley, 2016. URL https://books.google.es/books?id=-jXvCQAAQBAJ.

S. F. Lambert, K. C. Herman, M. S. Bynum, and N. S. Ialongo. Perceptions of racism and depressive symptoms in African American adolescents: The role of perceived academic and social control. *Journal of youth and adolescence*, 38(4):519–531, 2009. URL https://pubmed.ncbi.nlm.nih.gov/19636725/.

R. A. Lanham. *A Hypertext Handlist of Rhetorical Terms: For Macintosh Computers*. University of California Press, Berkeley, CA, USA, 1996.

N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2):434–443, 2012. URL https://doi.org/10.1016/j.chb.2011.10.014.

A. Laudanna, A. M. Thornton, G. Brown, C. Burani, and L. Marconi. Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. *III giornate internazionali di analisi statistica dei dati testuali*, 1:103–109, 1995. URL https://www.istc.cnr.it/sites/default/files/uploads/jadt95.pdf.

E. Lavergne, R. Saini, G. Kovács, and K. Murphy. TheNorth @ HaSpeeDe 2: BERT-based language model fine-tuning for Italian hate speech detection. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop, EVALITA - December 17th, 2020*, volume 2765, pages 142–147. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper135.pdf.

S. Lazzardi, V. Patti, and P. Rosso. Categorizing misogynistic behaviours in italian, english and spanish tweets. *Procesamiento del Lenguaje Natural*, 66(0):65–76, 2021. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6323.

J. Leader Maynard and S. Benesch. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal*, 9(3), 2016. URL https://digitalcommons.usf.edu/gsp/vol9/iss3/8/.

C. J. Lee and A. N. Katz. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, 13(1):1–15, 1998. URL https://doi.org/10.1207/s15327868ms1301_1.

H. Lee, Y. Yu, and G. Kim. Augmenting data for sarcasm detection with unlabeled conversation context. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 12–17, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.figlang-1.2.

A. Lees, J. Sorensen, and I. Kivlichan. Jigsaw @ AMI and HaSpeeDe2: Fine-tuning a pre-trained comment-domain BERT model. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper130.pdf.

J. Lemmens, I. Markov, and W. Daelemans. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nlp4if-1.2.

B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012. URL https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):1–16, 2019. URL https://doi.org/10.1371/journal.pone.0221152.

K. Manne. *Down girl: The logic of misogyny*. Oxford University Press, 2017. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190604981.001.0001/oso-9780190604981.

A. Marchetti, D. Massaro, and A. Valle. *Non dicevo sul serio. Riflessioni su ironia e psicologia*, volume 290. FrancoAngeli, 2007. URL https://www.francoangeli.it/Ricerca/scheda_libro.aspx?Id=14844.

S. Menini, A. P. Aprosio, and S. Tonelli. Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection. *CoRR*, abs/2103.14916, 2021. URL https://arxiv.org/abs/2103.14916.

S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021. URL https://doi.org/10.1145/3439726.

S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/S16-1003.

S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013a. URL https://doi.org/10.1111/j.1467-8640.2012.00460.x.

S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013b. URL https://doi.org/10.1111/j.1467-8640.2012.00460.x.

D. Molla and A. Joshi. Overview of the 2019 ALTA shared task: Sarcasm target identification. In *Proceedings of The 17th Annual Workshop of the Australasian Language Technology Association*, pages 192–196, Sydney, Australia, 4–6 Dec. 2019. Australasian Language Technology Association. URL https://www.aclweb.org/anthology/U19-1026.

C. Musto, G. Semeraro, M. de Gemmis, and P. Lops. Modeling community behavior through semantic analysis of social data: The Italian hate map experience. In J. Vassileva, J. Blustein, L. Aroyo, and S. K. D'Mello, editors, *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, Halifax, NS, Canada, July 13 - 17, 2016*, pages 307–308. ACM, 2016. URL https://doi.org/10.1145/2930238.2930274.

K. L. Nadal, K. E. Griffin, Y. Wong, S. Hamit, and M. Rasmus. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development*, 92(1):57–66, 2014. URL https://doi.org/10.1002/j.1556-6676.2014.00130.x.

U. Naseem, I. Razzak, P. Eklund, and K. Musial. Towards improved deep contextual embedding for the identification of irony and sarcasm. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–7. IEEE, 2020. URL https://doi.org/10.1109/IJCNN48605.2020.9207237.

D. Nikolaou. Does cyberbullying impact youth suicidal behaviors? *Journal of health economics*, 56:30–46, 2017. URL https://doi.org/10.1016/j.jhealeco.2017.09.009.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153. ACM, 2016. URL https://doi.org/10.1145/2872427.2883062.

D. Nozza, C. Volpetti, and E. Fersini. Unintended bias in misogyny detection. In P. M. Barnaghi, G. Gottlob, Y. Manolopoulos, T. Tzouramanis, and A. Vakali, editors, *2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019*, pages 149–155. ACM, 2019. URL https://doi.org/10.1145/3350546.3352512.

P. J. Ortiz Suárez, L. Romary, and B. Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.156.

E. W. Pamungkas and V. Patti. #NonDicevoSulSerio at SemEval-2018 task 3: Exploiting emojis and affective content for irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 649–654, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/S18-1106.

E. W. Pamungkas, A. T. Cignarella, V. Basile, and V. Patti. 14-ExLab@UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, pages 234–241. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2150/AMI_paper2.pdf.

E. W. Pamungkas, V. Basile, and V. Patti. Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management*, 57 (6):102360, 2020a. URL https://www.sciencedirect.com/science/article/pii/S0306457320308554.

E. W. Pamungkas, V. Basile, and V. Patti. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France, May 2020b. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.765.

H. Pan, Z. Lin, P. Fu, and W. Wang. Modeling the incongruity between sentence snippets for sarcasm detection. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2132–2139. IOS Press, 2020. URL https://doi.org/10.3233/FAIA200337.

P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma. Multi-label categorization of accounts of sexism using a neural framework. 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 1642–1652, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1174.

S. Pelosi, A. Maisto, P. Vitale, and S. Vietri. Mining offensive language on social media. In R. Basili, M. Nissim, and G. Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS, 2017. URL http://ceur-ws.org/Vol-2006/paper038.pdf.

J. Pennington, R. Socher, and C. Manning. GloVe: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014a. Association for Computational Linguistics. URL https://aclanthology.org/D14-1162.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014b. Association for Computational Linguistics. URL https://aclanthology.org/D14-1162.

P. M. Pexman and K. M. Olineck. Does sarcasm always sting? Investigating the impact of ironic insults and ironic compliments. *Discourse Processes*, 33(3):199–217, 2002. URL https://doi.org/10.1207/S15326950DP3303_1.

R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001. URL http://www.jstor.org/stable/27857503.

R. Plutchik and H. Kellerman. *Theories of emotion*, volume 1. Academic Press, 1980. URL https://books.google.it/books?id=TV99AAAAMAAJ.

B. Poland. *Haters: Harassment, abuse, and violence online.* U of Nebraska Press, 2016. URL https://books.google.it/books?id=Jd4nDwAAQBAJ.

F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco. Hate speech annotation: Analysis of an Italian Twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS, 2017. URL http://ceur-ws.org/Vol-2006/paper024.pdf.

F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55:477–523, 2021. URL https://rdcu.be/cCdaB.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian*

Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS, 2019. URL http://ceur-ws.org/Vol-2481/paper57.pdf.

J. P. Posadas-Durán, H. Gómez-Adorno, I. Markov, G. Sidorov, I. Z. Batyrshin, A. F. Gelbukh, and O. Pichardo-Lagunas. Syntactic n-grams as features for the author profiling task. In L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, editors, *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS, 2015. URL http://ceur-ws.org/Vol-1391/136-CR.pdf.

R. A. Potamias, G. Siolas, and A. Stafylopatis. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320, 2020. URL https://doi.org/10.1007/s00521-020-05102-3.

M. Quintilianus. *M. Fabi Quintiliani Institutionis oratoriae liber IX: Introduzione, testo, traduzione e commento a cura di Alberto Cavarzere e Lucio Cristante*. WEIDMANN, 2020. URL https://books.google.it/books?id=LQPWDwAAQBAJ.

V. M. Raghavan, P. Mohana Kumar, R. Sundara Raman, and S. Rajeswari. Emotion and sarcasm identification of posts from Facebook data using a hybrid approach. *ICTACT Journal on Soft Computing*, 7(2):1427–1435, 2017. URL http://ischolar.info/index.php/IJSC/article/view/138286.

I. D. Rangel, G. Sidorov, and S. S. Guerra. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, 29:31–46, 2014. URL https://www.cic.ipn.mx/~sidorov/sel_onomazein_2014_29.pdf.

F. M. Rangel-Pardo, M. Franco-Salvador, and P. Rosso. A low dimensionality representation for language variety identification. In A. F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016, Revised Selected Papers, Part II*, volume 9624 of *Lecture Notes in Computer Science*, pages 156–169. Springer-Verlag, LNCS(9624), 2016. URL https://doi.org/10.1007/978-3-319-75487-1_13.

A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin. Offensive language detection using multi-level classification. In A. Farzindar and V. Keselj, editors, *Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence, Canadian, AI 2010, Ottawa, Canada, May 31 - June 2, 2010. Proceedings*, volume 6085 of *Lecture Notes in Computer Science*, pages 16–27. Springer, 2010. URL https://doi.org/10.1007/978-3-642-13059-5_5.

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. URL http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf.

A. Reyes and P. Rosso. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4):754–760, 2012. URL https://doi.org/10.1016/j.dss.2012.05.027.

A. Reyes and P. Rosso. On the difficulty of automatically detecting irony: Beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614, 2014. URL https://riunet.upv.es/bitstream/handle/10251/40330/kaisFinal.pdf.

A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012. URL https://doi.org/10.1016/j.datak.2012.02.005.

A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in Twitter. *Language resources and evaluation*, 47(1):239–268, 2013. URL https://doi.org/10.1007/s10579-012-9196-x.

E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 704–714. ACL, 2013. URL https://aclanthology.org/D13-1066/.

M. J. Rodriguez Cisnero and R. Ortega Bueno. UO @ HaSpeeDe2: Ensemble model for Italian hate speech detection. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper136.pdf.

F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67(0):195–207, 2021. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

L. Rösner and N. C. Krämer. Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society*, 2(3):1–13, 2016. URL https://doi.org/10.1177/2056305116664220.

K. Russell-Brown. *The Color of Crime*. Critical America. NYU Press, 2009. URL https://books.google.es/books?id=FVNv9iPupkIC.

P. Saha, B. Mathew, P. Goyal, and A. Mukherjee. Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700, 2018. URL http://arxiv.org/abs/1812.06700.

J. Sánchez-Junquera, B. Chulvi, P. Rosso, and S. P. Ponzetto. How do you speak about immigrants? Taxonomy and StereoImmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8), 2021. URL https://www.mdpi.com/2076-3417/11/8/3610.

M. Sanguinetti, C. Bosco, A. Mazzei, A. Lavelli, and F. Tamburini. Annotating Italian social media texts in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 229–239, Pisa,Italy, Sept. 2017. Linköping University Electronic Press. URL https://aclanthology.org/W17-6526.

M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli, and F. Tamburini. PoSTWITA-UD: An Italian Twitter treebank in Universal Dependencies. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018a. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/636.html.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2798–2895, Miyazaki, Japan, May 2018b. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1443.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018c. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1443.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018d. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1443.

M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2765/paper162.pdf.

A. Santilli, D. Croce, and R. Basili. A kernel-based approach for irony and sarcasm detection in Italian. In T. Caselli, N. Novielli, V. Patti, and P. Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, 2018. CEUR-WS. URL http://ceur-ws.org/Vol-2263/paper023.pdf.

S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright. Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, 62(5):578–598, 2020. URL https://doi.org/10.1177%2F1470785320921779.

S. Scamuzzi, M. Belluati, M. Caielli, C. Cepernich, V. Patti, S. Stecca, and G. Tipaldo. Fake news e hate speech. I nodi per un'azione di policy efficace. *Problemi dell'informazione, Rivista quadrimestrale*, 46(1):49–81, 2021. URL https://www.rivisteweb.it/doi/10.1445/100129.

H. Schmid. Probabilistic Part-of-Speech tagging using Decision Trees. In *International Conference on New Methods in Language Processing*, 1994. URL https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-1101.

K. Sentamilselvan, P. Suresh, G. Kamalam, S. Mahendran, and D. Aneri. Detection on sarcasm using machine learning classifiers and rule based approach. *IOP Conference Series: Materials Science and Engineering*, 1055(1):012105, Feb 2021. URL https://doi.org/10.1088/1757-899x/1055/1/012105.

S. Sharifirad, S. Matwin, and J. Duffy. Classification of different types of sexist languages on twitter and the gender footprint on each of the classes. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Vietnam, 2018*. CICLing, 2018. URL https://dalspace.library.dal.ca/bitstream/handle/10222/76331/sharifirad-sima-PhD-CSCI-Aug-2019.pdf?sequence=5.

C. Shelley. The bicoherence theory of situational irony. *Cognitive Science*, 25(5):775–818, 2001. URL https://doi.org/10.1016/S0364-0213(01)00053-2.

G. Sidorov, S. Miranda-Jiménez, F. V. Jiménez, A. F. Gelbukh, N. A. Castro-Sánchez, F. Velasquez, I. Díaz-Rangel, S. S. Guerra, A. Treviño, and J. Gordon. Empirical study of machine learning based approach for opinion mining in tweets. *Advances in Artificial Intelligence - 11th Mexican International Conference on Artificial Intelligence, MICAI*

2012, San Luis Potosí, Mexico, October 27 - November 4, 2012. Revised Selected Papers, Part I, 7629:1–14, 2012. URL https://doi.org/10.1007/978-3-642-37807-2_1.

M. Simi, C. Bosco, and S. Montemagni. Less is more? Towards a reduced inventory of categories for training a parser for the Italian Stanford dependencies. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 83–90, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/818_Paper.pdf.

E. R. Smith. Chapter 13 - social identity and social emotions: Toward new conceptualizations of prejudice. In D. M. Mackie and D. L. Hamilton, editors, *Affect, Cognition and Stereotyping*, pages 297–315. Academic Press, San Diego, 1993. URL https://www.sciencedirect.com/science/article/pii/B978008088579750017X.

C. P. Snow and S. Collini. *The two cultures.* Cambridge University Press, 1998. URL http://hdl.handle.net/2027/heb.03176.0001.001.

A. Søgaard, A. Johannsen, B. Plank, D. Hovy, and H. Martínez Alonso. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. URL https://aclanthology.org/W14-1601.

V. Speranza. Hate speech nei giornali italiani: sviluppo ed analisi di un corpus di titoli annotato. Master's thesis, Department of Foreign Languages, University of Torino, 2018.

D. Sperber and D. Wilson. Irony and the use-mention distinction. *Radical Pragmatics*, pages 295–318, 1981. URL http://www.dan.sperber.fr/wp-content/uploads/IronyAndTheUseMentionDistinction.pdf.

E. Spertus. Smokey: Automatic recognition of hostile messages. In B. Kuipers and B. L. Webber, editors, *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97, IAAI 97, July 27-31, 1997, Providence, Rhode Island, USA*, pages 1058–1065. AAAI Press / The MIT Press, 1997. URL http://www.aaai.org/Library/IAAI/1997/iaai97-209.php.

M. Straka and J. Straková. Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. URL https://aclanthology.org/K17-3009.

M. Stranisci, C. Bosco, D. I. Hernández Farías, and V. Patti. Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2892–2899, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1462.

E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, and G. Ruffo. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132–143, 2016. URL https://doi.org/10.1016/j.knosys.2016.05.035. New Avenues in Knowledge Bases for Natural Language Processing.

F. Tamburini. How "BERTology" changed the state-of-the-art also for Italian NLP. In J. Monti, F. Dell'Orletta, and F. Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS, 2020. URL http://ceur-ws.org/Vol-2769/paper_79.pdf.

M. Taulé, A. Ariza, M. Nofre, E. Amigó, and P. Rosso. Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural*, 67:209–221, 2021. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6390.

Y. Tay, A. T. Luu, S. C. Hui, and J. Su. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/P18-1093.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010. URL https://doi.org/10.1002/asi.21416.

G. Turner. *Ordinary people and the media: The demotic turn.* SAGE Publications, 2010.

X. S. Urizar and I. S. V. Roncal. Elhuyar at TASS 2013. In A. D. Esteban, I. A. Loinaz, and J. V. Román, editors, *Proceedings of the Workshop on Sentiment Analysis (TASS 2013)at XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013)*, pages 143–150, 2013. URL http://www.sepln.org/workshops/tass/2013/papers/tass2013-submission3-Elhuyar.pdf.

A. Utsumi. A unified theory of irony and its computational formalization. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2 of *COLING '96*, page 962–967, USA, 1996. Association for Computational Linguistics. URL https://doi.org/10.3115/993268.993334.

A. Utsumi. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806, 2000. URL https://www.sciencedirect.com/science/article/pii/S0378216699001162.

B. Van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5105.

T. A. Van Dijk. *Racism and the Press*. Routledge, 1991. URL http://www.discourses.org/OldBooks/Teun%20A%20van%20Dijk%20-%20Racism%20and%20the%20Press.pdf.

T. A. Van Dijk. New(s) racism: A discourse analytical approach. *Ethnic minorities and the media*, 37:33–49, 2000. URL http://www.discourses.org/OldArticles/New%28s%29%20racism%20-%20A%20discourse%20analytical%20approach.pdf.

C. Van Hee, E. Lefever, and V. Hoste. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. URL https://aclanthology.org/S18-1005.pdf.

C. Van Hee, E. Lefever, and V. Hoste. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. URL https://aclanthology.org/S18-1005.

B. Vidgen and L. Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32, 12 2021. URL https://doi.org/10.1371/journal.pone.0243300.

F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on Facebook. In A. Armando, R. Baldoni, and R. Focardi, editors, *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, volume 1816 of *CEUR Workshop Proceedings*, pages 86–95. CEUR-WS, 2017. URL http://ceur-ws.org/Vol-1816/paper-09.pdf.

P. Vijayaraghavan, I. Sysoev, S. Vosoughi, and D. Roy. DeepStance at SemEval-2016 task 6: Detecting stance in tweets using character and word-level CNNs. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 413–419, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/S16-1067.

B. C. Wallace. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 43(4):467–483, 2015. URL https://doi.org/10.1007/s10462-012-9392-5.

B. C. Wallace, D. K. Choe, and E. Charniak. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages

1035–1044, Beijing, China, July 2015. Association for Computational Linguistics. URL https://aclanthology.org/P15-1100.

P. A. Wang. #irony or #sarcasm - A quantitative and qualitative study based on Twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation, PACLIC 27, Taipei, Taiwan, November 21-24, 2013*, pages 349–356. National Chengchi University, Taiwan, 2013. URL https://aclanthology.org/Y13-1035/.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N16-2013.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, Aug. 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3012.

S. Weaver. A rhetorical discourse analysis of online anti-muslim and anti-semitic jokes. *Ethnic and Racial Studies*, 36(3):483–499, 2013. URL https://doi.org/10.1080/01419870.2013.734386.

M. Weber. Economy and society: An outline of interpretive sociology. *Bedminster Press*, 3, 1968.

C. M. Whissell. Chapter 5 - The dictionary of affect in language. In R. Plutchik and H. Kellerman, editors, *The Measurement of Emotions*, pages 113–131. Elsevier, 1989. URL https://www.sciencedirect.com/science/article/pii/B9780125587044500116.

M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg. Inducing a lexicon of abusive words – A feature-based approach. 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 1046–1056, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1095.

M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. Detection of abusive language: The problem of biased datasets. 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 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1060.

M. Wiegand, J. Ruppenhofer, and E. Eder. Implicitly abusive language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 576–587, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-main.48.

J. P. Williams. Does mention (or pretense) exhaust the concept of irony? *Journal of Experimental Psychology: General*, 113(1):127—-129, 1984. URL https://doi.org/10.1037/0096-3445.113.1.127.

A. Wimmer. *Ethnic boundary making: Institutions, power, networks.* Oxford University Press, 2013. URL https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199927371.001.0001/acprof-9780199927371.

C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang. THU_NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/S18-1006.

T. Xiong, P. Zhang, H. Zhu, and Y. Yang. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2115–2124. ACM, 2019. URL https://doi.org/10.1145/3308558.3313735.

M. Yoshioka, M. Jang, J. Allan, and N. Kando. Visualizing polarity-based stances of news websites. In D. Albakour, D. P. A. Corney, J. Gonzalo, M. Martinez-Alvarez, B. Poblete, and A. Valochas, editors, *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018*, volume 2079 of *CEUR Workshop Proceedings*, pages 6–8. CEUR-WS, 2018. URL http://ceur-ws.org/Vol-2079/paper2.pdf.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/375c71349b295fbe2dcdca9206f20a06-Paper.pdf.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/S19-2010.

M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings* *of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. URL https://aclanthology.org/2020.semeval-1.188.

G. Zarrella and A. Marsh. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclanthology.org/S16-1074.

S. Zhang, X. Zhang, J. Chan, and P. Rosso. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644, 2019. URL https://doi.org/10.1016/j.ipm.2019.04.006.

Y. Zhou, Z. Chen, and H. Yang. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. URL https://arxiv.org/abs/2011.12870.