HAL
open science

Automatic Hate Speech Detection on Social Media
Patricia Chiril

▶ To cite this version:

HAL Id: tel-03599458
https://theses.hal.science/tel-03599458

Submitted on 7 Mar 2022

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Patricia CHIRIL

Le 16 novembre 2021

Détection Automatique des Messages Haineux sur les Réseaux Sociaux

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Farah BENAMARA et Véronique MORICEAU

Jury
Mme Elena CABRIO, Maître de Conférences, Université Côte d'Azur, HDR, Rapporteure
M. Leon Strømberg DERCZYNSKI, Associate Professor, Université de Copenhague, Rapporteur
Mme Delphine BATTISTELLI, Professeure, Université Paris Nanterre, Examinatrice
Mme Béatrice DAILLE, Professeure, Université de Nantes, Examinatrice
M. Emiliano LORINI, Directeur de Recherche CNRS à l'IRIT, Examinateur
Mme Viviana PATTI, Associate Professor, Université de Turin, Invitée
Mme Farah BENAMARA, Maître de Conférences, Université Toulouse III, HDR, Directrice de thèse
Mme Véronique MORICEAU, Maître de Conférences, Université Toulouse III, Co-directrice de thèse
Mme Marlène COULOMB-GULLY, Professeure émérite, Université Toulouse II, Co-encadrante de thèse

_____ **Résumé** _____

Cette thèse se concentre sur deux objectifs : (I) *la détection des discours haineux* et plus particulièrement (II) *la détection du sexisme* dans les réseaux sociaux.

(I) Le discours de haine et le harcèlement sont très répandus dans la communication en ligne, en raison de la liberté d'expression, de l'anonymat des utilisateurs et de l'absence de réglementation fournie par les réseaux sociaux. Le discours de haine est axé sur des thèmes précis (misogynie, sexisme, racisme, xénophobie, homophobie, etc.) et cible différents groupes en fonction de caractéristiques telles que le sexe (misogynie, sexisme), l'ethnie, la race, la religion (xénophobie, racisme, islamophobie), l'orientation sexuelle (homophobie), etc. La plupart des approches de détection automatique des discours de haine traitent le problème comme une tâche de classification binaire sans tenir compte de leur orientation thématique ou de leur nature ciblée. Dans cette thèse, nous proposons d'aborder, pour la première fois, la détection des discours de haine dans une perspective multi-cibles. Nous utilisons des ensembles de données annotées manuellement, afin d'étudier le problème du transfert de connaissances à partir de différents ensembles de données ayant des centres d'intérêt et cibles différents.

(II) Le sexisme est un type de discours de haine. Il exprime un préjugé ou une discrimination fondée sur le sexe d'une personne. Il est fondé sur la croyance qu'un sexe ou un genre est supérieur à un autre. Nous pensons qu'il est important non seulement de pouvoir détecter automatiquement les messages à contenu sexiste postés sur les réseaux sociaux mais aussi de distinguer les véritables messages sexistes des messages qui relatent ou dénoncent le sexisme. En effet, alors que les messages pourraient être signalés et modérés dans le premier cas comme le recommandent les lois européennes, les messages relatant des expériences de sexisme ne devraient pas être modérés. Dans ce but, nous avons expérimenté différents modèles neuronaux, notamment des modèles permettant de détecter la présence de stéréotypes de genre dans le but d'améliorer la détection des contenus sexistes.

Nos résultats, d'une part, sont encourageants et constituent un premier pas vers la modération automatique des contenus sexistes et, d'autre part, démontrent que la détection multi-cibles des discours haineux à partir des ensembles de données existants, préalablement annotés, est possible.

Institut de Recherche en Informatique de Toulouse - UMR 5505

Université Paul Sabatier, 118 route de Narbonne, 31062 TOULOUSE cedex 4

──────── **Abstract** ────────

This dissertation is focused on two objectives: (I) *Hate Speech detection* and (II) *Sexism detection* in social media.

(I) Hate Speech and harassment are widespread in online communication, due to users' freedom and anonymity and the lack of regulation provided by social media platforms. Hate speech is *topically-focused* (misogyny, sexism, racism, xenophobia, homophobia, etc.) and each specific manifestation of hate speech *targets different vulnerable groups* based on characteristics such as gender (misogyny, sexism), ethnicity, race, religion (xenophobia, racism, Islamophobia), sexual orientation (homophobia), and so on. Most automatic hate speech detection approaches cast the problem into a binary classification task without addressing either the topical focus or the target-oriented nature of hate speech. In this dissertation, we propose to tackle, for the first time, hate speech detection from a multi-target perspective. We leverage manually annotated datasets, to investigate the problem of transferring knowledge from different datasets with different topical focuses and targets.

(II) Sexism is a type of hate speech. It can be defined as prejudice or discrimination based on a person's gender. It is based on the belief that one sex or gender is superior to another. We believe that it is important not only to be able to automatically detect messages with a sexist content but also to distinguish between real sexist messages and messages which relate sexism. Indeed, whereas messages could be reported and moderated in the first case as recommended by European laws, messages relating sexism experiences should not be moderated. We experimented with different neural models, in particular models that are able to detect the presence of gender stereotypes in order to improve sexism detection.

Our results are encouraging and constitute a first step towards automatic sexist content moderation and demonstrate that multi-target hate speech detection from existing datasets is feasible, which is a first step towards hate speech detection for a specific topic/target when dedicated annotated data are missing.

Acknowledgements

Thank you to Monica, although you did not *per se* help me write this dissertation (and will likely never read this), you have always been a great friend, and I appreciate you for that. Sometimes life gets in the way, but in the end we both know we will always have our *sister* to call.

It is a pleasure to thank my friends who have always been a major source of support. To my childhood friends, Alexandra and Lidia, I am grateful for how you have helped me grow over the years. To Kristell, Augustin, JB and Walid, thank you for the wonderful times we shared, especially the Sunday game nights. A warm word for my great friend Walid, the 4th Musketeer of my book club, with whom I had some of the best conversations over bubble tea breaks. To all my friends in Toulouse: Yara, Hadi, Hussein, Joseph, Mohamad and Tohme, thank you guys for always being there for me. You gave me the necessary distractions from my research and made my stay memorable. To the first friend I made at IRIT, Mohamad, I will never forget our late nights in the office, especially the parts in which we were trembling with fear because of the *'ghosts'* hunting that place. You will always be one of the original three Musketeers.

Finally, my deep and sincere gratitude to my family for their unconditional love, help and support. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am. This journey would not have been possible if not for you. You selflessly encouraged me to explore new directions in life and seek my own destiny. Words cannot express how much *I love you!*

In loving memory of my grandparents.

Contents

List of Figures

List of Tables

Introduction

Context and Motivations

Nowadays, people increasingly use social networking sites, not only as their main source of information, but also as media to post content, sharing their feelings and opinions. Social media are convenient, as sites allow users to reach people worldwide, which could potentially facilitate a positive and constructive conversation between users. However, this phenomenon has a downside, as there are more and more episodes of hate speech and harassment in online communication (Burnap and Williams, 2015). This is due especially to the freedom and anonymity given to users and to the lack of effective regulations provided by the social network platforms.

Hate speech may have different topical focuses: misogyny, sexism, racism, xenophobia, homophobia, Islamophobia, etc. which we refer to as *topics*. For each topic, hateful content is directed towards specific *targets* that represent the community (individuals or groups) receiving the hatred. For example, black people and white people are possible targets when the topical focus is *racism* (Silva et al., 2016), while women are the targets when the topical focus is *misogyny* or *sexism* (Manne, 2017). Hate speech is thus, by definition, *target-oriented*, as shown in the following tweets taken from (Davidson et al., 2019; Waseem and Hovy, 2016; Basile et al., 2019), where the targets are underlined.[1] These examples also show that different targets involve different ways of linguistically expressing hateful content such as references to racial or sexist stereotypes, the use of negative and positive emotions, swearing terms, and the presence of other phenomena such as envy and ugliness.[2]

(1)   *Women who are feminist are the ugly bitches who cant find a man for themselves*

(2)   *Islam is 1000 years of contributing nothing to mankind but murder and hatred.*

(3)   *Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack ! Stop Allowing illegals to Abuse the Taxpayer #Immigration*

(4)   *Seattle Mayoral Election this year. A choice between a bunch of women, non-whites, and faggots/fag lovers.*

---

[1]*N.B. In this dissertation we include examples of tweets that use vulgarity, degrading terms and hate speech.*

[2]See (Mathew et al., 2018) for an interesting lexical, linguistic and psycho-linguistic analysis of hateful accounts on Twitter.

The rise of online hatred and fake news have created a media climate that is sometimes hostile to its users. As such, new legislation to better regulate companies owning digital social networks (e.g., Google, Facebook, Twitter, etc.) have been put in place. However, social media networks have also offered a space were women feel brave enough for reporting their experiences (see for example *#meToo* or *#balanceTonPorc*). We argue that within this regulatory framework, standard approaches for hate speech automatic detection may unfortunately moderate denunciations of hateful acts.

Methodology and Contributions

In this dissertation we propose to undertake the following challenges:

- (C1) Experiment with the *development of models able to detect different types of sexism experiences in French tweets.*

- (C2) Investigate whether *gender stereotype detection can improve sexism detection.*

- (C3) Investigate the problem of *transferring knowledge from different datasets with different topical focuses and targets.*

To this end, we propose three main contributions.

From Binary Sexism Classification to the Detection of Sexism Experiences

As far as we are aware, the distinction between reports/denunciations of sexism experience and *'real'* sexist messages has not been addressed. In previous work, sexism detection is casted as a binary classification problem (*sexist* vs. *non-sexist*) or a multi-label classification by identifying the type of sexist behaviours (Jha and Mamidi, 2017; Sharifirad et al., 2018; Fersini et al., 2018c; Karlekar and Bansal, 2018; Parikh et al., 2019). We argue that casting the task of sexism detection as a binary classification problem is not sufficient. We believe that it is important not only to be able to automatically detect messages with a sexist content but also to distinguish between *'real'* sexist messages that target women (cf. (5) and (6)) and messages which relate sexism experiences (cf. (7)). Indeed, whereas messages could be reported and moderated in the first case as recommended by European laws, messages relating sexism experiences should not be moderated.

(5) *The goalkeeper has no merit in stopping this pregnant woman shooting*

(6) *She swims fast for a woman*

(7) *He said "who's gonna take care of your children when you are at ACL?"*

Our contributions include:

(1) A novel characterization of sexist content-force relation inspired by speech acts theory (Austin, 1962) *and discourse studies in gender* (Lazar, 2007; Mills, 2008). In collaboration with Dr. Alda Mari and Dr. Gloria Origgi from Institut Jean Nicod (Paris, France), we created a novel characterization which distinguishes different types of sexist content depending on the impact on the addressee (called *'perlocutionary force'*): sexist hate speech *directly addressed* to a target, sexist *descriptive assertions* not addressed to the target, or *reported assertions* that relate a story of sexism experienced by a woman. Our guiding hypothesis is that indirect acts establish a distancing effect with the reported content and are thus less committal on behalf of the addressee (Giannakidou and Mari, 2021).

(2) The first French dataset of about $12,000$ *tweets annotated for sexism detection* according to this new characterization and that is freely available for the research community.[3] The development of the annotation guidelines for the sexism corpus has been carried out in collaboration with Prof. Marlène Coulomb-Gully from Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales (LERASS, Toulouse, France). The characterization of sexist content, annotation guidelines and dataset description were published at the The 12th Language Resources and Evaluation Conference (LREC) (Chiril et al., 2020a).

(3) A pilot study in which we experiment with the development of models for (i) automatically detecting hate speech towards *two different targets* (immigrants and women) and (ii) automatically detecting sexism *from a multilingual perspective*, namely in English and French tweets. We propose both features-based models (relying on both language-dependent and language independent features) and a neural model to *measure to what extent hate speech detection is target-dependent*. We also experiment with multilingual embeddings by training on one language and testing on the other in order to *measure how the proposed models are language dependent*. This work was published at the French Conference TALN (Chiril et al., 2019a). Part of this work has also been published within the HatEval shared task (Basile et al., 2019)

---

[3]https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets

in The 13th International Workshop on Semantic Evaluation (Chiril et al., 2019b).

(4) The first approach to detect different types of reports/denunciations of sexism experiences in French tweets. We rely on deep learning architectures trained on top of a combination of several tweet's vectorial representations: word embeddings built from different sources (tweets, Wikipedia), complemented with both linguistic features, and various generalization strategies to account for sexist stereotypes and the way sexist contents are linguistically expressed, which is a first step before moving to a real scenario where gender stereotypes are automatically detected. This work was published at The 58th Annual Meeting of the Association for Computational Linguistics (ACL) (Chiril et al., 2020b).

Gender Stereotypes for Sexism Classification

Gender stereotypes defined by the Office of the High Commissioner for Human Rights as *′a generalised view or preconception about attributes, or characteristics that are or ought to be possessed by women and men or the roles that are or should be performed by men and women′* have been widely studied in psychology, communication studies and social science (Allport et al., 1954; Crawford et al., 2002; Beike and Sherman, 2014; Biscarrat et al., 2016). Stereotypes have been shown to represent the root cause of inter-group tensions (e.g., sexism, racism, etc.) (Fiske, 1998). For example, (8) (Fersini et al., 2018b) contains a stereotype based on a woman's place being in the kitchen, while in (9) (Basile et al., 2019), immigrants are perceived as relying heavily on the welfare system, highlighting the racial divide in the American society.

(8)   *Who makes the sandwiches at a feminist rally?*

(9)   *Illegals Cross Border Just in Time to Have #AnchorBabies for Welfare and Medicaid Ripping off U.S Taxpayers #RedNationRising #Trump #MAGA #SendThemBack ASAP*

Although several studies suggest that there is a significant correlation between the usage of stereotypes and hate speech (García-Sánchez et al., 2019; Francesconi et al., 2019), no one has empirically measured the impact of gender stereotype detection for sexist hate speech classification. To this end we:

(5) Present the first dataset annotated for gender stereotype detection. This dataset contains about 9,200 tweets in French annotated according to different stereotype aspects and is freely available for the research community.[4]

---

[4] https://github.com/patriChiril/An-Annotated-Corpus-for-Gender-Stereotype-Det

(6) Conduct a set of experiments designed to detect gender stereotypes and then, to use this prediction for sexism detection. We rely on several deep learning architectures leveraging various sources of linguistic knowledge to account for gender stereotypes and the way sexist contents are expressed in language.

Our results suggest that sexism classification can benefit from gender stereotypes detection. This work has been published in the Findings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Chiril et al., 2021a).

From Sexism Classification to Hate Speech Detection

Most of the existing systems designed for hate speech detection share two common characteristics. First, they are trained to predict the presence of general, target-independent hate speech, without addressing either the *topical focus* or the *target-oriented* nature of hate speech. Second, these systems are built, optimized, and evaluated based on a single dataset (be it *topic-generic* or *topic-specific*). Thus, it has become difficult to measure the generalization power of such systems and, more specifically, their ability to adapt their predictions in the presence of novel or different topics and targets (Yin and Zubiaga, 2021).

To address these final challenges, we propose a novel multi-target hate speech detection approach for handling a new emerging target by leveraging existing manually annotated datasets. This will enable a model to transfer knowledge from different datasets with different topics and targets. In the context of offensive content moderation, identifying the topical focus and the targeted community of hateful contents would be of great interest as it will allow us to detect hate speech for specific topics/targets when dedicated data are missing.

Our contribution is threefold:

(7) We explore the ability of hate speech detection models to capture common properties from generic hate speech datasets and to transfer this knowledge to recognize specific manifestations of hate.

(8) We experiment with the development of models for detecting both the topics (racism, xenophobia, sexism, misogyny) and the targets (gender, ethnicity) of hate speech going beyond standard binary classification. We investigate (a) *how to detect hate speech at a finer level of granularity* and (b) *how to transfer knowledge across different types of hate speech.* We rely on multiple topic-specific datasets and develop, in addition to the deep learning models designed to address

`ection-in-French-Tweets`

point *(7)*, a multitask architecture that has been shown to be quite effective in cross-domain sentiment analysis (Zhang et al., 2019; Cai and Wan, 2019).

(9) We study the impact of affective semantic resources in determining specific manifestations of hate speech. In this work, we also want to explore the affective characteristics of the language used in hate speech, continuing the very recent work by Rajamanickam et al. (2020), which suggests a strong relationship between abusive behavior and the emotional state of the speaker. We experiment with three affect resources as extra-features on top of several deep learning architectures: sentic computing (Cambria and Hussain, 2015) resources (SenticNet (Cambria et al., 2018), EmoSenticNet (Poria et al., 2013)) and semantically structured hate lexicons (HurtLex (Bassignana et al., 2018)). SenticNet has not, to the best of our knowledge, been used in hate speech detection. For each resource, we propose a systematic evaluation of the emotional categories that are the most productive for our tasks.

Our results show that multi-target hate speech detection from existing datasets is feasible, which is a first step towards hate speech detection for a specific topic/target when dedicated annotated data are missing. Moreover, we prove that domain-independent affective knowledge, injected into our models, helps finer-grained hate speech detection.

This work has been carried out in collaboration with Dr. Viviana Patti and Endang Wahyu Pamungkas from the University of Turin (Turin, Italy) and was published in the Cognitive Computation Journal (A Decade of Sentic Computing) (Chiril et al., 2021b).

Dissertation Outline

The dissertation is organized in four parts that can be read independently from each other, and each part focuses on one of the aforementioned contributions.

As one of the critical aspects that arises when discussing hate speech lies in its definition (although widely used, there is no agreement on its meaning and scope), in Part I we will examine the concept of hate speech through definitions employed by either international organizations or scholars by considering the abounding elements that are intertwined. In this part we also present an overview of the main works on hate speech and sexism detection.

In Part II we detail the data, the characterization of sexism content we propose and the annotation scheme. We then present the experiments that were carried out for detecting

sexist contents, as well as the pilot study in which we investigate whether the models that were developed are capable of detecting target agnostic hate speech.

In Part III we focus on the detection of sexist hate speech against women in tweets, studying for the first time the impact of gender stereotype detection on sexism classification. We begin this part by detailing the data and the annotation process of the first dataset annotated for gender stereotype detection, and then present the experiments that were carried out.

In Part IV we tackle, for the first time, hate speech detection from a multi-target perspective. We begin this part by presenting an overview of the main works on hate speech detection, and then we present the experiments carried out for investigating the problem of transferring knowledge from different datasets with different topical focuses and targets.

Finally, we provide an overview of this work and emphasise its contributions and limitations. We highlight ethical issues, potential applications, as well as our perspectives for future work.

Part I

Hate Speech Detection in Online Communication

1 What is Hate Speech?

Hate Speech (HS hereafter) and harassment are widespread in online communication, due to users' freedom and anonymity and the lack of regulation provided by social media platforms. We begin this chapter by analyzing different definitions employed by international organizations. Then, we examine the different topical focuses of hate speech (e.g., *misogyny*, *sexism*, *racism*, *xenophobia*, *homophobia*, etc.), as well as other related concepts (e.g., *abusive*, *offensive* or *aggressive* language). We continue our discussion in Chapter 2 and 3 where we provide, respectively: an overview of the main works on HS detection and sexism detection, including the available corpora and approaches employed for these tasks.

1.1 Legal Definitions

The Council of Europe, an international intergovernmental organisation deeply involved in the fight against HS through a variety of initiatives,[5] is the first and only institution to have adopted an official definition of HS. An exhaustive definition of HS was published by the European Commission against Racism and Intolerance (ECRI, 2015) because of an increased concern related to the spread of HS in Europe and the negative effects on the society:

---

[5] https://www.coe.int/en/web/no-hate-campaign/coe-work-on-hate-speech

> " *[...] hate speech is to be understood [...] as 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 in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of 'race',[a] colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status. [...] the Recommendation specifically excludes from the definition of hate speech any form of expression – such as satire or objectively based news reporting and analysis - that merely offends, hurts or distresses.* "

[a]ECRI rejects theories based on the existence of different races (i.e., all humans belong to the same race). In this case, the term 'race' is used to ensure that the persons who are generally (and erroneously) perceived as belonging to another race are not excluded from the protection provided under this definition.

This definition outlines HS in exhaustive detail and it provides clarification concerning the individual facets related to it. By adopting this definition and recognizing that HS can be based on manifestations not listed in its characterization, the cases to which the concept can be applied is hereby broadened. Another relevant elucidation in the context of this study pertains to the definition of *expression* in which the use of new technologies is included as a possible catalyst for hateful messages:

> " *'Expression' is understood [...] to cover speech and publications in any form, including through the use of electronic media, as well as their dissemination and storage. Hate speech can take the form of written or spoken words, or other forms such as pictures, signs, symbols, paintings, music, plays or videos. It also embraces the use of particular conduct, such as gestures, to communicate an idea, message or opinion.* "

Despite having to deal with the problem of HS several times, the European Court of Human Rights (ECHR) refrained from providing any clarification regarding the boundaries of

the term, and rather adopted the definition provided by the Council of Europe,[6] where HS was described as covering *'all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism 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'*. This comes as a result of ECHR preferring to keep a flexible framework that could be more easily adapted to the evolution of the HS phenomenon.

According to the Universal Declaration of Human Rights:[7]

> *"*
> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.*
> *"*

As a form of expression, HS ineluctably clashes with one of the main rights on which the European Union is built on: *freedom of expression*; and by steering towards discrimination and/or violence, it also contradicts the European values of respect and tolerance. As evidence indicates that free speech often results in hateful speech, in order to solve the conflict between the rights of an individual employing HS against the same rights of the others, ECRI affirms that:

> *"*
> *freedom of expression and opinion is not an unqualified right and that it must not be exercised in a manner inconsistent with the rights of others*
> *"*

As a direct consequence of having to assure all rights for all citizens, some restrictions to freedom of expression need to be applied in order to guarantee the respect of human dignity by setting in place different tools capable of countering this problem. The massive growth of user generated web content, along with the interactivity and anonymity the internet provides, poses many obstacles for the regulation of hateful content in the cyberspace.

---

[6]Recommendation No. R (97) 20 of the Committee of Ministers to Member States on 'Hate Speech': https://rm.coe.int/1680505d5b

[7]https://www.un.org/en/about-us/universal-declaration-of-human-rights

Moreover, as Internet Service Providers and Web-Hosting Services have a key responsibility in keeping their platforms safe (Cohen-Almagor, 2017), the European Commission has put in place a Code of Conduct for countering illegal HS online. Under this policy, major IT companies[8] committed to having in place *'clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content'* as well as *'providing Rules or Community Guidelines clarifying that they prohibit the promotion of incitement to violence and hateful conduct'* (Jourová, 2016). A set of HS definitions included by different social media platforms (Twitter,[9] Youtube,[10] Facebook,[11] Instagram[12]) in their conduct policies[13] are presented in Table 1.1.

Given the rapid development of technology and (in particular) the impact of online platforms on the society, the Digital Services Act,[14] a new legislation to better regulate tech giants, with particular focus on data management, disinformation and HS was enacted. This resolution seeks to protect users' fundamental rights online (including the freedom of speech) by setting in place a set of rules aimed at establishing a higher standard of *'fairness, transparency and accountability on how the providers of such platforms moderate content, on online advertising and on algorithmic processes'*. To assure that the rights of everyone are respected and that the internet is not diverted from its intended purpose for illicit ones, the French government (France being one of the many countries that adhered to the Digital Service Act) has made available a portal for reporting illegal contents or behaviours that one may have encountered while using the Internet. The latest attempt to moderate hateful content on the Internet, a bill introduced by the deputy Laetitia Avia, aimed to strengthen the contribution of digital operators in the fight against online HS. The key requirement of the law was to remove *'manifestly illegal'* HS and a broad range of other types of content within 24 hours of notice and a possibility of fines for *'systemic failure to cooperate with the authorities'*. However, the French Constitutional Court deemed this deadline as being too

---

[8]https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[9]https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[10]https://support.google.com/youtube/answer/2801939?hl=en

[11]https://www.facebook.com/communitystandards/hate_speech

[12]https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram

[13]Note that in this study we only included definitions adopted by platforms that are widely studied in the NLP literature for the task of HS detection.

[14]https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

Patricia Chiril

Table 1.1 – Comparison of HS policies across different social media platforms.

Platform	Hate Speech policy
	You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.
	Don't post content on YouTube if the purpose of that content is to incite hatred or encourage violence against individuals or groups based on the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status. We don't allow threats on YouTube, and we treat implied calls for violence as real threats.
	We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics such as occupation, when they're referenced along with a protected characteristic. We recognise that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If intention is unclear, we may remove content.

short and pointed out that these obligations could *'encourage the operators of online platforms to remove the content that is reported to them, whether or not they are clearly illegal'* at the expense of freedom of expression.

1.2 Types of Hate Speech

In spite of no universally accepted definition of HS and the way it differs from offensive language, there are some common elements that seem to arise. In particular, these messages may express threats, harassment, intimidation or *'disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic'* (Nockleby, 2000). As such, studies deal with different areas of online HS.

The most important research topics in the HS literature are described in the following.

1.2.1 Ethnicity-based Hate Speech

Racism. Despite its ubiquity in everyday language, racism is an important issue which is not easily defined in the scientific literature, as racist ideas can be expressed in numerous ways. Clark et al. (1999) define racism as *'beliefs, attitudes, institutional arrangements, and acts that tend to denigrate individuals or groups because of phenotypic characteristics or ethnic group affiliation'*. More recent studies show that racism is no longer strictly limited to physical or ethnic attributes and expand this definition by including insults, negative utterances and negative generalizations concerning social and cultural aspects (Tulkens et al., 2016a).

While there are many forms of HS, Silva et al. (2016) show that the most prevalent one is racism, with HS related to behavioural and physical aspects coming in second and third place, respectively. Several studies consider this social phenomenon as a key factor in causing unfair and avoidable inequalities in terms of power, resources and opportunities across racial or ethnic groups.

Racism can be expressed through negative and inaccurate stereotypes (one-word epithets, phrases, concepts, metaphors and juxtapositions), prejudice or discrimination and as a form of oppression, it is intrinsically linked to privilege, which results in providing unfair opportunities to dominant social groups (e.g., whites) (Berman and Paradies, 2010).

According to Berman and Paradies (2010), in practice, racism co-occurs at three conceptual levels which contribute to maintaining or amplifying the inequity in the distribution of opportunity across social groups:

- internalized racism (i.e., attitudes, beliefs or ideologies);

- interpersonal racism (i.e., human interactions);

- systemic/institutional racism (i.e., the production and allocation of resources within society).

However, this generalization included in its definition (i.e., different phenotypes or ethnic group affiliations) fails to account for the particularities of individual types of racism, which differ not only in the way they are conveyed, but also in their historical significance (e.g., anti-Semitism, Islamophobia, xenophobia, racism against African Americans, etc.).

Xenophobia. The Latin term xenophobia denotes the fear of foreigners, and although hostility towards outsiders is often a reaction to fear, Sundstrom and Kim (2014) argue that adhering to this etymology is insufficient and misleading as it conceals other affects (e.g., envy, resentment) associated with this phenomenon. As such, xenophobia is rooted in civic ostracism (i.e., beliefs, attitudes and affects about social exclusion), this discerning it from racism.

Although most migration is intra-continental, since 2015, with the so-called 'refugees and migrant crisis', the number of asylum seekers has significantly increased in Europe, migration becoming more diverse in terms of origin of migrants.[15] As a result, this phenomenon stimulated an increase in the number of hate crimes targeting migrants and refugees, making the development of tools for the identification of xenophobic behaviour extremely useful (Bosco et al., 2017; Basile et al., 2019). Moreover, the concerns raised by society's attitude regarding immigration, immigrant integration and social integration resulted in the development of European policies for effectively integrating migrants into their new societies (OECD, 2018). European Commission efforts for combating and preventing online HS against migrants and refugees include the development of campaigns[16] for building counter-narratives on migration. Nonetheless, the negative attitude towards Islam go back to before the 'refugees and migrant crisis'. For example, a report presented by the European Monitoring Centre on Racism and Xenophobia (EUMC)[17] shows that Islamic communities (and other vulnerable groups) have become targets of increased hostility in the wake of the the terrorist attacks in the United States on 11 September 2001.

Islamophobia. Despite the vast amount of research, there is a lack of terminological consensus amongst academics as to the core features of islamophobia (e.g., different theoretical concerns, political, geographical, and historical contexts). Another source of ambiguity is that in many studies, the term is not defined at all. Consequently, the readers might have a divergent understanding of the phenomenon, which in turn, might result in an increased difficulty in interpreting/synthesizing findings. The All-Party Parliamentary Group (APPG) on British Muslims[18] argue that a working definition is vital for taking the appropriate steps

---

[15]https://migrationdataportal.org/regional-data-overview/europe
[16]http://www.silencehate.eu/about-the-project/
[17]https://fra.europa.eu/sites/default/files/fra_uploads/199-Synthesis-report_en.pdf
[18]https://static1.squarespace.com/static/599c3d2febbd1a90cffdd8a9/t/5bfd1ea3352f531a6170ceee/1543315109493/Islamophobia+Defined.pdf

in response to inequalities faced by Muslim citizens and its absence has resulted in Islamophobia being overlooked in policy initiatives. However, the definition proposed by APPG (i.e., *'Islamophobia is rooted in racism and is a type of racism that targets expressions of Muslimness or perceived Muslimness'*) created confusion and was deemed as not being sufficient (due to its broadness) as it had the potential of limiting freedom of speech. As race and culture are closely intertwined with religion, Islamophobic HS must involve an attack against the religious identity and additionally it can also include a racial or cultural component. As such, a new definition was proposed:[19]

> *"*
> *A fear, prejudice and hatred [...] that leads to provocation, hostility and intolerance by means of threatening, harassment, abuse, incitement and intimidation [...] motivated by institutional, ideological, political and religious hostility that transcends into structural and cultural racism which targets the symbols and markers of being a Muslim.* *"*

Anti-semitism. According to the International Holocaust Remembrance Alliance,[20] an organization focused only on Holocaust-related issues,

> *"*
> *anti-Semitism is a certain perception of Jews, which may be expressed as hatred toward Jews. Rhetorical and physical manifestations of anti-Semitism are directed toward Jewish or non-Jewish individuals and/or their property, toward Jewish community institutions and religious facilities.* *"*

This phenomenon has been widely studied in the social science literature, Brustein and King (2004) stating that the *'Jew hatred is more multifaceted than other kinds of prejudice'*, this making anti-Semitism different from other forms of xenophobia. In addition to racial based discrimination, anti-Semitism also incorporates:

---

[19] https://www.ohchr.org/Documents/Issues/Religion/Islamophobia-AntiMuslim/Civil%20Society%20or%20Individuals/ProfAwan-2.pdf

[20] https://www.holocaustremembrance.com/resources/working-definitions-charters/working-definition-antisemitism?focus=antisemitismandholocaustdenial

- religious (Christian anti-Semitism (i.e., antipathy towards practices of Judaism) and Islamic anti-Semitism (often denied and illustrated as political polemic revolving around the Israeli-Palestinian conflict)),

- economic (from Judas to Rothschild, Jews are seen as wealthy and greedy people that control the business world and use their power for their own benefit), and

- political (e.g., polemic in connection with the Israel-Palestine conflict, *'Protocols of the Elders of Zion'* and Jewish world supremacy) prejudice.

Although both anti-Semitism and Islamophobia position Jews and Muslims as threatening outsiders (e.g., controlling global banks vs. 'stealing' jobs and 'burdening' welfare), Fastenbauer (2020) argues that there is a substantial difference in between Islamophobia and anti-Semitism. While the former is a type of xenophobia *'populistically stirred up by the extreme right wing'*, the reasons and the history of development of the latter are much more complex. In addition, anti-Semitism has an *'eliminatory character'* as it was driven by a desire for racial purity.

1.2.2 Gender-based Hate Speech

Sexism can be defined as prejudice or discrimination based on a person's gender. It is based on the belief that one sex or gender is superior to another. It can take several forms from sexist remarks, gestures, behaviours, practices, insults to rape or murder. Sexist HS is a message of inferiority directed against a historically oppressed group (usually directed against women at least in part because they are women), that is persecutory, hateful and degrading (Langton, 2012), some authors referring to it as: *'words that wound'* (Matsuda et al., 1993; Waldron, 2012; Delgado et al., 2015).

According to the Council of Europe:[21]

> *"*
> *The aim of sexist HS is to humiliate or objectify women, to undervalue their skills and opinions, to destroy their reputation, to make them feel vulnerable and fearful, and to control and punish them for not following a certain behaviour.*
> *"*

---

[21] https://rm.coe.int/1680651592

As such, its psychological, emotional and/or physical impacts can be severe.

Although in some countries HS is legally protected (e.g., in the United States HS is protected under the First Amendment as freedom of expression (Massaro, 1990)), many other countries have laws prohibiting it. For instance, for the five-year period mandate of French president Emmanuel Macron, gender equality has been declared a *major national cause*[22] and since the French law of 27 January 2017 related to equality and citizenship,[23] penalties due to discrimination are doubled (sexism being now considered an aggravating factor). Moreover, the High Council for gender equality (HCEfh)[24] is asked to make available an annual report on the state of sexism in France.

Both misogyny and sexism are common occurrences on all social media platforms (from the way the media portrays women, expectations about how and if they should share their opinions, to male dominance, violence and more)[25] and it raises concerns due to the fact that it may discourage or even prevent women from participating in social media. One of the first studies that attempted a manually misogyny detection on Twitter (Hewitt et al., 2016) stated that the misogynist abuse intensifies due to other users joining in the harassment of the targeted user.

Although overall **misogyny** and sexism share the common purpose of maintaining or restoring a patriarchal social order, Manne (2017) illustrates the contrast between the two ideologies. A sexist ideology (which often *'consists of assumptions, beliefs, theories, stereotypes and broader cultural narratives that represent men and women'*) will tend to discriminate between men and women and has the role of justifying these norms via an ideology that involves believing in men's superiority in highly prestigious domains (i.e., represents the *'justificatory'* branch of a patriarchal order). A misogynistic ideology does not necessarily rely on people's beliefs, values, and theories, and can be seen as a mechanism that has the role of upholding the social norms of patriarchies (i.e., represents the *'law enforcement'* branch of a patriarchal order) by differentiating between good women and bad women and punishing those who take (or attempt to take) a man's place in society.

Other phenomena that warrant attention are **homophobia** and **transphobia**. With increased concerns expressed over discrimination on the grounds of sexual orientation, this

---

[22]http://www.egalite-femmes-hommes.gouv.fr/marlene-schiappa-presente-ses-priorites-en-conseil-des-ministres/
[23]https://www.legifrance.gouv.fr/loda/id/JORFTEXT000033934948/
[24]http://www.haut-conseil-egalite.gouv.fr/hce/presentation-et-missions/
[25]https://rm.coe.int/1680590587

type of hostility has gained increasing attention. Weinberg (1971) defines homophobia as *'the dread of being in close quarters with homosexuals - and in the case of homosexuals themselves, self-loathing'*. Although this definition served as a model for defining a variety of negative attitudes towards sexual minorities, Herek (2004) argues that a more nuanced vocabulary is needed to understand all its underlying psychological, social, and cultural processes:

- *sexual stigma* (i.e., stigma attached to any nonheterosexual behavior, identity, relationship, or community);

- *heterosexism* (i.e., beliefs about gender, morality, and dangers (posed by sexual minorities) that perpetuate sexual stigma);

- *sexual prejudice* (i.e., negative attitudes based on sexual orientation, be it homosexual, bisexual, or heterosexual);

- *internalized homophobia* (i.e., self-loathing/an internal conflict between what one should be (i.e., heterosexual) and their sexual orientation).

1.2.3 Hate Speech and Other Related Concepts

Various other concepts related to HS exist, as shown in the following tweets taken from (Sanguinetti et al., 2018; Zampieri et al., 2019a) and Hate Speech Hackathon:[26]

- **aggressive language**: posts in which the user's *'intention is to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target'* (Sanguinetti et al., 2018)

   (1.1)   *tutto tempo danaro e sacrificio umano sprecato senza eliminazione fisica dei talebani e dei radicali musulmani e tutto inutile*
   (*it's all a waste of time, money and human lives without the extermination of Taliban and radical Muslims it's all useless*)

- **offensive language**: posts containing *'any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.'* (Zampieri et al., 2019a)

   (1.2)   *@USER Figures! What is wrong with these idiots? Thank God for @USER*

---

[26]https://www.swisstext.org/2018/workshops/Hackathon.html

- **abusive language**: *'any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion'* (Nobata et al., 2016)

(1.3)   *An Asshole That's Better Than You In Every Way.*

Several attempts at classifying these overlapping phenomena are found in the literature. Davidson et al. (2017) highlight that offensive language is often misclassified as HS due to an overly broad definition of the phenomenon as both concepts include frequent use of profanities. However, the use of offensive language can occur in contexts other than HS: trying to fit in with the others (i.e., conversational habits), using swear words for expressing a wide range of emotions (e.g., anger, joy, frustration, surprise) and it can achieve a positive social outcome by using swear words in jokes, ironic sarcasm, storytelling and even by replacing violence with swearing (Jay, 2009).

As previous works grouped different phenomena under the same umbrella term of *abusive language*, Waseem et al. (2017) propose a topology that synthesizes all these concepts by considering whether (i) the language is directed towards a specific target or towards a generalized group and (ii) the degree to which it is explicit. To further clarify these concepts and their relationships with each other, a classification of the overlapping abusive phenomena is presented in Figure 1.1 (Poletto et al., 2021). According to this framework, HS is an instance of abusive language, however manifestations of hatred that do not (necessarily) instigate a violent action are not categorized as HS under this definition.

In the following chapter we analyze the characteristics of different HS corpora representative of this phenomena and survey the main approaches used to detect HS online using Natural Language Processing (NLP) techniques.

Figure 1.1 – Relations between HS and other related concepts (Poletto et al., 2021).
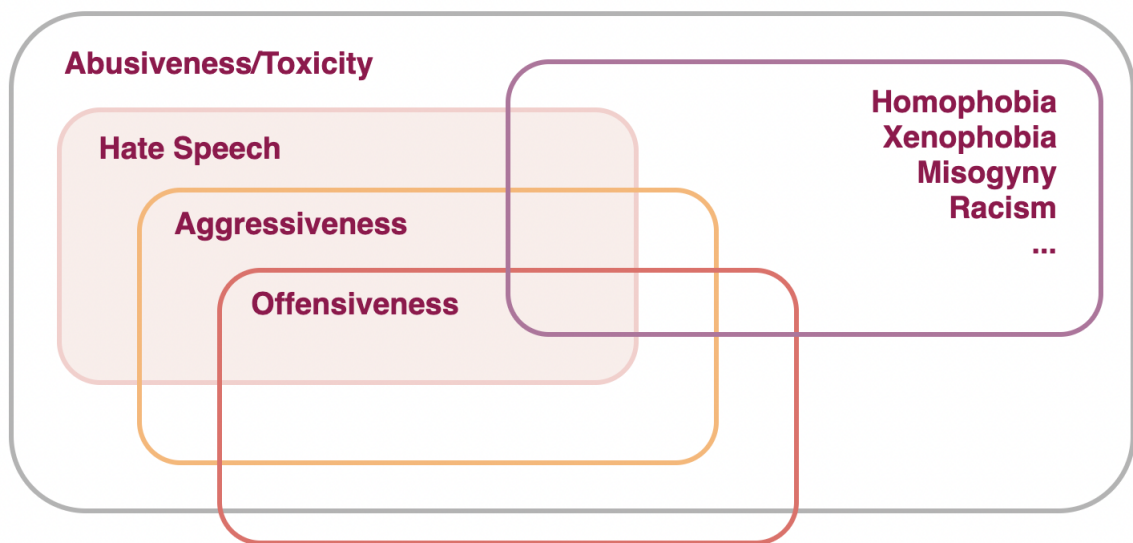
2 Hate Speech in Natural Language Processing

Due to the massive growth of user generated web content, there has been a growing interest in using Artificial Intelligence (AI) and NLP to address social and ethical issues. Let us mention the latest trends on *AI for social good* (Floridi et al., 2018, 2020), where the emphasis is on developing applications to maximize *'good'* social impacts while minimizing the likelihood of harm and disparagement to those belonging to vulnerable categories. See, for example, the literature on suicidal ideation detection, devoted to early intervention (Gaur et al., 2019). There are also recent works on the prevention of sexual harassment (Khatua et al., 2018), sexual discrimination (Khatua et al., 2019), cyberbullying and trolling (Menini et al., 2019), devoted to contrasting different kinds of abusive behavior targeting different groups and preventing unfair discrimination. Please note that in this chapter we focus on the task of HS detection. Related work for the tasks of cyberbullying and harassment detection are beyond the scope of this dissertation, as they require analysing, among others, the history of conversations and how the information is spread over social media networks.

Given the vast amount of social media data produced every minute,[27] manually monitoring social media content is impossible. It is, instead, necessary to detect HS automatically. To this end, many studies in the field exploit supervised approaches generally casting HS detection as a binary classification problem (i.e., abusive/hateful vs. not abusive/not hateful) (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Jurgens et al., 2019) relying on several manually annotated datasets that can be grouped into one of these categories:

- *Topic-generic* datasets, with a broad range of HS without limiting it to specific targets (Golbeck et al., 2017; Chatzakou et al., 2017; Founta et al., 2018). For example, Chatzakou et al. (2017) consider aggressive and bullying in their annotation scheme, while Founta

---

[27] https://www.internetlivestats.com/twitter-statistics/

et al. (2018) looks, in addition, for other expressions of online abuse such as offensive, abusive and hateful speech.

- *Topic-specific* datasets, where the HS category (racism, sexism, etc.) is known in advance (i.e., drives the data gathering process) and is often labelled. The HS targets, either person-directed or group-directed,[28] can be considered as *oriented*, containing, as they do, hateful content towards groups of targets or specific targets. For example, in (Waseem et al., 2017) scholars sampled data for multiple targets, that is racism and sexism for, respectively, religious/ethnic minorities HS and sexual/gender (male and female) HS. Others focus on single targets including, for instance, sampling for the misogyny topic, targeting women (Fersini et al., 2018b,a; Chiril et al., 2020b). Similarly, for the xenophobia and racism topics the target are groups discriminated against on the grounds of ethnicity (e.g., immigrants (Basile et al., 2019), ethnic minorities (Waseem and Hovy, 2016; Tulkens et al., 2016b), religious communities (Vidgen and Yasseri, 2020), Jewish communities (Zannettou et al., 2020), etc.).

Independently from the datasets that are used, all existing systems share two common characteristics. First, they are trained to predict the presence of general, target-independent HS, without addressing the problem of the variety of aspects related to both the topical focus and target-oriented nature of HS. Second, systems are built, optimized, and evaluated based on a single dataset, one that is either topic-generic or topic-specific. In order to address this issue and in order to improve the performance of the models, recent studies propose cross-domain classification, where the domain is used synonymously with dataset (Wiegand et al., 2018a; Waseem et al., 2018; Karan and Šnajder, 2018; Pamungkas and Patti, 2019). The idea consists in using a one-to-one configuration by training a system on a given dataset and testing the system on another one, using domain adaptation techniques. Most existing works map between fine-grained schemes (that are specific for each dataset) and a unified set of tags, usually composed of a positive and negative label to account for the heterogeneity of labels across datasets. Again, this binarization fails to discriminate among the multiple HS targets. Thus, it has become difficult to measure the generalization power of such systems and, more specifically, their ability to adapt their predictions in the presence of novel or different topics and targets (Vidgen and Derczynski, 2020).

---

[28]We do not make any distinction between HS directed towards a person/individual or a group, as done in previous studies (Waseem et al., 2017; Zampieri et al., 2019b,a).

In the following section, we first detail the characteristics of each of the datasets considered in this study. Then we present relevant prior works specifically related to HS detection, in order to provide readers with a broader context for NLP literature related to the analysis and to the recognition of hateful content in texts.

2.1 Hate Speech Datasets

In Table 2.1 we summarize the main existing corpora for HS. In particular, we highlight the platform from where the data were retrieved (i.e., Source), the annotation scheme, as well as the different covered languages.[29] Note that these datasets vary not only in terms of size and scope, but also in terms of HS characteristics that are considered (i.e., they capture different types of information).

Although the focus of this dissertation being HS, Table 2.1 also presents works on related phenomena such as abusive (Pavlopoulos et al., 2017), aggressive (Álvarez-Carmona et al., 2018) and offensive language (de Pelle and Moreira, 2017; Bretschneider and Peters, 2017; Zampieri et al., 2019a; Çöltekin, 2020; Pitenis et al., 2020; Sigurbergsson and Derczynski, 2020).

As it can be seen from Table 2.1, the size of the existing datasets ranges from a few hundreds (Ross et al., 2017) to a few million (Pavlopoulos et al., 2017) instances and the most exploited data source is Twitter. In terms of annotation scheme (although common annotation schemes are used in benchmark corpora for shared tasks), most of the presented works assume different levels of granularity.

Regarding the language, the majority of the resources are in English, however, the growing interest of the research community towards HS detection (and other related phenomena) has enabled a greater linguistic diversity.

---

[29] https://hatespeechdata.com provides an overview of the existing resources for studying online HS, and additionally supplies links to the data.

Table 2.1 – Hate Speech datasets.

Dataset	Language	No. of instances	Source	Annotation
(de Pelle and Moreira, 2017)	Portuguese	1,250	g1.globo.com	- not offensive vs. offensive - within offensive: sexism, racism, homophobia, xenophobia, religious intolerance, cursing
(Fortuna et al., 2019)	Portuguese	5,668	Twitter	not hate vs. hate (hierarchical annotation schema with 81 HS categories)
(Ljubešić et al., 2018)	Slovene	7,596,686	news articles comments (MMC RTW website)	retained vs. deleted comments (i.e., inappropriate comments containing insults, swearing, irony, etc.)
	Croatian	17,042,965	news articles comments (24sata website)	
(Sigurbergsson and Derczynski, 2020)	Danish	800	Facebook comments (Ekstra Bladet)	- offensive vs. not offensive
		1,400	r/Denmark sub-reddit	- within offensive: targeted vs. not targeted
		1,400	r/DANMAG sub-reddit	- within target: individual, group, other
(Ptaszynski et al., 2019)	Polish	11,041	Twitter	non-harmful vs. cyberbullying vs. HS and other harmful contents
(Çöltekin, 2020)	Turkish	36,232	Twitter	- offensive vs. not offensive - within offensive: targeted vs. not targeted - within target: individual, group, other
(Pavlopoulos et al., 2017)	Greek	1,450,000	Gazetta	abusive comments (reject) vs. not abusive (accept)
		115,000	Wikipedia	personal attacks (reject) vs. not (accept)
		159,686		toxic comments (reject) vs. not toxic (accept)
(Pitenis et al., 2020)	Greek	10,287	Twitter	offensive vs. not offensive
(Mubarak et al., 2017)	Arabic	1,100	Twitter	obscene vs. offensive (but not obscene) vs. normal
		32,000	news article comments (Aljazeera.net)	
(Albadi et al., 2018)	Arabic	6,000	Twitter	non-hateful vs. hate speech against six religious groups
(Mulki et al., 2019)	Arabic (Levantine dialect)	5,846	Twitter	hate vs. abusive (i.e., offensive, aggressive, insulting or profanity) vs. normal
(Ousidhoum et al., 2019)	Arabic	3,353	Twitter	- directness (direct vs. indirect) - hostility (abusive vs. hateful vs. offensive vs. disrespectful vs. fearful vs. normal) - target (race vs. gender vs. sexual orientation vs. religion vs. disability vs. other (e.g., political ideologies, social classes) - the name of the target group (individual vs. women vs. special needs vs. African descent vs. other) - annotator feeling (disgust vs. shock vs. anger vs. sadness vs. fear vs. confusion vs. indifference)
	English	5,647		
	French	4,014		
(Alshalan and Al-Khalifa, 2020)	Arabic	9,316	Twitter	non-hateful vs hateful (religious, racist, ideological, tribal and regional HS)
(Sanguinetti et al., 2018)	Italian	6,009	Twitter	- hate speech vs. not hate speech - agressiveness (no vs. weak vs. strong) - offensiveness (no vs. weak vs. strong) - irony (yes vs. no) - stereotype (yes vs. no) - five point intensity degree
(Bosco et al., 2018)	Italian	4,000	Facebook	- hate speech vs. non hate speech - agressiveness (no vs. weak vs. strong) - offensiveness (no vs. weak vs. strong) - irony (yes vs. no) - stereotype (yes vs. no) - five point intensity degree
		4,000	Twitter	
(Sprugnoli et al., 2018)	Italian	14,600 tokens (10 chats)	WhatsApp	- cyberbullying role (harasser vs. victim vs. bystander defender vs. bystander-assistant) - cyberbullying type (13 different classes of insults, discrimination, sexual talk and agressive statements) - sarcasm (yes vs. no) - offensive vs. non-offensive (i.e., joke)
(Chung et al., 2019)	Italian	1.071	Twitter	hate speech / counter-narrative pairs
	English	1,288		
	French	1,719		
(Sanguinetti et al., 2020)	Italian	8,012	Twitter	- hate speech vs. not hate speech towards a given target (i.e., muslims, Roma and immigrants) - stereotype (presence vs. absence) - presence of nominal utterances
		500	news headlines (online newspapers: Il Giornale, Liberoquotidiano, La Stampa, La Repubblica)	
(Kumar et al., 2018b)	Hindi-English	21,000	Facebook	- verbal aggression (overt vs. covert aggression) - target of aggression: physical threat vs. sexual threat vs. identity threat vs. non-threatening aggression - within identity threat/aggression: gendered vs. geographical vs. political vs. casteist vs. communal vs. racial
		18,000	Twitter	

Table 2.2 – Hate Speech datasets (cont.).

DATASET	LANGUAGE	NO. OF INSTANCES	SOURCE	ANNOTATION
(Bohra et al., 2018)	Hindi-English	4,575	Twitter	hate speech vs. normal speech
(Mathur et al., 2018)	Hindi-English	3,189	Twitter	non-offensive vs. abusive vs. hate-inducing
(Ross et al., 2017)	German	541	Twitter	- hate speech vs. non hate speech - the tweet should be banned (yes vs. mo) - offensiveness of the tweet on a six point Likert scale
(Bretschneider and Peters, 2017)	German	2,649	Facebook (Pegida page)	- offensive vs. non-offensive
		2,641	Facebook (Ich bin Patriot aber kein Nazi page)	- severity of offensiveness (two point scale)
		546	Facebook (Kriminelle Auslander raus page)	- target (foreigner vs. government vs. press vs. community vs. other vs. unknown)
(Wiegand et al., 2018b)	German	8,541	Twitter	- offensive vs. other - within offensive: abuse vs. insult vs. profanity vs. other
(Mandl et al., 2019)	German	4,669	Twitter	- hate and offensive vs. non-hate and offensive
	English	7,005		- the post contains hate speech (yes vs. no)
	Hindi	5,983	Facebook	- the post contains offensive content (yes vs. no) - the post contains profane words (yes vs. no)
(Ibrohim and Budi, 2018)	Indonesian	2,016	Twitter	not abusive vs. abusive but not offensive vs. offensive
(Ibrohim and Budi, 2019)	Indonesian	13,169	Twitter	- hate speech and abusive language vs. not hate speech - within hate speech: weak vs. moderate vs. strong - hate speech target: religion vs. ethnicity vs. physical disability vs. gender/sexual orientation vs. other/slander
(Álvarez-Carmona et al., 2018)	Mexican Spanish	10,856	Twitter	aggressive vs. non-aggressive
(Waseem, 2016)	English	6,909	Twitter	racism vs. sexism vs. both vs. neither
(Bretschneider and Peters, 2016)	English	16,975	World of Warcraft	harassment: (offender, victim) tuple
		17,354	League of Legends	
(Wulczyn et al., 2017)	English	115,737	Wikipedia comments	personal attacks (blocked vs. random)
(Golbeck et al., 2017)	English	35,000	Twitter	harassing (racist/misogynystic/homophobic/bigoted, threats, hate speech, direct harassement, potentially offensive) vs. non-harassing
(Gao and Huang, 2017)	English	1,528	Fox News comments	hateful vs. non-hateful
(Rezvan et al., 2018)	English	75,000	Twitter	non-harassing vs. harassing (sexual, racial, intellectual, political, appearance-related)
(Ribeiro et al., 2018)	English	4,972	Twitter	hateful vs. non-hateful users
(ElSherief et al., 2018b)	English	27,330	Twitter	hate speech tweets
		25,278		instigator accounts
		22,287		target accounts
(de Gibert et al., 2018)	English	10,568	Stormfront	- hate speech vs. not hate speech - relation label given separately to sentences that need each other to be understood as hate speech
(Zampieri et al., 2019a)	English	14,100	Twitter	- offensive vs. not offensive - within offensive: targeted vs. untargeted - within target: individual vs. group vs. other
(Qian et al., 2019)	English	33,776	Gab	- hate speech vs. not hate speech
		22,324	Reddit	- if the post is hateful, how would the annotator respond to interfere
(Gomez et al., 2020)	English	2,435	Twitter	non-hateful vs. hateful (racist vs. sexist vs. homophobic vs. religion based attacks vs. other)

2.2 Datasets Used in this Study

In this dissertation we propose to undertake the following challenges:

1. Experiment with the *development of models able to detect different types of sexism experiences in French tweets*.

2. Investigate the problem of *transferring knowledge from different datasets with different topical focuses and targets* (cf. Part IV).

For the first challenge, as there are no existing dataset available, in Part II we will present the first French dataset annotated for sexism detection according to a novel characterization of sexist content-force relation inspired by speech acts theory (Austin, 1962) and discourse studies in gender (Lazar, 2007; Mills, 2008).

For tackling the second challenge, we leverage seven manually annotated HS corpora from previous studies. We selected these datasets as they are freely available to the research community. Among them, two are topic-generic (`Davidson` (Davidson et al., 2017) and `Founta` (Founta et al., 2018)), and four are topic-specific about four different topics: *misogyny* (the Automatic Misogyny Identification (AMI) dataset collection from both `IberEval` (Fersini et al., 2018b) and `Evalita` (Fersini et al., 2018a)), *misogyny and xenophobia* (the `HatEval` dataset (Basile et al., 2019)), and *racism* and *sexism* (the `Waseem` dataset (Waseem and Hovy, 2016)). Each of these topics targets either gender (sexism and misogyny) and/or ethnicity, religion or race (xenophobia and racism). In the following, we detail the characteristics of each of the datasets that were considered in this study:

- **Davidson.** The dataset has been built by Davidson et al. (2017) and contains 24,783 English tweets[30] manually annotated with three labels including *hate speech*, *offensive*, and *neither*. These tweets were sampled from a collection of 85.4 million tweets gathered using the Twitter search API, focusing on tweets containing keywords from HateBase.[31] The dataset was manually labeled by using the CrowdFlower platform,[32] where at least three annotators annotated each tweet. With an inter-annotator agreement of 92%, the final label for each instance was assigned according to a majority vote. Only 5.8% of the total tweets were labeled as *hate speech* (cf. (2.1)) and 77.4% as *offensive* (cf. (2.2)), while the remaining 16.8% were labelled as *not offensive*.

  (2.1)   *#DTLA is trash because of non-Europeans are allowed to live there*

  (2.2)   *What would y'all lil ugly bald headed bitches do if they stop making make-up & weave?*

---

[30] Although in the original paper the authors mention that the dataset consists of 24,802 annotated tweets, we only found this number of instances in the shared GitHub repository: https://github.com/t-davidson/hate-speech-and-offensive-language

[31] A multilingual repository, which allows for the identification of HS terms by region: https://hatebase.org

[32] Now Figure Eight https://www.figure-eight.com/

- **Founta**. The dataset consists of 80,000 tweets in English[33] annotated with four mutually exclusive labels including *abusive*, *hateful*, *spam* and *normal* (Founta et al., 2018). The original corpus of 30 millions tweets was collected from 30 March 2017 to 9 April 2017 by using the Twitter Stream API. For each tweet, the authors also extracted the meta-information and linguistic features in order to facilitate the filtering and sampling process. Annotation was done by five crowdworkers and the final dataset was composed of 11% tweets labeled as *abusive* (cf. (2.3)), 7.5% as *hateful* (cf. (2.4)), 59% as *normal*, and 22.5% as *spam* (cf. (2.5)).

  (2.3)   *Benedict Cumberbatch is a damn stupid name. I hope history doesn't remember him fondly. I hope his legacy becomes trash.*

  (2.4)   *Niggas worst than your side bitch always questioning they position*

  (2.5)   *Beats by Dr. Dre urBeats Wired In-Ear Headphones - White https://t.co/9tREpqfyW4 https://t.co/FCaWyWRbpE*

- **Waseem**. It consists of English tweets collected over a period of two months by using representative keywords (common slurs) that target religious, sexual, gender and ethnic minorities (Waseem and Hovy, 2016). The authors manually annotated the dataset with a third expert annotator reviewing their annotations. The final dataset consists of 16,914 tweets, with 3,383 instances from Sexism$_{\text{Waseem}}$ targeting gender minorities (cf.(2.6)), 1,972 from Racism$_{\text{Waseem}}$ with racist instances (cf. (2.7)), and 11,559 tweets that were judged to be neither sexist nor racist.[34]

  (2.6)   *Sounds like we've got a well good ref' today, bloody women should just stay in the kitchen!*

  (2.7)   *It's not about any specific individuals, but about an ideology that will always produce terrorists.*

- **HatEval**. The dataset consists of 19,600 tweets distributed across two different targets: immigrants (cf. (2.8)) and women (cf. (2.9)) (Basile et al., 2019). Most of the tweets that

---

[33]At the moment of collecting the data, from the original dataset (http://ow.ly/BqCf30jqffN) we were able to retrieve only 44,898 tweets, though in a recent shared task (https://sites.google.com/view/icwsm2020datachallenge/home) the full dataset was made available.

[34]When collecting the data, we were able to retrieve only 16,488 instances (3,216 targeting gender minorities, 1,957 racist and 11,315 that were neither racist nor sexist).

target women were derived from the `AMI corpora` (cf. Section 3.3), while the remainder of the dataset was collected over a period of three months (from July to September 2018) by employing the same approaches as AMI. The dataset was annotated by using the Figure Eight crowdsourcing platform. In each instance, the annotators were asked to specify whether a tweet conveys HS or not towards any given targets. The annotators were also asked to indicate whether the author of the tweet was aggressive and to identify the target of the tweet (i.e., a specific individual or a group of people). The inter-annotator agreement obtained for each category is 0.83, 0.73, and 0.70 for English, 0.89, 0.47, 0.47 for Spanish respectively. The final label was assigned based on a majority vote by adding two expert annotations to the crowd-annotated data. The final distribution of the dataset includes 13,000 English tweets (6,500 for each target) and 6,600 Spanish tweets.

(2.8)   *Your boats shall drown in the Mediterranean Sea and the rest of you, which had not assimilated into our society will leave immediately. #RefugeesNotWelcome #IllegalAliens*

(2.9)   *Its a good thing I always wear a glove on my left hand because if I EVER had to touch hands with a woman my IQ would totally drop to 0 Lol*

Table 2.3 provides a general overview of these datasets,[35] along with the labels used in their annotation schemes. We can observe that the classes are imbalanced in most datasets, where the majority class is the negative class (non-HS), except for the AMI collection (`AMI-IberEval` and `AMI-Evalita`) and `Davidson`.

2.3   Dataset Bias

Olteanu et al. (2019) define data bias as *'a systematic distortion in the sampled data that compromises its representativeness'* (i.e., the data that is available is not capable of properly capturing the studied phenomenon). Considering that HS corpora are increasingly used for building automated detection systems, the understanding of the limitations of these datasets is critical.

Davidson et al. (2019) tested the existence of racial bias by training classifiers on five widely used annotated for hate speech and abusive language Twitter datasets (Waseem,

---

[35]The `AMI corpora` is further detailed in Section 3.3.

Table 2.3 – General overview of the datasets used in this study.

Dataset	Labels	No. of instances	Topic	Target
Davidson	hate speech	1,430		
	offensive	19,190	generic	none
	neither	4,163		
Founta	abusive	27,037		
	hateful	4,948	generic	none
	spam	14,024		
	normal	53,790		
Waseem	racism	1,957		race
	sexism	3,216	specific	women
	none	11,315		
Evalita	misogyny	2,245	specific	women
	not misogyny	2,755		
IberEval	misogyny	1,851	specific	women
	not misogyny	2,126		
HatEval	immigrant	2,427		women
	women	2,608	specific	ethnicity
	not hate speech	6,936		

Note: The "No. of instances" totals are: Davidson 24,783; Founta 99,799; Waseem 16,488; Evalita 5,000; IberEval 3,977; HatEval 11,971.

2016; Waseem and Hovy, 2016; Davidson et al., 2017; Golbeck et al., 2017; Founta et al., 2018) and comparing their performance on tweets labeled by race (Blodgett et al., 2016). The authors argue that the bias is likely to be dependent on the data itself and not on the classifier. As the main objective consists in detecting existing bias and not on in improving the predictive performance, the choice of classifier was rather standard, a regularized logistic regression with Bag of Words features for each dataset, that was later used for predicting the class for unseen tweets. Although the contributions brought by using features like word embeddings are significant, in order to avoid including any additional bias in the models the authors chose not to use them. After performing a basic preprocessing for each of the datasets and testing the classifier's performance on tweets written in African American English (black-aligned corpus) with tweets written in Standard American English (white-aligned corpus), substantial racial disparities were observed in the performance of all classifiers. Although African Americans are often targeted with racism and HS (i.e., it is expected

that the white aligned tweets are more likely to use racist language or HS), Davidson et al. (2019) observe that negative labels (racism, HS) are assigned more often to the tweets in the black-aligned corpus. Davidson et al. (2019) were able to find evidence of substantial racial bias in all of the tested datasets, their findings suggesting that by using these datasets a system will penalize African Americans at a higher rate, resulting in racial discrimination.

Sap et al. (2019) conduct a similar test in order to determine if the existing approaches for performing HS detection contain racial bias and assess how the bias is acquired and how it propagates throughout the predictive models. For both language corpora (i.e., (Davidson et al., 2017; Founta et al., 2018)), the authors trained a classifier for predicting the toxicity label of a tweet. Although the models were able to achieve high accuracies, they also inferred strong associations between African American English (AAE) and various HS categories, corroborating the existence of dialect bias in these corpora and that text alone does not determine offensiveness. As bias often derives from the training data, meaning that the individual biases of the annotators accumulate into systematic training data biases, the authors also propose a method for mitigating annotator bias through *'dialect and race priming'* (i.e., by designing tasks that explicitly highlight the inferred dialect of a tweet or likely racial background of its author, the annotators were asked to take into consideration the likely racial background of a tweet author as well as specify if the tweet was offensive to them or anyone else).

Kim et al. (2020) go a step further and study the interaction between race (black/white) and gender (male/female) to study the influence of the intersection of racial and gender bias upon the distribution of hateful and abusive labels in the (Founta et al., 2018) dataset. Their results show that a tweet is more likely to be labeled as abusive if it presents features associated to African American language. Moreover, features more closely associated with African American male language are more likely to be labeled as hateful (this trend is not as prominent for the female counterpart).

Dixon et al. (2018) argue that every machine learning model is designed to express a bias towards the task that needs to be solved (i.e., the bias of a model trained for identifying toxic comments consists in attributing higher scores to the toxic comments than the others). The authors define unintended bias as a model that expresses bias towards a different task than the one that needs solving (i.e., the case in which the model in the above example would also express bias towards the gender of the people) and address one specific subcase

of this definition, which they call *'identity term bias'*. Dixon et al. (2018) found that machine learning models tend to attribute higher toxicity scores to innocuous statements like *I am a gay man* or *You are a good woman* due to unintended bias probably introduced by the use of the words *gay* and *woman*. This was introduced as *'false positive bias'*, caused by the model overgeneralizing from the training data, as identity terms affected by the false positive bias are disproportionately used. In order to mitigate the data imbalance which causes this form of unintended model bias, the authors manually created a set of common identity terms for which they added additional data, in order to have similar overall distribution of toxic/non-toxic comments across the dataset. By testing a Convolutional Neural Network (CNN) built in order to identify toxicity in comments from Wikipedia Talk Pages, the authors were able to prove that the proposed bias mitigation technique reduces unintended bias of the model's real-valued scores.

Wiegand et al. (2019) examined the issue of data bias on abusive language detection datasets (Razavi et al., 2010; Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Wulczyn et al., 2017; Founta et al., 2018; Kumar et al., 2018a) and analyzed the relation with the way the data have been sampled. The first step in their methodology consists in computing the proportion of instances that contain explicit abusive language according to a lexicon of abusive words (Wiegand et al., 2018a), as well as the proportion of instances that contain implicit abusive language (i.e., a message that becomes abusive through the use of sarcasm, irony, jokes, negative stereotypes, etc.). Their results show that datasets that apply a biased sampling for corpus collection (i.e., instances matching query words that are likely to occur in abusive language) contain a high degree of implicit abuse which may lead to a decrease in performance due to the difficulty of learning lexical cues that convey implicit abuse. Wiegand et al. (2019) illustrated that datasets with a high degree of implicit abuse can be more affected by data bias and by trying to remove biased words (i.e., the words having the highest Pointwise Mutual Information towards abusive messages) and query words, the performance is much lower than originally reported.

Finally, Tsvetkov (2020) argues that even when having *'perfect'* annotations, the current HS classifiers may still learn and amplify correlations between a protected attribute (e.g., AAE) and abusive/offensive/hate speech. The proposed approach for dealing with the annotation bias, relies on using an adversarial objective for discouraging a model from encoding information related to the protected attribute.

2.4 Hate Speech Detection in Online Communication

Detecting hateful content, as well as its author, still raises difficulties for the social media platforms, both Facebook and Twitter facing criticism for not doing enough to prevent it.[36] As a consequence, in recent years, because of the difficulty of the task, research focused on HS detection related to race, religion and gender became a point of interest in NLP. The focus on gendered and ethnicity-based HS is due, in part, to the wide availability of English corpora developed by the computational linguistics community for those targets. But it also depends on the fact that most monitoring exercises by institutions countering online HS in different countries and territories (e.g., European Commission (EU Commission, 2016)) report ethnic-based hatred (including anti-migrant hatred) and gender-based hatred as the most common type of online HS (Chetty and Alathur, 2018).

With the increasing user generated content, it is impractical to rely on the manual detection of abusive posts. The gravity of the problem requires the automatization of filtering inappropriate content; however, this is still an open problem.

The automatic detection of online HS is not a simple task, especially because of the thin line between abusive language and freedom of speech. For example, the use of swear words could become an issue in HS detection (Swamy et al., 2019; Pamungkas et al., 2020a), where their presence might lead to false positives: for instance, when they are used in a non-abusive way in humor, emphasis, catharsis, and when conveying informality. But they could also become a strong signal for spotting HS, when they are used in an abusive context.

Most studies that deal with automatic HS detection exploit supervised approaches to classify HS and non-HS content. First studies in the field relied on traditional machine learning approaches with hard-coded features. Several classifiers were used, such as:

- Naive Bayes (NB) (Kwok and Wang, 2013; Agarwal and Sureka, 2017)

- Logistic Regression (LR) (Djuric et al., 2015; Waseem and Hovy, 2016; Badjatiya et al., 2017; Davidson et al., 2017; Fehn Unsvåg and Gambäck, 2018; Mishra et al., 2019)

- Support Vector Machines (SVM) (Greevy and Smeaton, 2004; Warner and Hirschberg, 2012; Burnap and Williams, 2014, 2015, 2016; Tulkens et al., 2016a; Badjatiya et al., 2017)

---

[36]https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/

- Decision Tree (DT) (Burnap and Williams, 2014, 2015, 2016; Agarwal and Sureka, 2017)

- Random Forest (RF) (Burnap and Williams, 2014, 2015, 2016; Badjatiya et al., 2017; Agarwal and Sureka, 2017)

A wide range of features have been employed including: lexical features (e.g., n-grams, Bag of Words, Tf/IDf, lexicon-based); syntactic features (e.g., speech parts and typed dependency); stylistic features (e.g., number of characters, punctuation, text length); as well as some Twitter specific features (e.g., the number of user mentions, hashtags, URLs, social network information (Mishra et al., 2019); and other user features (Waseem and Hovy, 2016; Fehn Unsvåg and Gambäck, 2018; Qian et al., 2018)).

Recently, the task of automatic HS detection has focused on exploiting neural models such as Long Short-Term Memory (LSTM) (Vigna et al., 2017; Mishra et al., 2019), Bidirectional LSTM (BiLSTM) (Qian et al., 2018), Gated Recurrent Unit (GRU) (Mossie and Wang, 2019), and CNN (Badjatiya et al., 2017) coupled with word embedding models such as FastText,[37] word2vec,[38] and ELMo (Peters et al., 2018).

A fair amount of works that deal with HS detection have come from teams that participated in recently shared tasks such as HatEval (Basile et al., 2019), AMI (Fersini et al., 2018b,a), and Hate Speech and Offensive Content Identification (HASOC) (Mandl et al., 2019). HatEval was introduced at SemEval 2019 and focused on the detection of hateful messages on Twitter directed towards two specific targets: immigrants and women. This was done from a multilingual[39] perspective (English and Spanish). The best-performing system in English HatEval (Indurthi et al., 2019) exploited a straightforward SVM with a Radial Basis Function (RBF) kernel that uses Google's Universal Sentence Encoder (USE) (Cer et al., 2018a) feature representation. HASOC, an HS and offensive language identification shared task at FIRE 2019, covers three languages: English, German, and Hindi. For English, the best performance was achieved by an LSTM network with ordered neurons and an attention mechanism (Wang et al., 2019).

All the aforementioned shared tasks provided datasets in languages other than English: i.e., Italian, Spanish, Hindi, and German. Other languages used in shared tasks include Italian (HasSpeeDe (Bosco et al., 2018) which focuses on detecting HS towards immigrants) and

---

[37]https://fasttext.cc/
[38]https://code.google.com/archive/p/word2vec/
[39]In this case, 'multilingual' refers to the fact that two datasets were made available as part of the competition. The submitted systems were trained and tested separately on each language.

German (GermEval (Wiegand et al., 2018b) which focuses on offensive language identification). This has enabled the development of multilingual models (Aluru et al., 2020; Corazza et al., 2020; Pamungkas et al., 2020b, 2021a). For example, Corazza et al. (2020) propose a recurrent neural architecture for detecting HS in English, German, and Italian by leveraging monolingual datasets (Waseem and Hovy, 2016; Wiegand et al., 2018b; Bosco et al., 2018) annotated as containing HS/offensive language or not. Aluru et al. (2020) have conducted a multilingual HS detection analysis in nine languages over a corpora from 16 different sources. Their results show that BERT (Bidirectional Encoder Representations from Transformers) based models perform best in experimental settings where a larger amount of data is available, while simpler models such as LR coupled with LASER[40] embeddings achieve good performances in low resource settings.

Finally, d'Sa et al. (2020) explore label propagation semi-supervised learning, a technique that uses the labels of annotated instances to *'transduce'* the labels to unlabeled data for the task of HS classification. Their results show that this is an effective technique in very low resource scenarios, the performance gains decreasing with the increase of available annotated data.

---

[40]https://github.com/facebookresearch/LASER

3 Sexism in Natural Language Processing

In this chapter, we focus on HS that targets women (i.e., *sexism* and *misogyny*) and first present a short analysis of the nature and the effects of sexism. Then, we detail the characteristics of each of the existing sexism datasets and present relevant prior works specifically related to sexism detection, in order to provide readers with a broader context for NLP literature related to the analysis and to the recognition of sexist content in texts.

3.1 Sexism in Gender Studies

The nature and the effects of sexism have been deeply analyzed in fields such as social psychology. Sexism can be expressed at different linguistic granularity levels going from lexical to discursive (Cameron, 1992). For example, women are often designated through their relationship with men or motherhood (cf. (3.1)) or they are characterized through their physical characteristics (cf. (3.2)).

(3.1)   *A man killed in a shooting* vs. *Mother of 2 killed in a crash*

(3.2)   *The journalist who presents the news* vs. *The blonde who presents the news*

   Glick and Fiske (1996) view sexism as a *multidimensional construct* that encompasses two components: *hostile* and *benevolent sexism*. Hostile sexism covers the aspects of sexism that fit Allport et al. (1954) definition of prejudice: *'aversive or hostile attitude toward a person who belongs to a group, simply because he belongs to that group, and is therefore presumed to have the objectionable qualities ascribed to that group'* (cf. (3.3)). In contrast, benevolent sexism is subjectively positive and sexism is expressed in the form of a compliment (cf. (3.4)). Despite the positive feeling, benevolent sexism shares common assumptions with hostile sexism, where

women are the *'weaker sex'* and they are unfit to hold the same power and status as men (i.e., it endorses traditional gender roles). As such, benevolent sexism can be used to counteract or justify hostile sexism.

In addition, Glick and Fiske (1996) define *ambivalent sexism* as simultaneously holding both hostile and benevolent beliefs and viewing them as consistent with each other (cf. (3.5)). In their analysis, Glick and Fiske (1997) hypothesized that both hostile and benevolent sexism encompass three sources of male ambivalence: *paternalism*, *gender differentiation*, and *heterosexuality*. Each component has two aspects (hostile and benevolent), and their main role is to justify the critical issues that characterize relationships between the sexes.

(3.3)   *The world would be a better place without women*

(3.4)   *Many women have a quality of purity that few men have*

(3.5)   *Women are incompetent at work* <u>and</u> *Women must be protected*

In communication studies, the analysis of political discourse (Bonnafous, 2003; Coulomb-Gully, 2012), sexist abuse or media discourse (Dai and Xu, 2014; Biscarrat et al., 2016) show that political women presentations are stereotyped: use of physical or clothing characteristics, reference to private life, etc. From a sociological perspective, studies focus on social media contents (tweets) or SMS in order to analyze public opinion on gender-based violence (Purohit et al., 2016) or violence and sexist behaviours (Barak, 2005; Megarry, 2014).

3.2   Gender in Language Models

Word embeddings have become one of the most used types of features in many NLP models and are widely used for a variety of downstream tasks. However, these word representations have been proven to reflect social biases (such as race and gender) inherited from data used to train them (Caliskan et al., 2017).

Bolukbasi et al. (2016) found that the embeddings contain stereotypical analogies such as:

$$\overrightarrow{man} - \overrightarrow{woman} = \overrightarrow{computer\,programmer} - \overrightarrow{homemaker}$$

where the word *programmer*, although gender neutral by definition, is going to be closer to *man* than *woman*.

Due to their wide-spread usage, the need of removing this kind of bias arises.

In order to prevent these type of analogies from existing in the vector space, Bolukbasi et al. (2016) propose subtracting the *gender bias subspace* (i.e., the principal components for ten gender pair difference vectors; e.g., $\overrightarrow{woman} - \overrightarrow{man}$, $\overrightarrow{she} - \overrightarrow{he}$, $\overrightarrow{her} - \overrightarrow{his}$, etc.) from each biased word.

Park et al. (2018) built upon this study by using the identity term template method (Dixon et al., 2018) to study gender bias in performance across two HS and abusive language detection datasets. The authors experimented with three neural networks (CNN, GRU, Bidirectional GRU with self-attention) previously used for the task of abusive language detection. A first test in which the debiased word embeddings proposed by Bolukbasi et al. (2016) were used, shows that debiased word embeddings alone are not capable of effectively correcting the bias of the system. In order to improve the models, Park et al. (2018) propose a framework based on combining three methods: debiased word embeddings (Bolukbasi et al., 2016), gender swap and bias fine-tuning. Gender swapping consists in augmenting the training data by identifying and swapping entities with their equivalent gendered entity, based on the intuition that given a pair of sentences having only the identity terms different (e.g., *He is happy* and *She is happy*) the model should predict the same label for the task of abusive language detection, while bias fine-tuning consists in a model being trained on a less biased corpus and fine-tuned on the target corpus that contains a larger bias. Through the combination of these three methods the gender bias was significantly reduced, although for all the methods some performance loss was reported.

Zhao et al. (2018b) take a different approach by trying to remove the bias during the training phase. The authors train Glove embeddings (Pennington et al., 2014) from scratch with a modified loss function that clusters the information related to a *'protected attribute'* (gender, in this case) to a specific coordinate of the embedded vector. Although the initial results seemed promising, Gonen and Goldberg (2019) argue that the methods proposed by both Bolukbasi et al. (2016) and Zhao et al. (2018b) are rather hiding the bias instead of removing it. By testing these debiasing algorithms on the top 500 male and female more biased words, Gonen and Goldberg (2019) find that despite the change in relation to the gender direction, these words tend to cluster with other words that are affected by implicit gender stereotypes, meaning that the biases in word embeddings persist even after attempting to remove them.

More recently, researchers have started quantifying, analyzing and mitigating the gender bias exhibited by contextualized embeddings (Zhao et al., 2019; Kurita et al., 2019; Tan and Celis, 2019; Guo and Caliskan, 2021). Their results show that contextualized word models inherit human-like biases, which are then propagated to downstream tasks.

3.3   Sexism Datasets

Compared to (generic) HS, fewer datasets have been developed for the task of detecting HS towards women. In the following, we first detail the characteristics of each of the existing corpora for misogyny detection, then present current approaches for the task.

- **AMI corpora.** Two AMI datasets were made available to the research community: `IberEval` (Fersini et al., 2018b) containing 3,977 English tweets and 4,138 Spanish tweets collected over a period of four months (from 20th of July until 30th of November 2017) and `Evalita` (Fersini et al., 2018a) that comprises 5,000 tweets for English and 5,000 tweets for Italian. The main goal of the AMI task consists in identifying tweets that convey hate or prejudice against women while categorizing forms of misogynous behaviour:

    – 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* (cf. (3.9));

    – dominance - *to assert the superiority of men over women to highlight gender inequality* (cf. (3.8));

    – 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* (cf. (3.7));

    – sexual harassment & threats of violence - *to describe actions as sexual advances, requests for sexual favours, harassment of a sexual nature; intent to physically assert power over women through threats of violence.* (cf. (3.6));

    – discredit - *slurring over women with no other larger intention* (cf. (3.10)).

    In addition, the annotation includes the target of a given instance (specific individual or a generic group). The datasets used in these tasks were collected by employing three

different approaches: representative keywords and hashtags; monitoring potential victims; as well as by downloading the history of users that have explicitly misogynistic behavior on their Twitter profiles.

Below we present five examples of tweets annotated as misogynous taken respectively from `IberEval` and `Evalita`.

(3.6)   *I kinda want to see you again just so I can punch you in the kidney. #WomenSuck*

(3.7)   *Yes yes Ann, lets continue to perpetuate the hysterical woman stereotype. Such a shame. You dont deserve your position of power. A disgrace to fellow women.*

(3.8)   *Don't get offended ladies, men are always right*

(3.9)   *Only places my wife can drive are to restaurants and grocery stores*

(3.10) *When I start spending money on you, then you mean something to me cause Ian investing my money in no hoe*

- **Zeinert**. It consists of 27,900 Danish comments collected from multiple platforms (Twitter, Facebook, Reddit) using representative keywords (common slurs, hashtags) as well as terms that do not appear exclusively in a misogynistic context. For annotating the corpus, Zeinert et al. (2021) propose a misogyny labeling scheme embedded within a taxonomy for labeling abusive language (cf. Figure 3.1). In addition to the five types of misogynous behaviour present in the `AMI corpora`, the authors discovered that the most frequently represented type of misogynous behaviour in this corpus is an implicit form of misogyny, *neosexism*, which describes the direct denial that misogyny exists (cf. (3.11) in which the existence of discrimination is questioned).

(3.11) *Can you point to research showing that childbirth is the reason why mothers miss out on promotions ?*

- **Jha**. The dataset has been built by Jha and Mamidi (2017) and contains 10,095 English tweets manually annotated with three labels including *benevolent* if the tweet exhibits subjectively positive sentiment but is sexist, *hostile* if the tweet exhibits explicitly negative emotion and is sexist, and *others* if the tweet is non-sexist. The tweets belonging to class *benevolent* were gathered using the Twitter search API, focusing on tweets containing keywords and hashtags that are generally used when exhibiting benevolent sexism

Figure 3.1 – The annotation scheme of the `Zeinert` corpus ([Zeinert et al., 2021](#)).
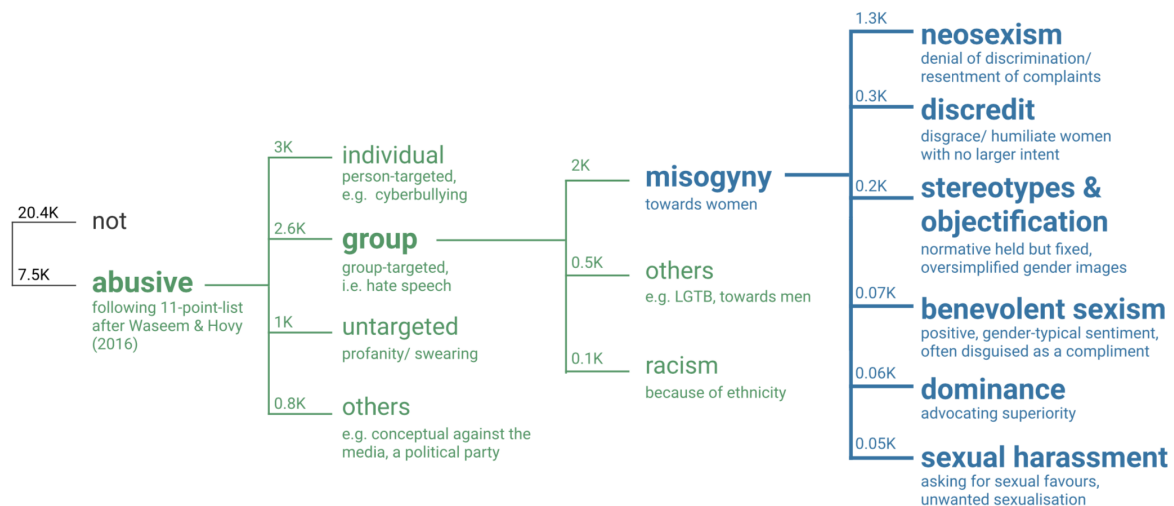
(e.g., *as good as a man*, *for a girl*, *#adaywithoutwomen*), while the tweets labelled as *hostile* and *others* come from the `Waseem` dataset (classes *sexist* and *neither*).

- **Parikh**. The dataset has been built by [Parikh et al. (2019)](#) and contains 13,023 accounts of sexism extracted from the Everyday Sexism Project website[41] manually annotated with 23 non mutually exclusive labels (cf. Table 3.1). The types of sexism identified in this corpus include: role stereotyping (cf. (3.12)), attribute stereotyping (cf. (3.13)), sexual harassment (excluding assault) (cf. (3.15)), body shaming and internalized sexism (cf. (3.14)).

(3.12) *Cool story babe. Now go make me a sandwich.*

(3.13) *Why is the economist still under 'Mens Interests' in my local supermarket?*

(3.14) *The weight will be hard to loose afterwards and my husband will find me less attractive.*

(3.15) *i can't even walk in town with my best friend without being wistled at, stared at, and have comments made to us.*

- **Guest**. It consists of 6,567 (primarily English) Reddit posts and comments collected from 24 subreddits that were either identified as misogynistic or, if not misogynistic,

---

[41] https://everydaysexism.com

Table 3.1 – Descriptions of the categories of sexism used in (Parikh et al., 2019).

CATEGORY	DESCRIPTION
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a coworker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats
Rape	FBI's expanded definition of rape
Sexual assault (excluding rape)	Any sexual contact without consent; unwanted touching
Sexual harassment (excluding assault)	Any sexually objectionable behaviour
Tone policing	Comments or actions that cause or aggravate restrictions on how women communicate
Moral policing (excluding tone policing)	The promotion of discriminatory codes of conduct for women in the guise of morality; also applies to statements that feed into such codes and narratives
Victim blaming	The act of holding the victim responsible (fully or partially) for sexual harassment, violence, or other sexism perpetrated against her
Slut shaming	Inappropriate comments made about women 1) deviating from conservative expectations relating to sex or 2) dressing in a certain way when it gets linked to sexual availability
Motherhood-related discrimination	Shaming, prejudices, or other discrimination or misconduct related to the notion of motherhood; also applies to the violation of reproductive rights
Menstruation-related discrimination	Shaming, prejudices, or other discrimination or wrongdoings related to periods
Religion-based sexism	Sexist discrimination or prejudices stemming from religious scriptures or constructs
Physical violence (excluding sexual violence)	Domestic abuse, murder, kidnapping, confinement, or other physical acts of violence linked to sexism
Mansplaining	A woman being condescendingly talked down to by a man; also applies when a man gives an unsolicited advice or explanation to a woman related to something she knows well that she disapproves of
Gaslighting	Sexist manipulation of the victim through psychological means into doubting her own sanity
Other	Any type of sexism not covered by the above categories

they discuss women or are related to misogyny. For annotating the corpus, Guest et al. (2021) developed a hierarchical taxonomy with three levels. After making the distinction between misogynistic and not misogynistic content, four subtypes of misogyny were elaborated: *misogynistic pejoratives*, *descriptions of misogynistic treatment* (threatening language and disrespectful actions), *acts of misogynistic derogation* (intellectual and moral inferiority, sexual and/or physical limitations) and *gendered personal attacks against women*.

Table 3.2 summarizes the available corpora for the task of detecting HS that targets women.[42] Note that we only present the distribution of instances annotated as conveying (or not) HS towards women (the number of instances for datasets that include in their an-

---

[42]A new shared task consisting in the identification of misogynous memes, MAMI, was recently proposed. For the moment, however, we do not have any information regarding dataset statistics. https://competitions.codalab.org/competitions/34175

notation scheme HS towards other targets are not provided).

Table 3.2 – Misogyny datasets.

Dataset	Language	Labels	No. of instances		Source	Availability
(Waseem and Hovy, 2016)	English	sexism	3,216	14,531	Twitter	
		none	11,315			✓
(Jha and Mamidi, 2017)	English	benevolent sexism	712	10,095	Twitter	
		hostile	2,254			
		other	7,129			✓
(Fersini et al., 2018a)	English	misogyny	2,245	5,000	Twitter	
		not misogyny	2,755			✓
(Fersini et al., 2018a)	Italian	misogyny	2,340	5,000	Twitter	
		not misogyny	2,660			✓
(Fersini et al., 2018b)	English	misogyny	1,851	3,977	Twitter	
		not misogyny	2,126			✓
(Fersini et al., 2018b)	Spanish	misogyny	2.064	4,138	Twitter	
		not misogyny	2,074			✓
(Basile et al., 2019)	English	HS towards women	2,608	9,544	Twitter	
		not hate speech	6,936			✓
(Basile et al., 2019)	Spanish	HS towards women	1,664	4,009	Twitter	
		not hate speech	2,345			✓
(Lynn et al., 2019)	English	misogyny	1,034	2,285	Urban Dictionary definitions	
		not misogyny	1,251			✓
(Parikh et al., 2019)	English	23 categories of sexism		13,023	Everyday Sexism Project	✓
(Sharifirad et al., 2019)	English	indirect harassment	260	3,240	Twitter	
		sexual harassment	417			
		physical harassment	123			
		non-sexist	2,440			✗
(Bhattacharya et al., 2020)	Hindi, Bangla, English	gendered/misogynystic	3,000	25,000	YouTube comments	
		non-gendered/non-misogynystic	23,000			✓
(Grosz and Conde-Cespedes, 2020)	English	sexist	627	1,142	Twitter	
		non-sexist	515			✓
(Guest et al., 2021)	English	misogyny	699	6,567	Reddit	
		not misogyny	5,868			✓
(Zeinert et al., 2021)	Danish	misogyny	2,000	22,400	Twitter	
		not abusive	20,400			✓

As it can be seen from Table 3.2, the size of the existing datasets ranges from 1,142 (Grosz and Conde-Cespedes, 2020) to 22,400 (Zeinert et al., 2021) instances and the most exploited data source is Twitter. In terms of annotation scheme (although common annotation schemes are used in benchmark corpora for shared tasks, e.g., `AMI corpora`), most of the presented works assume different levels of granularity.

Regarding the language, most of the resources are in English, or are part of a multilingual corpus (Bhattacharya et al., 2020). Other languages that are represented in the misogyny corpora include Danish (Zeinert et al., 2021), Italian (Fersini et al., 2018a) and Spanish (Fersini et al., 2018b; Basile et al., 2019). As far as we are aware, no such resource exists for French.

3.4 Sexism Detection in Social Media

The growing interest of the research community towards HS detection in recent years led to the development of a wide array of methods targeting the problem (cf. Section 2.4). However, the automatic detection of misogynistic content is still an open problem and computational approaches dealing with this phenomena are not as abundant.

Bartlett et al. (2014) analyze misogyny in terms of evolution over time by collecting English tweets (from UK based accounts) containing the word *'rape'*[43] or tweets containing terms that are broadly considered to be used in a misogynistic way (the most prevalent terms were *'slut'* and *'whore'*).[44] The authors estimate that 12% of the tweets containing the word *'rape'* appear to be threatening, and in 29 % of the tweets the term was used in a casual/metaphorical way. For the tweets containing the terms *'slut'* and *'whore'*, 18 % are estimated to be misogynistic, and 35 % tweets appeared to use the terms in a casual/metaphorical way.

Hardaker and McGlashan (2016) investigate the abuse directed towards the feminist campaigner, Caroline Criado-Perez. In order to study how the online discourse communities are formed, the authors relied on a corpus of 76,275 English tweets collected during a three month period (from July to September 2013). For the task at hand, Hardaker and McGlashan (2016) combine quantitative approaches from the fields of corpus linguistics (to detect emerging discourse communities) and qualitative approaches from discourse analysis (to analyse how these communities construct their identities).

A preliminary result of a study conducted by Fulper et al. (2014) suggests that the increase of misogynistic language on social media is associated with an increase in the number of sexual violence cases. In another study, Farrell et al. (2019) explore the evolution of misogynistic ideas within and across seven different Reddit communities. In particular, their results show that violence and hostility towards women online are increasing. In this context, it is important to automatically detect messages with a misogynistic content on social platforms and possibly to prevent its wide-spreading.

To better understand the context in which misogynistic language is used, Hewitt et al. (2016) used several terms related to slurs against women to gather 5,500 tweets over the span of a week. (Hewitt et al., 2016) is one of the first studies that highlights the challenges

---

[43]The tweets were collected over a period of two months, from 26 December 2013 to 9 February 2014.

[44]The tweets were collected over a period of one month, from 9 January to 4 February 2014.

of detecting online misogyny.

To our knowledge, the automatic detection of sexist messages currently deals only with English, Italian and Spanish and a fair amount of works that tackle this problem have come from teams that participated in recently shared tasks such as HatEval (Basile et al., 2019) (cf. Section 2.4) and AMI (Fersini et al., 2018b,a). AMI, a shared task in two different evaluation campaigns in 2018 (IberEval and Evalita), focuses on detecting HS that targets women. In English, the best results were achieved by traditional models for both AMI-IberEval (SVM with several handcrafted features (Pamungkas et al., 2018)) and AMI-Evalita (LR coupled with vector representation that concatenates sentence embedding, Tf/IDf and average word embeddings (Saha et al., 2018)).

A few notable neural network techniques include (Jha and Mamidi, 2017) who employed an LSTM model to classify messages as: benevolent, hostile and non-sexist. Zhang and Luo (2019) implement two deep neural network models (CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction) in order to classify social media texts as racist, sexist, or non-hateful. Karlekar and Bansal (2018) use a single-label CNN-LSTM model with character-level embeddings to classify three forms of sexual harassment: commenting, ogling/staring, and touching/groping. Sharifirad et al. (2018) focus on diverse forms of sexist harassment (indirect, information threat, sexual, physical) using LSTM and CNN on augmented dataset obtained via ConceptNet *is-a* relationships and Wikidata. Finally, Parikh et al. (2019) consider messages of sexism experienced by women in the *Everyday Sexism Project* web site and classify them according to 23 non mutually exclusive categories using BiLSTM, CNN, and CNN-BiLSTM models trained on top of several distributional representations (character, subwords, words and sentence) along with additional linguistic features. Both distributional word vectors (FastText (Grave et al., 2018), Glove (Pennington et al., 2014), ELMo (Peters et al., 2018)) and sentence embeddings were explored (BERT (Devlin et al., 2019), USE (Cer et al., 2018b), InferSent (Conneau et al., 2017a)).

Finally, Parikh et al. (2019) employed a wide range of features from the work of Recasens et al. (2013) on biased language detection including the presence (or absence) of: assertive verbs, hedges, factive verbs, entailment, implicative verbs, report verbs, strong subjective, weak subjective, positive words and negative words. The authors exploited emotion signals: eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) from the NRC Emotion Lexicon (Mohammad and

Turney, 2013) and affect scores (valence, arousal, dominance) (Mohammad, 2018). In addition, Parikh et al. (2019) also considered PERMA (Positive Emotions (positively valenced emotions, e.g., joy, contentment, excitement), Engagement (multi-dimensional construct that includes behavioral, cognitive, and affective components), Relationships, Meaning, and Accomplishments) features (Schwartz et al., 2016).

Conclusion

In this part we examined the concept of HS from two perspectives: definitions employed by either international organizations or scholars. As the main focus of this work is sexist HS detection in French tweets, we only addressed the European legislation. In addition, we also investigated the different topical focuses (e.g., misogyny, sexism, racism, etc.) and targets (i.e., the community receiving the hatred) of HS.

The study of HS detection is multifaceted, and available datasets feature different focuses and targets. Despite limitations, some works have tried to bridge this range by proposing a domain adaptation approach to transfer knowledge from one dataset to other datasets with different topical focuses (cf. Section 1.2). Regarding the language, the majority of the resources are in English, however, the growing interest of the research community towards HS detection (and other related phenomena) has enabled a greater linguistic diversity.

Despite the plethora of research dealing with HS detection, as far as we are aware, no work has addressed sexism detection in French, although recent years have shown increased efforts aimed at dealing with this problem. Moreover, previous work considers sexism either as a type of HS, along with racism, homophobia, or HS against immigrants (Schrading et al., 2015; Waseem and Hovy, 2016; Golbeck et al., 2017; Davidson et al., 2017; Basile et al., 2019) or study it as such. In this latter case, detection is casted as a binary classification problem (*sexist* vs. *non-sexist*) or a multi-label classification by identifying the type of sexist behaviours (Jha and Mamidi, 2017; Sharifirad et al., 2018; Fersini et al., 2018c; Karlekar and Bansal, 2018; Parikh et al., 2019).

Although social media and web platforms have provided a space in which sexist HS thrives, they have also offered a space in which women can finally report the sexist behaviours they experience (see hashtags such as *#metoo* or *#balancetonporc*). In this context we believe that it is important not only to be able to automatically detect messages with

a sexist content but also to make the distinction between reports/denunciations of sexism experiences and *'real'* sexist messages. This will constitute the objective of Part II of this dissertation.

Part II

Sexism Detection in French Tweets

Motivation

Social media and web platforms have offered a large space for spreading sexist hate speech (in France, 10% of sexist abuses come from social media (Bousquet et al., 2019)) but also allow sharing stories of sexism experienced by women (see *The Everyday Sexism Project*[45] available in many languages, *Paye ta shnek*[46] in French, or hashtags such as #*metoo* or #*balancetonporc*). In this context, it is important to automatically detect sexist messages on social platforms and possibly to prevent the wide-spreading of gender stereotypes, especially towards young people, which is a first step towards offensive content moderation (see the recommendations of the European Commission).[47]

In this dissertation, as we address the problem of French sexist messages detection, we consider sexism in its common French usage, i.e., *discrimination or hate speech against women*.

We propose to undertake the following challenges:

- (1) Introduce *a novel characterization of sexist content-force relation* inspired by speech acts theory (Austin, 1962) and discourse studies in gender (Lazar, 2007; Mills, 2008).

- (2) Develop *the first French dataset of about* 12, 000 *tweets annotated for sexism detection* according to this new characterization.

- (3) Experiment with the *development of models able to detect different types of sexism experiences in French tweets*, based on their impact on the target, going beyond standard binary classification.

The reminder of this part is organized as follows. In the next chapter, we detail the data and the annotation process for the new French sexism corpus. Chapter 2 presents the experiments carried out when investigating whether different models are able to distinguish

---

[45]https://everydaysexism.com/
[46]https://payetashnek.tumblr.com/
[47]https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC0555

between reports of sexism experiences and *'real'* sexist messages, as well as a pilot study in which we investigate whether these models are capable of detecting target agnostic HS. We end this part by discussing our main findings.

1 Data and Annotation

Before detailing our annotation scheme and the result of the annotation procedure, we present the theoretical backgrounds on which we based our study.

1.1 Characterizing Sexist Content

Sexism may be expressed explicitly or implicitly (see the following tweets from our French data) using different pragmatic devices, including:

- Negative opinion, abusive message:

  (1.1) *Meuf tu connais rien au foot. Tais toi. Contente de fan girler sur les joueurs et de mouiller sur MBappé & Neymar. Merci bien. Une fille qui connait le foot..*
  (Girl, you know nothing about football. Shut up. Happy to be a fangirl of the players and get wet because of MBappé & Neymar. Thanks a lot. A girl who knows football.)

- Stereotype:

  (1.2) *C'est bon t'es une femme forte, te manque que la cuisine pour atteindre la perfection*
  (It's ok you're a strong woman, you only need the kitchen to reach perfection)

- Humor, irony:

  (1.3) *Le fait maison c'est toujours mieux. La preuve, on préfère toujours sa femme à sa prostituée. #humour.*
  (Homemade is always better. The proof, we always prefer the wife to the prostitute. #humor)

- Benevolent sexism (i.e., a sexist comment expressed positively, in the form of a compliment (Glick and Fiske, 1996)):

(1.4)   *Elle court vite pour une femme.*

         (She runs fast for a woman.)

Propositional content can be introduced in discourse by acts of varying forces (Austin, 1962): it can be asserted (e.g., *Paul is cleaning up his room*), questioned (e.g., *Is Paul cleaning up his room*?), or asked to be performed as with imperatives (e.g., *Paul, clean up your room*!). In philosophy of language on the one hand and feminist philosophy on the other, speech acts have already been advocated in a variety of manners. Most accounts however either focus on the type of act (assault-like, propaganda, authoritative, etc.) that derogatory language performs (Langton, 2012; Bianchi, 2014) or concentrate on the analytical level at which the derogatory content is interpreted, whether it provides meaning at the level of the presupposition (or more largely non at-issue content (Potts, 2005)) or of the assertion (Cepollaro, 2015).

Our study pursues a different line of analysis, whereby speech acts bearing on derogatory content are ranked according to their perlocutionary force and assertions are classified as more or less direct. Specifically, in order to make emerge different degrees of downgrading tones, we have chosen to distinguish cases where the addressee is directly addressed from those in which she is not, as done in HS analysis (ElSherief et al., 2018a; Ousidhoum et al., 2019). ElSherief et al. (2018a) consider that directed HS is explicitly directed at a person while generalized HS targets a group. For (Ousidhoum et al., 2019), a hateful tweet is direct when the target is explicitly named, or indirect when it is *'less easily discernible'*. In this respect our categorization overlaps for the direct category, but differs from previous approaches in that it casts it in a classification of perlocutionary forces and thus of potential impact on the target of the sexist act.

Unlike these approaches and the definitions of target used in (Fersini et al., 2018a; Basile et al., 2019), we do not consider the number of targets of a sexist message (it can be either a woman, a group of women or all women) but rather distinguish the target from the addressee. Our use of the notions of *'directness'* and *'indirectness'* is also transverse to the ones used in (Lazar, 2007; Chew and Kelley-Chew, 2007) or Mills (2008), who resort to the label indirectness for subtle forms of sexism that perpetuate gender stereotypes through humor, presuppositions, metaphors, etc.

Our notions of *directedness* and *indirectedness* target the force of the speech act type as immediately addressing the target of the sexist content (*direct act*), describing the target (*in-*

direct act) or that the target is reporting (*reported sexism*). All these three types of acts can contain subtle and non-subtle sexist content. The main goal of our classification is thus to focus on the impact of the content by resorting to the force of the act and not only to its content. Directedness or absence of it, is a new categorization that allow us to envisage sexism not only from the perspective of the content of the assertion, but from the perspective of the impact of the assertion on the victim. We have thus established a new correlation between three types of speech acts of assertion and sexist messages.

Sexist content in **directed assertions** is explicitly directed at a woman but contrary to both approaches cited above, it can also be directed at a group of women or all women. Across the different classifications of speech acts (Portner, 2009) there is a basic distinction that cuts across different types of speech acts: directedness and indirectedness. *Direct* speech acts such as imperatives are addressee-oriented and they require that the addressee performs an action (responding (with questions) or acting (with imperatives)). Indirect speech acts are not addressee-oriented.

Assertions themselves can be direct or indirect. They are direct when they are in the second person (i.e., *you*). They require that the addressee be committed to the truthfulness of their content. Since a direct sexist assertion is a type of speech act that immediately involves the addressee and triggers a request of commitment, direct assertions of sexism have been ranked as most prominent expressions of sexism with the greater impact on the victim. Most prominently, with assertions, directedness is the trigger of perlocutionary content, rendering the assertion an *insult*.

Descriptive assertions are not directed to the addressee: the target can be a woman, a group of women, or all women, it can be named but is not the addressee. Descriptive assertions are in the third person and thus may have a lower impact on the addressee in comparison with second person assertions, as they do not commit her to the truth of the content by soliciting a response. They report generic content and not *ad personam* content (Mari et al., 2012). Nonetheless, they convey sexist content and are downgrading for the target of the description.

Finally, in **reported assertions**, the sexist content is a report of an experience or a denunciation of a sexist behaviour. They may elicit an even lower commitment on behalf of the addressee (see (Portner, 2009; Giannakidou and Mari, 2021) for a general discussion on evidentiality and reportativity). The speaker is not committed to the truth of a reported content

(as in *I heard that you were coming too*). However, when reporting sexist content, the speaker is still conveying lack of commitment, and a general sense of disapproval or dismissal may emerge.

As it appears, the three types of assertions have a sexist content, but only the first two ones are really sexist. Indeed, direct and descriptive assertions are first-hand information, whereas reported ones are second-hand information. As such, they may trigger a different reaction from the receiver: in the first two cases, the addressee is immediately involved as the target of the sexist dismissal; in the third case, she is the witness of a sexist report. By reporting a sexist speech, the target can distance, comment, or even denounce. Our classification is one that allows to potentially investigate different reactions, both as psychological effects and action that the target can undertake.

1.2   Data Collection

Our corpus is new and it contains French tweets collected between October 2017 and May 2018. In order to collect sexist and non sexist tweets, we followed Anzovino et al. (2018) approach using:

- a set of representative keywords: *femme, fille (woman, girl)*, *enceinte (pregnant)*, some activities (*cuisine (cooking), football, journaliste*), insults (*pute, salope, conne, connasse (slut, bitch), hystérique*);

- the names of women/men potentially victims or guilty of sexism (mainly politicians) : *Theresa May, Angela Merkel, Nicolas Hulot, etc.*;

- specific hashtags to collect stories of sexism experiences: *#balancetonporc, #moiaussi, #sexisme, #sexiste, #SexismeOrdinaire, #EnsembleContreLeSexisme, #payetashnek, #payetontaf, #payetonsport*.

Thus, we collected around 115,000 tweets. After removing the duplicates, about 30,000 tweets contain the specific hashtags. The keyword distribution is presented in Table 1.1.

Table 1.1 – Keyword distribution in our French dataset.

Representative keywords		Persons		Specific hashtags	
femme	6,903	fillon	1,435	#payetonbahut	250
fille	621	hulot	1,134	#payetonsport	51
féminisme	1,591	darmanin	1,667	#payetonj-ournal	305
féministe	1,928	peneloppe fillon	1,110	#payetashnek	1,078
connasse	3,086	catherine deneuve	1,482	#balancetonporc	16,711
conne	2,668	angela merkel	1849	#moiaussi	3994
salope	2,840	aurore bergé	5,119	sexisme	3,006
pute	4,055	theresa may	2,900	#sexisme	1,739
hystérique	2,434	valérie pecresse	3,860	#sexiste	2,131
aussi bien qu'un homme	369	christiane taubira	3,133	#EnsembleContre	
aussi vite qu'un homme	74	christine lagarde	1,579	LeSexisme	516
comme un homme	1,311	segolene royal	2,309		
pour une fille	3,229	nadine morano	7,554		
drague	909				
viol	4,379				
cuisine	2,491				
enceinte	4,345				
foot	4,742				
bleues	1,146				
journaliste	1,424				
TOTAL	50,545		35,131		29,781

1.3 Annotation Guidelines

We used a set of 150 tweets to define the annotation guidelines. As previously stated, the novelty of our approach lies in the fact that we want to identify sexist content in tweets and also determine whether the tweet is really sexist (i.e., directly address to a target or describing a target) or is a story of sexism experienced by a woman. The annotation scheme of the French sexism corpus is presented in Figure 1.1.

Given a tweet, its annotation consists in assigning it one of the following three categories:

(i) Non-sexist. The tweets falling under this category have no sexist content, as in (1.5), (they may contain a specific hashtag but the content is not sexist) or they mention sexism-related topics, but do not comment on them (e.g., news reports and other messages that are

Figure 1.1 – The annotation scheme of the French sexism corpus.
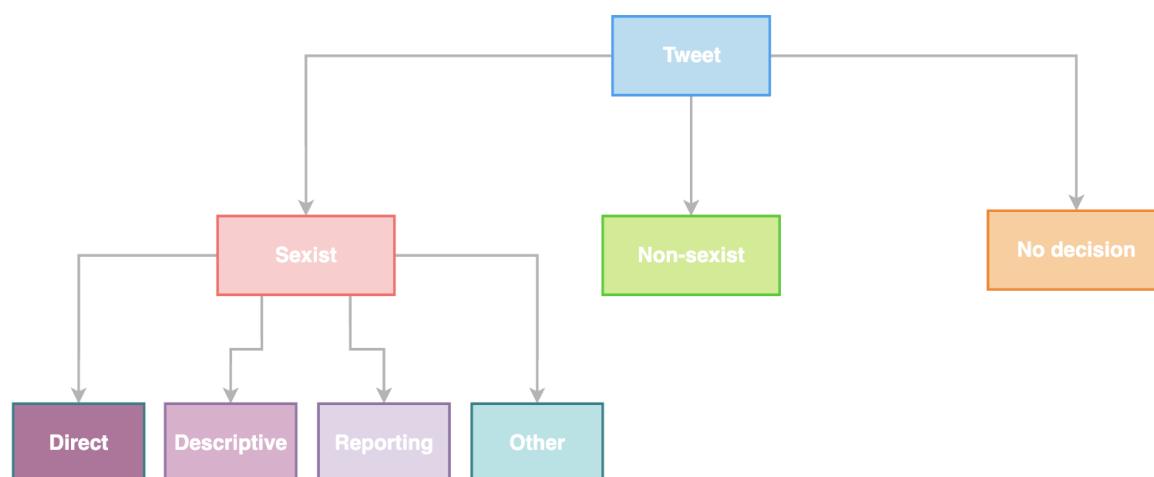
intended to be neutral, cf. (1.6)).

(1.5)   *Paris Match : journal d'investigation*
        (Paris Match: an investigative journal)

(1.6)   *La créatrice du #balancetonporc attaquée en justice pour diffamation*
        (The creator of #SquealOnYourPig sued for defamation)

(ii) No decision. This category is used to annotate tweets that are too ambiguous to be categorized as **sexist** or **non-sexist**. This can be due to a lack of context, either regarding the conversation the tweet is part of, or regarding the topic it addresses. The form of the message can also be the root cause of the problem: the tweet is unintelligible, unfinished or abstruse, notably because of too many spelling mistakes, expressions, or specific abbreviations (cf. (1.8), an ambiguous instance in which it is difficult to identify the position taken by the author). This category also includes tweets in foreign languages and tweets that incorporate an image (cf. (1.7)), video, or link that requires interpretation to determine if the tweet is sexist.

(1.7)   *J'ai envie de poster ça sur facebook mais j'ai peur des commentaires ...*
        pic.twitter.com/kMq(...)
        (I'd like to post this on Facebook but I fear comments... pic.twitter.com/kMq(...))

(1.8) 😭😭😭😭*Mais la vie elle est drole le gars le matin il s'est dit on va faire une affiche avec une femme, une fleche et le mot cuisine*

(😭😭😭😭*But life is funny, the guy said to himself in the morning we'll make a poster with a woman, an arrow and the word kitchen)*

(iii) Sexist content: it can be either **direct**, **descriptive** or **reporting**. The first two categories comprise real sexist messages, but not the last one, as reporting tweets must not be considered as sexist in the context of moderation.

Direct sexist content, directly addressed to a woman or a group of women, generally uses second person pronoun/verb and imperatives, as shown in the examples below (linguistic clues are underlined).

(1.9) *t'es une femme pq tu veux parler de foot?*

(*You're a woman why do you want to talk about football?*)

(1.10) *les femmes qui sont en plus Dijonnaise ne parlez pas de foot sivouplai c'est comme si un aveugle manchot parler de passer le permis*

(*women who are also from Dijon please don't talk about football it's as if a one-handed blind person was thinking about getting a driving license*)

In both (1.9) and (1.10) the same stereotype is employed (i.e., football is perceived as a male sport), but the target differs: whereas the first tweet is directly addressed to a woman (i.e., the second person singular is used), the second tweet targets a group of women (women from Dijon in particular).

In **descriptive sexist content** the tweet describes a woman (indirectly, by using the third person singular) or women in general. This type of tweet is particularly prone to stereotyping and the linguistic clues include the presence of a named entity as the target or use of generalizing terms, as shown in (1.11) and (1.12).

(1.11) *Anne Hidalgo est une femme. Les femmes aiment faire le ménage. Anne Hidalgo devrait donc nettoyer elle-même les rues de Paris*

(*Anne Hidalgo is a woman. Women love cleaning the house. Anne Hidalgo should clean the streets of Paris herself*)

(1.12) *une femme a besoin d'amour de remplir son frigo, si l'homme peut le lui apporter en con-
trepartie de ses services (ménages, cuisine, etc) j'vois pas elle aurait besoin de quoi d'autre*

*(A woman needs love, to fill the fridge, if a man can give this to her in return for her services
(housework, cooking, etc), I don't see whatelse she needs)*

When the sexist content is in fact a **report** of a sexism experience or a denunciation of
a sexist behaviour, we observe the presence of reporting verbs, quotation, locations (as re-
ports often mention public spaces where the experience happened) or specific hashtags (e.g.,
#balanceTonPorc, *#moiAussi*), as shown in (1.13), (1.14) and (1.15).

(1.13) *je m'assoupis dans le métro, je rouvre les yeux en sentant quelque chose de bizarre : la main
de l'homme assis à côté de moi sur ma cuisse. #balancetonporc*

*(I doze in the subway, I open my eyes feeling something weird: the hand of the man sat next
to me on my leg #SquealOnYourPig)*

(1.14) *Mon patron m'a demandé : "qui va cuisiner pour ton mari quand tu seras pas là ?"*

(My boss asked me: "who's going to cook for your husband when you're away?")

(1.15) *Je ne suis pas une grande fan de Theresa May mais pourquoi parler de "ses escarpins et ses
cuissardes vernies" et la traiter d'allumeuse ? #vincenthervouet #sexisme*

*(I am not a fan of Theresa May but why talking about "her shoes and varnished boots" and
call her a tease? #vincenthervouet #sexism)*

Note that it is possible to classify tweets in the sexist category without classifying them in
one of its subcategories (i.e., *direct*, *descriptive*, *reporting*) and these differ from the instances
annotated as **non-sexist** and **no decision** (i.e., tweets that seem to have a sexist character,
but are too ambiguous to be classified). This means that the tweet has a sexist character,
however it does not necessarily approve of the message and can, on the contrary, denounce
it. For example, (1.16) is similar to a reporting tweet but the term *Lol* (i.e., laughing out loud)
makes its interpretation ambivalent. A similar issue arises in (1.17) where the expression *The
sweet sound* is ambiguous and makes the position of the author undetermined: denunciation
or encouragement?

(1.16) *Mon père toutes les 5 min: "c'est quoi ce tir de femme enceinte" Mdrrrr #FRAPAR*

My dad every 5 min: "what's with the pregnant woman shooting" Lollll #FRAPAR

(1.17) *Le doux son de la culture du viol*

       The sweet sound of rape culture

1.4 Manual Annotation

300 tweets have been used for the training of five annotators, master degree's students (two female and two male) and a computational science researcher (female) in Communication and Gender, and then removed from the corpus. Subsequently, 1,000 tweets have been annotated by all annotators so that the inter-annotator agreement could be computed. Although the perception of sexism is often considered subjective, the average Cohen's Kappa is:

- **0.72** for the **sexist/non-sexist/no decision** categories

- **0.71** for the **direct/descriptive/reporting/non-sexist/no decision** categories

which indicates a strong agreement.

One example of disagreement in the annotation process is presented in (1.18), having annotators labeling the tweet as *sexist* or as *reporting*. This disagreement could be attributed to the annotators misunderstanding the tweet as *'Valérie Pécresse is a woman and should do the housework'*, whilst the video embedded in the text shows Valérie Pécresse ironically making the comment *'women should do the housework'*.

(1.18) *... Valérie Pécresse Rien de tel qu'1 femme pr faire le ménage* `https://vine.co/v/ee`

       `ZWtelZQlH`

       (*Valérie Pécresse Nothing is better than a woman for doing the housework* `https://vine`

       `.co/v/eeZWtelZQlH`)

We noticed that the Kappa scores between female annotators are very close to the ones between male annotators and the main disagreements are between the **non-sexist** and **descriptive** categories.

For these 1,000 tweets, the final labels have been assigned according to a majority vote.

Finally, a total of 12,274 tweets have been annotated according to the guidelines after removing the tweets annotated as **no decision**. Table 1.2 and Figure 1.2 show the distribution of the annotated corpus.

Table 1.2 – Tweet distribution in our French dataset.

Sexist content				Non-sexist	Total
4,487				7,787	12,274
direct	descriptive	reporting	other		
45	780	3,222	440		

Figure 1.2 – Tweet distribution in our French dataset.
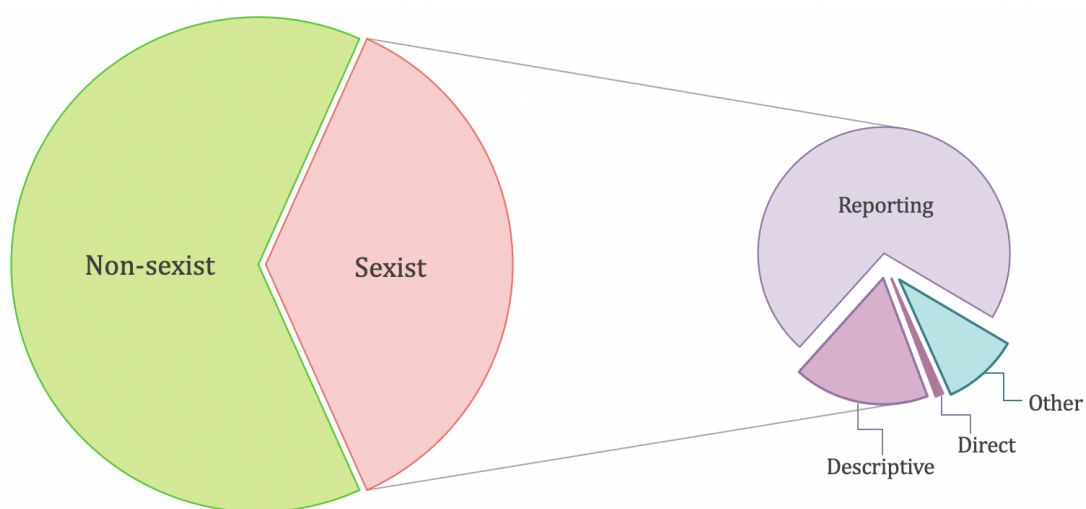
In this chapter, we have presented the first corpus of French tweets annotated for sexism detection. The novelty of our approach is that not only tweets with a sexist content are labelled but the type of content is also characterized: the tweet is either directly addressed to a target (a woman or all women), describes a target or reports/denounces sexism experienced by a woman. We think that it is important to distinguish between these usages in a context of offensive content moderation on social media since stories of sexism experiences should not be moderated.

2 Sexist Hate Speech Detection

Now that we presented the first French dataset annotated for sexism detection according to a novel characterization of sexist content-force relation inspired by speech acts theory and discourse studies in gender, in the first part of this chapter, we introduce a set of experiments to detect sexist content in three configurations: binary classification (*sexist* content vs. *non-sexist*), three classes (*reporting* content vs. *non-reporting* vs. *non-sexist*), and a cascade classifier (first sexist content and then reporting). These configurations go beyond standard binary classification of HS messages (cf. Section 3.4).

Then, we present a pilot study in which we investigate to what extent HS detection is target-dependent and whether the creation of models able to capture common properties of HS is feasible.

2.1 Sexism Detection

In this section we aim to answer two main research questions:

- *Are models able to distinguish between reports of sexism experiences and 'real' sexist messages?*

- *Can learning from generalized concepts improve the performance of a model?*

2.1.1 Methodology

In order to identify sexist assertions, we propose the following three configurations:

- *BIN*: *sexist* vs. *non-sexist* content

- *3-CLASS*: *sexist* (i.e., *direct* and *descriptive*) vs. *reporting* vs. *non-sexist* content

- *CASC*: a cascade classification with *sexist* vs. *non-sexist* content in the first stage, followed by *reporting* vs. *non-reporting* in the second stage.

For the task at hand, we propose several models ranging from standard Bag of Words (our baseline) to deep learning models. To this end, our annotated corpus has been divided into train and test sets. Table 2.1 shows the distribution of these sets. In the next sections, we detail our models, provide and discuss our results.

Table 2.1 – Tweet distribution in train/test datasets.

		Sexist content		Non sexist
TRAIN	3,564			6,255
	direct	**descriptive**	**reporting**	
	38	599	2,559	
TEST	923			1,532
	direct	**descriptive**	**reporting**	
	7	181	663	

2.1.2 Models

In this section, we first provide a detailed description of all the models we experiment with. We end this section by summarizing them.

2.1.2.1 Models Description

We experiment with the following architectures:

– **Baseline (SVM$_{BoW}$)**. The baseline is a SVM (linear kernel, $C = 0.1$) with unigrams, bigrams and trigrams Tf/IDf.

– **BiLSTM$_{attention}$**. The model uses a Bidirectional LSTM with an attention mechanism that attends over all hidden states and generates attention coefficients.[48] The hidden states were then averaged using the attention coefficients in order to generate the final state which

---

[48]We also experimented with other neural architectures, like CNN, but the results were lower.

was then fed to a one-layer feed-forward network for obtaining the final label prediction. We used pre-trained on Wikipedia and Common Crawl FastText French word vectors with an embedding dimension of 300 (Grave et al., 2018). We experimented with different hidden state vector sizes, dropout values and attention vector sizes. The results reported here were obtained by using 300 hidden units, an 150 attention vector, a dropout of 50% and the Adam optimizer with a learning rate of 10^{-3}. For the BiLSTM we used a ReLU activation function and we run all the experiments for maximum 100 epochs, with a patience of 10 and batch size of 64.[49]

– **CNN**. This model was inspired by (Badjatiya et al., 2017; Gambäck and Sikdar, 2017). It uses FastText French word vectors (with the dimension of 300) and three 1D convolutional layers, each one using 100 filters and a stride of 1, but with different window sizes (respectively 2, 3, and 4) in order to capture different scales of correlation between words, with a ReLU activation function. We further downsample the output of these layers by a 1D max-pooling layer and we feed its output into the final dense layer.

– **CNN-LSTM.** This model extends the previous CNN model by adding a LSTM layer[50] (capable of capturing the order of a sequence) that takes its input from the max pooling layer. Next, a global max pooling layer feeds the highest value in each timestep dimension to a final softmax layer.

– **BERT$_{\text{base}}$**. It uses the pre-trained BERT model (BERT-Base, Multilingual Cased) (Devlin et al., 2019) on top of which we added an untrained layer of neurons. We then used the HuggingFace's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for 3 epochs.

After an exploratory analysis of the data, we observed that about 47% of the tweets embed in their text at least one URL. Due to the short length of a tweet, incorporating URLs is useful for amplifying the message, while also minimizing the time it takes to compose the message. By ignoring the content present at a shared URL, an important part of the meaning of the message is lost, as it becomes harder to identify the context. In order to feed more

---

[49]The hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.

[50]We also experimented with GRU, but the results were lower.

information to the classifier, instead of removing or replacing the URLs with replacement tokens, we propose to substitute them with the title found at the given URL.[51] For example, the URL (2.1) will be replaced with the title (2.2).

(2.1)   *https://marieclaire.be/fr/liberte-dimportuner-sexisme-catherine-deneuve/*

(2.2)   *Tribune sur la « liberté d'importuner »: le sexisme expliqué à Catherine Deneuve*

Emotional content holds an important place in language, as sometimes, what people write may not actually reflect their feelings at the time of writing those words. Emojis have become very popular in social media and are interesting because they encode meaning that otherwise would require more than one word to convey (e.g., grinning face, smiling face with 3 hearts, beaming face with smiling eyes, etc.). Based on the assumption that word embeddings capture the meaning of words better than emoji embeddings capture the meaning of emojis, we followed the strategy proposed by Singh et al. (2019) and we replaced all the emojis with their detailed descriptions.[52]

For example, in the following tweet:

(2.3)   *La journée #EnsembleContreLeSexisme est une très bonne initiative de @MarleneSchiappa*
👏 *Il faut juste rappeler que le sexisme contre les hommes existe aussi. Le schéma femme =*
victime et homme = porc n'est pas aussi simple ✌️
(The #TogetherAgainstSexism day is a very good initiative by @MarleneSchiappa 👏 *We*
just need to remember that sexism against men also exists. The schema woman=victim and
man=pig is not so simple ✌️ *)*

- 👏 will be replaced with the emoji description: <Applaudissements> (*applause*)

- ✌️ will be replaced with the emoji description: <V de la victoire> (*victory hand*).

After replacing the URLs and emojis as described above, several deep learning models were also trained and evaluated on our dataset (the models adopting this replacement strategy will incorporate in their name ^{R}).

---

[51] In case a particular webpage is not available anymore, the URL is removed from the tweet.
[52] We relied on a manually built emoji lexicon that contains 1,644 emojis along with their polarity and detailed description.

- **SVM$^R_{BoW}$.** This is the same model as the baseline (i.e., SVM$_{BoW}$) but adopts in its preprocessing step the URL and emoji replacement strategies.

- **BERT^R.** This model uses the same architecture as **BERT$_{base}$**, but adopts in its pre-processing step the URL and emoji replacement strategies. Replacing URLs and emojis improved the results for all the models we have tested, therefore we adopted these replacement strategies for all the models.

- **BERT$^R_{own\_emb + base}$.** Following Parikh et al. (2019), we also experiment with stacking multiple embeddings. We tailored a pre-trained BERT model[53] for which we used the whole non annotated dataset (i.e., 205,000 tweets). The original BERT model uses a WordPiece tokenizer, which is not available in OpenSource. Instead, we used a SentencePiece[54] tokenizer in unigram mode. Training the model using the Google Cloud infrastructure with the default parameters for 1 million steps took approximately 3 days.

- **BERT$^R_{features}$.** We relied on state of the art features that have shown to be useful for the task of HS detection:

  - *Surface features* (tweet length in words, the presence of personal pronoun and third-person pronoun, punctuation marks, URLs, images, hashtags, @userMentions and the number of words written in capital);

  - *Emoji features* (number of positive and negative emojis from a manually built emoji lexicon that contains 1,644 emojis);

  - *Opinion features* (number of positive, negative and neutral words in each tweet relying on opinion (Benamara et al., 2014), emotion (Piolat and Bannour, 2009) and a manually built slang French lexicon containing 389 words[55]);

  - Hedges (negation and modality), reporting verbs, imperative verbs, and verbs used for giving advice (e.g., *advise, suggest, recommend*).

---

[53]We experimented with different configurations by incorporating different French pre-trained embeddings available: Glove (Pennington et al., 2014), FastText (Grave et al., 2018), Flair (Akbik et al., 2018) and CamemBERT (Martin et al., 2020) but none of the configurations were able to achieve results better than BERT$_{base}$.

[54]https://github.com/google/sentencepiece

[55]http://www.linternaute.com/dictionnaire/fr/usage/argot/1/

Sexism is often expressed by using gender stereotypes (i.e., ideas whereby women and men are arbitrarily assigned characteristics and roles determined and limited by their gender). In order to force the classifier to learn from generalized concepts rather than words which may be rare in the corpus, we adopt several replacement combinations extending Badjatiya et al. (2017)'s approach consisting in replacing some words/expressions that trigger sexist content by their generalized term. However, instead of using a flat list composed of most frequent words that appear in a particular class and then replace them by similarity relationships, we rather rely on manually built lists of words[56] often used in sexist language (hereafter `<SexistVocabulary>`): [57]

- **designations** (around 10 words such as *femme (woman), fille (girl), nana (doll), ...*);

- **insults** (around 400 words/expressions extracted from GLAWI, a machine-readable French Dictionary (Hathout and Sajous, 2016));

and 130 gender stereotyped words grouped according to the following taxonomy as usually defined in gender studies (see Section 2.1):

- **physical characteristics** (e.g. *petite (little), bouche (mouth), robe (dress), ...* for women; *petit (little), gros (fat), ...* for men);

- **behavioural characteristics** (e.g. *bavarde (gossipy), jalouse (jealous), tendre (loving), ...* for women; *macho, viril (virile), ...* for men);

- **type of activities** (e.g. *mère (mother), cuisine (cooking), infirmière (nurse), ...* for women; *football, médecin (doctor), ...* for men).

Please note that only 1% of all these words have been used as keywords for collecting the corpus.

In addition, we also built two other lists:

- **names** (952/832 female/male first names to detect named entities)

---

[56]Available at https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets/tree/master/generalization.

[57]Following Badjatiya et al. (2017), we also experiment with automatic word lists but the results were not conclusive as frequent words were too generic and not representative of the problem we want to solve.

- around 170 words/expressions for **places** as they are mainly useful for detection of reporting messages since they represent public spaces where sexist acts may occur.(e.g. *métro (subway), rue (street), bureau (office), ...*).

To this end, we experimented with distinct generalization strategies:

– BERT$^R_{\text{gen (X)}}$. This model is similar to **BERT^R** but the words/expressions that trigger sexist content are replaced by their generalized term (where **gen (X)** denotes the adopted replacement strategy):

1. Hypernym replacement:

   - **gen(Hypernym)** e.g., *petite/petit (little)* are replaced by `<Physical_Characteristics>`,

   - **gen(Hypernym_gendered)** e.g., *hysterique (hysterical)* is replaced by `<female_Behavioural_Characteristics>` as in (2.4), while *macho* is replaced by `<male_Behavioural_Characteristics>`),

   - **gen(SexistVocabulary)** e.g., both *petite (little)* and *poupée (doll)* are replaced by the same tag `<Sexist_Vocabulary>`

   (2.4) *Tu ne doit plus avoir de wifi lol, https://www.ncbi.nlm.nih.gov/pubmed/3455741 tien un peux de lecture ne ferai pas de mal a ton cerveaux <u>hysterique</u> feministe voyant le racisme partout*
   (You shouldn't have wifi anymore lol, https://www.ncbi.nlm.nih.gov/pubmed/3455741 here a little reading wouldn't hurt your <u><female_Behavioural_Characteristics></u> feminist brain seeing racism everywhere)

2. Named entities replacement:

   - **gen(Place)** where the public space in which the sexist act occurred is replaced by `<location>` as in (2.5):

   (2.5) *je m'assoupis dans le <u>métro</u>, je rouvre les yeux en sentant quelque chose de bizarre : la main de l'homme assis à côté de moi sur ma cuisse. #balancetonporc*
   (I doze in the <u><location></u>, I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig)

- **gen(Name)** where the named entity is replaced by `<Name>` (e.g., both *Yvonne* and *Pierre* are replaced by `<Name>`) and **gen(Name_gendered)** (e.g., *Yvonne* is replaced by `<female_Name>` and *Pierre* is replaced by `<male_Name>`)

2.1.2.2 Summary of the Proposed Models

Table 2.2 summarizes all the models that were employed for the task of sexism detection.

Table 2.2 – Models employed for the task of sexism detection. ‡: baseline models.

MODEL	DESCRIPTION
SVM‡	- inspired by Badjatiya et al. (2017)
SVM$^R_{BoW}$	- based on SVM‡ but adopts URL/emoji replacement strategies
BiLSTM$_{attention}$ ‡	- inspired by Parikh et al. (2019)
CNN ‡	- inspired by Badjatiya et al. (2017); Gambäck and Sikdar (2017)
CNN-LSTM ‡	- inspired by Karlekar and Bansal (2018)
BERT$_{base}$ ‡	- inspired by Parikh et al. (2019)
BERT^R	- based on BERT$_{base}$ ‡ but adopts URL/emoji replacement strategies
BERT$^R_{own\_emb+base}$ ‡	- inspired by Parikh et al. (2019)
BERT$^R_{features}$	- BERT^R that incorporates state of the art features that have been shown to be useful for the task of HS detection
BERT$^R_{gen}$	- BERT^R in which words/expressions that trigger sexist content are replaced by their generalized term

2.1.3 Results

2.1.3.1 Results for the *BIN* Configuration

Table 2.3 shows how the experiments were set up and presents the best results in terms of accuracy (*A*), macro-averaged F-score (*F1*), precision (*P*) and recall (*R*) in bold.

Among the seven models, BERT$_{base}$ represents our best performing one in terms of both accuracy and F-score. Replacing URLs and emojis with respectively the words within the title link and emoji description boosts the results of BERT^R by 3.6% in terms of F-score, while the SVM classifier applied to the dataset pre-processed with the same strategy (i.e., SVM$^R_{BoW}$) provides the highest precision amid all the models.

Table 2.3 – Results for *sexist* vs. *non sexist* content classification.

CLASSIFIER	A	P	R	$F1$
SVM$_{\text{BoW}}$	0.535	0.500	0.473	0.486
SVM$^R_{\text{BoW}}$	0.596	**0.818**	0.553	0.659
CNN	0.684	0.635	0.571	0.601
CNN+LSTM	0.676	0.623	0.657	0.640
BiLSTM$_{\text{attention}}$	0.695	0.501	0.554	0.527
BERT$_{\text{base}}$	0.773	0.726	0.721	0.723
BERT^R	**0.790**	0.762	**0.767**	**0.759**

Table 2.4 presents the detailed results for each class (*sexist/non sexist* content) obtained by our best baseline, BERT^R. We note that the results are lower for the sexist content class which leaves enough room for improvement.

Table 2.4 – Results per class with BERT^R.

CLASS	$F1$	P	R	$F1$
non sexist	0.843	0.832	0.856	0.762
sexist content	0.682	0.702	0.662	

As shown in Table 2.5, adding linguistic features to the embeddings increases the results for the BIN configuration, therefore we keep BERT$^R_{\text{features}}$ as basis for the rest of the models. The best results were obtained when employing the replacement strategy based on place and gendered names.

2.1.3.2   Results for the *3-CLASS* Configuration

Table 2.6 shows how the experiments were set up and presents the best results in bold.

Similarly to the BIN configuration, we observe that training BERT with stacked embeddings did not improve over BERT^R. Concerning the generalization strategies, all replacements were productive and outperformed all the previous models, observing that gendered

Table 2.5 – Results for the BIN classification.

CLASSIFIER	A	F1	P	R
BERT^R	0.790	0.762	0.767	0.759
$\text{BERT}^R_{\text{own\_emb + base}}$	0.768	0.751	0.712	0.795
$\text{BERT}^R_{\text{features}}$	0.795	0.787	0.819	0.761
$\text{BERT}^R_{\text{features + gen(Hypernym)}}$	0.806	0.804	0.835	0.776
$\text{BERT}^R_{\text{features + gen(Hypernym\_gendered)}}$	0.809	0.807	0.840	0.777
$\text{BERT}^R_{\text{features + gen(Name)}}$	0.790	0.796	0.830	0.766
$\text{BERT}^R_{\text{features + gen(Name\_gendered)}}$	0.815	0.806	0.841	0.775
$\text{BERT}^R_{\text{features + gen(SexistVocabulary\_gendered)}}$	0.801	0.807	0.836	0.781
$\text{BERT}^R_{\text{features + gen(Place)}}$	0.826	0.813	0.848	0.782
$\text{BERT}^R_{\text{features + gen(Place + Hypernym)}}$	0.803	0.799	0.836	0.766
$\text{BERT}^R_{\text{features + gen(Place + Hypernym\_gendered)}}$	0.819	0.811	0.846	0.779
$\text{BERT}^R_{\text{features + gen(Place + Name\_gendered)}}$	**0.837**	**0.824**	**0.865**	**0.787**
$\text{BERT}^R_{\text{features + gen(Place+Hypernym\_gendered+Name\_gendered)}}$	0.819	0.818	0.857	0.783

replacements are better. This shows that forcing the classifier to learn from general concepts is a good strategy for sexism content detection. In particular, we observe that the best replacement depends on the task (i.e., for 3-CLASS it is place, while for the BIN configuration it's place and gendered names). In both cases, replacing only public spaces with the generic `<location>` was one of the best strategy with 0.813 and 0.655 *F*1 for respectively *BIN* and *3-Class*. Multiple replacements (cf. last line in the table) were however, less productive.

2.1.3.3 Results for the *CASC* Configuration

Cascading models are known for being very accurate and can be used in the context of moderation as we cannot afford to take actions against users that are following the guidelines and policies. In the first stage we used the best performing model for *sexist* content vs. *non sexist* classification (i.e., $\text{BERT}^R_{\text{gen(Place+Name\_gendered)}}$). The instances classified as containing a sexist content by the first model were further used as the testing set for the second model (the best performing model for the 3-CLASS classification task in terms of F-score,

Table 2.6 – Results for the 3-CLASS classification.

CLASSIFIER	A	F1	P	R
BERT$_{\text{base}}$	0.714	0.540	0.572	0.515
BERT^R	0.726	0.567	0.609	0.531
BERT$^R_{\text{own\_emb + base}}$	0.708	0.526	0.605	0.513
BERT$^R_{\text{features}}$	0.754	0.588	0.625	0.556
BERT$^R_{\text{features + gen(Hypernym)}}$	0.763	0.614	0.649	0.598
BERT$^R_{\text{features + gen(Hypernym\_gendered)}}$	0.767	0.635	0.663	0.620
BERT$^R_{\text{features + gen(Name)}}$	0.755	0.620	0.656	0.606
BERT$^R_{\text{features + gen(Name\_gendered)}}$	0.760	0.643	0.665	0.630
BERT$^R_{\text{features + gen(SexistVocabulary\_gendered)}}$	0.764	0.635	0.654	0.627
BERT$^R_{\text{features + gen(Place)}}$	0.769	**0.655**	0.673	**0.646**
BERT$^R_{\text{features + gen(Place + Hypernym)}}$	0.758	0.622	0.654	0.610
BERT$^R_{\text{features + gen(Place + Hypernym\_gendered)}}$	**0.771**	0.652	**0.689**	0.630
BERT$^R_{\text{features + gen(Place + Name\_gendered)}}$	0.769	0.629	0.657	0.615
BERT$^R_{\text{features + gen(Place+Hypernym\_gendered+Name\_gendered)}}$	0.764	0.634	0.662	0.618

i.e., BERT$^R_{\text{gen(Place)}}$). In Table 2.7, the results corresponding to the non-sexist class of CASC classifier present the improvement brought by the second stage classifier, i.e., it was able to correct (predict as non-sexist) instances that were misclassified during the first stage. The last line of Table 2.7 presents the overall results obtained after the two stages of classification. The results show an improvement over the best system of 3-CLASS, proving the usefulness of a cascading approach with an increasing system complexity.

2.1.4   Error Analysis

A manual error analysis shows that misclassification cases are due to several factors, among which humor and satire (as in (2.6)) or the use of stereotypes (as in (2.7)), mainly because they are not expressed by a single word or expression but by metaphors. In the examples below, the underlined words highlight the leading cause of misclassification.

(2.6)   *Ma femme est hystorique. C'est comme hystérique, sauf que lorsqu'elle pète un câble elle me*

Table 2.7 – Results per class for the three tasks.

Task	Class	F1	P	R
Bin	non sexist	0.874	0.894	0.855
	sexist	0.773	0.836	0.719
	overall	0.824	0.865	0.787
3-Class	non sexist	0.849	0.855	0.842
	reporting	0.666	0.633	0.703
	sexist	0.452	0.532	0.392
	overall	0.655	0.673	0.646
Casc	non sexist	**0.882**	0.912	0.855
	reporting	**0.745**	0.719	0.775
	sexist	**0.518**	0.541	0.497
	$A = 0.831$			
	overall	0.717	0.724	0.709

sort des vieux dossiers.

(*My wife is hystorical. That's like hysterical, except that when she's angry she pulls out old files*)

(2.7)   *je demande pas ce qu'elle a fait sous le bureau pour arriver à se plateau*

(*I'm not asking what she did under the desk to be on this TV set*)

In particular for reporting tweets, we found many misclassified messages without any reporting verb or quotes as in (2.8), but also messages denunciating sexism using situational irony as in (2.9).

(2.8)   *Royal les rendrait elle tous fous? Alain Destrem (UMP): Ségolène Royal en boubou bleu, ça me rappelle ma femme de ménage !*

(*Does Royal make them all crazy? Alain Destrem (UMP): Ségolène Royal wearing a blue boubou, it reminds me my cleaning woman!*)

(2.9)   *Continuons à communier... Notre héros national avait des comptes en Suisse et n'était pas loin du #balancetonporc... Mais bon communions, rassemblons nous...*

(*Let's keep on be united... Our national hero had bank accounts in Switzerland and was not far from #SquealOnYourPig... But OK let's be united, let's get together...*)

2.2 From Sexism Detection to Hate Speech Detection: Preliminary Experiments

As we developed several models capable of achieving good performances for the task of sexism detection, we now investigate whether these models are able to detect target agnostic HS. To this end, we propose an experimental setting in which we train and test the models on several datasets independently.

2.2.1 Datasets

The data used in this experimental setting comes from two corpora. The first one, `HatEval`, is an already existing corpus containing English tweets annotated for HS against immigrants and women (cf. Section 2.1). The second corpus, `Sexism`, is a subset of the French dataset (cf. Section 1.3) comprised of 3,000 tweets[58] annotated for HS against women (i.e., *sexism* vs. *non-sexism*). Table 2.8 shows the distribution of the tweets for both tasks (i.e., HS and sexism detection).

Table 2.8 – Distribution of instances in both French and English datasets.

DATASET	LABELS		NO. OF INSTANCES	
`HatEval` (English)	hateful	5,512		13,071
	not-hateful	7,559		
`Sexsim` (French)	sexist	659		3,085
	not-sexist	2,426		

Similar to sexism (cf. Section 1.1), HS against immigrants may be expressed explicitly or implicitly using different pragmatic devices. See for example the following tweets taken from `HatEval` that illustrate a negative opinion (cf. (2.10)), a stereotype (cf. (2.11)) and sarcasm (cf. (2.12) and (2.13)):

(2.10) *I love how you are basically using this as an excuse to invade this country. I don't deny that your kind (Mestizos) are part Native American. But it doesn't mean shit. The Europeans built a far more advanced civilization than the Natives could ever dream of. #DeportThemAll.*

---

[58]As these experiments were carried out before the end of the annotation campaign, we only had access to 30% of the dataset.

(2.11) *Stop allowing Illegals to Dump their kids at the border like Road Kill Make them take they kids with them and not burden U.S Taxpayers for Medicaid Education and Food Stamps which is what their kids get Trump #MGA #RedNationRising #Immigration*

(2.12) *I'd say electrify the water but that would kill wildlife. #SendThemBack.*

(2.13) *Where is this? Brazil? Uganda? Sudan? Nope, it is France. Got to love that cultural enrichment thing going on. #openborders #refugeesnotwelcome #slums.*

For both corpora, several models have been built, all tested using 10-cross-validation to better compare our results in cross-lingual experiments. In the next sections, we detail the proposed models and then discuss the results.

2.2.2 Models

To measure to what extent HS detection is target-dependent, we propose several models ranging from standard Bag of Words (our baseline), features-based models to neural models. For all the models, due to the noise in the data, we performed standard text pre-processing: removing user mentions, URLs, RT, stop words, degraded stop words and the words containing less than 3 characters were filtered out. For `HatEval`, all the remaining words were stemmed using the Snowball Stemmer,[59] while for `Sexism`, tweets have been lemmatized using the French MSTParser.[60] We also experimented without stems and lemmas, but the results were not conclusive.

– Baseline. In all experiments, we used as our baseline unigrams, bigrams and trigrams Tf/IDf (we ignored the terms that appear in less than four tweets, as well as the terms that appear in more than 80% of the tweets).

– Feature-based models. We relied on state of the art features that have shown to be useful in HS detection. Our features include the following:

- *Surface features:* such as the tweet length in words, the presence or absence of punctuation marks (sequence of question/exclamation marks), the presence of URLs and @user mentions.

---

[59] http://snowballstem.org
[60] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_mst.html

- *Sentiment features*: The idea is to test whether identifying user's opinion can better classify his attitude as hateful or non-hateful. We took into consideration several existing lexicons: AFINN (Nielsen, 2011), SentiWordNet (Esuli and Sebastiani, 2006), Liu and Hu opinion lexicon,[61] HurtLex (a multilingual hate word lexicon divided in 17 categories) (Bassignana et al., 2018) and a lexicon containing 1,818 profanity English words created by combining a manually built offensive words list, the noswearing dictionary[62] and an offensive word list.[63] In the final models we chose to include only HurtLex and the lexicon we built, as none of the other models outperformed our baseline model. For the French corpus, we chose to use HurtLex, as it already contains hate words translated into French.

- *Emojis features*: We relied on a manually built emojis lexicon that contains 1,644 emojis along with their polarity among positive, negative and neutral.

We experiment with several combinations of the features above, and we finally keep the most relevant ones by applying the Chi2 feature selection algorithm. The best performing features have been used to train four classifiers (C_1, C_2 for the task of HS detection and C_3, C_4 for the task of sexism detection). For each classifier, we experimented with several machine learning algorithms (NB, LR, SVM, DT and RF) in order to evaluate and select the best performing one. Hereby, the HS baseline is a RF (the number of trees in the forest = 360 with a maximum depth of the tree = 600) and the sexism baseline is a SVM (linear kernel, C = 0.1). For C_2, best results have been obtained when using RF only for intermediate classification, whose output were then combined and passed onto a final Extreme Gradient Booster classifier. The four classifiers are as follows:

- C_1 : combines the length of the tweet with the number of words in the profanity lexicon with a baseline architecture as described above

- C_2 : on top of C_1 features we also used the number of positive and negative emojis and emoticons and we perform linear dimensionality reduction by means of truncated Singular Value Decomposition (latent semantic analysis on Tf/IDf matrices).

- C_3 : combines the length of the tweet with the number of words in the HurtLex lexicon on top of a baseline architecture

---

[61] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
[62] https://www.noswearing.com/dictionary
[63] http://www.cs.cmu.edu/~biglou/resources/bad-words.txt

- C_4 : the same features as C_3 but with a C_2 system architecture

– **BiLSTM$_{\textbf{attention}}$.** This model was previously presented in Section 2.1.2. For the task of HS detection however, we used pre-trained on tweets[64] Glove embeddings with an embedding dimension of 200 (Pennington et al., 2014).

– **BiLSTM$^M_{\textbf{attention}}$.** It uses the previously described **BiLSTM$_{\textbf{attention}}$** model in which we replaced the embeddings by multilingual embeddings: Glove bilingual word embeddings[65] obtained as described in (Ferreira et al., 2016) as well as French and English FastText word vectors mapped into the same embedding space following the alignment approach presented in (Smith et al., 2017).

2.2.3 Results

Since the number of sexist instances in the French corpus is relatively small, the results presented here were obtained by using 10-cross validation. Table 2.9 shows how the experiments were set up and presents the results in terms of accuracy (A), macro-averaged F-score (F), precision (P) and recall (R). The best results in terms of macro-averaged F-score (the evaluation metric used for ranking at SemEval) are presented in bold, while the columns left empty were intentionally left so, as we employed same system architectures with different features for the two tasks. Overall, our results show that when using the same model, the results achieved for the task of HS detection are better than the results for sexism detection.

Among the systems, C_1 represents our best performing one for the task of HS detection, while C_4 performed best for the task of sexism detection.

A manual error analysis of the instances for which our best performing model and manual annotation differ shows that in the misclassification of hateful instances intervene several factors: the presence of off-topic tweets, the lack of context (as some words that trigger hate in certain contexts may have different connotations in others) and implicit HS that employs stereotypes or metaphors in order to convey hatred. We also identified tweets for which we question the original label when taking into account the class definition. Below, we have provided some examples.

Example 1 (`HatEval`): Although in the first tweet (cf. (2.14) annotated as not hateful)

---

[64]We also experimented with pre-trained on Wikipedia word vectors, however the accuracy decreased by 3%.

[65]http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html

Table 2.9 – Hate speech and sexism detection results in both HATEVAL and SEXISM corpora.

	HATE SPEECH DETECTION				SEXISM DETECTION			
	A	$F1$	P	R	A	$F1$	P	R
Baseline	0.772	0.762	0.764	0.669	0.827	0.676	0.734	0.335
C_1	0.788	**0.780**	0.785	0.684	–	–	–	–
C_2	0.781	0.778	0.754	0.723	–	–	–	–
C_3	–	–	–	–	0.830	0.441	0.751	0.306
C_4	–	–	–	–	0.822	**0.688**	0.665	0.386
BiLSTM$_{\text{attention}}$	0.736	0.727	0.709	0.646	0.77	0.497	0.416	0.522

the user talks about Donald Trump, which doesn't fit in the targeted categories (immigrants or women), the annotation raises problems when trying to classify tweets such as the second one (cf. (2.15) annotated as hateful).

(2.14) *I love my religious brothers and sisters, but @realDonaldTrump, FUCK YOU, YOU'RE NOT EVEN A REAL THEOCRAT YOU FAT USLESS BITCH.*

(2.15) *Worse i have proof. A picture i took of you and one you took of me on the same night. Useless ungreatful kunt!*

Example 2 (`HatEval`): The first tweet (cf. (2.16) annotated as not hateful), containing the users opinion on Poland receiving immigrants, seems more hateful than the second tweet (cf. (2.17) annotated as hateful), in which the user depicts a series of events.

(2.16) *If Germans want rapefugees[66] they can keep them. Poland will not accept a single rapefugee. Not even one!!! Poland remains proud and firm!*

(2.17) *GERAMNY: African rapefugee climbs into house, steel expensive goods, rook a knife and abuse a girl of the family.*

Example 3 (`Sexism`): Both of the following tweets were misclassified due to the lack of context and knowledge about the world. In the first tweet (cf. (2.18)), as we don't have

---

[66]Accordind to Urban Dictionary, the term rapefugee is usually used when referring to the Muslim refugees coming into Europe in a derogatory way, as refugees are perceived as being more likely to raping people.

enough information about the *liberté d'importuner* movement, we aren't able to properly classify the disagreement of the user with Catherine Deneuve's statements. The same problem arises in the second tweet (cf. (2.19)), as the speech employs irony.

(2.18) *Ce que je pense de la "liberté d'importuner". #Sexisme #CatherineDeneuve #Tribune C'est pas parce que vous aimez la soumission qu'on doit toutes apprécier. L'avis des vieilles bourgeoises qui ne prennent plus le métro sur les frotteurs, on s'en passe.*
(What I think about "freedom to annoy". #Sexism #CatherineDeneuve #Tribune It's not because you like submission that we all have to like it. We don't need the opinion of old middle-class women who don't take the subway among rubbing men? anymore, we don't need it)

(2.19) *Merkel en Allemagne. Thatcher et maintenant #TheresaMay au Royaume-Uni. En France une femme présidente? Folie! Décadence!*
(Merkel in Germany. Thatcher and now #TheresaMay in the United Kingdom. A woman president in France? Madness! Decadence!)

Conclusion

In this second part, we have presented the first corpus of French tweets annotated for sexism detection. It is composed of about 115,000 tweets among which 12,274 are annotated according to a new characterization of sexist content inspired from both speech act theory and discourse studies in gender. The novelty of our approach is that not only tweets with a sexist content are labelled but the type of content is also characterized: the tweet is either directly addressed to a target (a woman or all women), describes a target or reports/denounces sexism experienced by a woman. We believe that it is important to distinguish between these usages in a context of offensive content moderation on social media since stories of sexism experiences should not be reported.

In addition, we have presented the first approach to detect reports/denunciations of sexism from real sexist content. We experimented with several deep learning models in binary, three classes and cascading configurations, showing that BERT trained on word embeddings, linguistic features and generalization strategies (i.e., place and hypernym replacements) achieved the best results for all the configurations, and that cascade classification allows to successfully correct misclassified non-sexist messages. These results are encouraging and demonstrate that detecting reporting assertions of sexism is possible, which is a first step towards automatic offensive content moderation. Error analysis shows that missclassifications may be due to the non-detection of gender stereotypes. We will analyze in Part III how gender stereotype detection can be used to improve sexism detection.

Finally, we have presented a pilot study for multi-target HS detection. This will be further investigated in Part IV. As part of the pilot study, we also experimented with multilingual embeddings by training on one language and testing on the other in order to measure how the proposed models are language dependent (Chiril et al., 2019a). The multilingual experiments results are somewhat comparable to the results obtained when training and

testing on the French data. This is very encouraging as one can rely on external annotated data for sexism in other languages to learn a model on a different language.

Gender Stereotype Detection to Improve Sexism Detection

Motivation

Gender stereotypes (GS hereafter) may be used as a way to express sexism. They are defined by the Office of the High Commissioner for Human Rights (OHCHR) as *'a generalised view or preconception about attributes, or characteristics that are or ought to be possessed by women and men or the roles that are or should be performed by men and women'*.

Although stereotypes can be positive or negative, the information that they provide (what a group is like, why group members are the way they are) is often linked to negative attitudes towards members of certain social groups (Fiske, 1998). As such, stereotypes represent the root cause inter-group tensions (e.g., sexism, racism, etc.) because they convey information which models ones behaviour towards stereotyped social group members. In addition, stereotypes also model the way in which the members of stereotyped social group perceive themselves.

GS have been widely studied in psychology, communication studies and social science (Allport et al., 1954; Crawford et al., 2002; Beike and Sherman, 2014; Biscarrat et al., 2016). In NLP, they have been studied mainly to detect or remove gender bias in word embeddings or word association graphs (Bolukbasi et al., 2016; Park et al., 2018; Madaan et al., 2018; Dev and Phillips, 2019; Du et al., 2019) as well as to identify disparity across gender in various applications like co-reference resolution (Zhao et al., 2018a) and sentiment analysis (Felmlee et al., 2019).

In addition to GS, other types of stereotypes have been investigated, such as in the HaSpeeDe 2 shared task (Sanguinetti et al., 2020) which focused on racist stereotypes with tasks for stereotypes and HS detection against minority groups. Francesconi et al. (2019) conducted an error analysis on the HaSpeeDe 2018 evaluation campaign (Bosco et al., 2018) concluding that there is a significant correlation between the usage of racist stereotypes and HS and that the false positive rate of hateful tweets is slightly higher for tweets that also

contain stereotypes. Although similar correlations have been observed between GS and HS from a psychological perspective (García-Sánchez et al., 2019), to our knowledge, no one has empirically measured the impact of GS detection for sexist HS classification.

In this part, we aim to bridge the gap by studying the link between GS and sexist HS by investigating whether GS detection can be helpful for detecting sexist messages on Twitter.

Our contributions include:

(1) The first dataset annotated for GS detection. This dataset contains about 9,200 tweets in French annotated according to different stereotype aspects (cf. Chapter 2).

(2) A set of experiments first to detect GS and then, to use this prediction for sexism detection (cf. Chapter 3). We rely on several deep learning architectures leveraging various sources of linguistic knowledge (label embeddings, generalization strategies based on both manual and automatically generated lexicons) to account for GS and the way sexist contents are expressed in language.

We end this part by discussing our main findings.

1 Related Work

Before detailing our dataset and annotation scheme, and the experiments carried out, we present the theoretical backgrounds on which we based our study. We start this chapter by analyzing what a stereotype (and in particular, GS) is, then we investigate how stereotypes are shared through language, and finally, we provide readers with a broader context for NLP literature related to the analysis of GS.

1.1 What is a Stereotype?

Stereotypes were originally defined by Lippmann (1946) as *'pictures in our heads'*, contending that our imagination is shaped by the pictures we see. This definition explains the way in which opinions are formed and manipulated because of what we trust, that in consequence *'leads to stereotypes that are hard to shake'*.

The information conveyed by stereotypes serves multiple functions:

- Stereotypes provide information about what a group is like (i.e., they are *descriptive*).

- Stereotypes provide information about why group members are the way they are (i.e., they are *explanatory*) (e.g., *Women are not good at math* is one of the stereotypes most often used to explain the low number of women pursuing a math-oriented career).

The information gained in this way allows one to rapidly assess and make sense of the surrounding complex social environment.

When interacting with a new individual, one tends to associate him as belonging to a certain social group as it becomes easier to infer information from previous knowledge and experiences with similar individuals (Allport et al., 1954). This simplification with regards to our beliefs and expectations towards an individual through the process of stereotyping

has significant downsides as it involves making generalizations based on characteristics attributed to a social group without taking into account individual and situational constraints of group members.

Although stereotypes can be both positive and negative, typically, these generalizations are linked to negative attitudes towards members of certain social groups (Fiske, 1998). As such, stereotypes represent the root cause of many problems in society such as sexism, racism and other inter-group tensions because they convey attributional information that model the way in which stereotyped social group members are being treated by others, as well as the way in which they perceive themselves. As we use stereotypes in order to make sense of the surrounding world, our stereotypical expectations help us in identifying and interpreting the things we learn about the others.

Beike and Sherman (2014) identify three levels of social information that roughly *'correspond to behavioural prediction, impression formation and stereotype development'*:

- Behavioural – this level of information directly links a specific individual to their behaviour in a specified situation (e.g., *This boy helped his sister with her math homework yesterday*). This type of information can be learned through direct observation or communication.

- Individual – represents a more abstract information about an individual and consists of generalized behavioural characteristics that are observed over time (e.g., *This boy is good at math*).

- Group-level information – this level of information refers to the qualities and characteristics of a social group, the information being separated from specific individuals and generalized across situations (e.g., *Boys are good at math*).

Beike and Sherman (2014) explain that the *inductive process* through which one infers information from the lower levels in order to draw inferences at higher levels corresponds to stereotype formation. Similarly, *deduction* occurs when one uses higher level information in order to draw conclusions about individual members of a social group (i.e., *stereotyping*). *Analogy*, another possible process, refers to using any level information in order to draw conclusions at the same level (e.g., *He is good at math, so he must be really smart*).

Crawford et al. (2002) showed that for social groups with a high number of members,

trait inference is drawn from group members (i.e., *induction*) and once a generic impression is formed, it is applied to all the other group members (i.e., *deduction*). In contrast, information about social groups with a low number of members is processed and learned individually. Once a behaviour information is drawn for one member, no further generalizations (to other members or to the group as a whole) are made, each member being treated as an unique individual.

Communication plays a crucial role in the formation of social groups stereotypes, as an observer can communicate its perception of a specific situation at different levels, and in turn, the receiver will have to draw its own inferences, albeit, generally, the communication of group-level information is most likely used to convey existing (shared) social group stereotypes, which in turn contributes to their consensualization and maintenance.

Although the content of stereotypes varies, Weiner (1986) argues that the attributional values communicated through stereotypes fall into three categories:

- *Locus of causality* – according to the attribution theory, the causes can be internal (e.g., traits, behaviours) or external (influenced by the environmental agents).

- *Controllability* – according to the attribution theory, the second dimension refers to causes that can be controlled (e.g., a behaviour that does not conform to the social norms) or not (e.g., congenital abilities).

- *Stability* – this dimension helps one predict future behaviours as the causes are perceived to be stable (i.e., high expectation that the behaviour will continue over time) or unstable.

Based on the dimensions of the attribution theory, stereotypes can be seen (in most of the cases) as causes internal to the social group that are relatively stable over time. As such, by stereotyping social groups and their members, low ability attribution plays a crucial role in the expectancy for success, which in turn might limit the opportunities allotted to them as they are not deemed capable (for a detailed review, the reader is invited to refer to (Reyna, 2000)).

Allport et al. (1954) created a five-stage hierarchical scale for measuring the manifestation of prejudice in society, ranked by the harm it produces, from verbal aggression to physical violence:

- *Antilocution* – using derogatory or HS against another social group. While antilocution is often considered harmless, it can negatively affect the self-esteem of the targeted group as this type of speech relies on negative stereotypes based on preconceived judgements rather than facts.

- *Avoidance* – members of a group are actively avoiding members belonging to a different social group, which leads to isolation and social exclusion.

- *Discrimination* – a social group is denied equal treatment, opportunities and services.

- *Physical attack* - when members of a group vandalize, burn, destroy another group's property or violently attack someone's physical integrity.

- *Extermination* – history has provided multiple examples where a group has been exterminated through genocide and ethnic cleansing (e.g., World War II, Rwanda, Bosnia War).

This suggests that language (i.e., *antilocution*) represents the first step towards mistreatment. Prejudice and stereotypes often lead to real life consequences, such as property destruction, physical assaults and various forms of discrimination, this raising the need of examining the motivation behind HS as well as the psychological consequences of this type of verbal abuse.

Often stimulated by feelings of fear, anger and ignorance or need of dominance, the targeted groups can experience hateful attitudes manifested in both language use (negative labels, derogatory terms, stereotypes) and discrimination.

The new opportunities offered by social media networks, both in terms of interaction and anonymity, has led to a massive growth of user generated web content which contains a large spreading of abusive messages that increase and cluster in time subsequent to a *'trigger'* event (e.g., conflict, economic crisis, terrorist attacks, migration). For instance, King and Sutton (2013) proved there is a connection between the terrorist attack of 9/11 in the USA and the increase of hate crimes with an anti-Islamic motive, 58% of them committed 2 weeks following the event.

1.2 Gender Stereotypes

Many differences exist in between men and women, *'Women are from Venus, men are from Mars'* being a phrase often used for illustrating the existing differences in the way of acting, thinking, feeling, etc., suggesting that men and women are so different that they might as well come from different planets. Men and women do, indeed, show gender differences in many life aspects, but these differences are much smaller than the Venus-Mars dichotomy suggests.

A study conducted by Bennett et al. (2000) provided strong evidence indicating that both children and adults categorize unknown individuals based on their gender, this being considered a primary feature in social information processing.

Stereotypes (and GS in particular) can be useful for making quick assertions, but the reader should keep in mind that by categorizing people only based on their gender one has an oversimplified view of reality, which reinforces the perceived boundaries between women and men and seemingly justifies the social implications of role differentiation and social inequality.

As gender continues being seen only as a binary categorization, GS not only reflect the differences between women and men, but also impose what men and women should be and how they should behave in regards to different life aspects.

One significant consequence of GS is the reinforcement of gender inequality, *agency* (i.e., traits such as competence and independence) and *communion* (i.e., concerns about the welfare of others and relationship with them) being the core dimensions used to characterize GS. Although biological attributes may impact one's behaviour and choice of occupational roles, research indicates that the gender differences develop over time, and that they influenced by family, friends and education. For example, women are communal, kind and family oriented, whereas men should be agentic, skilled and work oriented (Ellemers, 2018).

With the change of the gender landscape in the last century, women started having more representations and visibility than they had in the past in occupations that have been traditionally dominated by men (ranging from sports to education), which may allow them to obtain a higher prestige and status. Although the progress seems considerable, studies show

that women are treated differently or less favorably. [67] [68]

Based on this change in the positions occupied by women in society, as well as the broadening of opportunities presented to women, Haines et al. (2016) conducted a study in order to analyze to what extent GS changed over a period of 30 years (in between 1983 and 2014), with participants assessing the likeliness of gendered characteristics (e.g., traits, behaviours, occupations, physical characteristics) to belong to a typical man or woman. The authors collected the data in a similar way as Deaux and Lewis (1984) and assessed whether people's beliefs changed over time in order to match the actuality, the main difference in between the methodology of the two studies dwelling in the age of the participants. Haines et al. (2016) set as another goal understanding whether age has any relationship with gender stereotyping, hence their decision of testing a full-spectrum age range (19 to 73 years old) as opposed to only college students as in (Deaux and Lewis, 1984). The results of the study are surprising as the authors didn't find any indication of substantial change of basic stereotypes in spite of all the societal changes.

Xu et al. (2019) define the narrative in which a man represents a woman's way to a happy, fulfilling life as the *'Cinderella complex'* and they analyzed the female emotional dependency on male characters in a collection of books, movie synopses and movie scripts. By using pretrained word2vec models, the authors constructed a vector representing the dimension of *happy* vs. *unhappy* that was used for calculating the *'happiness scores'* of words surrounding specific female and male characters. The authors first selected the movie synopsis of Cinderella and by calculating the happiness scores they observed that the happiness of Cinderella depends on the prince, but not vice versa. Further testing on different movie genres showed that when appearing together in the same context, the happiness score of the female characters is higher than when they appear alone. Another important finding, consistent with previous research, shows that male characters are more likely to be described by using verbs, as opposed to women, who are described by using adjectives.

By analyzing the audience rating of the movies, Xu et al. (2019) observed that the movies highlighting male characters as adventure oriented have a higher acceptance rate than the movies highlighting them as romantic relationship oriented. These represent important findings as GS embedded into movies and books may maintain gender inequality and expose

---

[67]https://www.payscale.com/data/gender-pay-gap
[68]https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en

Patricia Chiril

more people to stereotyped narratives.

Given the extensive use of gender categories and the seeming utility of differentiating between women and men, we observe that people may be resistant to change their stereotypes to any significant degree, this contributing to stereotype stability and maintenance.

1.3   Stereotypes in Verbal Communication

Communication plays a crucial role in the emergence, reinforcement and change of shared social groups stereotypes. Stereotypes might be acquired through verbal communication as well as other co-occurring factors (e.g., direct observation or interaction, media depiction, etc.). Often in a conversation, labels having positive or negative connotations can be used in order to communicate the content of a set of stereotypical characteristics associated with a social group or its members. Slurs have the role of derogating the members of a social group by conveying hostile stereotypical expectancies, while simultaneously conveying a negative affect.

In terms of shared information, research shows that communication tends to revolve around information that is consistent with our stereotypical beliefs, as it facilitates the communication and is less likely to lead to misunderstandings or disagreements (Klein et al., 2008).

As verbal communication constitutes a major mean through which stereotypes are communicated, the words chosen for describing one's achievements reflect the stereotypical attributions we tend to make. Maass (1999) shows that when describing a behaviour that matches stereotypical expectations we tend to use more abstract terms (e.g., *That boy is smart*), as opposed to using more concrete terms for counter-stereotypical behaviours (e.g., *The girl did well on the math test*).

Accompanying the verbal communication, the non-verbal communication also plays an important role in conveying and reinforcing GS, as men tend to display more open and expansive postures (which relate to dominance and higher power), while women display more closed postures (which relate to submission and lower power) (Cashdan, 1998).

The way in which men and women express their emotions also play an important role in communicating and reinforcing GS, Plant et al. (2000) suggesting that men and women express differently similar emotional experiences (e.g., negative emotions are expressed by

men in the form of anger, while women express them in the form of sadness).

Through four studies of social classification (gender, age, education and political educa-tion), Carpenter et al. (2017) examined words and phrases that contribute to the language that makes up stereotypes. The methodology consists in collecting data and the ground truth labels from Twitter users and asking raters to correlate a tweet's author to its ground truth category (i.e., in order to assess the inaccurate stereotypes about women, one needs to an-alyze the phrases written by men that led the raters to rate the author as being a woman). The results show that the raters were generally able to correctly guess an individual's group membership, this suggesting that stereotypes are not always exaggerations, although mis-classifications appeared to be influenced by the exaggerated aspects of stereotypes (e.g., women were likely to be categorized as men if they were talking about technology or news).

1.4   Stereotypes in Social Media

As previously seen, language plays a crucial role in the way in which beliefs and expecta-tions about a specific social group are formed and shared within large groups of people.

The term *'stereotype threat'* was first introduced by Steele and Aronson (1995) and it refers to a situation in which a person belonging to a stereotyped social group behaves in a way that confirms the negative assumptions about that group when pressured by concerns about possibly confirming those negative stereotypes. Steele and Aronson (1995) showed through several experiments that when the race was emphasized, the performance on standardized tests of black students was lower than the performance of white students, as opposed to the case in which the stereotypes were not made salient before the test and the results were equivalent or even better.

In order to provide more insight into how the stereotype threat behind GS might affect boys and girls, Wille et al. (2018) examined the effect of short segments involving GS in a math television show. Based on previous studies that showed a short-term decrease in girl's performance and motivation when reminded that girls have a lower math performance than boys, the authors examined the effects of this stereotype on the performance, attitudes to-wards math and motivational dispositions of fifth graders. Their findings suggest that al-though these types of stereotypes might increase the children's stereotypes endorsement, due to their young age and short exposure to stereotyped material, there were almost no

effects on their motivational dispositions, attitudes, and performance.

Felmlee et al. (2019) use sentiment analysis in order to examine the degree of negativity of messages that include gendered insults as well as adjectives used for reinforcing feminine stereotypes. For the sentiment classifier, the authors use a combination of scores from an ensemble of three lexicons: AFINN (Nielsen, 2011), Bing (Hu and Liu, 2004) and a modified version of VADER (Gilbert and Hutto, 2014) that includes derogatory terms towards women found in a manual examination of a corpus of tweets collected over a period of $2^{1/2}$ years. Felmlee et al. (2019) also investigated the spread of a negative tweet in terms of retweets, replies and likes, as well as the social roles involved in a conversation: *aggressor* (i.e., the user responsible for the aggressive message), *reinforcer* (i.e., users who support, retweet or like the aggressive message), *bystander* (i.e., users that are aware of the aggressive message but do not interfere), *victim* (i.e., target of the aggressive message), *defender* (i.e., users defending the victim). The results show that by including insulting words that reinforce feminine stereotypes (especially references to physical characteristics) the degree of negativity of a message is significantly increased.

1.5 Stereotypes in Natural Language Processing

As previously seen, stereotypes can be useful for making quick assertions about other people. Inspired by the studies of stereotypes in psychology, one of the first works that exploits these characteristics clusters is (Rich, 1979), who introduced a stereotype recommender system tasked with suggesting novels that people might find interesting.

Racist stereotypes have been extensively investigated in NLP (Fokkens et al., 2018). For example, the dataset of the HaSpeeDe 2 shared task contains annotated tweets and newspaper headlines, with the main goal of identifying contents that convey hate or prejudice against a given target (immigrants, Muslims and Roma people) with an auxiliary task of determining the presence or absence of a stereotype towards that given target. Among participants, only Lavergne et al. (2020) consider the interaction between HS and stereotype detection by employing a multitask learning approach that achieves the best scores in the competition. The presence of stereotypes against immigrants has also been annotated in Italian (Sanguinetti et al., 2018) and Spanish political debates (Sánchez-Junquera et al., 2021), the latter being annotated according to a fine-grained taxonomy to capture the positive (threats) and negative dimensions (victims) of stereotypes.

Concerning GS, there are some datasets dedicated to sexist HS annotated with stereotype. Among them, Parikh et al. (2019) propose a dataset which contains 13,023 accounts of sexism extracted from the Everyday Sexism Project website manually annotated with 23 labels. The annotation scheme includes two categories for GS: *role stereotyping* (i.e., false generalizations about certain roles being more appropriate for women) and *attribute stereotyping* (i.e., linking women to some physical, psychological, or behavioural qualities). Parikh et al. (2019) classify these messages using LSTM, CNN, CNN-LSTM and BERT models trained on top of several distributional representations (characters, subwords, words and sentences) along with additional linguistic features.

The AMI shared task at IberEval and EvalIta 2018 consisted in detecting sexist tweets and then identifying the type of sexist behaviour according to a taxonomy defined by Anzovino et al. (2018): discredit, stereotype, objectification, sexual harassment, threat of violence, dominance and derailing. Most participants used SVM models and ensemble of classifiers for both tasks with features such as n-grams and opinions (Fersini et al., 2018b).

Besides shared tasks, few studies investigated GS detection. Among them, Felmlee et al. (2019) use sentiment analysis in order to examine the degree of negativity of messages that include gendered insults as well as adjectives used for reinforcing feminine stereotypes. The results show that by including insulting words that reinforce feminine stereotypes (especially references to physical characteristics) the degree of negativity of a message is significantly increased. Cryan et al. (2020) compare two methods for GS detection in job postings showing that a transformer (BERT) model outperforms a lexicon-based approach with adjectives and verbs that are potentially related to GS.

Fokkens et al. (2018) introduce *'microportraits'*, a collection of descriptions provided by a text with regards to a given entity (i.e., person, group, object, event), and investigate what choices the writers make when describing an entity. By targeting information about given entities (that shares certain characteristics), common patterns used for describing them could be identified and so, stereotypes could be investigated. When investigating the stereotyping of Muslims in the Dutch media, Fokkens et al. (2018) show that the *microportraits* provide a more detailed insight into the portrayal of a group.

Francesconi et al. (2019) conducted an error analysis on the HaSpeeDe 2018 evaluation campaign. The results suggest that there is a significant correlation between the usage of stereotypes and HS and the authors showed that the false positive rate of hateful tweets is

slightly higher for tweets that also contain stereotypes. However, in the HaSpeeDe 2 shared task, Lavergne et al. (2020) is the only team that considers the interaction between HS and stereotype detection by employing a multitask learning approach.

In the NLP field, the approaches to stereotype detection are very recent and to our knowledge, this is the first study that investigates the possible correlation between sexist tweets and tweets expressing stereotype ideas about women.

2 Data and Annotation

In this chapter, we first present our guidelines for the manual annotation of GS and the result of the annotation procedure. Since the resulting GS dataset is relatively small, we also propose a method for data augmentation based on sentence similarity with multilingual external resources.

2.1 Characterizing Gender Stereotypes

GS have been considered as an independent component with respect to sexism, so that the latter does not entail the former, and vice versa. When a stereotype is present, it can be expressed explicitly or implicitly (i.e., if one can paraphrase the message or infer a content such as *'(all) women are...'*).

- **Sexist tweets containing explicit stereotypes.** The underlined passage of (2.1) high-lights the stereotype.

  (2.1)   *Anne Hidalgo est une femme. Les femmes aiment faire le ménage. Anne Hidalgo devrait donc nettoyer elle-même les rues de Paris*
  (Anne Hidalgo is a woman. Women love cleaning the house. Anne Hidalgo should clean the streets of Paris herself)

- **Sexist tweets containing implicit stereotypes.** (2.2) implies that women should know how to cook, while (2.3) implies that women are hysterical and resentful.

  (2.2)   *C'est bon t'es une femme forte, te manque que la cuisine pour atteindre la perfection*
  (It's good you're a strong woman, you only need the cooking skills to reach perfection)

  (2.3)   *Ma femme est hystorique. C'est comme hystérique, sauf que lorsqu'elle pète un câble elle me sort des vieux dossiers*

(*My wife is hystorical. That's like hysterical, except that when she's angry she pulls out old files*)

- **Sexist tweets that do not contain stereotypes.** Although in (2.4) the user uses sexist humor to point out that the dress is considered too short, it does not generalize to all women.

(2.4)   *Mais nan! Elle a juste mis sa robe à 90 degrés en machine c'est tout*
         (*But no! She just put her dress in the washing machine at 90 degrees that's all*)

According to Haut Conseil à l'Égalité, GS are schematic and globalizing representations that attribute supposedly *'natural'* and *'normal'* characteristics (psychological traits, behaviours, social roles or activities) to women and men. This definition and previous research (Deaux and Lewis, 1984) which suggest that GS have different and independent components (i.e., trait descriptors, physical characteristics, role behaviours and occupational status) lead to the creation of three categories of stereotypes that include: *physical characteristics*, *behavioural characteristics* (i.e., compartmental and psychological traits) and *activities* (i.e., social roles, activities and occupational status).

In the instances presented below we underlined some passages in order to highlight the characteristics of women that bear stereotypes.

- The **physical characteristics** can be related to the physical strength (e.g., (2.5) contains the stereotype that women are weak) or to the physical aspect such as (2.6) (girls should have long hair) and (2.7) (girls who put on mini-skirts are easy girls (and deserve what happens to them: criticism, insults, rape, ...)).

(2.5)   *Femme je t'aime, surtout, enfin Pour ta <u>faiblesse</u> et pour tes yeux*
         (*Woman I love you, especially, finally For your <u>weakness</u> and for your eyes*)

(2.6)   *Les <u>cheveux courts</u> pour une fille c'est une mauvaise idée hein*
         (*<u>Short hair</u> for a girl is a bad idea*)

(2.7)   *T'as raison,si elle est vulgaire et en <u>tenue de callgirl</u>, c'est son choix...*
         (*You're right, if she's vulgar and wearing a <u>callgirl outfit</u>, it's her choice...*)

- **Behavioural characteristics** are related to intelligence (cf. (2.8)), emotions and sensibility such as in (2.9) (women are crazy/irrational) and (2.10) (women are hysterical and

resentful). (2.11) contains the stereotype that women always change their mind and are gold diggers.

(2.8)   *c'est femme est une grosse nulle comme toutes les journaliste(E)s ! inculte*
        (*this woman is a big loser like all the journalists! uneducated*)

(2.9)   *elle est dangereuse cette vieille folle j'imagine que c pas elle qui va aller combattre il faut la faire taire*
        (*she is dangerous this old madwoman I imagine that it's not her who will fight it is necessary to hush her up*)

(2.10) *Je suis censée me reconnaitre dans la pub "Just Fab" avec une conne hystérique qui hurle ?*
        (*Am I supposed to recognize myself in the "Just Fab" ad with a screaming hysterical bitch?*)

(2.11) *Que les âmes sensibles se bouchent les oreilles Y a pas + grosse CONNASSE que la femme qui veut divorcer de son notable de mari, te dde de lui consacrer un temps fou pr prendre sa décision, finit par retourner auprès de son porte monnaie sur pattes et oublie de payer ta facture*
        (*Sensitive souls cover your ears There isn't a bigger BITCH than the woman who wants to divorce her wealthy husband, asks you to allocate to her an insane amount of time for taking a decision, and ends up returning to her wallet with legs and forgets to pay your bill*)

- **Activities** (i.e., activities, jobs, hobbies) that are stereotypically assigned to women as in (2.12) which implies that a woman's place is in the kitchen, or (2.13) which implies that women don't understand football.

(2.12) *Faut jamais épouser une femme qui ne sait pas faire la cuisine*
        (*Never marry a woman who cannot cook*)

(2.13) *T'es une femme je serai jamais d'accord avec toi pour du foot. Va faire des videos de contouring pour chien ou des ongles de chaton et arrête de nous Peter les couilles.*
        (*You're a woman, I'll never agree with you on football. Go and make videos about contouring for dogs or nails and stop Busting our balls.*)

In addition to the three aforementioned stereotype categories, we include a fourth type, **designation**, for instances that contain implicit stereotypes such as in (2.14).

(2.14)   *Derrière chaque Femme sommeille une <u>Princesse</u>, et derrière toutes princesse Hélas, une connasse - #GaspardProust*

(In every woman there is a princess, and in every <u>princess</u>, there is unfortunately a bitch - #GaspardProust)

A tweet may contain stereotypes pertaining to more than one category, such as in (2.15) where designation (*woman*) is used in conjunction with behavioural (*women hold grudges*) and physical characteristics (*women are weak*) attributed to women. Additionally, present in many instances, the author of the tweet also uses insults.

(2.15)   *La <u>femme</u> est une <u>petite chose fragile</u> mais aussi une grosse <u>pute</u> <u>rancunière</u> c'est féministe et éclairé?*

(The <u>woman</u> is a <u>fragile little thing</u> but also a big <u>bitch</u> holding <u>grudges</u> it's feminist and enlightening?)

2.2   `Stereo`^{O}: The Original Dataset

As previously mentioned (cf. Section 1.5), all existing datasets labelled with GS are dedicated to sexist HS detection and GS are considered as a form of sexism/misogyny. But a message containing a GS is not necessarily sexist and vice-versa (e.g., the message *"football is not for girls": it's over now!* contains the stereotype *girls cannot/must not play football* but the meaning conveyed by the whole message is not sexist). This is why we decided to rely on two different datasets for both sexism and GS detection tasks.

To build the dataset for the GS detection, we used a non-annotated subset of 9,282 French tweets from the available corpus (cf. Section 1.2).

Given a tweet, its annotation consists in assigning it at least one of the following categories: *insult*, *designation*, *physical characteristic*, *behavioural characteristic*, *activity* and *non-stereotype* (note that these five categories are not mutually exclusive). A tweet is annotated as *non-stereotype* when it does not contain a stereotype. The annotation scheme of the French GS corpus is presented in Figure 2.1.

Subsequent to a training stage, 1,000 tweets have been annotated by two annotators (native French speakers, one male and one female, Master's Degree students in Communication and Gender) so that the inter-annotator agreement could be computed (Kappa = 0.79).

Figure 2.1 – GS annotation scheme.
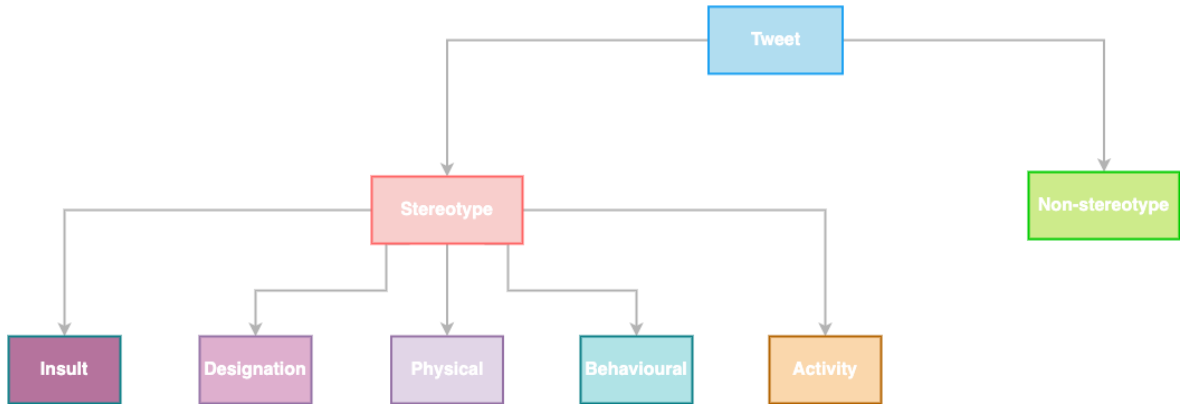
For these 1,000 tweets, the final labels have been assigned according to a majority vote.

Finally, a total of 9,282 have been annotated, among which 91.47% do not contain a stereotype and 8.53% contain a stereotype. This results in a highly imbalanced dataset which size is relatively the same as in other datasets (e.g., 9% of the tweets contain a gender stereotype in the `AMI corpora`). Among the instances containing a stereotype, 10% of the tweets are annotated with multiple GS labels. Table 2.1 shows the distribution of the dataset, hereafter called **`Stereo`**^{O}.

Table 2.1 – `Stereo`^{O} corpus distribution.

NonStereotype	Stereotype					Total
	insult	designation	physical	behaviour	activity	
8,490	792					9,282
	175 (1.88%)	67 (0.72%)	164 (1.76%)	202 (2.17%)	395 (4.25%)	

2.3 **`Stereo`**^{aug}: The Augmented Dataset

The corpus being quite small, with the non-stereotype class much more prevalent than the stereotype class, we decided to augment the training data to counter class imbalance. The rare class (i.e., *stereotype*) is the class of more interest in the sense that the cost of misclas-

sifying stereotypes (as non-stereotypes) is higher than misclassifying non-stereotypes (as stereotypes) as typically, the use of stereotypes is linked to negative attitudes towards members of certain social groups.

In the following sections we present existing strategies for dealing with imbalanced datasets, as well as our proposed strategy for dealing with the issue.

2.3.1 Strategies for Dealing with Imbalanced Datasets

Banko and Brill (2001) showed that very different Machine Learning (ML) algorithms performed almost identically for the task of Natural Language Disambiguation once they were fed enough training data. Based on these results, the authors suggest reconsidering the trade-off between focusing on algorithm development versus directing the efforts towards corpus development. The idea that for solving complex problems data matters more than algorithms was further popularized by Halevy et al. (2009). In reality, both these assertions are true only in a certain context. If the algorithm that is used is too complicated for the amount of data available, this will result in high variance problems (which lead to model *overfitting*) that can be addressed by increasing the number of instances in the corpus. However, if the algorithm that is used is too simple to explain the data, this will result in high bias models (*underfitting*), which will not benefit from adding more data, although they may benefit from adding more/better features (i.e., feature engineering). As much as data is needed, the quality of the data is very important: if the training data is full of errors, outliers and noise, the model is less likely to perform well, as detecting the underlying patterns becomes much harder.

Most ML algorithms typically need thousands of examples (even for simple problems) for the algorithm to work properly. In order to be able to generalize well, it is crucial for the training set to be representative of the cases one wants to generalize to. A small size of the sample will result in sampling noise (i.e., non-representative data as a result of chance). That is not to say that by having very large samples the issue of sampling noise will be resolved, as even this can be non-representative if the sampling method is flawed (i.e., sampling bias).

Throughout this section the terms *majority class* and *minority class* will be used, however, the proposed solutions can very well be applied to a multi-class problem, where several majority/minority classes could be found.

There are a number of strategies to counter class imbalance among which down-

sampling, oversampling, weighting the data and adapting the loss function. We review them below and explain why there are not suitable to augment our stereotype dataset.

2.3.1.1 Down-sampling (undersampling) the Majority Class

The amount of data needed depends on the application and generally, the more easily distinguishable the positive class is from the negative class, the less data is needed.

Undersampling is based on the idea that the dominant class has many redundant instances, and as such, a set of majority class instances can be discarded. Many different undersampling techniques exist depending on whether the method selects:

- *which instances from the majority class should be kept* (e.g., Condensed Nearest Neighbors (Hart, 1968), Near Miss (Mani and Zhang, 2003));

- *which instances from the majority class should be deleted* (e.g., random undersampling, Edited Nearest Neighbors (Wilson, 1972)), Tomek Links (Tomek, 1976));

- *a combination of which instances from the majority class should be kept and deleted* (e.g., One-Sided Selection (Kubat et al., 1997), Neighborhood Cleaning Rule (Laurikkala, 2001)).

However, these strategies do not use all the available information (i.e., all the annotated instances), which may lead to information loss. As such, undersampling is often a solution of little interest, rarely implemented, except in scenarios with large and complex datasets, case in which preparing/exploring the data and building pilot models is too expensive.

2.3.1.2 Oversampling (upsampling) the Minority Class

The drawback of undersampling could be overcome by oversampling the minority class by adding additional instances (to the minority class) and forcing the model to focus on the least represented examples.

Several approaches can be applied for obtaining new instances:

- *Random oversampling*, one of the earliest proposed methods, consists in randomly duplicating instances in the minority class. This method was shown to be an effective solution to the imbalance problem (Branco et al., 2015). However, this strategy may lead to model overfitting.

- *Collecting more data (finding a new data source).* For example, Rosenthal et al. (2021) propose using democratic co-training, a semi-supervised technique for collecting new offensive data using an offensive language identification dataset (Zampieri et al., 2019a) as seed. In order to diversify the models' rationales, the distant supervision is performed by an ensemble of models (PMI, FastText, LSTM, BERT). Each of the models has to be trained on the labeled dataset and has to predict the confidence of the positive class for each instance in a new unannotated tweet dataset.[69] Further, the average and the standard deviation of the confidences predicted by each of the models in the democratic co-training setup were used in order to create the semi-supervised dataset.

- Applying *data generation* techniques for generating slightly modified (or new) instances (from the already existing data) which will share the label of the original class of the instance from which they have been generated. Although common in Computer Vision, additional challenges are raised in NLP, as one needs to find semantically invariant transformations.

2.3.1.3 Data Augmentation Techniques

There are a number of strategies for data augmentation. Before reviewing the most used techniques, in Table 2.2 we provide an overview of the main existing NLP techniques for data augmentation.

Back-translation (cf. Figure 2.2) is a technique based on paraphrasing that relies on translating an instance (from the source language) to another language before translating it back into the source language (Yu et al., 2018). The major advantage of employing this method is that the overall semantics of the sentence are maintained while bringing more syntactical diversity to the newly generated data.

Techniques relying on replacing some words in the text while preserving its meaning (**lexical substitution**) for generating additional data:

- *replacing random words with one of their synonyms as given by a thesaurus* (e.g., WordNet (Miller, 1995), BabelNet (Navigli and Ponzetto, 2012), ConceptNet (Speer et al., 2017)) (Zhang et al., 2015; Mueller and Thyagarajan, 2016; Wei and Zou, 2019).

---

[69]The authors created the new tweets dataset by collecting instances containing the 20 most common English stopwords.

Table 2.2 – NLP techniques for data augmentation.

DATA AUGMENTATION TECHNIQUE		METHODOLOGY
Back-translation		- translate an instance (from the source language) to another language before translating it back into the source language (Yu et al., 2018)
Lexical substitution	Thesaurus-based substitution	- replace a random word with one of its synonyms as given by a thesaurus (Zhang et al., 2015; Mueller and Thyagarajan, 2016; Wei and Zou, 2019)
	Word-embeddings substitution	- replace a word with one of its nearest neighbors in the embedding space (Wang and Yang, 2015)
	Masked Language Model	- using transformer Masked Language Model predictions for replacing and inserting tokens in the previously masked portion of the text (Garg and Ramakrishnan, 2020)
	Tf/IDf based substitution	- replace uninformative words (i.e., the words having the lowest Tf/IDf scores) with other non-keywords (Xie et al., 2020)
Surface transformations (contractions and expansions)		- transform verbal forms from contraction to expansion (and vice versa) (Coulombe, 2018)
Syntax trees transformations		- the dependency tree of the original sentence is first generated, then transformed by using grammar rules (Coulombe, 2018)
Instance crossover		- randomly swap two halves of two random instances having the same label (Luque, 2019)
Noise injection	Random insertion	- insert a random synonym of a non stop word in a random position in the sentence (Wei and Zou, 2019)
	Random swap	- swap the position of two random words in the sentence (Wei and Zou, 2019)
	Random deletion	- randomly remove each word in the sentence with a probability p (Wei and Zou, 2019)
	Blank noising	- randomly replace a words in the sentence with a placeholder token (Xie et al., 2017)
	Spelling error injection	- inject spelling errors to a random word in the sentence
	Sentence shuffling	- shuffle the sentences of an instance
Mixup	wordMixup/senMixup	- generate new instances by linearly interpolating word/sentence embeddings (Guo et al., 2019)
Generative methods	Pre-trained Language Models	- finetune a pre-trained language model and generate new instances by using the class label and a few initial words as cue for the model (Anaby-Tavor et al., 2020)

- *leveraging pre-trained word embeddings for selecting the nearest neighbors in the embedding space as replacement for some of the words in the text* (Wang and Yang, 2015). In addition, in order to replace words with their context-specific synonyms, Hemker and Schuller (2018) proposed an approach based on selecting the words with the highest cosine similarity (using a large pre-trained word2vec model), while also checking the part-of-speech tag quality for disambiguating the word meaning (i.e., ensuring that an ambiguous word is not replaced with the most frequent meaning).

- *leveraging BERT* (or other transformer models) *Masked Language Model (MLM) predictions for replacing and inserting tokens in the previously masked portion of the text* (Garg and Ramakrishnan, 2020). Although seeming to work well in English, in French, the results are not as convincing (cf. Figure 2.4).

Figure 2.2 – Data augmentation through back translation.

C'est bon t'es une femme forte,
te manque que la cuisine pour atteindre la perfection

It's okay you're a strong woman,
you only need cooking to achieve perfection

C'est bon tu es une femme forte,
tu n'as besoin que de cuisiner pour atteindre la perfection
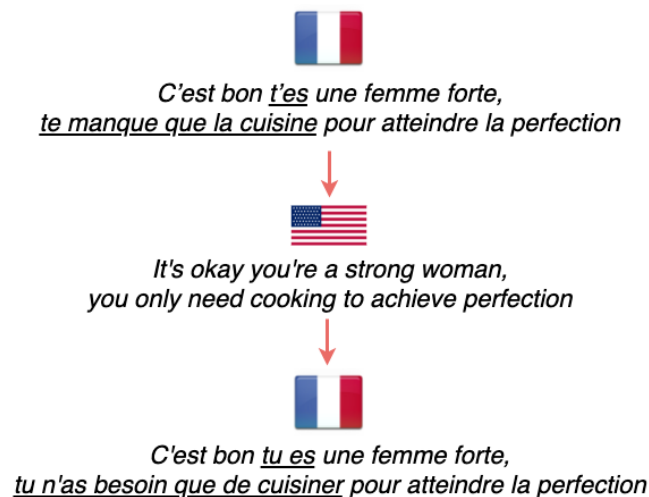
Figure 2.3 – Thesaurus-based synonym replacement.

Elle court vite pour une femme
(She runs fast for a woman) → BabelNet → *Elle court rapidement pour une femme*
(She runs fast for a woman)

Figure 2.4 – BERT Masked Language Model.[a]

A woman should be [MASK] →
A woman should be happy
A woman should be strong
A woman should be beautiful
A woman should be prepared
A woman should be careful
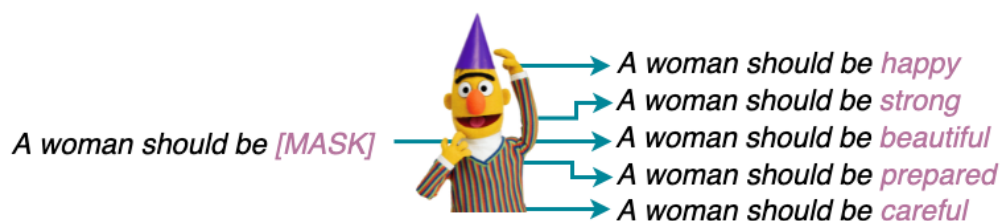
[a] The MLM predictions for the same prompt in French (i.e., *Une femme doit être [MASK]*) are: *considérée (considered), construite (built), élevée (raised), morale (moral), écrite (written).*

- Xie et al. (2020) argue that while back-translation is good at maintaining the overall semantics of a sentence, one can not choose the words to be replaced, which might be of interest for tasks where some keywords are more informative than others (i.e., they are decisive in determining the class membership). As such, the authors propose an approach for *identifying the uninformative words* (i.e., the words having the lowest

Tf/IDf scores) *which can then be replaced (with other non-keywords)* without affecting the ground-truth label of the instance.

Another augmentation technique relies on **surface transformations**, semantically invariant transformations that are language dependent and which rely on contractions and expansions (cf. Figure 2.5). In order to preserve the semantic invariance, Coulombe (2018) proposes to allow ambiguous contractions but avoid ambiguous expansions that can lead to misinterpretations.

Figure 2.5 – Surface transformations relying on contractions and expansions.

C'est bon *tu es* une femme forte,
te manque que la cuisine pour atteindre la perfection

Contraction →

← Expansion

C'est bon *t'es* une femme forte,
te manque que la cuisine pour atteindre la perfection

Coulombe (2018) proposes a second strategy using **syntax trees transformations**, where the dependency tree of the original sentence is first generated, then transformed by using grammar rules. Finally, the transformed dependency tree is used to generate a paraphrased sentence (e.g., the transformation from active voice to the passive voice of sentence (and vice versa) is a semantically invariant transformation).

Inspired by the chromosome crossover operation from genetic algorithms, Luque (2019) propose an **instance crossover** augmentation technique. In this approach, the samples are divided into two halves, and then two random instances having the same label (in this case, polarity) have their halves swapped. The authors hypothesize that the resulting instances will preserve the polarity of the sentiment, despite not being grammatically and semantically sound.

Generating new instances through **noise injection** relies on duplicating instances and injecting noise into them. The added parasitic noise will not change the semantic of the new instance, but rather introduce several variations of the same sample which will allow the model to better generalize when encountering instances having this kind of perturbations. Several noise injection techniques were proposed:

- *random insertion* relies on finding a random synonym for a random non stop word in the sentence and inserting it in a random position (Wei and Zou, 2019);

Figure 2.6 – Data augmentation through random insertion.

vite ——synonym——▶ rapidement

random word   random position insertion

Elle court <u>vite</u> pour une femme
(She runs <u>fast</u> for a woman)

Elle court rapidement vite pour une femme
(She runs <u>fast fast</u> for a woman)
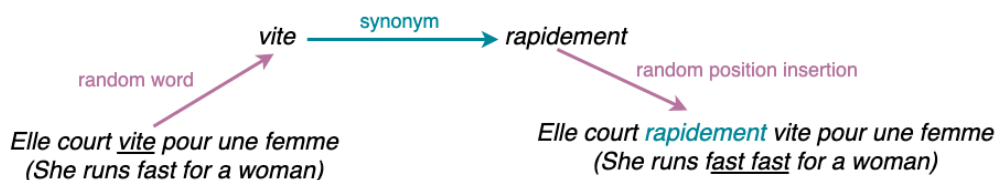
- *random swap* relies on randomly choosing two words in the sentence and swapping their position (Wei and Zou, 2019);

- *random deletion* relies on randomly removing each word in the sentence with a probability p (Wei and Zou, 2019);

Figure 2.7 – Data augmentation through random swap and random deletion.

Elle court vite <u>pour une</u> femme
(She runs fast <u>for a</u> woman)

random swap

Elle court vite <u>une pour</u> femme
(She runs fast <u>a for</u> woman)

Elle court vite <u>pour une</u> femme
(She runs fast <u>for a</u> woman)

random deletion

Elle court vite <u>pour</u> femme
(She runs fast <u>for</u> woman)

- *blank noising* is similar to the *random deletion* technique, but rather than deleting a word, it will replaced it with a placeholder token (e.g., '__') (Xie et al., 2017);

Figure 2.8 – Data augmentation through blank noising.

Elle court vite pour <u>une</u> femme
(She runs fast for <u>a</u> woman)

blank noising

Elle court vite pour __ femme
(She runs fast for __ woman)

- two other techniques (not referenced in literature) rely on *injecting spelling errors* (either to some random words in the sentence or by simulating typing errors i.e., replacing some letters in a word by letters found close by on a keyboard) and *shuffling the sentences of an instance*.

Initially introduced by Zhang et al. (2017), **Mixup** is an image augmentation technique where new instances are generated by linearly interpolating pixels of random image pairs. Contrary to other data augmentation techniques, the images can belong to different classes. Guo et al. (2019) adapted this technique for NLP tasks and propose two strategies of Mixup on sentence classification:

- in *wordMixup* (cf. Figure 2.9) the interpolation is performed on word embeddings (i.e., the two instances are zero-padded to the same length and their word embeddings are interpolated);

Figure 2.9 – wordMixup technique (Guo et al., 2019) (the added part to the standard sentence classification model is in the orange rectangle).
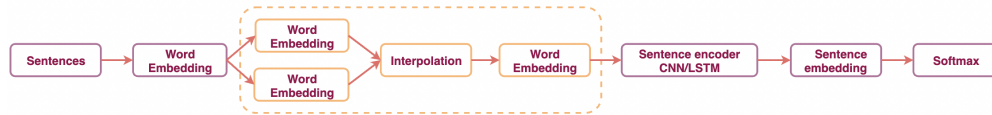
- in *senMixup* (cf. Figure 2.10) the interpolation is performed on sentence embeddings (i.e., the hidden embeddings for the two instances are generated by an encoder (e.g., CNN, LSTM) before being linearly interpolated).

Figure 2.10 – senMixup technique (Guo et al., 2019) (the added part to the standard sentence classification model is in the orange rectangle).
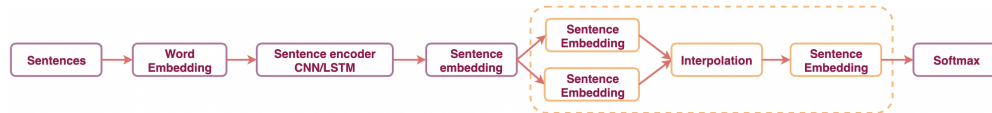
Hemker and Schuller (2018) proposed using Natural Language Generation models for auto-generating new semantically similar instances based on the training data. However, as the new instances may contain the same (or similar) words as the original training instance put into a different order, employing this technique may result in generating instances that do not make sense to humans.

Anaby-Tavor et al. (2020) propose finetuning a large pre-trained language model (e.g.,

BERT, GPT2, BART) and generate new instances by using the class label and a few initial words as cue for the model.

Data generation by *'perturbing'* existing instances in order to create new ones is a variation of oversampling via bootstrapping. By creating new similar instances to the ones belonging to the minority class, the algorithm could learn a richer set of information for building classification rules. One example is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) which finds an instance similar to the one being oversampled and creates a synthetic instance that is a randomly weighted average of the original and the neighboring instance, where the weight is separately generated for each predictor. In this case, the number of synthetic instances generated depends on the oversampling ratio required for balancing the classes.

Weighting the data. Weighting the data provides an alternative to both undersampling the majority class and oversampling the minority class as many classification algorithms take a weight argument that allows up/down weighting the data.

Adapting the loss function. As many classification algorithms optimize a certain criteria or loss function, there are studies in the literature that propose modifying the loss function in order to avoid the problems raised by the minority class.

2.3.1.4 Interim Conclusion

In the previous section we presented an overview of existing strategies for dealing with imbalanced datasets. Due to the relatively small amount of instances annotated as *stereotype* in our corpus (i.e., 792) we do not consider down-sampling as a valid strategy, as the risk of discarding useful information provided by the *non-stereotype* class increases.

Despite the plethora of data augmentation techniques, the new instances obtained through these methods may contain the same or similar words as the original instance but in a different order, which may result in generating instances that do not make sense to humans. In addition, these methods do not guarantee that the new generated instances belong to the same class as the original ones. To avoid this, we propose a new approach for data augmentation based on sentence similarity.

2.3.2 Data Augmentation via Sentence Similarity

2.3.2.1 Methodology

We aim to answer one main research question:

  - *Is sentence similarity an effective data augmentation strategy?*

To this end we followed five semantically motivated strategies for augmenting the stereotype corpus by extending the training dataset (i.e., Stereo$^{O}_{\texttt{train}}$) with:

1. All the instances from the AMI corpora annotated as *stereotype* (all the instances in the combined training and testing sets from both Evalita and IberEval). [70]

2. Instances from the AMI corpora annotated as: *discredit, derailing, dominance, stereotypes*. The distribution of the tweets across all the categories is presented in Table 2.3.

3. Instances from the Parikh corpus annotated as: *Role stereotyping, Attribute stereotyping, Body shaming, Hyper-sexualization (excluding body shaming), Internalized sexism, Slut shaming, Motherhood-related discrimination, Menstruation-related discrimination*. As the instances in this corpus are annotated with non mutually exclusive labels, an instance is selected as candidate if it is labeled with at least one of the aforementioned categories. To this end a subset of 4,321 instances were used.

4. *Sentence similarity*. We propose a new approach for data augmentation based on sentence similarity. We use SentenceBERT, a modification of BERT that derives semantically sentence embeddings that can be compared using cosine-similarity (Reimers and Gurevych, 2019), to extend our training dataset with the most similar sentences from multilingual corpora (i.e., AMI corpora, Parikh corpus).[71] In this experimental setting, a threshold was experimentally set[72] and the selected instances were automatically labeled as *stereotype* upon adding them to the training dataset.

---

[70]As the two datasets (i.e., Evalita and IberEval) used the same approach for collecting the data and annotation guidelines, the duplicate instances that were found were removed.

[71]As the Waseem dataset contains a set of instances that target gender minorities, although not annotated for stereotypes, we conducted a sentence similarity experiment in order to test whether we could find instances similar to the ones in our French stereotype corpus. However, a manual inspection showed that the quality of the most similar instances is not suitable for the task at hand.

[72]$T = 0.4$ for the Parikh corpus and $T = 0.45$ for the AMI corpora.

5. *A new collection of French tweets* (`French_new`) on which we apply the sentence similarity approach. The newly collected French dataset includes tweets collected in between 3 June 2018 and 20 November 2020 with a small set of keywords selected from the stereotype lexicon. These keywords are different from those used for the initial data collection (i.e., $Stereo^O$). Due to a high number of returned instances, we set a threshold of 50,000. After removing the duplicate instances, for the keywords that returned more tweets than our limit, we selected only the longest 50,000 instances. Finally, a total of 350,127 tweets containing the aforementioned keywords were used for computing the similarity scores. Table 2.4 presents the distribution of the newly collected data.

Table 2.3 – General overview of the datasets used for augmenting $Stereo^O$.

LABEL	DATASET							
	$Evalita_{EN}$		$Evalita_{IT}$		$IberEval_{EN}$		$IberEval_{ES}$	
Stereotype	179		668		137		151	
Dominance	148		71		49		302	
Derailing	92	1,785	24	1,828	29	1,568	20	1,649
Sexual harassment	352		431		410		198	
Discredit	1,014		634		943		978	
Non-misogynous	2,215		2,172		1,683		1,658	

2.3.2.2   Selecting the Best Augmentation Strategy

As several datasets are available (cf. Table 2.3 and 2.4), in the following we present the models used for investigating which is the dataset that works best for augmenting the initial corpus. To this end, we used two baseline models to perform GS detection on both the initial and the augmented datasets. Our models are as follows:

– **FlauBERT_base**. This is our baseline[73] that uses FlauBERT-Base Cased (Le et al., 2020) (without any additional inputs) on top of which we added an untrained layer of neurons. We then used the HuggingFace's PyTorch implementation of FlauBERT (Wolf et al., 2019) that we trained for 3 epochs.

---

[73]Note that when choosing the baseline model we experimented with different transformer architectures The results with CamemBERT (Martin et al., 2020) and Multilingual BERT (Devlin et al., 2019)) were lower.

Table 2.4 – Number of tweets containing the keyword in `French`$_{\text{new}}$

KEYWORD	NO. OF INSTANCES
moche	50,000
fesses	50,000
jupe	50,000
bavarde	28,602
dépensière	7,422
dévouée	14,103
infirmière	50,000
poupée	50,000
cuisine	50,000

– **BERT$_{\text{base}}$**. This model is similar to **FlauBERT$_{\text{base}}$** but as we perform multilingual augmentation we rely on the multilingual BERT (BERT-Base, Multilingual Cased) model (Devlin et al., 2019) instead.

Our aim is to investigate the effectiveness of sentence similarity as a data augmentation technique and test its performance against augmentation with manually annotated data. To this end, three experiments were carried out on datasets augmented: 1) with additional instances manually annotated as *stereotype*, 2) with additional instances manually annotated for sexism/misogyny, and 3) through sentence similarity.

Table 2.5 presents the results of GS detection when augmenting the training dataset (i.e., `Stereo`$^{O}_{\text{train}}$) with all the the instances from the `AMI corpora` annotated as *stereotype*, while Table 2.6 presents the results when the experiments were carried out with instances from multilingual corpora annotated for different types of misogynistic behaviours. In both experimental settings, the best results were obtained when augmenting the training dataset with instances from `Evalita`$_{\text{IT}}$.

Table 2.5 – Results for GS detection when training on additional data annotated as *stereotype*.

DATA	P	R	F1
Stereo^O	0.656	0.659	0.658
Stereo^O + Evalita$_{EN}$	0.691	0.665	0.677
Stereo^O + IberEval$_{EN}$	0.713	0.689	0.700
Stereo^O + IberEval$_{ES}$	0.739	0.670	0.697
Stereo^O + Evalita$_{IT}$	0.732	0.693	**0.710**
Stereo^O + all	0.739	0.624	0.659

Table 2.6 – Results for GS detection when training on additional data annotated with other categories.

DATA	P	R	F1
Stereo^O	0.656	0.659	0.658
Stereo^O + Evalita$_{EN}$	0.713	0.670	0.690
Stereo^O + IberEval$_{EN}$	0.715	0.666	0.689
Stereo^O + IberEval$_{ES}$	0.712	0.655	0.682
Stereo^O + Evalita$_{IT}$	0.739	0.697	**0.717**
Stereo^O + AMI corpora	0.717	0.648	0.680
Stereo^O + Parikh	0.726	0.665	0.694

Table 2.7 presents the results of GS detection when augmenting Stereo$^O_{train}$ with the most similar instances from multilingual corpora. For all sources of augmentation, a threshold T was set experimentally and the most similar instances were automatically labelled as *stereotype* and added to our training dataset.[74] This allows to select similar instances in terms of vocabulary (cf. (2.16)) but also of syntactic patterns (cf. (2.17)).

(2.16)   Initial tweet: *Je reconnais la cuisine comme territoire incontesté de la Femme*

---

[74]$T = 0.4$ for the Parikh corpus, $T = 0.45$ for the IberEval dataset and $T = 0.5$ for the newly collected French data as the number of similar instances returned was higher.

(*I admit that the kitchen is the uncontested territory of women.*

Similar English tweet (IberEval$_{EN}$, *T*=0.459): #*YesAllWomen belong in the kitchen*

(2.17) Initial tweet: *Pourquoi il y a toujours une fenêtre dans une cuisine ? C'est pour que la femme ait un point de vue*

(*Why is there always a window in the kitchen? So that women can have a point of view*)

Similar English tweet (IberEval$_{EN}$, *T*=0.496): *Why do women get married in white? So they match the kitchen appliances.*

Below we present a GS instance from Stereo^{O} (cf. (2.18)) and the instances with the highest similarity that were retrieved (cf. (2.19) - (2.22)):

(2.18) *Les femmes conduisent aussi bien qu'un homme*

(*Women drive as well as men*)

(2.19) IberEval$_{EN}$: *If women were supposed to be drivers, giving head to them while driving would be just as easy as it is for them to do to men* (*T*=0.6546, label: *discredit*)

(2.20) Evalita$_{EN}$: *a women without a man is like a car without an engine; it doesnt work.* (*T*=0.5547, label: *dominance*)

(2.21) Parikh: *Getting leered at by male drivers while stuck in traffic. Also: comments that women are bad drivers.* (*T*=0.5978, label: *Attribute stereotyping, Body shaming, Denial or trivialization of sexist misconduct, Internalized sexism, Moral policing (excluding tone policing)*)

(2.22) French$_{new}$: *La femme est beaucoup plus endurante que l'homme. La femme est plus dévouée que l'homme*

(*The woman is much more enduring than the man. The woman is more devoted than the man*) (*T*=0.5020)

Overall, a system trained on a dataset that contains additional instances obtained by computing the similarity outperformed or had similar results with a system trained on a dataset obtained by injecting instances annotated as *stereotype*. The best results were obtained when augmenting Stereo$^{O}_{train}$ with the most similar instances retrieved from IberEval$_{EN}$ and the newly collected French dataset.

Table 2.7 – Results for GS detection when training on additional data obtained through similarity.

Data	P	R	$F1$	No. of additional instances
Stereo^O	0.656	0.659	0.658	-
Stereo^O + Evalita$_{EN}$	0.723	0.685	0.703	1,674
Stereo^O + IberEval$_{EN}$	0.702	0.713	0.707	1,914
Stereo^O + IberEval$_{ES}$	0.725	0.662	0.692	1,257
Stereo^O + Evalita$_{IT}$	0.731	0.677	0.704	904
Stereo^O + all	0.719	0.653	0.684	5,852
Stereo^O + Evalita$_{EN}$ + IberEval$_{EN}$	0.735	0.680	0.706	3,691
Stereo^O + Evalita$_{IT}$ + IberEval$_{ES}$	0.692	0.677	0.684	2,161
Stereo^O + Parikh	0.690	0.679	0.684	4,321
Stereo^O + Parikh +IberEval$_{EN}$	0.712	0.697	0.704	6,338
Stereo^O + French$_{new}$	0.673	0.698	0.686	2,241
Stereo^O + French$_{new}$ +IberEval$_{EN}$	0.734	0.706	**0.719**	4,155

We also performed this augmentation strategy for each GS type. Since only 10% of the tweets contain more than one type of GS, we decided to keep the predominant conveyed stereotype as the gold label for the experiments.

Finally, the augmented training dataset (i.e., Stereo^{aug}) is now composed of 4,891 tweets (the initial 792 stereotype tweets in French, 1,914 additional tweets in English from IberEval$_{EN}$ and 2,241 additional tweets in French from French$_{new}$), which represents an augmentation of about 45% of the initial corpus (see distribution in Table 2.8).[75]

---

[75]When performing the augmentation strategy for instances with multiple labels, if the same instance was retrieved for more than one category, it was not included in the augmented dataset (this is the reason why in Table 2.8 the number of instances in Stereo^{aug} for the binary classification is different than for multi-label classification).

Table 2.8 – Stereotype corpus distribution in the initial and augmented datasets.

	Stereo^O			**Stereo**^{aug}		
Non Stereotype 8490	792			Initial French: 792		
				Eng IberEval: 1,914 / New Fr: 2,241		
	physical	behaviour	activity	physical	behaviour	activity
	164	202	395	689	473	1224

3 Automatic Detection of Gender Stereotypes

In this chapter we aim to answer two main research questions:

- *Are models able to capture common properties of gender stereotypes?*

- *Can gender stereotype prediction improve the performance of a model built for the task of sexism detection?*

3.1 Gender Stereotype Detection

For the task at hand, our annotated GS corpus has been divided into train (80%) and test (20%) sets. Table 3.1 shows the distribution of these sets.

Table 3.1 – French gender stereotype corpus ($\texttt{Stereo}^O$) - train/test distribution.

DATASET	LABELS	NO. OF INSTANCES	
$\texttt{Stereo}^O_{\texttt{train}}$	stereotype	633	7,433
	non-stereotype	6,800	
$\texttt{Stereo}^O_{\texttt{test}}$	stereotype	159	1,849
	not-stereotype	1,690	

For the experiments, all new instances obtained through augmentation techniques (cf. Section 2.3.2.2) are added to the train set, the test set being the same in all configurations and composed only of initial tweets from $\texttt{Stereo}^O$.

In the next sections, we detail our models, provide and discuss our results.

3.1.1 Methodology

Our main objective is to *identify the most appropriate deep learning architecture able to capture the linguistic characteristics of GS in short messages*. To this end, we propose several models relying on different contextualized pre-trained models as input:

- FlauBERT - when the training dataset consists only of instances belonging to the initial French stereotype corpus (i.e., $\texttt{Stereo}^O$);

- multilingual BERT - when the training dataset incorporates the most similar instances from the multilingual corpora (i.e., $\texttt{Stereo}^{aug}$).

In this way, we are comparing different methods employed for stereotype detection on both the original and augmented datasets.

3.1.2 Models

Our models are as follows:

– FlauBERT$^L_{\textbf{base}}$. This model is similar to FlauBERT$_{\text{base}}$, but it uses focal loss (Lin et al., 2017) instead.[76] Our aim here is to compare with one of the most effective approaches for handling imbalanced datasets based on loss function modification (Cui et al., 2019). This model has been only trained on $\texttt{Stereo}^O_{\text{train}}$ to better compare with the data augmentation strategy based on sentence similarity.

– FlauBERT$_{\textbf{lex}}$/BERT$_{\textbf{lex}}$. In order to force the classifier to learn from generalized concepts rather than words which may be rare in the corpus, we adopt several replacement combinations (cf. Section 2.1.2). We used the previously described French lexicon comprising 130 gender stereotyped words that we grouped according to our five categories (*physical characteristics*, *behavioural characteristics*, *activities*, *insults* and *designations*) and replaced these words/expressions when present in tweets by their category. Note that only 1% of these words overlap with the ones used to collect the initial and extended datasets. When applied on English inputs, we automatically translated the words by aligning French and English FastText word vectors (Conneau et al., 2017b) and selecting the nearest neighbor in the

---

[76]Results with dice loss Li et al. (2020) were lower.

target space.

– **FlauBERT_ConceptNet/BERT_ConceptNet**. Instead of relying solely on manually built lists of words, we try to automatically extend them with words extracted through Concept-Net (Speer et al., 2017), a multilingual knowledge graph for natural language words or phrases in their undisambiguated forms. Although similar knowledge bases exist (e.g., BabelNet (Navigli and Ponzetto, 2012)), our choice is motivated by the fact that for a given word, ConceptNet is focusing on common-sense relationships to other words, as opposed to BabelNet, which focuses on dictionary definitions of words (i.e., WordNet-style synsets). In addition, ConceptNet has a larger coverage for French. Lexicon extension works as follows:[77] given a word in the French lexicon, we extend it via the relations *SimilarTo* and *Synonym*.[78] For example, for *bavarde (talkative)*, the retrieved words include *jacasse (chatter)* and *commère (gossip girl)*.[79] After following this strategy, we obtained a total of 725 entries in French (used for FlauBERT) and 1,993 entries in French and English (used for BERT).

– **FlauBERT_label_emb/BERT_label_emb**. Our stereotype categories being relatively informative, another way to force the classifier to infer the correct link between a given message and the GS it may evoke is to leverage additional information as given by the labels themselves. We therefore propose to use label embedding (Wang et al., 2018), a technique that embeds both class labels and the text into a joint latent space, where the model can be trained to cross-attend the inputs and labels in order to improve the model performance. Our models are similar to (Si et al., 2020) who consider the joint representation of the tweet and its corresponding class token and incorporate label embeddings into the self-attention modules. The label embeddings for the class stereotype are initialized as the average of the corresponding keyword embeddings (here, we consider the words in the lexicon as keywords representative for the class stereotype), while the label embedding for the non-stereotype class is initialized at random. For $\texttt{Stereo}^{aug}$, the English keywords were obtained in the same manner as for BERT_lex.

---

[77]We also tried extending these lexicons by selecting only three seed words from each of the lexicon's categories, however we noticed that the results tend to decrease. Moreover, the selection of the seed words is an important factor, as some words can provide more and/or better relations.

[78]Extension via *RelatedTo* relation was not conclusive.

[79]https://conceptnet.io/c/fr/bavarde

Since current studies consider GS as a type of sexism/misogyny, we also compare with the best performing models for sexist HS detection (cf. Section 2.1.2):

– **CNN$_{\text{FastText}}$** (Karlekar and Bansal, 2018) that uses FastText pre-trained French word vectors (with the dimension of 300);

– **CNN-LSTM** (Karlekar and Bansal, 2018; Parikh et al., 2019) based on the previous CNN model by adding an LSTM layer (except that we used word-level embeddings instead of character/sentence-level as the results were lower);

– **BiLSTM$_{\text{attention}}$** (Parikh et al., 2019) which relies on a Bidirectional LSTM with an attention mechanism that attends over all hidden states and generates attention coefficients.

All the proposed models have been evaluated on $\texttt{Stereo}^O$ test set while the hyperparameters were tuned on the validation sets (20% of the training dataset), such that the best validation error was produced.

3.1.3   Results

Stereotype detection, and GS in particular, being a new task, there is no strong state of the art models to compare with apart (Sánchez-Junquera et al., 2021) and the winner system at HaSpeeDe2 by Lavergne et al. (2020) for binary stereotypes detection against immigrants and the one by Cryan et al. (2020) for binary gender bias classification in job postings. Both models are based on pre-trained contextualized embeddings which have been fine tuned on the task without accounting for any prior linguistic knowledge about GS. These models are thus similar to our **FlauBERT$_{\text{base}}$** and **BERT$_{\text{base}}$**.

Table 3.2 presents the results for the binary GS detection task in terms of macro-averaged F-score ($F1$), precision (P) and recall (R) with the best results presented in bold. We observe that best baselines are without surprise FlauBERT$_{\text{base}}$ and BERT$_{\text{base}}$ and more importantly, that data augmentation via sentence similarity as given by SentenceBERT is very effective. Indeed, the model trained on $\texttt{Stereo}^{aug}$ achieves better results than the one trained on $\texttt{Stereo}^O$, outperforming FlauBERT$^L_{\text{base}}$, the model designed to handle class imbalance in the original dataset. Another important finding is that all the models that incorporate GS knowledge improve over the baselines, the best strategy being the one based on ConceptNet. Also, the results for label embeddings are close to the one based on manual lexicon of GS.

These results suggest that in the absence of a lexicon, label embeddings could be a valid strategy.

Table 3.2 – Results for the most productive strategies for binary classification. ‡: baseline models.

CLASSIFIER	P	R	$F1$
CNN‡	0.619	0.630	0.624
CNN+LSTM‡	0.572	0.622	0.595
BiLSTM$_{attention}$‡	0.589	0.593	0.590
FlauBERT$_{base}$‡	0.656	0.659	0.658
FlauBERT$^L_{base}$	0.672	0.667	0.669
BERT$_{base}$‡	**0.734**	0.706	0.719
FlauBERT$_{lex}$	0.674	0.693	0.683
BERT$_{lex}$	**0.734**	0.718	0.725
FlauBERT$_{ConceptNet}$	0.711	0.704	0.708
BERT$_{ConceptNet}$	0.726	**0.731**	**0.729**
FlauBERT$_{label\_embeddings}$	0.685	0.680	0.682
BERT$_{label\_embeddings}$	0.729	0.717	0.724

Table 3.3, presents the results obtained through different system configurations for each of the three categories of GS (i.e., physical characteristics, behavioural characteristics and activities), while Table 3.4 presents the results for the multi-class GS classification. As the labels assigned to the tweets in the GS corpus are not mutually exclusive we attributed the final label according to the most predominant stereotype present in that instance. For both experimental settings, the results are consistent with the binary GS detection (i.e., *stereotype* vs. *non-stereotype*) where the best results were obtained by BERT$_{ConceptNet}$.

3.1.4 Model Explainability

As we further wanted to understand the reasons behind the predictions of the best performing models (based on ConceptNet generalizations), we used LIME (Ribeiro et al., 2016), a technique that *'explains the predictions of any classifier in an interpretable and faithful manner'*. The way the LIME algorithm works can be simplified through the following steps:

1. For a given data point, its features are repeatedly randomly perturbed (i.e., words are removed from the input and then observations on the impact on the model prediction

Table 3.3 – Results for binary stereotype type detection.

CLASSIFIER	ACTIVITIES			PHYSICAL			BEHAVIOURAL		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
FlauBERT$_{base}$	0.708	0.664	0.683	0.741	0.597	0.637	0.486	0.500	0.493
FlauBERT$_{lex}$	0.730	0.666	0.692	0.741	0.625	0.664	0.486	0.500	0.493
FlauBERT$_{ConceptNet}$	0.723	0.670	0.696	0.743	**0.629**	0.681	0.493	0.509	0.500
BERT$_{base}$	0.700	0.697	0.699	0.773	0.626	0.691	**0.606**	0.556	0.579
BERT$_{ConceptNet}$	**0.746**	**0.708**	**0.725**	**0.775**	0.628	**0.693**	**0.606**	**0.562**	**0.583**

Table 3.4 – Results for the multi-label GS classification.

CLASSIFIER	P	R	$F1$
FlauBERT$_{base}$	**0.573**	0.432	0.480
BERT$_{base}$	0.569	0.458	0.507
BERT$_{ConceptNet}$	0.571	**0.462**	**0.510**

are made).

2. Get predictions for each perturbed data instance.

3. Compute an approximate linear *'explanation model'* using predictions.

Below we provide some examples. The shades of blue and orange highlight the words that contributed towards the model predicting the instance as *non-stereotype* and *stereotype* respectively (the darker the shade, the higher the contribution of that particular word). Please note that we employ FlauBERT$_{ConeptNet}$ (instead of BERT$_{ConeptNet}$) in order to make sure that the results are not influenced by the augmentation technique.

The first tweet (cf. (3.1)), annotated as containing a GS related to activities stereotypically assigned to women (i.e., cooking) originally predicted as *non-stereotype* by FlauBERT$_{base}$ (cf. Figure 3.1) was correctly classified as containing a GS after following the generalization strategy (cf. Figure 3.2).

(3.1)   *Je viens de voir une émission qui fait cuisiner les maris sur les ordres de leur femme... En*
         mode les femmes ça fait la cuisine à la maison lol... Sur France 2 en plus... Tristesse
         (*I just saw a show that makes husbands cook based on their wives' orders ... In fact women*
         cook at home lol ... On France 2 on top of that ... Sadness)

Figure 3.1 – LIME explanations of FlauBERT$_{base}$ for (3.1).

Je viens de voir une émission qui fait cuisiner les maris sur les ordres de leur femme... En
mode les femmes ça fait la cuisine à la maison lol... Sur France 2 en plus... Tristesse

Figure 3.2 – LIME explanations of FlauBERT$_{ConceptNet}$ for (3.1)..

Je viens de voir une émission qui fait <caractéristique_Activité_Femme> les maris sur les
ordres de leur <designation_Femme>… En mode les <designation_Femme> ça fait la
<caractéristique_Activité_Femme> à la maison lol... Sur France 2 en plus... Tristesse

A similar behaviour can be observed in (3.2), a tweet which contains both a designation
(i.e., woman) and a GS related to activities stereotypically assigned to women (i.e., cooking).
Moreover, for both examples, we can observe that before performing the generalization, the
words that contributed the most towards the prediction of class *non-stereotype* are not very
informative (cf. Figure 3.1 and Figure 3.3).

(3.2)   *Une autre : Pourquoi il y a toujours une fenêtre dans une cuisine ?... C'est pour que la femme*
         ait un point de vue...
         (*Why is there always a window in the kitchen? So that women can have a point of view*)

Overall, we can conclude that coupling GS information as encoded in external lexicons
(either manually built or extended) with contextualized representation of words is a good
strategy, enabling the classifier to learn from generalized concepts rather than words them-
selves. However, even if this strategy relies on a manual list of seed words in a given lan-
guage, we show that it is generic enough since it is both (a) *language independent* thanks to

Figure 3.3 – LIME explanations of FlauBERT$_{base}$ for (3.2).

Une autre : Pourquoi il y a toujours une fenêtre dans une cuisine ?... C'est pour que la femme ait un point de vue...

Figure 3.4 – LIME explanations of FlauBERT$_{ConceptNet}$ for (3.2)..

Une autre : Pourquoi il y a toujours une fenêtre dans une <caractéristique_Activité_Femme> ?... C'est pour que la <designation_Femme> ait un point de vue...

knowledge graphs such as ConceptNet that was able to capture word similarity in a multilingual context, and (b) *target independent and transferable to other languages* because lists of representative stereotype words targeting other social groups can be easily built by extending existing compiled lists proposed in the literature (e.g., (Garg et al., 2018) for ethnic stereotypes and HurtLex (Bassignana et al., 2018) for negative stereotypes).

3.1.5 Error Analysis

A manual error analysis shows that misclassification cases are due to two main factors: the presence of a GS along with its contrary (denouncing tweets) leading to false negatives (58% of misclassifications) as in (3.3), and the presence of many words designating or describing women along with words usually used in GS leading to false positives as in (3.4).

(3.3) *Justin Trudeau se balade torse nu : il casse les codes. Une femme porte une robe courte : c'est insupportable. En France, les femmes ont gagné le droit de s'habiller comme elles le veulent. (Justin Trudeau is shirtless: he breaks the rules. A woman wears a short dress: it's unbearable. In France, women have the right to dress as they want)*

(3.4) *J'arrive pas a comprendre les gens qui supporte plusieur club t'aime qu'une femme normalement t'a qu'une mere normalement c'est la meme pour le foot t'aime qu'un club (I don't understand people who support several clubs. You love only one woman, you have only one mother. It's the same for football, you love only one club).*

3.2 Gender Stereotype Detection for Improving Sexism Detection

3.2.1 Methodology

We aim to show *how GS prediction* (considered as an auxiliary task) *can be used for sexism detection* (the main task). To this end, we used the only available resource in French, the sexism corpus presented in Section 1.3: 11,834 tweets annotated with the *sexist* tag if the tweet conveys a sexist content and *non-sexist* if not, the distribution being 34.2% for the positive class and 65.80% for the negative one. 20% of the data has been used for testing our models. It is important to note that as there is no overlap between this dataset and the GS one, this will prevent the models for sexism detection (which will integrate stereotype prediction) to be biased.

Several strategies for injecting the stereotype information in the sexism detection task were explored, ranging from using the predictions of the best stereotype model to multitask approaches (Ruder, 2017).

To this end we compare with: (1) the only existing model for French for detecting sexist HS (cf. Section 2.1), and (2) existing models that consider stereotypes as an auxiliary task to improve HS classification. Lavergne et al. (2020) is the only team in the recently shared task HaSpeeDe 2 that considers the interaction between HS towards immigrants and racial stereotype detection by employing a multitask learning approach.

3.2.2 Models

Our models are as follows:

–BERT$_{\text{gen}}$. It takes the best performing model for the task of sexism detection (cf. Section 2.1.2) which is based on BERT and trained on word embeddings, linguistic features (surface and opinion features) and generalization strategies (replacement of places and persons by an hypernym).

–BERT$_{\text{tag}}$. It uses the predictions of the best performing model for stereotype detection (i.e., BERT$_{\text{ConceptNet}}$ trained on the augmented dataset) for adding at the end of each tweet a tag indicating the presence of stereotypes (BERT$_{\text{tag\_binary}}$) or the type of stereotype (BERT$_{\text{tag\_type}}$).

–MT$_{\textbf{Lavergne}}$ (Lavergne et al., 2020). It is based on a BERT multitask architecture trained on a dataset annotated for both the presence of HS and stereotypes. However, in our case, since we rely on two different datasets (one for each task), we used the stereotype predictions of the best performing stereotype model (i.e., BERT$_{\text{ConceptNet}}$) to automatically label the sexism dataset with stereotype information. In this architecture, after transferring the text to contextual embeddings in the shared layers and retrieving the first token hidden state of the shared BERT model, we apply a dropout of 0.1 and connect it to two different layers (corresponding to the two classification tasks: sexism and stereotype).

–AngryBERT (Awal et al., 2021). This model was specifically designed to address the problem of imbalanced datasets by jointly learning HS detection with emotion classification and target identification as secondary tasks. It has been shown to outperform many strong existing multitask models, including MT-DNN (Liu et al., 2019). The overall architecture of AngryBERT is illustrated in Figure 3.5. In this architecture, the shared layer (i.e., a pre-trained BERT model) is used for learning the task-invariant features, while the private layers (i.e., BiLSTMs) are used for learning the task-specific representations. The aggregated representation of each task is then fed into its classification layer (i.e., a MLP follow by a Softmax layer for normalization).

In our case, the primary task of AngryBERT is sexism detection while the second being the detection of stereotypes. In addition to this initial configuration (**AngryBERT$_{\textbf{base}}$**), four models are newly proposed, depending on both (i) the number of labels to predict in the auxiliary task, and (ii) the dataset on which the generalization with hypernyms is performed. As previously shown (cf. Section 2.1.3), the generalization strategy performs well on the sexism dataset. In addition, we observed that a similar generalization can be employed for our task with good results. Based on these observations we are analyzing whether this generalization approach should be adopted in the sexism (i.e., **AngryBERT$_{\textbf{sexism}}$**) or in the stereotype dataset (i.e., **AngryBERT$_{\textbf{stereo}}$**).[80] In addition, as the GS dataset does not contain only instances annotated as *stereotype* vs. *non-stereotype*, but also different categories, we are analyzing whether the auxiliary task should be binary (i.e., **AngryBERT$^{\textbf{2}}$**) or multi-class (i.e., **AngryBERT$^{\textbf{4}}$**). For all the settings, the auxiliary task was trained on the augmented multilin-

---

[80]Note that we do not perform the generalization in both datasets as to not introduce bias.

Figure 3.5 – AngryBERT architecture (Awal et al., 2021).
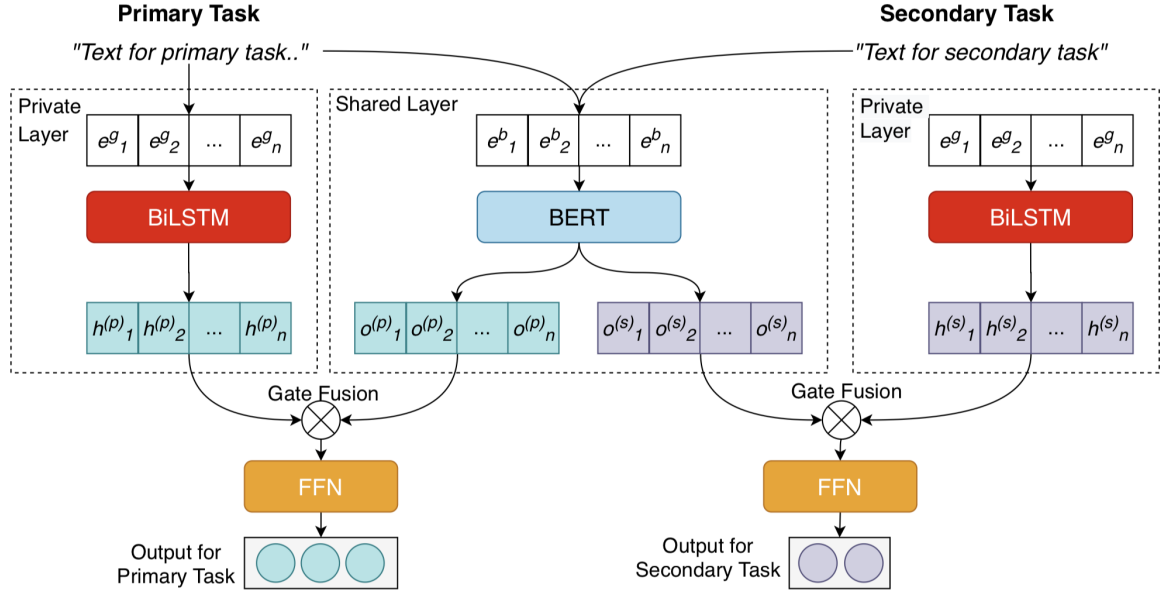
gual dataset and the generalization relies on ConcepNet, as it performed the best (cf. Section 3.1.3).

3.2.3 Results

Table 3.5 presents the multitask and the baselines results. We observe that injecting stereotypes labels as given by the automatic classifier (i.e., BERT_{tag}) outperforms both $\text{MT}_{\text{Lavergne}}$ and $\text{AngryBERT}_{\text{base}}$, the two multitask baselines. In particular, predicting the types of stereotypes is the most productive when compared to presence identification (F-score 0.796 vs. 0.776). However, when GS information is predicted jointly with sexist labels, the results tend to decrease for all AngryBERT configurations except for $\text{AngryBERT}^2_{\text{sexism}}$ and $\text{AngryBERT}^4_{\text{sexism}}$ in which we performed ConcepNet generalization on the sexism dataset only. Here again, GS types are the best with an F-score of 0.827, significally beating our strong baseline BERT_{gen} ($p < 0.05$ using the McNemar's Test statistic).

A closer look into the results per class shows that $\text{AngryBERT}^4_{\text{sexism}}$ was able to better predict sexist content (F-score=0.805 vs. 0.773 for BERT_{gen}). This suggests that GS information is definitively helpful for sexist content detection when it is injected as additional

knowledge on top of the primary task.

Table 3.5 – Results for sexist classification. ‡: baselines.

CLASSIFIER	P	R	F1
$BERT_{gen}$‡	**0.865**	0.787	0.824
$BERT_{tag\_binary}$‡	0.821	0.736	0.776
$BERT_{tag\_type}$ ‡	0.835	0.761	0.796
$MT_{Lavergne}$‡	0.803	0.749	0.775
$AngryBERT_{base}$ ‡	0.725	0.727	0.726
$AngryBERT^2_{stereo}$	0.730	0.728	0.729
$AngryBERT^4_{stereo}$	0.733	0.737	0.735
$AngryBERT^2_{sexism}$	0.836	0.813	0.824
$AngryBERT^4_{sexism}$	0.839	**0.816**	**0.827**

3.2.4 Error Analysis

An error analysis shows that 59% of missclassified instances are false negatives (sexist tweets detected as non sexist) and among them only 7% contain a GS (with a manual observation). This suggests that the majority of these sexist instances cannot benefit from the GS auxiliary task, confirming that sexist content does not necessarily entail the presence of stereotypes, as in (3.5).

(3.5) *La chance de #SégolèneRoyal, c'est qu'aux pôles ils ne mangent pas de dinde pour #Thanks-Giving ! #LaDindeSurvivante !*
(Ségolène Royal is lucky, they dont' eat turkey for Thanksgiving in the Poles! #TheSurvivor-Turkey)

Among the false positives (non sexist tweets detected as sexist), 93% are predicted as non stereotype and a manual observation confirms that only 4% contain a GS. This means that the classification errors are due to the sexism classifier. When looking at these instances, we note that 57% contain hashtags usually dedicated to sexism which are misused as in (3.6).

(3.6) *Pourquoi il n'y a jamais aucun pâtissier qui met des aliments improbables du style la tomate, du guacamole #TopChef #BalanceTonPorc*

(*Why isn't there any pastry chef who puts strange food like tomato, guacamole #TopChef #SquealOnYourPig*)

As shown with the above examples, error classifications are often due to humor, jokes, irony or puns, meaning that accounting for these phenomena for HS detection is still an open problem.

Conclusion

In this part, we proposed the first approach for GS detection in tweets as well as several deep learning strategies to inject appropriate knowledge about how stereotypes are expressed in language into sexist HS classification. Our main contributions include:

(1) a new dataset for GS detection;

(2) a method to counter class imbalance based on sentence similarity from multilingual external datasets;

(3) different strategies to incorporate GS triggers as input into the learning process based on automatically extended lexicon via a multilingual knowledge graph, and finally;

(4) an empirical evaluation of the positive impact of multiclass GS detection on improving HS against women based on multitask architectures, beating several strong state of the art baselines.

GS is an understudied problem and we believe it should not only be viewed as a type of sexism/misogyny but considered instead as an independent task to be used in other applications as well. Among them, education is a promising future direction for selecting which digital media/books are being given to children, as previous research has indicated that the stereotypes children encounter in their environment can impact their motivational dispositions and attitudes. In the future, we plan on addressing these issues, as well as developing approaches for leveraging the GS information in other datasets annotated for sexism.

Emotionally Informed Hate Speech Detection: a Multi-target Perspective

Motivation

In spite of there being no universally accepted definition of HS, this study employs the most common one. HS is defined here as *'any type of communication that is abusive, insulting, intimidating, and/or that incites violence or discrimination, and that disparages a person or a vulnerable group based on characteristics such as ethnicity, gender, sexual orientation and religion'* (Erjavec and Kovačič, 2012).

In this part, we consider different manifestations of HS with different topical focuses, including *sexism*, *misogyny*, *racism*, and *xenophobia*. Each specific instance targets different vulnerable groups based on characteristics such as gender (*sexism* and *misogyny*), ethnicity, religion and race (*xenophobia* and *racism*). The focus on gendered and ethnicity-based HS is due, in part, to the wide availability of English corpora developed by the computational linguistics community for those targets. But it also depends on the fact that most monitoring exercises by institutions countering online HS in different countries and territories (e.g., European Commission (EU Commission, 2016)) report ethnic-based hatred (including anti-migrant hatred) and gender-based hatred as the most common type of online HS (Chetty and Alathur, 2018).

As previously seen, an immediate but rather expensive solution for handling a new specific target is that of building new target-oriented datasets from scratch; as has been done in previous studies (Ibrohim and Budi, 2019). In this part, we propose instead a novel multi-target HS detection approach by leveraging existing manually annotated datasets. These will enable the model to transfer knowledge from different datasets with different topics and targets. In the context of offensive content moderation, identifying the topical focus and the targeted community of hateful contents would be of great interest for two important reasons. First, it will allow us to detect HS for specific topics/targets when dedicated data are missing. Second, it will prevent widespread stereotypes and help to develop social poli-

cies for protecting victims, especially in response to trigger events (King and Sutton, 2013). For example, with the recent outbreak of COVID-19, a spike in racist and xenophobic messages targeting Asians in Western countries was observed. A system specifically designed to detect HS that targets migrants in a pre-COVID-19 context would most likely have failed at picking out this post-COVID-19 HS. Indeed, most of the messages would not have been moderated as the type of language learned during training was for other groups, the most frequent targets of HS in pre-COVID times.

We propose to undertake the following challenges:

- (1) *Explore the ability of HS detection models to capture common properties from topic-generic datasets and transfer this knowledge to recognize specific manifestations of HS.*

- (2) Experiment with the development of models to detect both topics (racism, xenophobia, sexism, misogyny) and HS targets, going beyond standard binary classification, to investigate *how to detect HS at a finer level of granularity* and *how to transfer knowledge across different topics and targets*.

- (3) Study the impact of affective knowledge encoded in sentic computing resources (SenticNet, EmoSenticNet) and in semantically-structured hate lexicons (HurtLex) in determining specific manifestations of HS.

The remainder of this part is organized as follows. In the next chapter, we present an overview of the main works on HS detection. Chapter 2 details the experiments carried out and the results obtained when generalizing HS phenomena across multiple datasets, predicting multi-target HS, and building emotionally-informed models.

We end this part by discussing our main findings.

1 Related Work

We present the related work in four parts. First, we briefly introduce the affective computing and sentiment analysis research field, in order to provide readers with a broader context for NLP literature related to the analysis and to the recognition of affective states and emotions in texts. Second, relevant prior works specifically related to HS detection are presented. Third, we review the domain adaptation study in sentiment analysis and abusive language detection, something particularly important in bringing out the novelty of our contribution. Finally, we provide an overview of the few attempts to exploit affective information in improving abusive language detection.

1.1 Affective Computing and Sentiment Analysis

Affective computing, a development of the last decades, is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects: i.e., the experience of feelings or emotions. Today, identifying affective states from text is regarded as being fundamental for several domains, from human-computer interaction to artificial intelligence, from the social sciences to software engineering (Cambria et al., 2017). The wide popularity of social media, which facilitates users publishing and sharing contents – providing accessible ways for expressing feelings and opinions about anything, anytime – also gave a major boost to this research area. This was especially true within the NLP field. Here the abundance of data allowed the research community to tackle more in-depth, long-standing questions such as understanding, measuring and monitoring the sentiment of users towards certain topics or events, expressed in mere texts or through visual and vocal modalities (Poria et al., 2018). Indeed, robust and effective approaches are made possible by the rapid progress in supervised learning technologies and the huge amount of user-generated content available online. Such techniques are typically motivated by the need to

extract user opinions on a given product or, say, in surveying political views and they often exploit knowledge encoded in affective resources, such as sentiment and emotion lexicons and ontologies.

The interest in lexical knowledge about the multi-faceted and the fine-grained facets of affect encoded in such resources is, by no means, limited to sentiment analysis. The use of such affective resources has also recently been explored in other related tasks, such as personality (Mohammad and Kiritchenko, 2013; Mehta et al., 2020) and irony detection (Sulis et al., 2016; Farías et al., 2016) or author profiling (Pardo and Rosso, 2016). Concerning abusive language detection, which is the specific task of interest here, there are attempts at exploiting emotion signals to improve the detection of this kind of phenomena (cf. Section 1.3). No one has investigated the impact of emotion features on HS detection, which is one of the challenges tackled in this dissertation.

1.1.1   Supervised and Semi-Supervised Learning for Social Data Analysis

The field has recently been surveyed in (Benamara et al., 2017; Yadav and Vishwakarma, 2020). The vast majority of the analyzed papers describe approaches to sentiment analysis based on supervised learning, where there is a text classification task at the sentence or message level, focused mostly on detecting from text valence or *sentiment*, either using a binary value or with a strength/intensity component coupled with the sentiment (Thelwall et al., 2012). In particular, deep learning-based methods are becoming very popular due to their high performance, and they have been increasingly applied in sentiment analysis (Yadav and Vishwakarma, 2020; Minaee et al., 2021). Furthermore, there is an ever-increasing awareness of the need to take a holistic approach to sentiment analysis (Cambria et al., 2017) by handling the many finer-grained tasks involved in extracting meaning, polarity and specific emotions from texts. This includes the detection of irony and sarcasm (Sulis et al., 2016; Karoui et al., 2017; Hazarika et al., 2018).

Due to a large amount of available (but unlabeled) data, many studies have recently highlighted the importance of exploring unsupervised and semi-supervised machine learning techniques for sentiment analysis tasks. For example in (Hussain and Cambria, 2018), the authors exploited both labeled and unlabeled commonsense data. Their proposed affective reasoning architecture is based on SVM and the merged use of random projection scaling in a vector space model and was exploited for emotion recognition tasks.

1.1.2 Emotion Categorization Models and Affective Resources

Still, despite the maturity of the field, choosing the right model for operationalizing affective states is not a trivial task. Research in sensing sentiment from texts has put the major emphasis on recognizing polarities (positive, negative, neutral orientation). However, comments and opinions are usually directed toward a specific target or aspect of interest, and as such, finer-grained tasks can be envisioned. For instance, aspect-based sentiment analysis identifies the aspects of given target entities and the sentiment expressed for each aspect (Pontiki et al., 2014). At the same time, the stance detection emerging task focuses on detecting what particular stance a user takes toward a specific target, something that is particularly interesting in political debates (Mohammad et al., 2017).

Moreover, given the wide variety of affective states, recent studies advocate a finer-grained investigation of the role of *emotions*, as well as the importance of other affect dimensions such as emotional intensity or activation. Depending on the specific research goals addressed, one might be interested in issuing a discrete label describing the affective state expressed (frustration, anger, joy, etc.) in accordance with different contexts of interaction and tasks. Emotions are transient and typically episodic, in the sense that, over time, they can come and go. This depends, of course, on all sorts of factors, factors which researchers might be interested in understanding and modeling according to a domain or task-specific research objectives.

Both basic emotion theories, in the Plutchik-Ekman tradition (Plutchik, 1980; Ekman, 1999), and dimensional models of emotions (Russell, 1980) provide a precious theoretical grounding for the development of lexical resources and computational models for affect extraction. Sentiment-related information is, indeed, often encoded in lexical resources, such as affective lists and corpora, where different nuances of affect are captured, such as sentiment polarity, emotional categories, and emotional dimensions (Poria et al., 2013; Mohammad and Turney, 2013; Cambria et al., 2018). These kinds of lexicons are usually lists of words to which a positive or negative or/and an emotion-related label (or score) is associated. Besides flat lists of affective words, lexical taxonomies have also been proposed, enriched with sentiment and/or emotion information (Baccianella et al., 2010; Poria et al., 2013). However, there is a general tendency to go towards richer, finer-grained models. These will very possibly include complex emotions. This is especially the case in the context of data-driven and task-driven approaches, where restricting automatic detection to only a small set of

basic emotions is too limited, not least in terms of actionable affective knowledge. This general tendency is also reflected in the development of semantically richer resources. These include and model semantic, conceptual, and affective information associated with multi-word natural language expressions, by enabling the concept-level analysis of sentiment and emotions conveyed in texts, like the ones belonging to the SenticNet family (Cambria et al., 2018, 2020). Moreover, when the task addressed is related to a specific portion of the affective space, domain-specific affective resources and lexicons can be envisioned. This is the case with abusive language detection, where the use of lexicons of hateful words (Bassignana et al., 2018) can lead to interesting results.

1.1.3   Word Intensity and Polarity Disambiguation

All such resources represent a rich and varied lexical knowledge about affect, under different perspectives, and virtually all sentiment analysis systems may incorporate lexical information derived from them.[81] However, many opinion keywords carry varying polarities in different contexts, posing huge challenges for sentiment analysis research. Contextual polarity ambiguity is an important still little studied problem in sentiment analysis. This has recently been addressed in (Xia et al., 2015), where a Bayesian model is proposed that uses opinion-level features to solve the polarity problem of sentiment-ambiguous words: intra-opinion features (i.e., the information that helps in thoroughly conveying the opinion); and inter-opinion features (i.e., the information connecting two or more opinions). The intra-opinion features resolve the polarity of most sentiment words. The inter-opinion features usually play a secondary role, either by improving the confidence of a good prediction or by assisting in calculations when some of the features are missing.

Another interesting challenge for the field is related to the possibility of measuring sentiment and emotion intensity, which is of paramount importance in analyzing the finer-level details of emotions and sentiments (Mohammad et al., 2018) in real-world applications. A novel solution to this problem is proposed in (Akhtar et al., 2020), where, in order to leverage the various advantages of different supervised systems, a Multi-Layer Perceptron (MLP) based ensemble framework for predicting the intensity of sentiments (in financial microblog messages and news headlines) and emotions (in tweets) is proposed. The ensemble model combines the output of three deep learning models (CNN, LSTM and GRU) and a feature-

---

[81]For a comprehensive description and an evaluation of the different ways lexicons have been employed in sentiment analysis systems, see (Nissim and Patti, 2017).

based Support Vector Regression (SVR) model. The SVR model utilizes word and character Tf/IDf, Tf/IDf weighted word vectors, and a diverse set of lexicon features, such as the positive and negative word count (extracted from MPQA (Wiebe and Mihalcea, 2006) and Bing Liu (Ding et al., 2008)), the positive, negative, and aggregate scores of each word extracted from NRC Hashtag Sentiment and NRC Sentiment140 (Mohammad et al., 2013), as well as the sum of the positive, negative and aggregate scores of each word computed from SentiWordNet (Baccianella et al., 2010). For emotion intensity prediction, the authors also include: the word count of each of the emotions from NRC Word-Emotion Association lexicon (Mohammad and Turney, 2013); the sum of association scores for the words with the emotions extracted from NRC Hashtag Emotion (Mohammad, 2012); the aggregate of positive and negative word scores computed from AFINN (Nielsen, 2011); and the sentiment score of each sentence returned by VADER (Gilbert and Hutto, 2014). The proposed framework shows good results with comparatively better performance over state-of-the-art systems.

1.2 Domain Adaptation in Abusive Language Detection

The study of HS detection is multifaceted, and available datasets feature different focuses and targets. Despite limitations, some works have tried to bridge this range by proposing a domain adaptation approach to transfer knowledge from one dataset to other datasets with different topical focuses.

The first attempt to deal with this issue was reported in (Waseem et al., 2018). They used the multi-task learning (MTL) approach, arguing that it would be possible to share knowledge between two or more objective functions to leverage information encoded in one abusive language dataset to better-fit others. Karan and Šnajder (2018) proposed using a traditional machine learning approach for classifying abusive language in a cross-domain setting, in order to get better system interpretability. This work also explored the use of the *frustratingly simple domain adaptation* (FEDA) framework (Daumé III, 2007) to facilitate domain sharing between different datasets. The main finding of this work is that the model did not generalize well when applied to various domains, even when trained on a much bigger out-domain dataset. Rizoiu et al. (2019) adopted transfer learning as a domain adaptation approach by exploiting the LSTM network coupled with ELMo embeddings. LSTM has also been used by Pamungkas and Patti (2019), who employed it with a list of abusive keywords

from the Hurtlex lexicon (Bassignana et al., 2018), as a proxy for transferring knowledge across different datasets. Their main findings are: (i) that the model trained on more than one general abusive language dataset will produce more robust predictions; and (ii) that HurtLex is able to boost the system performance in the cross-domain setting.

BERT (Devlin et al., 2019) was also applied in cross-domain abusive language detection (Swamy et al., 2019). This work found that BERT can share knowledge between one domain dataset and other domains, in the context of transfer learning. They argue that the main difficulty in the cross-domain classification of abusive language is caused by dataset issues and their biases. It is consequently impossible for datasets to capture the phenomenon of abusive language in its entirety. Mozafari et al. (2019) also investigated BERT by using new fine-tuning methods based on transfer learning, relying on Waseem (Waseem and Hovy, 2016) and Davidson (Davidson et al., 2017) datasets in their experiments. Finally, HatEval, a recently shared task (Basile et al., 2019), also provided an HS dataset that covers two different targets, women and immigrants. Therefore, participants are required to build a target-agnostic model able to detect HS with more than one target (cf. Section 2.4).

Cross-domain classification approaches in abusive language detection share three common characteristics: (1) Dataset labels are aligned to deal with the varieties of annotation schemes. Hence, all datasets (be they topic-generic or topic-specific) share the same coarse-grained characterization of HS (i.e., *hateful* vs. *non-hateful*). (2) Systems follow a one-to-one configuration (i.e., they are trained on one dataset and tested on another) in order to analyze their robustness in generalizing the different phenomena contained in each dataset. (3) Predictions are binary, ignoring the target/topic nature of HS. In this work, we intend to focus on the different topics/targets in several datasets by proposing a multi-target HS classification task.

To this end, instead of using the typical one-to-one configuration, we propose to solve the problem using a many-to-many configuration capable of identifying a given topic/target when trained in topic-generic or topic-specific datasets. The many-to-many configuration has already been shown to be quite effective in cross-domain aspect-based sentiment analysis (Peng et al., 2018; Liu et al., 2018; Ganin and Lempitsky, 2015; Goodfellow et al., 2014; Zhang et al., 2019; Cai and Wan, 2019) and is used here for the first time in an HS detection task.

1.3 Affective Information in Abusive Language Detection Tasks

Recently, some works exploiting emotion signals to improve abusive language detection have been carried out. The study by Samghabadi et al. (2020) proposed an architecture that uses the Emotion-Aware Attention (EA) mechanism to quantify the importance of each word based on the emotion conveyed by the text. They used DeepMoji model (Felbo et al., 2017) and NRC Emotion Lexicon (Mohammad and Turney, 2013) to extract emotion information from the given texts. Their analysis of the results shows the importance of affective information in augmenting system performance. Similar conclusions have been drawn in (Pamungkas et al., 2020a) who exploited the NRC Emotion Lexicon (Mohammad and Turney, 2013) and EmoSenticNet (Poria et al., 2013). Finally, the most recent work by Rajamanickam et al. (2020) came up with a joint model of emotion and abusive language detection in a MTL setting. This led to significant improvements in abuse detection performance when evaluated in both the OffensEval 2019 (Zampieri et al., 2019b) and Waseem and Hovy datasets (Waseem and Hovy, 2016).

As far as we know, no previous work has explored the impact of emotion features in predicting HS targets in a multi-target setting. Moreover, most of the works listed here model their tasks as a binary classification, with the aim of predicting the abusiveness of a given utterance *per se* (i.e., without specifying either a topic or a target). In the next chapter (cf. Chapter 2), we classify a message as hateful or not-hateful. But we go further. We want also to detect the HS topic and the target to whom the message is addressed. We also propose to employ EmoSenticNet, HurtLex, and for the first time, SenticNet.

2

Towards Multi-target Hate Speech Detection

In this chapter, we focus on the detection of the HS topic and the target to whom the message is addressed. To the best of our knowledge, we are the first to address target-based computational HS detection, continuing recent corpus-based linguistic studies on categorizing HS and their associated targets (Silva et al., 2016).

We first present the existing datasets that were used. Then, we present the models, experiments and results for HS topic detection and for target detection. Finally, we explore the impact of emotion resources on target detection by identifying the emotion categories that are the most suitable for predicting a given topic/target of HS detection.

2.1 Datasets

We experiment with seven available HS corpora from previous studies among which two are topic-generic (`Davidson` (Davidson et al., 2017) and `Founta` (Founta et al., 2018)), and four are topic-specific about four different topics: *misogyny* (the English[82] AMI dataset collection from both `IberEval` (Fersini et al., 2018b) and `Evalita` (Fersini et al., 2018a)), *misogyny and xenophobia* (the `HatEval` dataset (Basile et al., 2019)), and *racism* and *sexism* (the `Waseem` dataset (Waseem and Hovy, 2016)). Each of these topics target either gender (sexism and misogyny) and/or ethnicity, religion or race (xenophobia and racism).

Table 2.1 provides a general overview of the datasets,[83] along with the labels used in their annotation schemes. We can observe that the classes are imbalanced in most datasets, where the majority class is the negative class (non-HS), except for the AMI collection

---

[82] As the majority of the resources annotated for different topics/targets are in English, we only selected the English instances from this multilingual corpora.

[83] For more details regarding the collection and annotation of the data, the reader is invited to refer to Section 2.1.

(`AMI-IberEval` and `AMI-Evalita`) and `Davidson`.

Table 2.1 – General overview of the datasets along with their topics and targets.

DATASET	LABELS	NO. OF INSTANCES	TOPIC	TARGET
Davidson	hate speech	1,430	24,783	generic	none
	offensive	19,190			
	neither	4,163			
Founta	abusive	27,037	99,799	generic	none
	hateful	4,948			
	spam	14,024			
	normal	53,790			
Waseem	racism	1,957	16,488	specific	race women
	sexism	3,216			
	none	11,315			
Evalita	misogyny	2,245	5,000	specific	women
	not misogyny	2,755			
IberEval	misogyny	1,851	3,977	specific	women
	not misogyny	2,126			
HatEval	immigrant	2,427	11,971	specific	women ethnicity
	women	2,608			
	not hate speech	6,936			

For our experiments, the corpora have been divided into train and test sets keeping the same tweet distribution as the original papers. This was done in order to make better comparisons with the state-of-the-art results.[84] Table 2.2 and Table 2.3 provide the distribution of instances in these two sets. As one of the research questions that we want to address involves the possibility of transferring knowledge from several topic-specific datasets into another topic-specific dataset where the topic is unseen, we decided to merge under the same topic (i.e., misogyny) both the `AMI corpora` and `HatEval dataset`.[85]

In the next three sections, we show how these datasets have been used to develop models that are able to generalize HS across multiple datasets (cf. Section 2.2); transfer knowledge

---

[84]The only difference with the original paper appears in the training set of the `HatEval` dataset as we found duplicate instances (already there in the `AMI corpora`).

[85]We recall that these two datasets used the same approach for collecting the data and for annotation guidelines.

Table 2.2 – Distribution of instances in topic-generic datasets (used as training).

DATASET	LABELS		NO. OF INSTANCES
Founta	hateful	1,930	39,700
	not-hateful	37,770	
Davidson	hateful	1,430	5,593
	not-hateful	4,163	

Table 2.3 – Distribution of instances in the train/test sets in topic-specific datasets.

TOPIC	Racism (**Waseem**)			Sexism (**Waseem**)		
	Racism	Non-racism	Total	Sexism	Non-sexism	Total
TRAIN	1,346	7,943	9,289	2,253	7,943	10,196
TEST	611	3,373	3,984	963	3,373	4,336

TOPIC	Misogyny (**AMI corpora + HatEval**)				Xenophobia (**HatEval**)		
		Misogyny	Non-misogyny	Total	Hateful	Non-hateful	Total
TRAIN	Evalita	1,785	2,215	4,000	1,988	3,012	5,000
	HatEval	1,305	1,396	2,701			
	IberEval	1,568	1,683	3,251			
	Total	4,658	5,294	9,952			
TEST	Evalita	460	540	1,000	629	870	1,499
	HatEval	623	849	1,472			
	IberEval	283	443	726			
	Total	1,366	1,832	3,198			

across topics and targets (cf. Section 2.3); and leverage emotions to improve multi-target HS detection (cf. Section 2.4). The various forms of bias introduced when building these datasets are discussed in Section 2.3, as they may have a strong impact on the multi-target experiments proposed in this dissertation.

2.2 Generalizing Hate Speech Phenomena Across Multiple Datasets

2.2.1 Methodology

We aim to answer two main research questions:

- *Are models able to capture common properties of HS and transfer this knowledge from topic-generic datasets to topic-specific datasets?*

- *How do these models compare with ones that are trained on topic-specific datasets?*

To this end, we propose the following two configurations:

- $Top^G \longrightarrow Top^S$: Train on topic-general HS datasets (i.e., `Davidson` and `Founta`)[86] and test on *all* topic-specific datasets (i.e., $Racism_{Waseem}$, $Sexism_{Waseem}$, $Misogyny_{Evalita}$, $Misogyny_{IberEval}$, $Misogyny_{HatEval}$, and $Xenophobia_{HatEval}$) without splitting them into train/test.

- $Top^S \longrightarrow Top^S$: Train on the combined training sets of all topic-specific datasets (i.e., `Waseem`, `HatEval`, `Evalita`, and `IberEval`) and test on the test set of each topic-specific dataset.

These two configurations are cast as a binary classification task, where the system needs to predict whether a given tweet is hateful (1) or not (0). To this end, we experiment with several performing state of the art models for HS detection. This is a necessary first step in measuring to what extent existing models are capable of transferring knowledge across different HS datasets, be they topic-generic or topic-specific.

2.2.2 Models

Our models are as follows:[87]

---

[86]We only use the *hateful* and *not-hateful* instances, although the data is annotated as *hate-speech*, *offensive* and *none* (for the `Davidson` dataset) and annotated as *hate-speech*, *abusive*, *normal* and *spam* (for the `Founta` dataset).

[87]In an exploratory attempt at finding the best way of representing the data, we included a standard pre-processing step (i.e., URLs and user mentions replacement with replacement tokens, RT removal) as well as emoji replacement with their detailed description (Singh et al., 2019). However, the results were inconclusive.

– **Baseline**. This model is straight-forward based on a Linear Support Vector Classifier (LSVC). The use of linear kernel is based on (Joachims, 1998), who argue that the linear kernel has an advantage for text classification. They observe that text representation features are frequently linearly separable. Hereby, the baseline is an LSVC with unigrams, bigrams, and trigrams Tf/IDf.

– **LSTM.** This model uses a LSTM network (Hochreiter and Schmidhuber, 1997) with an architecture consisting of several layers, starting with an embedding layer representing the input to the LSTM network (128 units), followed by a dense layer (64 units) with ReLU activation function. The final layer consists of a dense layer with sigmoid activation producing the final prediction. In order to get the best possible results, we optimized the batch size (16, 32, 64, 128) and the number of epochs (1-5). We used as input either randomly initialized embeddings (**LSTM**) or FastText[88] English word vectors with an embedding dimension of 300 (Grave et al., 2018) pre-trained on Wikipedia and Common Crawl (**LSTM$_{\text{FastText}}$**). LSTM, a type of Recurrent Neural Network (RNN), has already been proven as a robust architecture in HS detection (Badjatiya et al., 2017).

– **CNN$_{\text{FastText}}$**. This model was inspired by (Badjatiya et al., 2017; Gambäck and Sikdar, 2017). It uses FastText English word vectors (with the dimension of 300) and three 1D convolutional layers, each one using 100 filters and a stride of 1, but with different window sizes (respectively 2, 3, and 4) in order to capture different scales of correlation between words, with a ReLU activation function. We further downsample the output of these layers by a 1D max-pooling layer and we feed its output into the final dense layer. All the experiments run for a maximum of 100 epochs, with a patience of 10 and a batch size of 32.[89]

– **ELMo**. This model employs ELMo (Peters et al., 2018), a deep contextualized word representation, which shows a significant improvement in the study of HS (Rizoiu et al., 2019). Since we implement ELMo as a Keras layer,[90] we were able to add more layers after the word embedding layer. The latter is followed by a dense layer (256 units) and a dropout

---

[88] https://fasttext.cc/
[89] All the hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.
[90] https://keras.io/

rate of 0.1, before being passed to another dense layer (2 units) with a sigmoid activation function, which produces the final prediction. This architecture is fine-tuned based on the number of epochs (1-15) and batch-size (16, 32, 64, and 128), and optimized by using Adam optimizer.[91]

– BERT. This model uses the pre-trained BERT model (BERT-Base, Cased), (Devlin et al., 2019) on top of which we added an untrained layer of neurons. We then used the Hugging-Face's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for three epochs with a learning rate of 2e-5 and AdamW optimizer. It is based on (Swamy et al., 2019) where it achieved the best results for the task of abusive language detection.

2.2.3 Results

2.2.3.1 Results for the $Top^G \longrightarrow Top^S$ Configuration

Table 2.4 and Table 2.5 present our results when training respectively on `Founta` and `Davidson`. We provide our results in terms of accuracy (A), macro-averaged F-score (F_1), precision (P) and recall (R) with the best results in terms of F_1 presented in bold.

We recall here that we focus on learning topic-generic HS properties and test how neural models are able to extrapolate this information in order to detect topic-specific HS. The results show that **ELMo** outperformed other models in the `Waseem` dataset (`Racism`$_{\text{Waseem}}$, `Sexism`$_{\text{Waseem}}$) when trained on `Davidson`. When trained on `Founta`, **CNN$_{\text{FastText}}$** obtained the best results for `Sexism`$_{\text{Waseem}}$ and **BERT** for `Racism`$_{\text{Waseem}}$. For most of the topic-specific testing datasets (`AMI corpora` in particular), the results are comparable across the two general HS training datasets (`Davidson` and `Founta`), with higher disparities being observed in the `Waseem` results.

2.2.3.2 Results for the $Top^S \longrightarrow Top^S$ Configuration

Table 2.6 presents the results obtained when focusing on learning topic-specific HS properties by combining all training sets of all datasets. The overall picture of the results shows that our baseline (i.e., **LSVC**) performed quite well when compared to other models: it presents a decrease of anywhere in between 1% and 11% in terms of $F1$ score, when compared to the

---

[91]We use the default parameter of Adam optimizer as described in https://www.tensorflow.org/api _docs/python/tf/keras/optimizers/Adam

Table 2.4 – Results for $Top^G \longrightarrow Top^S$ configuration when training on `Founta`.

DATASET	Baseline				LSTM				LSTM$_{\text{FastText}}$			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.680	0.601	0.638	0.850	0.613	0.533	0.570	0.842	0.666	0.585	0.623	0.846
Sexism$_{\text{Waseem}}$	0.555	0.516	0.534	0.760	0.585	0.517	0.549	0.771	0.624	0.543	0.581	0.773
Xenophobia$_{\text{HatEval}}$	0.632	0.542	0.583	0.622	0.602	0.507	0.550	0.601	0.589	0.509	0.546	0.601
Misogyny$_{\text{Evalita}}$	0.627	0.582	0.603	0.612	0.692	0.634	0.662	0.661	0.679	0.649	**0.664**	0.669
Misogyny$_{\text{IberEval}}$	0.622	0.569	0.594	0.592	0.669	0.610	0.638	0.630	0.662	0.625	0.643	0.641
Misogyny$_{\text{HatEval}}$	0.615	0.584	0.599	0.615	0.632	0.616	0.624	0.636	0.636	0.631	0.633	0.642
Misogyny$_{\text{all}}$	0.645	0.584	0.613	0.616	0.655	0.619	0.636	0.643	0.651	0.632	**0.641**	0.649

DATASET	CNN$_{\text{FastText}}$				BERT				ELMo			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.700	0.627	0.661	0.855	0.705	0.742	**0.723**	0.840	0.584	0.568	0.575	0.806
Sexism$_{\text{Waseem}}$	0.622	0.563	**0.591**	0.767	0.528	0.501	0.514	0.712	0.543	0.524	0.533	0.736
Xenophobia$_{\text{HatEval}}$	0.624	0.517	0.565	0.607	0.651	0.652	**0.651**	0.611	0.581	0.520	0.548	0.604
Misogyny$_{\text{Evalita}}$	0.649	0.612	0.629	0.637	0.651	0.659	0.654	0.663	0.635	0.608	0.621	0.630
Misogyny$_{\text{IberEval}}$	0.629	0.590	0.609	0.609	0.661	0.639	**0.649**	0.661	0.602	0.571	0.586	0.590
Misogyny$_{\text{HatEval}}$	0.609	0.595	0.601	0.616	0.632	0.637	**0.634**	0.639	0.620	0.602	0.610	0.625
Misogyny$_{\text{all}}$	0.628	0.615	0.621	0.630	0.643	0.637	0.639	0.647	0.627	0.597	0.612	0.621

best-performing models for a specific topic. For most topics, the best results were obtained by **BERT**, with the only exception being for the Misogyny$_{\text{HatEval}}$ dataset, where **ELMo** obtained the best results (with a difference of almost 2% in terms of $F1$ score). We note that Misogyny$_{\text{HatEval}}$ is the only dataset for which **ELMo** achieved good results. For all the other datasets, the results are low, even lower than the baseline.[92] We also note that state of the art models achieved good results for both topics in the Waseem dataset, whereas they attain lower results when tested on the xenophobia topic from the HatEval dataset. However, our results are similar to the ones obtained by state-of-the-art baselines for Waseem ($F1$=0.739 (Waseem and Hovy, 2016)) and HatEval ($F1$=0.451 (Basile et al., 2019)).[93]

In order to assess whether training on topic-specific data improves the results beyond those achieved by training on topic-generic data, we compare our results with both the baselines and the best-submitted systems in the shared task competition where these data has been used (only available for AMI corpora). The comparison was made by training ei-

---

[92]The baseline achieved better results in all datasets, except the topics in the HatEval dataset.

[93]The baseline for the Waseem dataset is a LR coupled with character n-grams and the gender information of the tweet author, while the baseline for the HatEval shared task is a straightforward SVM with Tf/IDf features.

Table 2.5 – Results for $Top^G \longrightarrow Top^S$ configuration when training on `Davidson`.

DATASET	Baseline				ELMo				LSTM			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
`Racism`$_{\text{Waseem}}$	0.585	0.560	0.572	0.814	0.665	0.661	**0.663**	0.833	0.573	0.535	0.553	0.852
`Sexism`$_{\text{Waseem}}$	0.558	0.528	0.542	0.747	0.628	0.586	**0.606**	0.761	0.574	0.526	0.549	0.761
`Xenophobia`$_{\text{HatEval}}$	0.601	0.541	0.569	0.615	0.616	0.544	0.577	0.620	0.604	0.517	0.557	0.605
`Misogyny`$_{\text{Evalita}}$	0.668	0.666	0.667	0.672	0.623	0.624	0.624	0.626	0.680	0.681	**0.680**	0.682
`Misogyny`$_{\text{IberEval}}$	0.638	0.633	0.635	0.639	0.632	0.631	0.631	0.635	0.678	0.676	**0.677**	0.680
`Misogyny`$_{\text{HatEval}}$	0.635	0.636	0.635	0.630	0.621	0.622	0.621	0.619	0.638	0.636	0.637	0.623
`Misogyny`$_{\text{all}}$	0.653	0.654	0.654	0.657	0.623	0.617	0.620	0.628	0.657	0.658	0.657	0.656

DATASET	LSTM$_{\text{FastText}}$				CNN$_{\text{FastText}}$				BERT			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
`Racism`$_{\text{Waseem}}$	0.613	0.656	0.634	0.775	0.622	0.617	0.619	0.812	0.605	0.561	0.582	0.819
`Sexism`$_{\text{Waseem}}$	0.544	0.540	0.542	0.699	0.586	0.557	0.571	0.744	0.544	0.531	0.537	0.741
`Xenophobia`$_{\text{HatEval}}$	0.635	0.547	0.588	0.624	0.641	0.551	**0.592**	0.628	0.635	0.527	0.575	0.607
`Misogyny`$_{\text{Evalita}}$	0.635	0.620	0.627	0.602	0.652	0.653	0.652	0.652	0.676	0.678	0.677	0.673
`Misogyny`$_{\text{IberEval}}$	0.649	0.635	0.643	0.623	0.653	0.653	0.653	0.654	0.663	0.661	0.662	0.661
`Misogyny`$_{\text{HatEval}}$	0.619	0.593	0.606	0.562	0.659	0.647	**0.652**	0.626	0.639	0.644	0.641	0.624
`Misogyny`$_{\text{all}}$	0.633	0.614	0.623	0.594	0.658	0.657	**0.658**	0.648	0.654	0.654	0.654	0.649

ther on a topic-general dataset (i.e., $Top^G \longrightarrow Top^S$) or on all topic-specific datasets (i.e., $Top^S \longrightarrow Top^S$), and testing the test data provided by the organizers of AMI-IberEval and AMI-Evalita. Table 2.7 shows our results.

When compared to the AMI `Misogyny`$_{\text{Evalita}}$ and `Misogyny`$_{\text{IberEval}}$ baselines[94] provided in terms of accuracy (respectively 0.605 and 0.783), we observe that using a topic-specific training approach, **BERT** achieved more than a 10% increase for both datasets, while for the topic-generic training approach the only improvement of (0.5%) is brought by **BERT** trained on the `Davidson` dataset (for `Misogyny`$_{\text{Evalita}}$). When comparing the results with the best-submitted systems (0.704 and 0.913[95]) we still observe a small improvement achieved by **BERT** trained on topic-specific data for the `Misogyny`$_{\text{Evalita}}$ task, though all the other system results were lower. These results confirm that a model trained with a combination of several datasets with different topical focuses is more robust than a model trained on a topic-generic dataset.

---

[94]SVM with linear kernel trained on the unigram representation of the tweets.

[95]The best-submitted system for the AMI Evalita competition is an LR with a vector representation that concatenates sentence embedding, Tf/IDf and average word embeddings, while for the AMI IberEval competition it was an SVM with a combination of structural, stylistic and lexical features.

Table 2.6 – Results for $Top^S \longrightarrow Top^S$ when training on `Waseem`, `HatEval` and `AMI` train sets.

DATASET	Baseline				LSTM				LSTM$_{\text{FastText}}$			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.786	0.798	0.792	0.889	0.796	0.765	0.779	0.878	0.783	0.783	0.783	0.887
Sexism$_{\text{Waseem}}$	0.815	0.790	0.801	0.868	0.787	0.795	0.791	0.857	0.758	0.807	0.775	0.855
Xenophobia$_{\text{HatEval}}$	0.572	0.546	0.470	0.497	0.530	0.560	0.427	0.471	0.546	0.589	0.447	0.488
Misogyny$_{\text{Evalita}}$	0.645	0.646	0.645	0.646	0.652	0.652	0.648	0.648	0.661	0.660	0.657	0.658
Misogyny$_{\text{IberEval}}$	0.803	0.732	0.742	0.778	0.709	0.754	0.717	0.750	0.739	0.793	0.749	0.779
Misogyny$_{\text{HatEval}}$	0.659	0.551	0.421	0.487	0.613	0.688	0.534	0.561	0.564	0.665	0.447	0.502
Misogyny$_{\text{all}}$	0.630	0.624	0.601	0.602	0.650	0.654	0.631	0.631	0.636	0.644	0.612	0.614

DATASET	CNN$_{\text{FastText}}$				BERT				ELMo			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.764	0.800	0.782	0.827	0.775	0.844	**0.802**	0.884	0.616	0.833	0.651	0.874
Sexism$_{\text{Waseem}}$	0.793	0.798	0.795	0.816	0.807	0.829	**0.817**	0.869	0.589	0.815	0.599	0.810
Xenophobia$_{\text{HatEval}}$	0.492	0.471	0.481	0.462	0.619	0.543	**0.578**	0.577	0.562	0.596	0.543	0.609
Misogyny$_{\text{Evalita}}$	0.673	0.684	0.678	0.684	0.704	0.705	**0.704**	0.706	0.562	0.672	0.496	0.594
Misogyny$_{\text{IberEval}}$	0.713	0.742	0.727	0.735	0.841	0.840	**0.840**	0.848	0.538	0.774	0.460	0.639
Misogyny$_{\text{HatEval}}$	0.603	0.532	0.565	0.553	0.694	0.523	0.596	0.573	0.618	0.643	**0.615**	0.649
Misogyny$_{\text{all}}$	0.671	0.640	0.655	0.651	0.703	0.697	**0.676**	0.677	0.583	0.646	0.557	0.630

Table 2.7 – Comparison with related work in terms of accuracy.

SYSTEM	**Misogyny**$_{\text{Evalita}}$	**Misogyny**$_{\text{IberEval}}$
	A	A
Competition Baseline	0.605	0.783
Competition Best System	0.704	**0.913**
Best $Top^G(\textbf{Founta}) \longrightarrow Top^S$ (**ELMo/BERT**)	0.597	0.697
Best $Top^G(\textbf{Davidson}) \longrightarrow Top^S$ (**BERT/ELMo**)	0.610	0.658
Best $Top^S(all) \longrightarrow Top^S$ (**BERT**)	**0.706**	0.848

2.3 Multi-target Hate Speech Detection

2.3.1 Methodology

Now that we have established that the topic-generic datasets are not adequate for capturing specific instances of HS using state of the art HS detection models, the next step is to evaluate how topically focused datasets can be used to detect multi-target HS. This implies answering two main research questions:

- *Is combining topic-specific datasets better for predicting HS towards a given seen topic/target?*

- *What happens when the models are tested on a topic-specific dataset where the topic and/or the target are unseen?*

Let T be either a topic (Top) or a target (Tag). We propose the following configurations:

- $T^S \longrightarrow T^S_{seen}$: We model the task as a multi-label classification problem with two sub-configurations:

  (a) $Top^S \longrightarrow Top^S_{seen}$: Detect the hatefulness of a given tweet and the topic to which the HS belongs. Each tweet is thus classified into eight different classes, representing the combination of the four topics (racism, sexism, misogyny, xenophobia) and two HS classes (hate speech vs. non hate speech). As in the previous experiments (cf. Section 2.2.1), we combine all the training sets of the topic-specific datasets for training. Then, all the models are tested on the test set of each topic-specific datasets.

  (b) $Tag^S \longrightarrow Tag^S_{seen}$: It is similar to (a), except that it concerns the multi-label classification of targets. Therefore, we merge topic-specific train and test sets that share the same target (i.e., *women*: $Sexism_{Waseem}$ and $Misogyny_{all}$ and *ethnicity*: $Racism_{Waseem}$ and $Xenophobia_{HatEval}$).

- $T^S \longrightarrow T^S_{unseen}$: We model the task as a binary classification task to predict the topic/-target not previously seen during training time. We also design two experiments here:

  (c) $Top^S \longrightarrow Top^S_{unseen}$: It uses three out of the four topic datasets for training and the remaining topic dataset for testing (i.e., the dataset left out at training time).

For example, to detect the hatefulness of misogynistic messages, we train on the following topics: racism (Racism$_{\text{Waseem}}$), sexism (Sexism$_{\text{Waseem}}$) and xenophobia (Xenophobia$_{\text{HatEval}}$), then we test on the misogyny topic (i.e., comprising AMI corpora and Misogyny$_{\text{HatEval}}$).

(d) $Tag^S \longrightarrow Tag^S_{unseen}$: It is similar to (c), except that it concerns targets. For example, to detect the hateful messages that target women, we train by using the datasets related to the target race (i.e., Racism$_{\text{Waseem}}$ and Xenophobia$_{\text{HatEval}}$) and test on the four datasets related to the target *women* (i.e., Sexism$_{\text{Waseem}}$, the two AMI corpora and Misogyny$_{\text{HatEval}}$).

Both $T^S \longrightarrow T^S_{seen}$ (multi-label classification) and $T^S \longrightarrow T^S_{unseen}$ (binary classification) rely on the six models presented in Section 2.2.1 (i.e., **LSVC**, **LSTM**, **LSTM$_{\text{FastText}}$**, **CNN$_{\text{FastText}}$**, **ELMo**, and **BERT**). In addition, for $T^S \longrightarrow T^S_{seen}$ we propose a multi-task setting that consists of two classifiers that are trained jointly by multi-task objectives. The first classifier predicts whether the tweet is hateful or not (0 and 1), while the second one the topic of HS (racism (0), sexism (1), misogyny (2), and xenophobia (3)). The final label prediction is broken down into eight classes (cf. Table 2.8). The multi-task systems are compared to the previous six models used here as strong baselines.

Table 2.8 – Label combination in multi-task setting.

Target Label	Hate Speech Label	Final Label
Racism (0)	Not Hate Speech (0)	Not Racism (0)
	Hate Speech (1)	Racism (1)
Sexism (1)	Not Hate Speech (0)	Not Sexism (2)
	Hate Speech (1)	Sexism (3)
Misogyny (2)	Not Hate Speech (0)	Not Misogyny (4)
	Hate Speech (1)	Misogyny (5)
Xenophobia (3)	Not Hate Speech (0)	Not Hate Speech towards immigrants (6)
	Hate Speech (1)	Hate Speech towards immigrants (7)

MTL has already been successfully applied in cross-domain aspect-based sentiment analysis (cf. Sections 1.1 and 1.2 for related work in the field) and is used here for the first

time in an HS detection task, making a parallel between the sentiment domain (e.g., restaurant, book, hotel, etc.) and the topic/target of HS. Indeed, the main problem in sentiment analysis is the big performance decline in the out-domain setting (when a system is trained and tested with different dataset domains) compared to the in-domain setting (when a system is trained and tested on dataset within the same domain). Similar challenges also arise in the abusive language detection task, where a system is struggling to obtain a robust performance when trained and tested with different datasets. These usually have different focuses on the phenomena they want to capture.

2.3.2   Models

We experiment with state of the art models (i.e., **LSVC**, **LSTM**, **LSTM$_{\text{FastText}}$**, **CNN$_{\text{FastText}}$**, **ELMo**, and **BERT**, as described in Section 2.2.2) and extend them with a multi-task architecture, as described below:

   –LSTM$_{\text{multi-task}}$. First, we investigate successful approaches in multi-domain sentiment analysis, a research area that is more mature in dealing with multi-domain classification. For example, (Liu et al., 2018) used BiLSTM networks with adversarial training (Ganin and Lempitsky, 2015; Goodfellow et al., 2014) for learning general representation from all domains data. Peng et al. (2018) proposed a co-training approach for jointly learning the representation from both domain-invariant and domain-specific representations, while Zhang et al. (2019); Cai and Wan (2019) adopted a MTL approach. Among existing models, we decided to re-implement the system proposed in Cai and Wan (2019), as it has been shown to outperform existing models in one of the most used multi-domain sentiment classification benchmark dataset (Liu et al., 2017). This system consists of two BiLSTM classifiers, each of them classifying the domain (domain classifier) and the sentiment (sentiment classifier) of the tweets at the same time, with the loss of both tasks being added up. The output of the BiLSTM domain classifier is concatenated to the word embedding layer of the sentiment classifier to acquire a domain-aware representation. Then, the output of average pooling (after BiLSTMs) of the domain classifier is also concatenated to the sentiment classifier to obtain domain-aware attention.

   We extend the architecture proposed in (Cai and Wan, 2019). The first BiLSTM predicts whether a given tweet is hateful or not, while the second one predicts the topic/target of

HS. In this way, we obtain both topic/target-aware representation and topic/target-aware attention when predicting whether the tweet is hateful or not. For experiments, we fine-tune this model by varying the number of epochs (1-15) and batch-sizes (16, 32, 64, and 128) while keeping the same configurations as in (Cai and Wan, 2019). The model input is either embeddings randomly initialized (**LSTM$_{\text{multi-task}}$**) or FastText pre-trained embeddings, (**LSTM$_{\text{multi-task (FastText)}}$**)[96].

–ELMo$_{\text{multi-task}}$. We also modify our **ELMo** system (cf. Section 2.2.1) in order to be able to use it in multi-task setting. Therefore, we built two ELMo-based architectures to predict the hatefulness and topic/target of tweets. Each architecture starts with the ELMo embedding layer, followed by a dense layer with a ReLU activation function, before being passed into another dense layer with a sigmoid activation function to produce the final prediction. Since ELMo embeddings are not trainable, we could not get the topic/target-aware representation as in the previous BiLSTMs model. We can only transfer knowledge by concatenating the output of the first dense layer of the topic/target classifier to the dense layer of the hateful classifier. In this way, we expect to get meaningful information about the topic/target to classify the hatefulness of tweets. Again, we only tune the systems by optimizing the number of epochs and batch-sizes.

–BERT$_{\text{multi-task}}$. This model is similar to (Liu et al., 2019), where all tasks share and update the same low layers (i.e., **BERT** layers), except for the task-specific classification layer. In this architecture, after transferring the text to contextual embeddings in the shared layers and retrieving the first token hidden state of the shared **BERT** model, we apply a dropout of 0.1 and connect it to two different layers (corresponding to the two classification tasks: topic/target and hatefulness). To preserve individual task-specific loss functions and to perform training at the same time, we defined the losses for the two tasks separately and optimized them jointly (by backpropagating their sum through the model). This model was trained for three epochs with a learning rate of 2e-5 and AdamW optimizer.

---

[96]GloVe used in the original paper gives lower results.

2.3.3 Results

2.3.3.1 Results for the $T^S \longrightarrow T^S_{seen}$ Configurations

Table 2.9 and Table 2.10 present the results obtained in the $Top^S \longrightarrow Top^S_{seen}$ configuration in which the testing topic was previously seen during training. Table 2.9 presents the baseline results while Table 2.10 the multi-task results. We can observe that multi-task models are the best, outperforming all the baselines, the best systems being **LSTM$_{\text{multi-task (FastText)}}$** and **BERT$_{\text{multi-task}}$**. The results obtained on the `Waseem` dataset surpass all the others, which could be a consequence of the higher number of instances in this particular dataset when compared to the others. Overall, the best performance for the multi-topic HS detection task is achieved by **BERT$_{\text{multi-task}}$**, which attains the best result in eight out of nine test datasets.

Table 2.9 – Baseline results for $Top^S \longrightarrow Top^S_{seen}$.

DATASET	LSVC				LSTM				LSTM$_{\text{FastText}}$			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.701	0.844	0.766	0.610	0.841	0.827	0.834	0.856	0.816	0.856	**0.835**	0.855
Sexism$_{\text{Waseem}}$	0.694	0.852	0.765	0.545	0.781	0.859	0.818	0.827	0.782	0.869	**0.826**	0.832
Xenophobia$_{\text{HatEval}}$	0.474	0.544	0.507	0.404	0.459	0.601	0.521	0.387	0.496	0.651	0.563	0.421
Misogyny$_{\text{Evalita}}$	0.614	0.653	0.633	0.612	0.598	0.657	0.626	0.599	0.609	0.661	0.634	0.604
Misogyny$_{\text{IberEval}}$	0.642	0.841	0.728	0.643	0.504	0.716	0.592	0.502	0.607	0.782	0.684	0.582
Misogyny$_{\text{HatEval}}$	0.518	0.578	0.546	0.452	0.595	0.644	0.618	0.551	0.536	0.662	0.592	0.468
Misogyny$_{\text{all}}$	0.576	0.638	0.605	0.545	0.574	0.638	0.604	0.555	0.573	0.645	0.607	0.536

DATASET	CNN$_{\text{FastText}}$				BERT				ELMo			
	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A
Racism$_{\text{Waseem}}$	0.703	0.754	0.727	0.855	0.847	0.597	0.701	0.791	0.819	0.840	0.829	0.859
Sexism$_{\text{Waseem}}$	0.841	0.810	0.825	0.826	0.876	0.666	0.757	0.812	0.675	0.854	0.754	0.788
Xenophobia$_{\text{HatEval}}$	0.532	0.491	0.510	0.422	0.667	0.527	**0.588**	0.516	0.356	0.567	0.437	0.312
Misogyny$_{\text{Evalita}}$	0.653	0.586	0.618	0.595	0.723	0.672	**0.697**	0.670	0.427	0.650	0.516	0.431
Misogyny$_{\text{IberEval}}$	0.865	0.725	0.788	0.724	0.857	0.783	**0.818**	0.780	0.484	0.738	0.585	0.531
Misogyny$_{\text{HatEval}}$	0.602	0.563	0.582	0.505	0.681	0.581	**0.627**	0.632	0.529	0.624	0.573	0.488
Misogyny$_{\text{all}}$	0.656	0.612	0.633	0.643	0.702	0.654	**0.677**	0.657	0.488	0.634	0.551	0.479

Table 2.11 presents the results obtained for the $Tag^S \longrightarrow Tag^S_{seen}$ experiments in which the testing target was previously seen during training. The best result for the target women was obtained by **CNN$_{\text{FastText}}$**, while for the target race **LSTM$_{\text{multi-task (FastText)}}$** outperformed all the other models. Our results confirm our assumption that the multi-task approach is

Table 2.10 – Multi-task results for $Top^S \longrightarrow Top^S_{seen}$.

DATASET	LSTM$_{\textbf{multi-task}}$				LSTM$_{\textbf{multi-task (FastText)}}$			
	P	R	F_1	A	P	R	F_1	A
Racism$_{\texttt{Waseem}}$	0.787	0.851	0.818	0.877	0.839	0.811	0.825	0.828
Sexism$_{\texttt{Waseem}}$	0.774	0.867	**0.818**	0.848	0.763	0.842	0.801	0.797
Xenophobia$_{\texttt{HatEval}}$	0.475	0.534	0.503	0.407	0.495	0.621	0.551	0.422
Misogyny$_{\texttt{Evalita}}$	0.573	0.639	0.604	0.560	0.621	0.687	**0.653**	0.605
Misogyny$_{\texttt{IberEval}}$	0.556	0.774	**0.647**	0.542	0.644	0.792	**0.710**	0.621
Misogyny$_{\texttt{HatEval}}$	0.551	0.650	0.597	0.489	0.554	0.682	**0.612**	0.489
Misogyny$_{\texttt{all}}$	0.560	0.651	0.602	0.523	0.597	0.684	**0.637**	0.555

DATASET	ELMo$_{\textbf{multi-task}}$				BERT$_{\textbf{multi-task}}$			
	P	R	F_1	A	P	R	F_1	A
Racism$_{\texttt{Waseem}}$	0.677	0.862	0.758	0.827	0.835	0.667	**0.742**	0.865
Sexism$_{\texttt{Waseem}}$	0.599	0.862	0.707	0.764	0.870	0.703	**0.777**	0.874
Xenophobia$_{\texttt{HatEval}}$	0.356	0.617	**0.451**	0.340	0.650	0.585	**0.616**	0.513
Misogyny$_{\texttt{Evalita}}$	0.457	0.594	**0.517**	0.472	0.725	0.685	**0.704**	0.684
Misogyny$_{\texttt{IberEval}}$	0.479	0.714	0.573	0.541	0.865	0.774	0.817	0.774
Misogyny$_{\texttt{HatEval}}$	0.580	0.615	**0.597**	0.580	0.701	0.598	**0.646**	0.642
Misogyny$_{\texttt{all}}$	0.520	0.613	**0.563**	0.538	0.721	0.648	**0.682**	0.683

capable of a robust performance in a multi-topic experiment, proving its ability in transferring knowledge between different topics, as reported in previous cross-domain sentiment analysis studies.

2.3.3.2 Results for the $T^S \longrightarrow T^S_{unseen}$ Configuration

We begin by presenting the results in the $Top^S \longrightarrow Top^S_{unseen}$ experiments in which the testing topic was unseen during training. As shown in Table 2.12, we observe that in the absence of data annotated for a specific type of HS, one can use (already existing) annotated data for different kinds of HS.

As this experiment is cast as a binary classification task, we compare the results with the ones presented in Table 2.6 that concern $Top^S \longrightarrow Top^S$ when training on Waseem, HatEval

Table 2.11 – Baselines and multi-task results for $Tag^S \longrightarrow Tag^S_{seen}$.

SYSTEM	WOMEN				ETHNICITY			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.530	0.704	0.605	0.431	0.548	0.632	0.587	0.457
LSTM	0.678	0.713	0.695	0.711	0.650	0.608	0.628	0.728
LSTM$_{\text{FastText}}$	0.677	0.721	0.698	0.707	0.656	0.621	0.638	0.737
CNN$_{\text{FastText}}$	0.732	0.716	**0.724**	0.731	0.580	0.435	0.497	0.613
BERT	0.772	0.660	0.712	0.681	0.652	0.638	0.645	0.651
ELMo	0.582	0.654	0.616	0.657	0.588	0.656	0.620	0.710
LSTM$_{\text{multi-task}}$	0.667	0.719	0.692	0.710	0.631	0.649	0.640	0.774
LSTM$_{\text{multi-task (FastText)}}$	0.680	0.725	0.701	0.694	0.667	0.673	**0.670**	0.717
ELMo$_{\text{multi-task}}$	0.559	0.678	0.613	0.668	0.516	0.694	0.592	0.694
BERT$_{\text{multi-task}}$	0.772	0.671	0.718	0.692	0.649	0.642	0.645	0.657

and AMI train sets and where topics are seen in the test sets. We noticed that **CNN$_{\text{FastText}}$** was able to achieve a similar performance for the topic misogyny (0.655 in both $Top^S \longrightarrow Top^S_{unseen}$ and $Top^S \longrightarrow Top^S$), improving almost 2% for the target xenophobia (moving from 0.578 in $Top^S \longrightarrow Top^S$ with **BERT** to 0.595 in terms of F_1). However, lower results were obtained for the Waseem dataset, where the drop in terms of F_1 is between 15% and 20%. The overall results also show that **CNN$_{\text{FastText}}$** was the best in predicting unseen topics for the four topics we experiment on. By capturing different scales of correlation between words (i.e., bigrams, trigrams, and unigrams), the CNN model can detect different patterns in the sentence, regardless of their position (Shirbandi and Moradi, 2019).

Finally, Table 2.13 presents the results obtained when the models are trained on all the available data belonging to a target and tested on all the available data belonging to a different target (i.e., $Tag^S \longrightarrow Tag^S_{unseen}$). In line with the previous experiment, the best results were achieved by **CNN$_{\text{FastText}}$**. In order to better interpret these results, we conducted another experiment in which a model is trained only on data belonging to a target and tested on data belonging to a topical focus on a different target (e.g., training on the target women and testing on the topic xenophobia belonging to the target race). When comparing these results (cf. Table 2.14) with the ones presented in Table 2.12, one can observe the importance for the system of having learned some information regarding the target, even if the data

belongs to a different topical focus. In the absence of such information, a drop of anywhere in between 1% and 12% can be observed for the best-performing models.

Table 2.12 – Results for $Top^S \longrightarrow Top^S_{unseen}$.

SYSTEM	Racism$_{\mathtt{Waseem}}$				Sexism$_{\mathtt{Waseem}}$			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.458	0.490	0.474	0.820	0.491	0.498	0.494	0.761
LSTM	0.481	0.462	0.471	0.790	0.525	0.543	0.534	0.731
LSTM$_{\text{FastText}}$	0.489	0.460	0.473	0.787	0.507	0.518	0.513	0.740
ELMo	0.492	0.489	0.491	0.769	0.502	0.506	0.504	0.745
CNN$_{\text{FastText}}$	0.742	0.506	**0.602**	0.853	0.882	0.545	**0.674**	0.798
BERT	0.507	0.500	0.504	0.842	0.693	0.537	0.605	0.785

SYSTEM	Misogyny$_{\mathtt{all}}$				Xenophobia$_{\mathtt{HatEval}}$			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.580	0.581	0.581	0.577	0.629	0.536	0.579	0.603
LSTM	0.562	0.563	0.562	0.545	0.541	0.557	0.549	0.583
LSTM$_{\text{FastText}}$	0.564	0.572	0.568	0.535	0.508	0.560	0.535	0.583
ELMo	0.510	0.556	0.532	0.583	0.511	0.542	0.526	0.573
CNN$_{\text{FastText}}$	0.659	0.652	**0.655**	0.638	0.598	0.593	**0.595**	0.617
BERT	0.634	0.628	0.631	0.639	0.617	0.531	0.571	0.614

Table 2.13 – Results for $Tag^S \longrightarrow Tag^S_{unseen}$.

SYSTEM	WOMEN				ETHNICITY			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.399	0.491	0.440	0.676	0.438	0.491	0.463	0.753
LSTM	0.423	0.489	0.453	0.670	0.500	0.500	0.500	0.744
LSTM$_{\text{FastText}}$	0.445	0.487	0.465	0.659	0.476	0.489	0.482	0.722
ELMo	0.420	0.486	0.451	0.665	0.437	0.486	0.460	0.743
CNN$_{\text{FastText}}$	0.579	0.513	**0.544**	0.660	0.665	0.543	**0.598**	0.773
BERT	0.514	0.501	0.507	0.656	0.596	0.506	0.548	0.766

To conclude, the results confirm that the multi-task approach is able to achieve a robust performance, especially for the multi-topic HS detection task. These results are encouraging

Table 2.14 – Results for $Tag^S \longrightarrow Top^S_{unseen}$.

SYSTEM	Train on target: **women** and test on:							
	$Racism_{Waseem}$				$Xenophobia_{HatEval}$			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.446	0.488	0.466	0.819	0.494	0.499	0.497	0.577
LSTM	0.432	0.478	0.451	0.805	0.469	0.486	0.478	0.548
LSTM$_{FastText}$	0.434	0.475	0.451	0.798	0.480	0.492	0.486	0.557
ELMo	0.445	0.481	0.462	0.805	0.510	0.501	0.505	0.577
CNN$_{FastText}$	0.716	0.504	**0.592**	0.852	0.563	0.534	**0.548**	0.600
BERT	0.553	0.502	0.526	0.849	0.547	0.505	0.525	0.597

SYSTEM	Train on target: **ethnicity** and test on:							
	$Sexism_{Waseem}$				$Misogyny_{all}$			
	P	R	F_1	A	P	R	F_1	A
LSVC	0.391	0.486	0.431	0.756	0.498	0.470	0.484	0.569
LSTM	0.395	0.484	0.431	0.753	0.500	0.500	0.500	0.571
LSTM$_{FastText}$	0.403	0.479	0.431	0.741	0.474	0.495	0.484	0.560
ELMo	0.419	0.479	0.436	0.737	0.452	0.495	0.472	0.565
CNN$_{FastText}$	0.843	0.504	**0.631**	0.780	0.576	0.532	**0.553**	0.570
BERT	0.446	0.498	0.470	0.774	0.483	0.498	0.490	0.546

as they can constitute the first step towards targeted HS detection. This would be especially true for languages that lack annotated data for a particular target or in the aftermath of a triggering event.

2.4 Emotion-aware Multi-target Hate Speech Detection

2.4.1 Methodology

In this section, we focus on investigating the following questions:

- *To what extent does injecting domain-independent affective knowledge encoded in sentic computing resources and in semantically structured hate lexicons improve the performance for the two finer-grained tasks (i.e., detecting the hatefulness of a tweet and its topical focus)?*

- *Which emotional categories are the most productive?*

We experiment with several affective resources that have been proven useful for tasks related to sentiment analysis, including abusive language detection (cf. Section 1.3). Psychological studies suggest that abusive language is often deeply linked to the emotional state of the speaker, and that this is reflected in the affective characteristics of the haters' language. Our intuition, then, was that it would be reasonable to inject knowledge about emotions into our models as a domain-independent signal that might help to detect HS at a finer-grained level of granularity across different topical focuses and targets. In particular, we rely on:

- two concept-level resources from the sentic computing framework, where affective knowledge about basic and complex emotions is encoded, concerning different psychological models of emotions: SenticNet[97] (Cambria et al., 2018) and EmoSenticNet[98] (Poria et al., 2013), where emotional labels are related to the Plutchik (Plutchik, 1980) and Ekman's (Ekman, 1992) models of emotions.

- a hate lexicon (Hurtlex), where lexical information is structured in different categories depending on the nature of the hate expressed, to see whether this multifaceted affective information, specifically related to the hate domain, helps multi-topic and multi-target detection.

As discussed in Section 1, emotion features have already been used in several NLP tasks (e.g., sentiment analysis (Nissim and Patti, 2017) and figurative language detection (Sulis et al., 2016; Farías et al., 2016)). However, to the best of our knowledge, no one has investigated the impact of emotion features on HS detection. In particular, we make use of several affective resources (HurtLex and, for the first time, Sentic resources) and identify the emotion categories that are the most productive in detecting HS towards a given topic/target. To this end, we designed the following two experiments (we recall that T refers either to a topic (Top) or a target (Tag)):

- $(T^S \longrightarrow T^S_{seen})^{Hurt}$ and $(T^S \longrightarrow T^S_{seen})^{Sentic}$ where we respectively add features extracted from HurtLex and SenticNet (both from SenticNet and EmoSenticNet) on top of the models presented in Sections 2.2.1 and 2.3.1.

- $(Top^S \longrightarrow Top^S_{unseen})^{Sentic}$ where we explore the impact of general affect lexica on topically focused datasets.

---

[97] https://sentic.net
[98] https://www.gelbukh.com/emosenticnet/

The models developed for each experiment are detailed below.

2.4.2 Models

2.4.2.1 Sentic-based Models

SenticNet consists of a collection of commonly-used concepts with polarity (i.e., common-sense concepts with relatively strong positive or negative polarity), where each concept is associated with emotion categorization values expressed in terms of the Hourglass of emotions model (Cambria et al., 2012), which organizes and blends 24 emotional categories from Plutchik's model into four affective dimensions (*pleasantness*, *attention*, *sensitivity*, and *aptitude*). Each of these four dimensions is characterized by six *sentic levels* that measure the strength of an emotion. In this dissertation, we use SenticNet 5 that includes over 100,000 natural language concepts.

EmoSenticNet is another concept-based lexical resource and was automatically built by merging WordNet-Affect (Strapparava and Valitutti, 2004) and SenticNet, with the main aim of having a complete resource containing not only quantitative polarity scores associated with each SenticNet concept but also qualitative affective labels (Poria et al., 2013). In particular, it assigns WordNet-Affect emotion labels related to Ekman's six basic emotions (disgust, sadness, anger, joy, fear, and surprise) to SenticNet concepts. The whole list currently includes 13,189 annotated entries.

Several approaches for representing the affective information included in these two resources were tested by creating feature vectors composed of:

- 24 basic emotions extracted from SenticNet (six basic emotions for each of the four dimensions);

- 16 second level emotions extracted from SenticNet (these emotions are the result of combining the 'sentic levels' pairwise);

- all the affective information extracted from SenticNet (i.e., basic emotions and second level emotions);

- six emotions extracted from EmoSenticNet;

- emotions extracted from both SenticNet and EmoSenticNet;

- 24 basic emotions extracted from SenticNet only for the concepts present in Hurtlex.

All these additional features are concatenated with the previously described systems (cf. Section 2.2.1 and Section 2.3.1). The concatenation procedure depends on the architecture of the model, as follows:

- For the **LSTM**-based and **CNN** models, we concatenate the feature representation on the dense layer after the **LSTM/CNN** network.

- For the **ELMo** model, the feature representation is injected in the dense layer, after the **ELMo** embedding layer.

- After padding the feature vector to a size equal to the **BERT** model input size, these additional features are passed to a linear layer. The output of the features linear layer is then concatenated with the output of the **BERT** model, which will then be treated as input for the final linear layer.

2.4.2.2 Hurtlex-based Models

HurtLex is a multilingual hate word lexicon, which includes a wide inventory of about 1,000 hate words (originally compiled in a manual fashion for Italian by the linguist Tullio De Mauro (De Mauro, 2016)[99]) organized into 17 categories grouped in different macro-levels (Bassignana et al., 2018):

(a) *Negative stereotypes*: ethnic slurs (PS); locations and demonyms (RCI); professions and occupations (PA); physical disabilities and diversity (DDF); cognitive disabilities and diversity (DDP); moral and behavioral defects (DMC); and words related to social and economic disadvantage (IS).

(b) *Hate words and slurs beyond stereotypes*: plants (OR); animals (AN); male genitalia (ASM); female genitalia (ASF); words related to prostitution (PR); and words related to homosexuality (OM).

(c) *Other words and insults*: descriptive words with potential negative connotations (QAS); derogatory words (CDS); felonies and words related to crime and immoral behavior (RE); and words related to the seven deadly sins of Christian tradition (SVP).

---

[99]The list of hate words has been included in the Final Report (2017) issued by the "Joe Cox" Committee on intolerance, xenophobia, racism and hate, of the Italian Chamber of Deputies.

The lexicon has been translated into over 50 languages (English included) semi-automatically, by extracting all the senses of all the words from BabelNet (Navigli and Ponzetto, 2012). We were relying on the English version of Hurlex.[100] Out of the 17 categories, the following were selected for the two vulnerable categories targeted in the four specific manifestations of hate that we address in this dissertation.

- *misogyny* and *sexism*: male genitalia, female genitalia, words related to prostitution, physical disabilities and diversity, cognitive disabilities and diversity

- *xenophobia* and *racism*: animals, felonies and words related to crime and immoral behavior, ethnic slurs, moral and behavioral defects

We included this specific selection of the HurtLex categories features since a preliminary manual inspection of hateful contents targeting the two vulnerable groups suggests that different subsets of the HurtLex categories can be relevant in detecting any hateful speech against those targets. Moreover, concerning misogyny, we already have some positive experimental evidence about this selection from previous exploitation of Hurtlex for detecting HS targeting women (Pamungkas et al., 2018; Pamungkas and Patti, 2019).

We experimented with a number of representations of the selected features to train several classifiers:

- each of the selected Hurtlex categories is used as an independent feature (binary or frequency);

- all the selected Hurtlex categories (keeping in mind the choices made for the different targets) are combined in a single feature (i.e., at least one word from at least one of the categories is present) (binary or frequency).

2.4.3   Results

In the following, we present our results on injecting affective features in our models for all the configurations considered in Section 2.3 (i.e., $Top^S \longrightarrow Top^S_{seen}$, $Tag^S \longrightarrow Top^S_{seen}$ and $Top^S \longrightarrow Top^S_{unseen}$). In all the tables below, the models for which the results in terms of F_1 score outperformed the models without affective features are presented in bold. Moreover,

---

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

all the tables present an additional column Δ, to highlight the improvements due to the inclusion of the affective features based on Sentic computing resources and Hurtlex. (i.e., $\Delta = \textbf{Model} +_{\textbf{AffectiveFeatures}} \textbf{F1 - Model F1}$).

2.4.3.1   Results for Sentic computing emotion features

Table 2.15 presents the results obtained for the multi-label classification task by incorporating the sentic features (as described in the previous section and summarized below):[101]

(1) Basic emotions extracted from SenticNet.

(2) Basic emotions extracted from SenticNet only for the concepts present in Hurtlex.

(3) Second level emotions extracted from SenticNet.

(4) All SenticNet affective information (basic emotions + second level emotions).

(5) Emotions extracted from EmoSenticNet.

(6) Merging the affective information extracted from both SenticNet and EmoSenticNet.

As to the different representation strategies and combinations of sentic resources, we observed that the best results were obtained when integrating either the EmoSenticNet emotions, the first level emotions of SenticNet, or merging the SenticNet and EmoSenticNet emotions. In most cases, when including only the second level emotions of SenticNet, we see a drop in the performance of the model. The last results presented in Table 2.16 concern the $(Top^S \longrightarrow Top^S_{unseen})^{Sentic}$ setting in which we added sentic features for measuring the impact of general affective knowledge in predicting unseen topics. Three groups of features improve previous models for all the tested topics:

(1) Basic emotions extracted from SenticNet.

(2) Emotions extracted from EmoSenticNet.

(3) Merging the affective information extracted from both SenticNet and EmoSenticNet.

---

[101]We only report the results achieved by the multi-task models as they performed better (cf. Table 2.10).

Table 2.15 – Results for $(Top^S \longrightarrow Top^S_{seen})^{Sentic}$ and $(Tag^S \longrightarrow Tag^S_{seen})^{Sentic}$.

DATASET	LSTM$_{\text{multi-task + sentic}}$					LSTM$_{\text{multi-task (FastText) + sentic}}$				
	P	R	F_1	Δ	A	P	R	F_1	Δ	A
Racism$_{\text{Waseem}}$	0.776	0.855	0.814 (1)	- 0.004	0.865	0.834	0.838	**0.836** (5)	+ 0.011	0.855
Sexism$_{\text{Waseem}}$	0.771	0.882	**0.823** (6)	+ 0.005	0.851	0.792	0.854	**0.822** (5)	+ 0.015	0.832
Xenophobia$_{\text{HatEval}}$	0.459	0.500	0.479 (5)	- 0.024	0.398	0.504	0.575	0.537 (6)	- 0.014	0.435
Misogyny$_{\text{Evalita}}$	0.605	0.682	**0.641** (6)	+ 0.037	0.593	0.599	0.682	0.638 (5)	- 0.015	0.581
Misogyny$_{\text{IberEval}}$	0.573	0.752	**0.650** (6)	+ 0.003	0.562	0.639	0.815	**0.716** (5)	+ 0.006	0.615
Misogyny$_{\text{HatEval}}$	0.581	0.656	**0.616** (5)	+ 0.019	0.527	0.561	0.670	0.611 (6)	- 0.001	0.499
Misogyny$_{\text{all}}$	0.586	0.666	**0.624** (6)	+ 0.022	0.553	0.579	0.680	0.626 (5)	- 0.011	0.514
Racism + Xenophobia	0.616	0.620	0.618 (6)	- 0.022	0.741	0.659	0.656	0.658 (5)	- 0.012	0.734
Sexism + Misogyny	0.679	0.742	**0.709** (6)	+ 0.017	0.725	0.686	0.731	**0.707** (5)	+ 0.006	0.706
DATASET	ELMo$_{\text{multi-task + sentic}}$					BERT$_{\text{multi-task + sentic}}$				
	P	R	F_1	Δ	A	P	R	F_1	Δ	A
Racism$_{\text{Waseem}}$	0.702	0.851	**0.769** (5)	+ 0.011	0.830	0.855	0.666	**0.749** (3)	+ 0.007	0.863
Sexism$_{\text{Waseem}}$	0.623	0.867	**0.725** (1)	+ 0.018	0.789	0.870	0.717	**0.786** (6)	+ 0.009	0.798
Xenophobia$_{\text{HatEval}}$	0.377	0.604	**0.464** (1)	+ 0.013	0.365	0.617	0.532	0.571(1)	- 0.045	0.468
Misogyny$_{\text{Evalita}}$	0.458	0.611	**0.523** (6)	+ 0.006	0.471	0.714	0.664	0.688 (6)	- 0.016	0.661
Misogyny$_{\text{IberEval}}$	0.501	0.765	**0.605** (5)	+ 0.032	0.564	0.866	0.766	0.813 (1)	- 0.004	0.771
Misogyny$_{\text{HatEval}}$	0.576	0.613	0.594 (5)	- 0.003	0.575	0.705	0.592	0.644 (4)	- 0.002	0.633
Misogyny$_{\text{all}}$	0.522	0.612	**0.563** (5)	+ 0.001	0.539	0.705	0.652	0.677 (6)	- 0.005	0.624
Racism + Xenophobia	0.539	0.686	**0.604** (5)	+ 0.012	0.700	0.696	0.594	0.641 (3)	- 0.004	0.676
Sexism + Misogyny	0.572	0.676	**0.619** (5)	+ 0.006	0.671	0.765	0.685	**0.723** (1)	+ 0.005	0.668

2.4.3.2 Results for Hurtlex emotion features

Table 2.17 reports the results achieved by the best performing models for the $Top^S \longrightarrow Top^S_{seen}$ experiment (cf. Table 2.9) (i.e., **BERT$_{\text{multi-task}}$** and **CNN$_{\text{FastText}}$**) when incorporating the following most productive Hurtlex features:

(1) Hurtlex categories used as binary independent features.

(2) Hurtlex categories used as independent features (count).

(3) Single binary feature incorporating the selected Hurtlex categories.

(4) Single feature incorporating the selected Hurtlex categories (count).

In Table 2.17, the models for which the results in terms of F_1 surpassed the previ-

Table 2.16 – Results $(Top^S \longrightarrow Top^S_{unseen})^{Sentic}$.

SYSTEM	Racism_Waseem					Sexism_Waseem				
	P	R	F_1	Δ	A	P	R	F_1	Δ	A
LSTM_sentic	0.486	0.467	**0.476** (2)	+ 0.005	0.799	0.525	0.541	0.533 (2)	- 0.001	0.727
LSTM_FastText + sentic	0.495	0.482	**0.488** (3)	+ 0.004	0.818	0.510	0.530	**0.520** (2)	+ 0.007	0.748
ELMo_sentic	0.499	0.499	**0.499** (1)	+ 0.008	0.771	0.502	0.508	**0.505** (2)	+ 0.001	0.745
CNN_FastText + sentic	0.751	0.514	**0.610** (1)	+ 0.008	0.854	0.885	0.539	0.670 (2)	- 0.004	0.794

SYSTEM	Misogyny_all					Xenophobia_HatEval				
	P	R	F_1		A	P	R	F_1		A
LSTM_sentic	0.558	0.584	**0.571** (1)	+ 0.009	0.603	0.567	0.567	**0.567** (1)	+ 0.018	0.554
LSTM_FastText + sentic	0.542	0.569	0.555 (2)	- 0.013	0.592	0.593	0.592	**0.593** (1)	+ 0.060	0.588
ELMo_sentic	0.516	0.574	**0.543** (1)	+ 0.011	0.587	0.511	0.538	0.524 (2)	- 0.002	0.572
CNN_FastText + sentic	0.660	0.654	**0.657** (1)	+ 0.002	0.640	0.596	0.598	**0.597** (2)	+ 0.002	0.617

Table 2.17 – Results for $(Top^S \longrightarrow Top^S_{seen})^{Hurtlex}$ and $(Tag^S \longrightarrow Tag^S_{seen})^{Hurtlex}$.

DATASET	CNN_FastText + Hurtlex					BERT_multi-task + Hurtlex				
	P	R	F_1	Δ	A	P	R	F_1	Δ	A
Racism_Waseem	0.863	0.802	**0.831** (4)	+ 0.104	0.863	0.852	0.753	**0.799** (4)	+ 0.057	0.874
Sexism_Waseem	0.857	0.833	**0.845** (4)	+ 0.020	0.846	0.858	0.660	0.746 (2)	- 0.031	0.692
Xenophobia_HatEval	0.644	0.509	**0.569** (2)	+ 0.059	0.438	0.649	0.583	0.614 (2)	- 0.002	0.509
Misogyny_all	0.668	0.618	**0.642** (4)	+ 0.009	0.606	0.734	0.652	**0.690** (4)	+ 0.008	0.696
Misogyny_Evalita	0.656	0.615	**0.635** (3)	+ 0.017	0.611	0.738	0.695	**0.716** (4)	+ 0.012	0.693
Misogyny_IberEval	0.848	0.718	0.778 (1)	- 0.010	0.728	0.879	0.785	**0.829** (1)	+ 0.012	0.782
Misogyny_HatEval	0.658	0.642	**0.650** (4)	+ 0.068	0.616	0.705	0.613	**0.656** (4)	+ 0.010	0.659
Racism + Xenophobia	0.695	0.641	**0.667** (1)	+ 0.170	0.734	0.711	0.646	**0.677** (4)	+ 0.032	0.672
Sexism + Misogyny	0.741	0.701	0.720 (4)	- 0.004	0.740	0.756	0.653	0.701 (2)	- 0.017	0.643

ous models are presented in bold.[102] We observe that almost all the additional features were productive and outperformed the previous models. The improvements brought by **CNN_fastText+HurtLex** were higher compared to **BERT_multi-task + Hurtlex**: ranging from anywhere in between 1% and 17% (respectively Misogyny_all, and Racism + Xenophobia) vs. 1% and 5% (respectively Misogyny_HatEval and Racism_Waseem). The results of this experiment confirm our original assumption that including affective information and making use of specific lexicons leads to significant improvements in $Top^S \longrightarrow Top^S_{seen}$ experiments.

---

[102]An additional experiment consisted in combining the best Hurtlex feature representation with the best sentic feature representation for each of the datasets. However, the results did not improve.

2.5 Discussion

2.5.1 Error Analysis

In this section, we provide an error analysis focusing on the instances for which the predictions of our best performing model (**BERT$_{\text{multi-task}}$**) and manual annotation differ. We observe that misclassification is affected by several factors, including the absence of context within the utterance and the use of irony, stereotypes, and metaphors. Another relevant factor is the contextual similarities between the topical focuses in those datasets where the vulnerable category target is basically the same, e.g., *misogyny* and *sexism* (see (2.1) and (2.2) below[103]) and *xenophobia* and *racism* (see example (2.3)). In the examples provided below, we underlined some portions of the text in order to highlight the main source, in our view, of misclassification.

(2.1)   *I don't see why drinking and driving is such a big deal. <u>Letting women drive is just as hazardous!</u>* (gold label: *misogynistic*, predicted: *sexist*)

(2.2)   *<u>HYSTERICAL woman</u>. Not just woman. And, she didnt say he won.* (gold label: *misogynistic*, predicted: *sexist*)

(2.3)   *A piece at a time. Start by <u>outlawing new Mosques</u> and <u>stoping Muslim immigration.</u>* (gold label: *racist*, predicted: *xenophobia*)

Misogyny and sexism are closely-related notions, and the way in which they are related has been the object of investigation in philosophical literature in the last years (Manne, 2017; Richardson-Self, 2018). In order to take into account relatedness among those and other HS categories, we will consider, in the future, a strategy for putting fewer penalties for errors in predicting closely-related topics.

The use of irony is another important source of error. For example, in (2.4) the underlying stereotype, implying that there is no place for women as TV sportscasters, leads to the message being classified as *non-sexist*.

---

[103]Notice that in these two examples the users also rely on stereotypes: *'women can't drive'* and *'women are hysterical'*.

(2.4)   *They have to concentrate in the 2nd half of this half". <u>Wise words from our female commen-</u>*
        <u>tator."</u> (gold label: *sexist*, predicted: *non-sexist*)

In both (2.5) and (2.6) the users express their religious views on Islam. The model is
not able to correctly predict that these utterances are racist. Complex inference or logical
reasoning are needed to understand their point of views.

(2.5)   *The fact that I have a brain prevents me from accepting Islam.* (gold label: *racist*, predicted:
        non-racist)

(2.6)   *If you don't want to read a pedo, you have to stop reading the Quran.* (gold label: *racist*,
        predicted: *non-racist*)

Finally, although in (2.7) the user reports on a series of events, the model predicts the
message as conveying hate towards immigrants, most probably because of the use of the
word 'rapefugee'. This is a self-explanatory and derogatory term used for Muslim refugees
entering Europe.

(2.7)   *Westminster terror attack suspect named as '<u>Sudanese Rapefugee</u> who drove around London*
        looking for targets' before driving car into cyclists (gold label: *not-hateful against immi-*
        grants, predicted: *hateful against immigrants*)

2.5.2   Impact of Bias in Multi-target Hate Speech Detection

As observed in (Vidgen et al., 2019), HS datasets might contain systematic biases towards
certain topics and targets. In the context of automatic content moderation, the danger posed
by bias is considerable, as bias can unfairly penalize the groups that the automatic modera-
tion systems were designed to protect.

In line with previous works, we observed that bias has a strong impact on target-based
HS detection. Based on the results obtained in the cross-topic (i.e., $Top^S \longrightarrow Top^S_{unseen}$ con-
figuration, cf. Table 2.12), we noted a big performance drop in both Racism$_{Waseem}$ and
Sexism$_{Waseem}$ when compared to the $Top^S \longrightarrow Top^S_{seen}$ classification setting, as presented
in Table 2.6. One possible explanation for this drop is the bias problems characterizing the
Waseem dataset. As shown in (Wiegand et al., 2019), the Waseem dataset contains both au-
thor and topic bias, mostly because of their approach to data sampling. The methodology

adopted in (Wiegand et al., 2019) for studying this issue was also based on the experience of conducting cross-domain experiments (i.e., training on a dataset different from the one used for testing), in order to make the existing bias in abusive language datasets evident. Their results show that datasets that apply a biased sampling for corpus collection (instances matching query words that are likely to occur in abusive language) contain a high degree of implicit abuse. This might lead to a performance decrease due to the difficulty of learning lexical cues that convey implicit abuse. Wiegand et al. (2019) illustrated how datasets with a high degree of implicit abuse could be more affected by data bias. They observed that when query words and biased words (i.e., the words having the highest Pointwise Mutual Information towards abusive messages) are removed, the performance is much poorer than originally reported.

We draw the same observations in the $Top^G \longrightarrow Top^S$ experiments (cf. Section 2.2.3.1), where each model is trained on one of the two topic-generic datasets (i.e., `Founta` and `Davidson`) and tested on the topic-specific datasets. As previously mentioned, when comparing the results obtained in Table 2.4 and Table 2.5 with the ones presented in Table 2.6, the biggest performance drop is observed for the `Waseem` dataset. Again, the sampling biases characterizing that dataset may be a contributing factor.

Finally, let us mention the peculiarity of the results that we obtained for the `HatEval` dataset, especially the *xenophobia* portion; this is the only dataset where we observed a definite increase when training on topic-generic datasets, concerning the performances from training on topic-specific data. This counter-trend outcome needs to be further investigated. If possible, it should be investigated in relation to data sampling strategies adopted for `HatEval`, where training and test data were collected in different time frames (Florio et al., 2020).

Conclusion

In this part we investigate, for the first time, HS detection from a multi-target perspective, leveraging existing manually annotated datasets with different topical focuses (including sexism, misogyny, racism, and xenophobia) and different targets (gender, ethnicity, religion, and race). Several neural models have been proposed for transferring specific manifestations of hate across topics and targets, while also exploring multi-task approaches and additional affective knowledge. The main findings are:

Conclusion 1: Training on topic-generic datasets generally fails to account for the linguistic properties specific to a given topic. First, we experimented with several HS datasets with different topical focuses in a binary classification setting. This was done in order to capture general HS properties regardless of the dataset type (i.e., topic-generic or topic-specific). We investigated two experimental scenarios: the first one in which a system was trained on a topic-generic dataset and tested on topic-specific data; and a second one in which a given system was trained on a combination of several topic-specific datasets and tested on topic-specific data. The results show that by training a system on a combination of several (training sets from several) topic-specific datasets the system outperforms a system trained on a single topic-generic dataset. This finding partially confirms the assumption made by Swamy et al. (2019) according to which merging several abusive language datasets could assist in the detection of abusive language in non-generalizable (unseen) problems.

Conclusion 2: Combining topically focused datasets enabled the detection of multi-target HS even if the topic and/or target are unseen. Second, we proposed a classification setting which allows a given system to detect not only the hatefulness of a tweet, but also its topical focus in the context of a multi-label classification approach. Our findings show that a multi-task approach in which the model learns two or more tasks simultaneously, does better, in performance terms, than a single-task system, and the best model is the $\text{BERT}_{\text{multi-task}}$.

In the same way, we also proposed a cross-topic and cross-target experimental setting for the task of HS detection, where a system is trained on several sets of data with different topical focuses and targets and, then, tested on another dataset where its topical focus and target are unseen during training. Results show that $\mathbf{CNN_{FastText}}$ outperformed all the other systems in all the experimental scenarios. We believe that this is an important finding, which will pave the way for targeted HS manifestations, stimulated by a triggering event and which will solve the problem of a lack of annotated data for a particular topic/target.

Conclusion 3: Affective knowledge encoded in sentic computing resources and semantically structured hate lexicons improve finer-grained HS detection. Finally, when injecting domain-independent affective knowledge on top of deep learning architectures, multi-target HS detection improves in both settings where topic/target is seen and unseen at training time. The most useful group of features differ greatly on both topic/target and in terms of the model architectures. In most cases, the models incorporating EmoSenticNet emotions, the first level emotions of SenticNet, a blend of SenticNet and EmoSenticNet emotions or affective features based on Hurtlex, obtained the best results. However, when merging both the affective features based on Hurtlex and sentic computing resources, we observed a decline in the quality of the results.

Conclusion and Future Work

Main Contributions

This dissertation had a triple objective: (1) *develop models able to detect different types of sexism experiences in French tweets*, (2) investigate whether *gender stereotype detection can improve sexism detection*, and (3) investigate the problem of *transferring knowledge from different datasets with different topical focuses and targets*.

In order to achieve our objectives we started our study by providing an overview of the state of the art concerning the concept of hate speech. We examined the multifaceted aspects of hate speech from the perspective of both international organizations and scholars. This study has revealed three main shortcomings:

1. Despite the plethora of research dealing with hate speech detection, no work has addressed sexism detection in French. Moreover, the detection of sexism is casted as a binary classification problem (i.e., *sexist* vs. *non-sexist*) or a multi-label classification by identifying the type of sexist behaviours and no current method for discerning between reports of sexism and real sexist messages exists.

2. Although several studies suggest that there is a significant correlation between the usage of stereotypes and hate speech no one has empirically measured the impact of gender stereotype detection for sexist hate speech classification.

3. It has become difficult to measure the generalization power of systems designed for hate speech detection, more specifically, their ability to adapt their predictions in the presence of novel or different topics and targets.

In this dissertation, we bridge the gap by proposing solutions to each of the aforementioned shortcomings.

Firstly, we have presented the first corpus of French tweets annotated for sexism detection. The novelty of our approach is that not only tweets with a sexist content are labelled but the type of content is also characterized: the tweet is either directly addressed to a target, describes a target or reports/denounces sexism experienced by a woman. Several deep learning models have been proposed for distinguishing between reports/denunciations of sexism from sexist contents directed at or describing a target. These results are encouraging and demonstrate that detecting reporting assertions of sexism is possible.

Secondly, we proposed a new corpus for gender stereotype detection and the first approach for gender stereotype detection in tweets, as well as several deep learning strategies to inject appropriate knowledge about how stereotypes are expressed in language into sexist hate speech classification. An empirical evaluation of several multitask architectures shows the positive impact of multi-class gender stereotype detection on improving sexism detection.

Finally, we investigated, for the first time, hate speech detection from a multi-target perspective, leveraging existing manually annotated datasets with different topical focuses (including sexism, misogyny, racism, and xenophobia) and different targets (gender, ethnicity, religion, and race). Several neural models have been proposed for transferring specific manifestations of hate across topics and targets, while also exploring multi-task approaches and additional affective knowledge. Our results demonstrate that multi-task architectures are the best-performing models and that emotions encoded in sentic computing sources and hate lexicons are important features for multi-target hate speech detection. These results thereby show that multi-target hate speech detection from existing datasets is feasible. This is the first step towards hate speech detection for specific topics/targets when dedicated annotated data are missing.

This work offers several positive societal benefits. Hate speech is a well-known problem, and countering it via automatic methods can have a big impact on people's lives. This challenge is meant to spur innovation and encourage new developments for both hate speech/sexism detection and stereotype detection which can have positive effects for an extremely wide variety of tasks and applications.

Ethical Considerations

This dissertation does not contain any studies with human participants. In addition, the data that was used is composed of textual content from the public domain taken from datasets publicly available to the research community. These datasets also conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of the numeric identification number of each tweet.

For the corpora that has been developed within this study, the data have been annotated with respect to certain types of sexist or stereotypical language, however, we are not making

any claims about the authors of the tweets, neither share a large numbers of tweets from the same users. Additionally, if any of the users want to opt out from having their data being used for research, they can request that they be removed from the corpora by sending an email to the author of this dissertation.

Hate Speech Detection for Social Good

The corpora developed in this study are not intended to be used for collecting user information which could potentially raise ethical issues. Relying on models flagging posts as hateful/sexist/conveying stereotypes based on user statistics might be biased towards certain users which eventually could limit freedom of speech on the platform (Ullmann and Tomalin, 2020).

The desire to combat online hate speech cannot be done without automatic moderation tools, at the risk of increasing cases of algorithmic discrimination. However, the deployment of such algorithms should be done with care. Algorithmic discrimination results from the introduction of biases at the time of the design of the system. These biases consist in the transposition of general (often stereotyped) or statistical observations into systematic algorithmic conditions. The ethical aspects that should be considered include the choice of the performance metric to be optimized (Corbett-Davies et al., 2017), as well as the fairness of the model across a variety of conditions (e.g., different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, etc.)) (Mitchell et al., 2019). Finally, one should keep in mind that hate speech could be paired with promotions of positive online interactions (e.g., counter-speech (Chung et al., 2019), emphasis of moral ideals (Does et al., 2011)).

Future Work

In addition to the possible directions for future work discussed at the end of each part of the dissertation (cf. Conclusions 2.2.3, Conclusion 3.2.4 and Conclusion 2.5.2), below we provide some other interesting future directions.

Towards robust hate speech detection systems. Concerning the development of a robust system able to generalize hate speech towards different topical focuses and targets,

there is still room for improvement (Yin and Zubiaga, 2021). In further work, we want to explore other domain adaptation strategies, such as adversarial training. Adversarial training has been shown to be an effective method of learning representations in cross-domain classification in several tasks, including sentiment analysis and image classification (Ganin et al., 2016; Han et al., 2019; Xu et al., 2020).

Another path to explore is the **impact of bias in multi-target hate speech detection.** Bias in abusive language datasets is an open problem already observed by several previous studies (Park et al., 2018; Wiegand et al., 2019; Davidson et al., 2019; Mozafari et al., 2019), in which different variants of bias, such as topic bias, author bias, gender and racial bias were explored. As no further investigation on developing an approach in debiasing abusive language datasets has been offered, we also plan to examine this direction in the future in the interests of keeping hate speech detection fair and compliant.

Towards emotionally informed hate speech. Concerning the role of affective knowledge in detecting hateful contents, we observed that feeding our multi-label classification models with structured knowledge included in a hate lexicon like Hurtlex, where hate words are categorized according to different semantic areas, boosts the performance of the classifiers. This also suggests possible lines of future work.

According to the psychological literature hate words and, in particular, gendered and racial slurs, have evolved to the point that they are used, and perceived, to express negative emotions towards targets, therefore providing important information about the speaker's emotional state or his or her attitude toward the targeted entity (Hedger, 2013), even when they refer to descriptive qualities. We, therefore, think that it could be interesting to investigate the link between hateful language and the negative portions of the multifaceted emotion spectrum covered in sentic computing resources. In particular, we plan to test the effectiveness of the new version of the Hourglass model (Susanto et al., 2020), that provides a better understanding of neutral emotions and their association with other polar emotions and that includes some polar emotions that were previously missing (including self-conscious and moral emotions). The revisited Hourglass model calculates the polarity of a concept with higher accuracy. It also provides a new mechanism for classifying unknown concepts by finding the antithetic emotion of a missing concept and by flipping its

polarity. SenticNet 6 (Cambria et al., 2020) actually contains 200,000 words and multiword expressions. We believe it may prove a valuable resource for improving multi-topic and multi-target hate speech detection.

Towards multi-lingual hate speech detection. Though most of the available hate speech corpora are in English, the problem of hateful speech is not limited to one language. Given language diversity and the enormous amount of social media data produced in different regions of the world, the task of detecting hate speech from a multi-lingual perspective is also a significant challenge.[104] We, therefore, plan, in future, to explore the possibility of developing language-agnostic models capable of identifying hate speech in online communication.

Towards multi-modal hate speech detection. The models designed for detecting hate speech are trained using only the textual features; we did not account for pictures or videos included in the tweets. A new shared task in 2022, Multimedia Automatic Misogyny Identification (MAMI),[105] will be dedicated to the identification of misogynous memes, taking advantage of both text and images available as source of information. Multi-modality (Vijayaraghavan et al., 2021), as well as the detection of irony and sarcasm suggest possible lines of future work.

---

[104]See (Pamungkas et al., 2021b) for a survey regarding the available corpora and approaches employed in multilingual settings.

[105]https://competitions.codalab.org/competitions/34175

Appendix

Introduction

Contexte et Motivations

L'utilisation des réseaux sociaux est désormais très importante et les utilisateurs l'en servent non seulement pour avoir accès à l'information mais aussi pour partager leurs opinions et sentiments sur différents sujets. Etant pratique à utiliser, ces réseaux attirent des millions d'utilisateurs qui à travers le partage de contenu peuvent atteindre des personnes partout dans le monde. Ceci pourrait potentiellement faciliter une conversation positive et constructive entre eux. Cependant, à cause de cette grande ouverture sur le monde, certains utilisateurs initient des discours haineux et du harcèlement à travers leurs interactions (Burnap and Williams, 2015). Cela est dû notamment à la liberté d'expression et à l'anonymat accordés aux utilisateurs et au manque de réglementation efficace imposée par ces plateformes pour contrôler les contenus de leurs échanges.

Le discours de haine peut avoir des thématiques différentes : la misogynie, le sexisme, le racisme, la xénophobie, l'homophobie, l'islamophobie, etc. qui peuvent être désignés comme *thème*. Pour chaque thème, le contenu haineux est dirigé vers des *cibles* spécifiques qui représentent la communauté (individus ou groupes) recevant la haine. Par exemple, les personnes de peau noire ou blanche sont des cibles potentielles quand nous parlons de *racisme* (Silva et al., 2016), alors que les femmes sont ciblées quand nous parlons de *misogynie* ou de *sexisme* (Manne, 2017). Le discours haineux est, par définition, *orienté vers une cible* comme le montrent tweets suivants (Waseem and Hovy, 2016; Davidson et al., 2019; Basile et al., 2019), où les cibles sont soulignées. Ces exemples montrent également que différentes cibles impliquent différentes manières d'exprimer linguistiquement des contenus haineux tels que des références à des stéréotypes raciaux ou sexistes, l'utilisation d'émotions négatives et positives, des jurons et la présence d'autres phénomènes tels que l'envie et la

laideur.[106]

(2.8) *Women who are feminist are the ugly bitches who cant find a man for themselves*
(*Les femmes qui sont féministes sont des putes moches qui ne peuvent pas se trouver un homme*)

(2.9) *Islam is 1000 years of contributing nothing to mankind but murder and hatred.*
(*Depuis 1000 ans l'islam n'a contribué à rien pour l'humanité à part les meurtres et la haine*)

(2.10) *Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack ! Stop Allowing illegals to Abuse the Taxpayer #Immigration*
(*Les sans-papiers larguent leurs enfants ici pour qu'ils puissent obtenir des aides sociales et des écoles américaines Ils arnaquent les contribuables américains #RenvoyezLes ! Arrêtez d'autoriser les sans-papiers à abuser des contribuables #Immigration*)

(2.11) *Seattle Mayoral Election this year. A choice between a bunch of women, non-whites, and faggots/fag lovers.*
(*Election du maire de Seattle cette année. Un choix entre un tas de femmes, de non-blancs et de pédés*)

L'augmentation de la haine en ligne et les fausses informations ont créé un climat médiatique parfois hostile à ses utilisateurs. À ce titre, de nouvelles lois visant à mieux réglementer les entreprises propriétaires de réseaux sociaux numériques (par exemple, Google, Facebook, Twitter, etc.) ont été mises en place. Par ailleurs, ces réseaux sociaux ont également offert un espace où les femmes osent maintenant témoigner de leurs expériences (voir par exemple #*meToo* ou #*balanceTonPorc*). Nous pensons que dans ce cadre réglementaire, les approches standards de détection automatique des discours de haine sont malheureusement susceptibles de modérer ces dénonciations d'actes haineux.

Méthodologie et Contributions

Dans cette thèse nous proposons de relever les défis suivants :

---

[106]Voir (Mathew et al., 2018) pour une intéressante analyse lexicale, linguistique et psycho-linguistique des comptes haineux sur Twitter.

- (C1) Expérimenter le *développement de modèles capables de détecter différents types de sexisme dans les tweets français*.

- (C2) Étudier si *la détection des stéréotypes de genre peut améliorer la détection du sexisme*.

- (C3) Étudier *le problème du transfert de connaissances à partir de différents ensembles de données ayant des thèmes et des cibles différents*.

À cette fin, nous proposons trois contributions principales.

De la Classification Binaire du Sexisme à la Détection des Expériences de Sexisme

Pour autant que nous le sachions, la distinction entre les témoignages/dénonciations de sexisme et les messages sexistes *"réels"* n'a pas été abordée. Dans des travaux antérieurs, la détection du sexisme est présentée comme un problème de classification binaire (*sexiste* vs. *non sexiste*) ou une classification multi-classes en identifiant le type de comportements sexistes (Jha and Mamidi, 2017; Sharifirad et al., 2018; Fersini et al., 2018c; Karlekar and Bansal, 2018; Parikh et al., 2019). Nous estimons que présenter la détection du sexisme comme un problème de classification binaire n'est pas suffisant. En effet, il est important non seulement de pouvoir détecter automatiquement les messages à contenu sexiste, mais aussi de faire la distinction entre les *"vrais"* messages sexistes qui visent les femmes (cf. (2.12) et (2.13)) et les messages qui relatent des expériences de sexisme (cf. (2.14)). Alors que les messages peuvent être signalés et modérés dans le premier cas, comme le recommandent les lois européennes, les messages relatant des expériences de sexisme ne doivent pas être modérés.

(2.12)   *Le gardien de but n'a aucun mérite à arrêter ce tir de femme enceinte*

(2.13)   *Elle nage vite pour une femme*

(2.14)   *Il a dit: "qui va s'occuper de tes enfants quand tu seras à ACL ?"*

Nos contributions comprennent :

(1) Une nouvelle caractérisation du contenu sexiste mettant en valeur la relation contenu-force inspirée de la théorie des actes de langage (Austin, 1962) *et des études discursives sur le genre* (Lazar, 2007; Mills, 2008). En collaboration avec Alda Mari et Gloria Origgi de l'Institut Jean Nicod (Paris, France), nous avons créé une nouvelle caractérisation qui distingue différents

types de contenus sexistes en fonction de leur impact sur le destinataire (appelé *"force per-locutionnelle"*) : discours de haine sexiste *directement adressé* à une cible, *assertions descriptives* sexistes non adressées à la cible, ou *assertions rapportées* qui racontent une histoire de sexisme vécue par une femme. Notre hypothèse est que les actes indirects établissent un effet de distanciation avec le contenu rapporté et sont donc moins engageants de la part de l'allocutaire (Giannakidou and Mari, 2021).

(2) Le premier corpus français d'environ 12 000 tweets annotés pour la détection du sexisme selon cette nouvelle caractérisation et qui est librement disponible pour la communauté des chercheurs.[107] Le développement du guide d'annotation a été réalisé en collaboration avec Marlène Coulomb-Gully du Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales (LERASS, Toulouse, France). La caractérisation du contenu sexiste, les directives d'annotation et la description du corpus ont été publiées lors de la 12e Conférence sur les ressources linguistiques et l'évaluation (LREC) (Chiril et al., 2020a).

(3) Une étude préliminaire dans laquelle nous expérimentons le développement de modèles pour (i) la détection automatique des discours de haine envers *deux cibles différentes* (les immigrants et les femmes) et (ii) la détection automatique du sexisme *dans une perspective multilingue*, à savoir dans des tweets anglais et français. Nous proposons des modèles basés à la fois sur des caractéristiques dépendantes et indépendantes de la langue, et un modèle neuronal pour *déterminer dans quelle mesure la détection du discours haineux dépend de la cible*. Nous expérimentons également avec des « embeddings multilingues » en entraînant le réseau de neurones sur une langue et en le testant sur une autre afin de *mesurer à quel point les modèles proposés sont dépendants de la langue*. Ce travail a été publié à la conférence francophone TALN (Chiril et al., 2019a). Une partie de ce travail a également été publiée dans le cadre de la campagne d'évaluation HatEval (Basile et al., 2019) dans The 13th International Workshop on Semantic Evaluation (Chiril et al., 2019b).

(4) La première approche pour détecter les témoignages/dénonciations d'expériences de sexisme dans les tweets français. Nous nous appuyons sur des architectures d'apprentissage automatique entraînées à partir d'une combinaison de plusieurs représentations vectorielles de tweets : des « word embeddings » construits à partir de différentes sources (tweets, Wikipedia), complétés par des caractéristiques linguistiques et diverses stratégies de généralisation pour tenir compte des stéréotypes sexistes et de la manière dont les con-

---

[107]https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets

tenus sexistes sont exprimés linguistiquement. Ceci est une première étape avant de passer à un scénario où les stéréotypes de genre sont détectés automatiquement. Ce travail a été publié dans The 58th Annual Meeting of the Association for Computational Linguistics (ACL) (Chiril et al., 2020b).

Stéréotypes de Genre pour la Classification du Sexisme

Les stéréotypes de genre, définis par le Haut-Commissariat aux droits de l'homme comme *"une opinion ou une idée préconçue généralisée sur les attributs ou les caractéristiques que possèdent ou devraient être possédés par les femmes et les hommes ou sur les rôles que jouent ou devraient jouer les hommes et les femmes"* ont été largement étudiés en psychologie, dans les études de communication et en sciences sociales (Allport et al., 1954; Crawford et al., 2002; Beike and Sherman, 2014; Biscarrat et al., 2016). Il a été démontré que les stéréotypes représentent la cause principale des tensions entre les groupes (e.g., le sexisme, le racisme, etc.) (Fiske, 1998). Ainsi, l'exemple (2.15) (Fersini et al., 2018b) présente un stéréotype fondé sur le fait que la place de la femme est dans la cuisine, tandis que dans (2.16) (Basile et al., 2019), les immigrants sont perçus comme dépendant fortement du système d'aide sociale.

(2.15)   *Who makes the sandwiches at a feminist rally?*
         (Qui fait les sandwichs lors d'un rassemblement féministe ?)

(2.16)   *Illegals Cross Border Just in Time to Have #AnchorBabies for Welfare and Medicaid Ripping off U.S Taxpayers #RedNationRising #Trump #MAGA #SendThemBack ASAP*
         (Les Clandestins Traversent la Frontière Juste à Temps pour avoir des #AnchorBabies pour l'aide Sociale et Medicaid qui Dépouille les Contribuables Américains #RedNationRising #Trump #MAGA #SendThemBack ASAP)

Bien que plusieurs études suggèrent l'existence d'une corrélation significative entre l'utilisation de stéréotypes et les discours de haine (García-Sánchez et al., 2019; Francesconi et al., 2019), personne n'a mesuré empiriquement l'impact de la détection des stéréotypes de genre pour la classification des discours de haine sexiste. À cette fin, nous proposons :

(5) Le premier corpus annoté pour la détection des stéréotypes de genre. Ce corpus contient environ 9 200 tweets en français annotés selon différents types de stéréotypes et est librement disponible pour la communauté des chercheurs.[108]

---

[108] https://github.com/patriChiril/An-Annotated-Corpus-for-Gender-Stereotype-Det

(6) Un ensemble d'expériences destinées à détecter les stéréotypes de genre, puis à utiliser cette prédiction pour la détection du sexisme. Nous nous appuyons sur plusieurs architectures d'apprentissage profond exploitant diverses sources de connaissances linguistiques pour rendre compte des stéréotypes de genre et de la façon dont les contenus sexistes sont exprimés dans le langage.

Nos résultats suggèrent que la classification du sexisme peut bénéficier de la détection des stéréotypes de genre. Ce travail a été publié dans les Findings of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Chiril et al., 2021a).

De la Classification du Sexisme à la Détection des Discours Haineux

La plupart des systèmes existants conçus pour la détection des discours haineux ont deux caractéristiques communes. Premièrement, ils sont entrainés pour prédire la présence de discours haineux généraux, indépendants de la cible, sans tenir compte de *l'orientation thématique* ou de *la nature ciblée* du discours haineux. Deuxièmement, ces systèmes sont construits, optimisés et évalués sur la base d'un seul ensemble de données (qu'il soit *générique* ou *spécifique* à un thème). Il est donc difficile de mesurer le pouvoir de généralisation de ces systèmes et, plus précisément, leur capacité à adapter leurs prédictions en présence de thèmes et de cibles nouveaux ou différents (Yin and Zubiaga, 2021).

Pour relever ces derniers défis, nous proposons une nouvelle approche de détection des discours haineux multi-cibles pour traiter une cible nouvelle en exploitant les ensembles de données annotés manuellement déjà existants . Cela permettra à un modèle de transférer des connaissances de différents ensembles de données avec différents thèmes et cibles. Dans le contexte de la modération des contenus offensifs, l'identification du thème principal et de la communauté ciblée par les contenus haineux serait d'un grand intérêt car elle nous permettrait de détecter les discours haineux pour des thèmes/cibles spécifiques lorsque les données dédiées sont absentes.

Nos contributions portent sur trois points :

(7) Nous explorons la capacité des modèles de détection des discours haineux à saisir les propriétés communes à partir d'ensembles de données génériques sur les discours haineux et à transférer ces connaissances pour reconnaître les manifestations spécifiques de haine.

(8) Nous expérimentons le développement de modèles pour détecter à la fois les différents thèmes

ection-in-French-Tweets

(racisme, xénophobie, sexisme, misogynie) et les cibles (sexe, ethnicité) des discours de haine, au-delà de la classification binaire standard. Nous étudions (a) *comment détecter les discours haineux à un niveau de granularité plus fin* et (b) *comment transférer les connaissances entre différents types de discours haineux*. Nous nous appuyons sur plusieurs ensembles de données spécifiques à un thème. Nous développons également, en plus des modèles d'apprentissage profond conçus pour répondre au point *(7)*, une architecture multitâche qui s'est avérée assez efficace dans l'analyse des sentiments inter-domaines (Zhang et al., 2019; Cai and Wan, 2019).

(9) Nous étudions l'impact des ressources sémantiques affectives dans la détermination des manifestations spécifiques du discours de haine. Dans ce travail, nous voulons également explorer les caractéristiques affectives de la langue utilisée dans les discours de haine, dans la continuité des travaux très récents de Rajamanickam et al. (2020), qui suggèrent une relation forte entre le comportement abusif et l'état émotionnel du locuteur. Nous expérimentons trois ressources en tant que traits supplémentaires en complément de plusieurs architectures d'apprentissage : des ressources pour les sentiments et émotions (Cambria and Hussain, 2015) (SenticNet (Cambria et al., 2018), EmoSenticNet (Poria et al., 2013)) et des lexiques de haine sémantiquement structurés (HurtLex (Bassignana et al., 2018)). SenticNet n'a pas, à notre connaissance, été utilisé pour la détection de discours de haine. Pour chaque ressource, nous proposons une évaluation systématique des catégories émotionnelles qui sont les plus productives pour nos tâches.

Nos résultats montrent que la détection de discours haineux multi-cibles à partir d'ensembles de données existants est possible, ce qui constitue un premier pas vers la détection de discours haineux pour un thème/une cible spécifique lorsque des données annotées dédiées manquent. De plus, nous montrons que les connaissances affectives indépendantes du domaine, injectées dans nos modèles, permettent une détection plus fine des discours haineux.

Ce travail a été réalisé en collaboration avec Viviana Patti et Endang Wahyu Pamungkas de l'Université de Turin (Turin, Italie) et a été publié dans Cognitive Computation Journal (A Decade of Sentic Computing) (Chiril et al., 2021b).

Plan du Manuscrit

Le manuscrit est organisé en quatre parties qui peuvent être lues indépendamment les unes des autres, et chaque partie se concentre sur l'une des contributions présentées ci-dessus.

L'un des aspects critiques qui se pose lorsqu'on traite du discours de haine réside dans sa définition (bien qu'elle soit largement utilisée, il n'y a pas d'accord sur sa signification et sa portée). Dans la Partie I nous examinons le concept de discours de haine à travers les définitions employées par les organisations internationales ou les universitaires, en considérant les nombreux éléments qui s'entrecroisent. De plus, nous présentons également un aperçu des principaux travaux sur le discours de haine et la détection du sexisme.

Dans la Partie II nous détaillons les données, la caractérisation du contenu sexiste que nous proposons et le schéma d'annotation. Nous présentons ensuite les expériences réalisées pour détecter les contenus sexistes, ainsi que l'étude préliminaire dans laquelle nous cherchons à savoir si les modèles développés sont capables de détecter les discours de haine agnostiques ciblés.

Dans la Partie III nous nous concentrons sur la détection des discours de haine sexiste contre les femmes dans les tweets, en étudiant pour la première fois l'impact de la détection des stéréotypes de genre sur la classification du sexisme. Nous commençons cette partie en détaillant les données et le processus d'annotation du premier jeu de données annoté pour la détection des stéréotypes de genre, puis nous présentons les expériences qui ont été menées.

Dans la Partie IV nous abordons, pour la première fois, la détection des discours de haine d'un point de vue multi-cibles. Nous commençons cette partie en présentant une vue d'ensemble des principaux travaux sur la détection des discours haineux, puis nous présentons les expériences menées pour étudier le problème du transfert de connaissances à partir de différents corpus avec différents thèmes et cibles.

Nous finissons par présenter une synthèse de ce travail en soulignant ses contributions et ses limites. Nous soulignons également les questions éthiques, les applications potentielles, ainsi que nos perspectives pour les travaux futurs.

Conclusion

Contributions Principales

Cette thèse avait trois objectifs : (1) *développer des modèles capables de détecter différents types d'expériences de sexisme dans des tweets français*, (2) étudier si *la détection des stéréotypes de genre peut améliorer la détection du sexisme*, et (3) étudier le problème du *transfert de connaissances à partir de différents corpus avec des sujets et des cibles différents.*.

Afin d'atteindre nos objectifs, nous avons commencé notre étude en fournissant un aperçu de l'état de l'art concernant le concept de discours de haine. Nous avons examiné les multiples aspects du discours de haine du point de vue des organisations internationales et des universitaires. Cette étude a révélé trois lacunes principales :

1. Malgré la multitude de recherches portant sur la détection des discours de haine, aucun travail n'a été consacré à la détection du sexisme en français. De plus, la détection du sexisme est présentée comme un problème de classification binaire (i.e., *sexiste* vs. *non-sexiste*) ou une classification multi-classes en identifiant le type de comportements sexistes et aucune méthode actuelle ne permet de distinguer entre les rapports de sexisme et les vrais messages sexistes.

2. Bien que plusieurs études suggèrent qu'il existe une corrélation significative entre l'utilisation de stéréotypes et les discours de haine, personne n'a mesuré empiriquement l'impact de la détection des stéréotypes de genre pour la classification des discours haineux sexistes.

3. Il est devenu difficile de mesurer le pouvoir de généralisation des systèmes conçus pour la détection des discours haineux, plus précisément, leur capacité à adapter leurs prédictions en présence de sujets et de cibles nouveaux ou différents.

Dans cette thèse, nous comblons cette lacune en proposant des solutions à chacune des limitations mentionnées.

Tout d'abord, nous avons présenté le premier corpus de tweets français annoté pour la détection du sexisme. La nouveauté de notre approche est que non seulement les tweets ayant un contenu sexiste sont étiquetés mais aussi le type de contenu est également caractérisé : le tweet est soit directement adressé à une cible, soit décrit une cible, soit rapporte/dénonce le sexisme vécu par une femme. Plusieurs modèles d'apprentissage profond ont été proposés pour distinguer les rapports/dénonciations de sexisme des contenus sexistes adressés à une cible ou décrivant une cible. Ces résultats sont encourageants et démontrent que la détection des rapports/dénonciations de sexisme est possible.

Ensuite, nous avons proposé un nouveau corpus pour la détection des stéréotypes de genre et la première approche pour les détecter dans les tweets, ainsi que plusieurs stratégies d'apprentissage profond pour injecter des connaissances appropriées sur la façon dont les stéréotypes sont exprimés dans le langage dans la classification des discours de haine sexiste. Une évaluation empirique de plusieurs architectures multitâches montre l'impact positif de la détection multi-classe des stéréotypes de genre sur l'amélioration de la détection du sexisme.

Enfin, nous avons étudié, pour la première fois, la détection des discours haineux d'un point de vue multi-cibles, en tirant parti d'ensembles de données existants annotés manuellement avec différents sujets (notamment le sexisme, la misogynie, le racisme et la xénophobie) et différentes cibles (selon sexe, ethnicité, religion et race). Plusieurs modèles neuronaux ont été proposés pour transférer les manifestations spécifiques de haine entre les sujets et les cibles, tout en explorant également les approches multitâches et les connaissances affectives supplémentaires. Nos résultats démontrent que les architectures multitâches sont les modèles les plus performants et que les émotions encodées dans les sources informatiques sémantiques et les lexiques de haine sont des caractéristiques importantes pour la détection des discours de haine multi-cibles. Ainsi, la détection de discours de haine multi-cibles à partir d'ensembles de données existants est réalisable. Il s'agit d'une première étape vers la détection de discours de haine pour des sujets/cibles spécifiques lorsque des données annotées spécifiques sont absentes.

Ce travail présente plusieurs avantages pour la société. Le discours de haine est un problème bien connu, et le contrer par des méthodes automatiques peut avoir un impact impor-

tant sur la vie des gens. Ce défi a pour but de stimuler l'innovation et d'encourager de nouveaux développements tant pour la détection des discours haineux/sexistes que pour la détection des stéréotypes, ce qui peut avoir des effets positifs pour une très grande variété de tâches et d'applications.

Considérations Éthiques

Cette thèse ne contient pas d'études avec des participants humains. En outre, les données utilisées sont composées de contenu textuel du domaine public provenant d'ensembles de données publiquement disponibles pour la communauté des chercheurs. Ces ensembles de données sont également conformes à l'accord et à la politique du développeur Twitter, qui autorise la distribution illimitée du numéro d'identification numérique de chaque tweet.

Pour les corpus qui ont été développés dans le cadre de cette étude, les données ont été annotées par rapport à certains types de langage sexiste ou stéréotypé, cependant, nous ne faisons aucune déclaration sur les auteurs des tweets, ni ne partageons un grand nombre de tweets provenant des mêmes utilisateurs. En outre, si l'un des utilisateurs souhaite que ses données ne soient pas utilisées à des fins de recherche, il peut demander à être retiré du corpus en envoyant un courriel à l'auteur de cette thèse.

Détection des Discours Haineux pour le Bien Social

Les corpus développés dans cette étude ne sont pas destinés à être utilisés pour collecter des informations sur les utilisateurs, ce qui pourrait soulever des problèmes éthiques. Le fait de s'appuyer sur des modèles qui signalent les messages comme étant haineux, sexistes ou véhiculant des stéréotypes en se basant sur les statistiques des utilisateurs pourrait être biaisé en faveur de certains utilisateurs, ce qui pourrait éventuellement limiter la liberté d'expression sur la plateforme (Ullmann and Tomalin, 2020).

La volonté de lutter contre les discours de haine en ligne ne peut se faire sans outils de modération automatique, au risque d'augmenter les cas de discrimination algorithmique. Cependant, le déploiement de tels algorithmes doit être fait avec précaution. La discrimination algorithmique résulte de l'introduction de biais au moment de la conception du système. Ces biais consistent en la transposition d'observations générales (souvent stéréotypées) ou statistiques en conditions algorithmiques systématiques. Les aspects éthiques à

prendre en compte incluent le choix de la métrique de performance à optimiser (Corbett-Davies et al., 2017), ainsi que l'équité du modèle dans diverses conditions (par exemple, différents groupes culturels, démographiques ou phénotypiques, race, emplacement géographique, sexe, etc.) (Mitchell et al., 2019). Enfin, il faut garder à l'esprit que les discours de haine pourraient être associés à des promotions d'interactions en ligne positives (par exemple, contre-discours (Chung et al., 2019), mise en avant d'idéaux moraux (Does et al., 2011)).

Travaux Futurs

En plus des orientations possibles pour les travaux futurs discutées à la fin de chaque partie de la thèse (cf. Conclusion 2.2.3, Conclusion 3.2.4 et Conclusion 2.5.2), nous présentons ci-dessous quelques autres orientations futures intéressantes.

Vers des systèmes robustes de détection des discours de haine. En ce qui concerne le développement d'un système robuste capable de généraliser les discours de haine vers différents sujets et cibles, il y a encore place à l'amélioration (Yin and Zubiaga, 2021). Dans le cadre de travaux ultérieurs, nous souhaitons explorer d'autres stratégies d'adaptation au domaine, telles que l'entraînement contradictoire. L'entraînement contradictoire s'est avéré être une méthode efficace d'apprentissage des représentations dans la classification inter-domaines dans plusieurs tâches, notamment l'analyse des sentiments et la classification des images (Ganin et al., 2016; Han et al., 2019; Xu et al., 2020).

Une autre voie à explorer est **l'impact des biais dans la détection des discours de haine multi-cibles.** Le biais dans les ensembles de données de langage abusif est un problème ouvert déjà observé par plusieurs études antérieures (Wiegand et al., 2019; Davidson et al., 2019; Park et al., 2018; Mozafari et al., 2019), dans lesquelles différentes variantes de biais, telles que le biais de sujet, le biais d'auteur, le biais de genre et le biais racial ont été explorées. Étant donné qu'aucune autre enquête sur le développement d'une approche de débiaisage des ensembles de données de langage abusif n'a été proposée, nous prévoyons également d'examiner cette direction à l'avenir dans l'intérêt de garder la détection des discours de haine juste et conforme.

Vers un discours haineux émotionnellement informé. En ce qui concerne le rôle des connaissances affectives dans la détection des contenus haineux, nous avons observé que l'alimentation de nos modèles de classification multi-classes avec des connaissances structurées incluses dans un lexique de la haine tel que Hurtlex, où les mots haineux sont classés en fonction de différents domaines sémantiques, améliore les performances des classificateurs. Cela suggère également des pistes de travail pour l'avenir.

Selon la littérature psychologique, les mots haineux et, en particulier, les insultes sexistes et raciales, ont évolué au point d'être utilisés et perçus pour exprimer des émotions négatives envers des cibles, fournissant ainsi des informations importantes sur l'état émotionnel du locuteur ou son attitude envers l'entité ciblée (Hedger, 2013), même lorsqu'ils font référence à des qualités descriptives. Nous pensons donc qu'il pourrait être intéressant d'étudier le lien entre le langage haineux et les parties négatives du spectre d'émotions à multiples facettes couvertes par les ressources informatiques sémantiques. En particulier, nous prévoyons de tester l'efficacité de la nouvelle version du modèle du sablier (Susanto et al., 2020), qui permet de mieux comprendre les émotions neutres et leur association avec d'autres émotions polaires et qui inclut certaines émotions polaires qui étaient auparavant absentes (notamment les émotions conscientes de soi et morales). Le modèle revisité du sablier calcule la polarité d'un concept avec une plus grande précision. Il fournit également un nouveau mécanisme pour classer les concepts inconnus en trouvant l'émotion antithétique d'un concept manquant et en inversant sa polarité. SenticNet 6 (Cambria et al., 2020) contient actuellement 200 000 mots et expressions multi-mots. Nous pensons qu'il s'avére être une ressource précieuse pour améliorer la détection des discours de haine multi-sujets et multi-cibles.

Vers une détection multilingue des discours de haine. Bien que la plupart des corpus de discours de haine disponibles soient en anglais, le problème des discours de haine ne se limite pas à une seule langue. Compte tenu de la diversité des langues et de l'énorme quantité de données sur les médias sociaux produites dans différentes régions du monde, la détection des discours de haine dans une perspective multilingue constitue également un défi de taille.[109] Nous prévoyons donc, à l'avenir, d'explorer la possibilité de développer des modèles agnostiques sur le plan linguistique, capables d'identifier les discours de haine

---

[109]Voir (Pamungkas et al., 2021b) pour une enquête concernant les corpus disponibles et les approches employées dans des contextes multilingues.

dans les communications en ligne.

Vers une détection multimodale des discours de haine. Les modèles conçus pour détecter les discours de haine sont entraînés en utilisant uniquement les caractéristiques textuelles ; nous n'avons pas tenu compte des images ou des vidéos incluses dans les tweets. Une nouvelle tâche partagée en 2022, Multimedia Automatic Misogyny Identification (MAMI),[110] sera consacrée à l'identification des « memes » misogynes, en tirant parti à la fois du texte et des images disponibles comme source d'information. La multi-modalité (Vijayaraghavan et al., 2021), ainsi que la détection de l'ironie et du sarcasme suggèrent des pistes de travail future.

---

[110]https://competitions.codalab.org/competitions/34175

Bibliography

Agarwal, S. and Sureka, A. (2017). Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website. *CoRR*, abs/1701.04931.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Akhtar, M. S., Ekbal, A., and Cambria, E. (2020). How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Computational Intelligence Magazine*, 15(1):64–75.

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Allport, G. W., Clark, K., and Pettigrew, T. (1954). The nature of prejudice.

Alshalan, R. and Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences*, 10(23):8614.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In

Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain, volume 6.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. (2020). Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018*, pages 57–64.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

Awal, M. R., Cao, R., Lee, R. K., and Mitrovic, S. (2021). Angrybert: Joint learning target and emotion for hate speech detection. In *Advances in Knowledge Discovery and Data Mining - 25th Pacific-Asia Conference, PAKDD 2021, Virtual Event, May 11-14, 2021, Proceedings, Part I*, volume 12712 of *Lecture Notes in Computer Science*, pages 701–713. Springer.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In Barrett, R., Cummings, R., Agichtein, E., and Gabrilovich, E., editors, *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 759–760. ACM.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33.

Barak, A. (2005). Sexual harassment on the Internet. *Social Science Computer Review*, 23(1).

Bartlett, J., Norrie, R., Patel, S., Rumpel, R., and Wibberley, S. (2014). Misogyny on twitter.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against

Immigrants and Women in Twitter. In May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., and Mohammad, S. M., editors, *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63. Association for Computational Linguistics.

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In Cabrio, E., Mazzei, A., and Tamburini, F., 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*. CEUR-WS.org.

Beike, D. R. and Sherman, S. J. (2014). Social inference: Inductions, deductions, and analogies. *Handbook of social cognition*, pages 209–285.

Benamara, F., Moriceau, V., and Mathieu, Y. Y. (2014). Fine-grained semantic categorization of opinion expressions for consensus detection (in French). In *DEFT 2014 Workshop: Text Mining Challenge*, pages 36–44.

Benamara, F., Taboada, M., and Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Bennett, M., Sani, F., Hopkins, N., Agostini, L., and Malucchi, L. (2000). Children's gender categorization: An investigation of automatic processing. *British Journal of Developmental Psychology*, 18(1):97–102.

Berman, G. and Paradies, Y. (2010). Racism, disadvantage and multiculturalism: towards effective anti-racist praxis. *Ethnic and Racial Studies*, 33(2):214–232.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 158–168. European Language Resources Association (ELRA).

Bianchi, C. (2014). The speech acts account of derogatory epithets: some critical notes. *Liber Amicorum Pascal Engel*.

Biscarrat, L., Coulomb-Gully, M., and Méadel, C. (2016). One is not born a female CEO and...won't become one! *Gender equality and the media - a challenge for Europe. Routledge, ECREA Book Series.*

Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Bonnafous, S. (2003). "Femme politique" : une question de genre ? *Réseaux*, 120.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., and Tesconi, M. (2018). Overview of the EVALITA 2018 Hate Speech Detection Task. In Caselli, T., Novielli, N., Patti, V., and Rosso, P., 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.org.

Bosco, C., Patti, V., Bogetti, M., Conoscenti, M., Ruffo, G. F., Schifanella, R., and Stranisci, M. (2017). Tools and resources for detecting hate and prejudice against immigrants in social media. In *SYMPOSIUM III. SOCIAL INTERACTIONS IN COMPLEX INTELLIGENT SYSTEMS (SICIS) at AISB 2017*, pages 79–84. AISB.

Bousquet, D., Vouillot, F., Collet, M., and Oderda, M. (2019). 1er état des lieux du sexisme en France. Technical report, Haut Conseil à l'Egalité entre les femmes et les hommes. http://www.haut-conseil-egalite.gouv.fr/IMG/pdf/hce_etatdeslieux-sexisme-vf-2.pdf.

Branco, P., Torgo, L., and Ribeiro, R. P. (2015). A survey of predictive modelling under imbalanced distributions. *CoRR*, abs/1505.01658.

Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities. In *24th European Conference on Information Systems, ECIS 2016, Istanbul, Turkey, June 12-15, 2016*, page Research Paper 61.

Bretschneider, U. and Peters, R. (2017). Detecting offensive statements towards foreigners in social media. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL).

Brustein, W. I. and King, R. D. (2004). Anti-semitism in europe before the holocaust. *International Political Science Review*, 25(1):35–53.

Burnap, P. and Williams, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In *Proocedings of Conference on Internet, Policy & Politics*, pages 1–18.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11.

Cai, Y. and Wan, X. (2019). Multi-domain sentiment classification based on domain-aware embedding and attention. In Kraus, S., editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4904–4910. ijcai.org.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). *A Practical Guide to Sentiment Analysis*. Socio-Affective Computing. Springer International Publishing.

Cambria, E. and Hussain, A. (2015). Sentic computing. *Cognitive Computation*, 7(2):183–185.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., and Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 105–114, New York, NY, USA. Association for Computing Machinery.

Cambria, E., Livingstone, A., and Hussain, A. (2012). The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer.

Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. (2017). Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Cambria, E., Poria, S., Hazarika, D., and Kwok, K. (2018). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Cameron, D. (1992). *Feminism and Linguistic Theory.* Palgrave Macmillan.

Carpenter, J., Preotiuc-Pietro, D., Flekova, L., Giorgi, S., Hagan, C., Kern, M. L., Buffone, A. E., Ungar, L., and Seligman, M. E. (2017). Real men don't say "cute" using automatic language analysis to isolate inaccurate aspects of stereotypes. *Social Psychological and Personality Science*, 8(3):310–322.

Cashdan, E. (1998). Smiles, speech, and body posture: How women and men display sociometric status and power. *Journal of Nonverbal Behavior*, 22(4):209–228.

Cepollaro, B. (2015). In defence of a presuppositional account of slurs. *Language Sciences*, 52.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018a). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174. Association for Computational Linguistics.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018b). Universal sentence encoder. *CoRR*, abs/1803.11175.

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. D., Stringhini, G., and Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. In Fox, P., McGuinness,

D. L., Poirier, L., Boldi, P., and Kinder-Kurlanda, K., editors, *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 13–22. ACM.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40:108 – 118.

Chew, P. K. and Kelley-Chew, L. K. (2007). Subtly sexist language. *Colum. J. Gender & L.*, 16.

Chiril, P., Benamara, F., and Moriceau, V. (2021a). "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844.

Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019a). Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, pages 351–360. ATALA.

Chiril, P., Benamara, F., Moriceau, V., and Kumar, A. (2019b). The binary trio at semeval-2019 task 5: Multitarget hate speech detection in tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 489–493.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020a). An annotated corpus for sexism detection in french tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020b). He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online. Association for Computational Linguistics.

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., and Patti, V. (2021b). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate

speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Clark, R., Anderson, N. B., Clark, V. R., and Williams, D. R. (1999). Racism as a stressor for african americans: A biopsychosocial model. *American psychologist*, 54(10):805.

Cohen-Almagor, R. (2017). Balancing freedom of expression and social responsibility on the internet. *Philosophia*, 45(3):973–985.

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017a). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017b). Word translation without parallel data. *CoRR*, abs/1710.04087.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806.

Coulomb-Gully, M. (2012). *Présidente : le grand défi - femme, politique et medias*. Paris, Payot/Éd. Rivages.

Coulombe, C. (2018). Text data augmentation made simple by leveraging NLP cloud apis. *CoRR*, abs/1812.04718.

Crawford, M. T., Sherman, S. J., and Hamilton, D. L. (2002). Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of personality and social psychology*, 83(5):1076.

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., and Zhao, B. Y. (2020). Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dai, H. and Xu, X. (2014). Sexism in News: A Comparative Study on the Portray of Female and Male Politicians in The New York Times. *Open Journal of Modern Linguistics*, 4.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). 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*, pages 512–515. AAAI Press.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.

De Mauro, T. (2016). Le parole per ferire. *Internazionale*. 27 settembre 2016.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Deaux, K. and Lewis, L. L. (1984). Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of personality and Social Psychology*, 46(5):991.

Delgado, R., Wing, A. K., and Stefancic, J. (2015). Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling. In *Law Unbound!*, pages 223–228. Routledge.

Dev, S. and Phillips, J. M. (2019). Attenuating Bias in Word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, pages 879–887.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., 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.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate Speech Detection with Comment Embeddings. In Gangemi, A., Leonardi, S., and Panconesi, A., editors, *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 29–30. ACM.

Does, S., Derks, B., and Ellemers, N. (2011). Thou shalt not discriminate: How emphasizing moral ideals rather than obligations increases whites' support for social equality. *Journal of Experimental Social Psychology*, 47(3):562–571.

d'Sa, A. G., Illina, I., Fohr, D., Klakow, D., and Ruiter, D. (2020). Label propagation-based semi-supervised learning for hate speech classification. In *Insights from Negative Results Workshop, EMNLP 2020*.

Du, Y., Wu, Y., and Lan, M. (2019). Exploring human gender stereotypes with word association test. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6132–6142. Association for Computational Linguistics.

ECRI (2015). Ecri general policy recommendation no. 15 on combating hate speech, 8 december 2015. *URL: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01*.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Ekman, P. (1999). *Basic Emotions. Handbook of Cognition and Emotion.* John Wiley & Sons Ltd.

Ellemers, N. (2018). Gender stereotypes. *Annual review of psychology*, 69:275–298.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. M. (2018a). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018*, pages 42–51.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018b). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Erjavec, K. and Kovačič, M. P. (2012). "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, 15(6):899–920.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 417–422.

EU Commission (2016). Code of conduct on countering illegal hate speech online.

Farías, D. I. H., Patti, V., and Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Techn.*, 16(3):19:1–19:24.

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96.

Fastenbauer, R. (2020). Islamic antisemitism: Jews in the qur'an, reflections of european antisemitism, political anti-zionism: Common codes and differences. In *Confronting Antisemitism from the Perspectives of Christianity, Islam, and Judaism*, pages 279–300. De Gruyter.

Fehn Unsvåg, E. and Gambäck, B. (2018). The Effects of User Features on Twitter Hate Speech Detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 75–85, Brussels, Belgium. Association for Computational Linguistics.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1615–1625. Association for Computational Linguistics.

Felmlee, D., Rodis, P. I., and Zhang, A. (2019). Sexist slurs: Reinforcing feminine stereotypes online. *Sex Roles*, pages 1–13.

Ferreira, D. C., Martins, A. F., and Almeida, M. S. (2016). Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2019–2028.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Caselli, T., Novielli, N., Patti, V., and Rosso, P., 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.org.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., and de Albornoz, J. C., 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.org.

Fersini, E., Rosso, P., and Anzovino, M. (2018c). 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). volume 2150 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, 2(4):357–411.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2018). AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4):689–707.

Floridi, L., Cowls, J., and King, T. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 26:1771–1796.

Florio, K., Basile, V., Polignano, M., Basile, P., and Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Science*, 10(12):4180.

Fokkens, A., Ruigrok, N., Beukeboom, C., Sarah, G., and Van Atteveldt, W. (2018). Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

Francesconi, C., Bosco, C., Poletto, F., and Sanguinetti, M. (2019). Error analysis in a hate speech detection task: The case of haspeede-tw at evalita 2018. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Menczer, F., Lewis, B., and Rowe, K. (2014). Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, pages 57–64.

Gambäck, B. and Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Ganin, Y. and Lempitsky, V. S. (2015). Unsupervised Domain Adaptation by Backpropagation. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. S. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17:59:1–59:35.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266.

García-Sánchez, R., Almendros, C., Aramayona, B., Martín, M. J., Soria-Oliver, M., López, J. S., and Martínez, J. M. (2019). Are sexist attitudes and gender stereotypes linked? a critical feminist approach with a spanish sample. *Frontiers in psychology*, 10:2410.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Garg, S. and Ramakrishnan, G. (2020). Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R., and Pathak, J. (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, WWW '19, page 514–525, New York, NY, USA. Association for Computing Machinery.

Giannakidou, A. and Mari, A. (2021). *(Non) Veridicality in grammar and thought. Mood, Modality and Propositional Attitudes*. The University of Chicago Press.

Gilbert, C. and Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, volume 81, page 82.

Glick, P. and Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology*, 70(3):491.

Glick, P. and Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly*, 21(1):119–135.

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gnanasekaran, R. K., Gunasekaran, R. R., et al. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Greevy, E. and Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. In Sanderson, M., Järvelin, K., Allan, J., and Bruza, P., editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 468–469. ACM.

Grosz, D. and Conde-Cespedes, P. (2020). Automatic detection of sexist statements commonly used at the workplace. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 104–115. Springer.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., and Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Guo, H., Mao, Y., and Zhang, R. (2019). Augmenting data with mixup for sentence classification: An empirical study. *CoRR*, abs/1905.08941.

Guo, W. and Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Haines, E. L., Deaux, K., and Lofaro, N. (2016). The times they are a-changing… or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12.

Han, J., Zhang, Z., and Schuller, B. (2019). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine*, 14(2):68–81.

Hardaker, C. and McGlashan, M. (2016). "real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80–93.

Hart, P. (1968). The condensed nearest neighbor rule (coresp.). *IEEE transactions on information theory*, 14(3):515–516.

Hathout, N. and Sajous, F. (2016). Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1369–1376.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., and Mihalcea, R. (2018). CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of*

the 27th International Conference on Computational Linguistics, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hedger, J. A. (2013). Meaning and racial slurs: Derogatory epithets and the semantics/pragmatics interface. *Language & Communication*, 33(3):205 – 213.

Hemker, K. and Schuller, B. (2018). Data augmentation and deep learning for hate speech detection. *Imperial College London*.

Herek, G. M. (2004). Beyond "homophobia": Thinking about sexual prejudice and stigma in the twenty-first century. *Sexuality Research & Social Policy*, 1(2):6–24.

Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Hussain, A. and Cambria, E. (2018). Semi-supervised learning for big social data analysis. *Neurocomputing*, 275:1662–1673.

Ibrohim, M. O. and Budi, I. (2018). A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Ibrohim, M. O. and Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., and Varma, V. (2019). FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on psychological science*, 4(2):153–161.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nedellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer.

Jourová, V. (2016). Code of conduct on countering illegal hate speech online: First results on implementation. *European Commission.[cit. 8. brezen 2018]*.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Karan, M. and Šnajder, J. (2018). Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Karlekar, S. and Bansal, M. (2018). SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811.

Karoui, J., Benamara, F., Moriceau, V., Patti, V., Bosco, C., and Aussenac-Gilles, N. (2017). 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, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Khatua, A., Cambria, E., Ghosh, K., Chaki, N., and Khatua, A. (2019). Tweeting in Support of LGBT? A Deep Learning Approach. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 342–345, New York, NY, USA. Association for Computing Machinery.

Khatua, A., E., C., and Khatua, A. (2018). Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400.

Kim, J., Ortiz, C., Nam, S., Santiago, S., and Datta, V. (2020). Intersectional bias in hate speech and abusive language datasets. *CoRR*, abs/2005.05921.

King, R. D. and Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology*, 51(4):871–894.

Klein, O., Tindale, S., and Brauer, M. (2008). The consensualization of stereotypes in small groups. *Stereotype dynamics: Language-based approaches to the formation, maintenance, and transformation of stereotypes*, pages 263–292.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Citeseer.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of hindi-english code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Kwok, I. and Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. In desJardins, M. and Littman, M. L., editors, *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA*. AAAI Press.

Langton, R. (2012). Beyond belief: Pragmatics in hate speech and pornography. *Speech and harm: Controversies over free speech*, pages 72–93.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 63–66. Springer.

Lavergne, E., Saini, R., Kovács, G., and Murphy, K. (2020). Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection. In *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020*, volume 2765. CEUR-WS.

Lazar, M. M. (2007). Feminist critical discourse analysis: Articulating a feminist discourse praxis. *Critical Discourse Studies*, 4(2).

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. (2020). Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lippmann, W. (1946). *Public opinion*, volume 1. Transaction Publishers.

Liu, P., Qiu, X., and Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In Barzilay, R. and Kan, M., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1–10. Association for Computational Linguistics.

Liu, Q., Zhang, Y., and Liu, J. (2018). Learning Domain Representation for Multi-Domain Sentiment Classification. 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 541–550, New Orleans, Louisiana. Association for Computational Linguistics.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Ljubešić, N., Erjavec, T., and Fišer, D. (2018). Datasets of slovene and croatian moderated news comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131.

Luque, F. M. (2019). Atalaya at TASS 2019: Data augmentation and robust embeddings for sentiment analysis. 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*, volume 2421 of *CEUR Workshop Proceedings*, pages 561–570. CEUR-WS.org.

Lynn, T., Endo, P. T., Rosati, P., Silva, I., Santos, G. L., and Ging, D. (2019). Data set for automatic detection of online misogynistic speech. *Data in brief*, 26:104223.

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. *Advances in experimental social psychology*, 31:79–122.

Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., and Saxena, M. (2018). Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 92–105.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandalia, C., and Patel, A. (2019). Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Majumder, P., Mitra, M., Gangopadhyay, S., and Mehta, P., editors, *FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019*, pages 14–17. ACM.

Mani, I. and Zhang, I. (2003). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126. ICML United States.

Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.

Mari, A., Beyssade, C., and Del Prete, F. (2012). *Genericity*, volume 43. Oxford University Press.

Martin, L., Müller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7203–7219. Association for Computational Linguistics.

Massaro, T. M. (1990). Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.*, 32:211.

Mathew, B., Kumar, N., Ravina, Goyal, P., and Mukherjee, A. (2018). Analyzing the hate and counter speech accounts on twitter. *CoRR*, abs/1812.02712.

Mathur, P., Sawhney, R., Ayyar, M., and Shah, R. (2018). Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

Matsuda, M. J., Lawrence III, C. R., Delgado, R., and Crenshaw, K. W. (1993). Words that wound: Critical race theory. *Assaultive Speech, and the First Amendment*, 5.

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47.

Mehta, Y., Majumder, N., Gelbukh, A. F., and Cambria, E. (2020). Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Menini, S., Moretti, G., Corazza, M., Cabrio, E., Tonelli, S., and Villata, S. (2019). A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mills, S. (2008). *Language and sexism*. Cambridge University Press Cambridge.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3):62:1–62:40.

Mishra, P., Tredici, M. D., Yannakoudakis, H., and Shutova, E. (2019). Author profiling for hate speech detection. *CoRR*, abs/1902.06734.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Mohammad, S. (2012). # emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad, S. and Kiritchenko, S. (2013). Using nuances of emotion to identify personality. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-13)*, Boston, MA.

Mohammad, S., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Mohammad, S. and Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3).

Mossie, Z. and Wang, J.-H. (2019). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, page 102087.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In Cherifi, H., Gaito, S., Mendes, J. F., Moro, E., and Rocha, L. M., editors, *Complex Networks and Their Applications VIII - Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019*, volume 881 of *Studies in Computational Intelligence*, pages 928–940. Springer.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In Rowe, M., Stankovic, M., Dadzie, A.-S., and Hardey, M., editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98.

Nissim, M. and Patti, V. (2017). Semantic aspects in sentiment analysis. In Pozzi, F. A., Fersini, E., Messina, E., and Liu, B., editors, *Sentiment Analysis in Social Networks*, chapter 3, pages 31 – 48. Morgan Kaufmann.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Nockleby, J. T. (2000). Hate speech. In *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.*, pages 1277–1279.

OECD (2018). *Working Together for Local Integration of Migrants and Refugees*.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of EMNLP-IJCNLP*.

Pamungkas, E. W., Basile, V., and Patti, V. (2020a). 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.

Pamungkas, E. W., Basile, V., and Patti, V. (2020b). Misogyny detection in twitter: a multilingual and cross-domain study. *Inf. Process. Manag.*, 57(6):102360.

Pamungkas, E. W., Basile, V., and Patti, V. (2021a). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Inf. Process. Manag.*, 58(4):102544.

Pamungkas, E. W., Basile, V., and Patti, V. (2021b). Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, pages 1–27.

Pamungkas, E. W., Cignarella, A. T., Basile, V., and Patti, V. (2018). *14-ExLab@UniTo* for AMI at IberEval2018: Exploiting Lexical Knowledge for Detecting Misogyny in English and Spanish Tweets. In Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., and de Albornoz, J. C., 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 234–241. CEUR-WS.org.

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and Cross-lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Pardo, F. M. R. and Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing & Management*, 52(1):73–92.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., and Varma, V. (2019). 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. Association for Computational Linguistics.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35.

Peng, M., Zhang, Q., Jiang, Y.-g., and Huang, X. (2018). Cross-Domain Sentiment Classification with Target Domain Specific Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M. A., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Piolat, A. and Bannour, R. (2009). An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. *Current psychology letters*, 25.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020). Offensive language identification in greek. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5113–5119. European Language Resources Association.

Plant, E. A., Hyde, J. S., Keltner, D., and Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24(1):81–92.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Plutchik, R. and Kellerman, H., editors, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pages 3–33. Academic press, New York.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., and Bandyopadhyay, S. (2013). Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., and Hussain, A. (2018). Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intelligent Systems*, 33(6):17–25.

Portner, P. (2009). *Modality*. Oxford University Press.

Potts, C. (2005). *The logic of conventional implicatures*. Oxford Studies in Theoretical Linguistics.

Ptaszynski, M., Pieciukiewicz, A., and Dybała, P. (2019). Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.

Purohit, H., Banerjee, T., Hampton, A., Shalin, V. L., Bhandutia, N., and Sheth, A. P. (2016). Gender-Based Violence in 140 characters or Fewer: A #BigData Case Study of Twitter. *First Monday*, 21(1).

Qian, J., Bethke, A., Liu, Y., Belding, E. M., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4754–4763. Association for Computational Linguistics.

Qian, J., ElSherief, M., Belding, E., and Wang, W. Y. (2018). Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.

Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

Reyna, C. (2000). Lazy, dumb, or industrious: When stereotypes convey attribution information in the classroom. *Educational Psychology Review*, 12:85–110.

Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., and Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, pages 33–36.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3(4):329–354.

Richardson-Self, L. (2018). Woman-hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2):256–272.

Rizoiu, M., Wang, T., Ferraro, G., and Suominen, H. (2019). Transfer Learning for Hate Speech Detection in Social Media. *CoRR*, abs/1906.03829.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., and Nakov, P. (2021). Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Saha, P., Mathew, B., Goyal, P., and Mukherjee, A. (2018). Hateminers : Detecting hate speech against women. *CoRR*, abs/1812.06700.

Samghabadi, N. S., Hatami, A., Shafaei, M., Kar, S., and Solorio, T. (2020). Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88.

Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., 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 EVALITA*.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Schmidt, A. and Wiegand, M. (2017). 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. Association for Computational Linguistics.

Schrading, N., Ovesdotter Alm, C., Ptucha, R., and Homan, C. (2015). An Analysis of Domestic Abuse Discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., Blanco, E., Dziurzynski, L., Park, G., Stillwell, D., et al. (2016). Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 516–527. World Scientific.

Sharifirad, S., Jacovi, A., Univesity, I. B. I., and Matwin, S. (2019). Learning and understanding different categories of sexism using convolutional neural network's filters. In *WNLP@ ACL*, pages 21–23.

Sharifirad, S., Jafarpour, B., and Matwin, S. (2018). Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs. In *Proceedings of the Second Workshop on Abusive Language Online*.

Shirbandi, A. and Moradi, B. (2019). Comparative Study of Combination of Convolutional and Recurrent Neural Network for Natural Language Processing. Technical report, Easy-Chair.

Si, S., Wang, R., Wosik, J., Zhang, H., Dov, D., Wang, G., and Carin, L. (2020). Students need more attention: Bert-based attention model for small data with application to automatic patient message triage. In *Machine Learning for Healthcare Conference*, pages 436–456. PMLR.

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive language and hate speech detection for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 687–690. AAAI Press. 10th International Conference on Web and Social Media, ICWSM 2016 ; Conference date: 17-05-2016 Through 20-05-2016.

Singh, A., Blanco, E., and Jin, W. (2019). Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2096–2101.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Steele, C. M. and Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, 69(5):797.

Strapparava, C. and Valitutti, A. (2004). WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sulis, E., Farías, D. I. H., Rosso, P., Patti, V., and Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge Based Systems*, 108:132–143.

Sundstrom, R. R. and Kim, D. H. (2014). Xenophobia and racism. *Critical philosophy of race*, 2(1):20–45.

Susanto, Y., Livingstone, A. G., Ng, B. C., and Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Sánchez-Junquera, J., Chulvi, B., Rosso, P., and Ponzetto, S. (2021). How Do You Speak about

Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Applied Sciences, 11(8), 3610.*

Tan, Y. C. and Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772.

Tsvetkov, M. X. A. F. Y. (2020). Demoting racial bias in hate speech detection. *SocialNLP 2020*, page 7.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016a). A Dictionary-based Approach to Racism Detection in Dutch Social Media. *CoRR*, abs/1608.08738.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016b). The automated detection of racist discourse in Dutch social media. *Computational Linguistics in the Netherlands Journal*, 6:3–20.

Ullmann, S. and Tomalin, M. (2020). Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

Vidgen, B. and Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Vidgen, B. and Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In Armando, A., Baldoni, R., and Focardi, R., 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.org.

Vijayaraghavan, P., Larochelle, H., and Roy, D. (2021). Interpretable multi-modal hate speech detection. *CoRR*, abs/2103.01616.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wang, B., Yunxia Ding, S., and Zhou, X. (2019). YNU Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)*.

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., and Carin, L. (2018). Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

Wang, W. Y. and Yang, D. (2015). That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Warner, W. and Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In Waseem, Z., Chung, W. H. K., Hovy, D., and Tetreault, J. R., editors, *Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, pages 78–84. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

Waseem, Z., Thorne, J., and Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online Harassment*, pages 29–55. Springer.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Weinberg, G. (1971). Words for the new culture. gay.

Weiner, B. (1986). *An Attribution Theory of Motivation and Emotion*, volume 92.

Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). 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.

Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018a). 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. Association for Computational Linguistics.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018b). Overview of the germeval 2018 shared task on the identification of offensive language.

Wille, E., Gaspard, H., Trautwein, U., Oschatz, K., Scheiter, K., and Nagengast, B. (2018). Gender stereotypes in a children's television program: Effects on girls' and boys' stereo-

type endorsement, math performance, motivational dispositions, and attitudes. *Frontiers in psychology*, 9:2435.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Xia, Y., Cambria, E., Hussain, A., and Zhao, H. (2015). Word polarity disambiguation using bayesian model and opinion-level features. *Cognitive Computation*, 7(3):369–380.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., and Ng, A. Y. (2017). Data noising as smoothing in neural network language models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Xu, H., Zhang, Z., Wu, L., and Wang, C.-J. (2019). The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PloS one*, 14(11).

Xu, Z., von Ritter, L., and Serra, G. (2020). Hierarchical Adversarial Training for Multi-domain Adaptive Sentiment Analysis. In Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., and Ras, Z. W., editors, *Complex Pattern Mining - New Challenges, Methods and Applications*, volume 880 of *Studies in Computational Intelligence*, pages 17–32. Springer.

Yadav, A. and Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). 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. Association for Computational Linguistics.

Zannettou, S., Finkelstein, J., Bradlyn, B., and Blackburn, J. (2020). A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 786–797.

Zeinert, P., Inie, N., and Derczynski, L. (2021). Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online*.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., and Chen, E. (2019). Interactive Attention Transfer Network for Cross-Domain Sentiment Classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5773–5780. AAAI Press.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. (2019). Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018b). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.