University of Turin

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

PhD Dissertation
Sohail Akhtar

Perspective-aware Models for Hate Speech Detection towards Vulnerable Communities

Advisors
Valerio Basile
Università degli Studi di Torino, Italy
Viviana Patti
Università degli Studi di Torino, Italy

PhD Coordinator
Marco Grangetto

A thesis submitted in fulfilment of the requirements for the degree of PhD in Computer Science(INF/01)

Academic Year 2020-2021

Abstract

In recent years, hand in hand with the massive rise of online abuse, the research community has shown great interest in abusive language and hate speech detection, also due to the considerable social impact of these phenomena on the well-being of a society. In online discussion on controversial topics in social media platforms, the users have shown a stronger tendency to express their opinions, and this resulted in the spread of hate against vulnerable communities including immigrants and migrants. The efforts to regulate hate speech against minorities are extremely challenging and fail due to the fact that the consequences of hate speech on the victim's physical and mental well-being are not considered. In order to recognize the complexity of the situation, it is important to understand the perspective of the victims of online hate and physical abuse. The issue needs considerable attention from researchers and policy makers in order to protect the disadvantaged social groups.

It has been observed that in real life hate incidents, when the members of any vulnerable community are targeted, the victims have an opportunity to express their views and opinions about the hate incident. In such cases, the background of people involved in these crimes is often available. However, when we consider detecting online hate where we typically rely on the annotation process in a crowd-sourcing scenario, in most cases the available platforms do not provide any background information (culture, ethnicity, social background etc.) about the workers annotating the datasets. This is a limit in general, and especially in the case of HS detection. Indeed, abusive language and HS against different communities often contain stereotypical words which might arouse positive or negative sentiments or different reactions in annotators with different backgrounds. Due to the constant unavailability of such information on the annotators, there is the challenge of observing high polarization among the annotator's judgments about the same potentially abusive messages, which might result in low inter-annotator agreements which also affect the quality of training datasets for HS detection.

This thesis aims to investigate the phenomenon of abusive language and its various forms on social media with a particular focus on modeling the perspectives of individual annotators. We plan to investigate the divergence of opinions between different annotator groups including the victims of abusive language and hate speech. Ideally, involving the victims and targets of hate speech, such as migrants, in the annotation process would potentially help us to understand their views about online hate incidents. The driving force behind our work is based on the assumption that a group of annotators can be divided into sub-groups based on some personal characteristics such as cultural background, common social behavior and other similar factors. The victims are the ones who mostly suffer emotionally and psychologically in online abuse. However, the fine-grained information about the cultural, ethnic, or demographic background of the annotators is usually not available, or it is not a primary factor when selecting expert or-volunteers as annotators. We therefore proposed a methodology to automatically model the different perspectives that annotators may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. In our method, supervised learning models are trained to learn different point of views of the human annotators. This is achieved by dividing the annotators into groups by using polarization index (P-index), a metric that automatically divides the annotators into groups. This division of the annotators of each dataset into groups helped us in modeling the polarization of their annotations at the message level. We then create a separate gold standard for each of the

groups and perform classification tasks to measure the performance of perspective-aware supervised models in a multi-lingual setting. We then proposed an ensemble classifier that considers all the learned perspectives in an inclusive fashion. Such method aims to "give voice" to all the existing perspectives on a certain phenomenon equally. To understand the point of views of victims, we develop a multi-perspective BREXIT dataset which is annotated by the victims of HS and perform deep qualitative analysis as well to understand the perspectives of groups with known personal background. We extended our methodology further to tackle issues like handling large numbers of annotators, problems of exhaustive search with polarization index and unavailability of the annotators other than crowd-sourcing platforms. A quantitative method is developed to mine perspectives from annotators groups. We used unsupervised clustering techniques (e.g. k-means clustering) to divide the annotators into more than two groups by finding the optimal number of partitions.

We also extend our datasets to measure the effectiveness of our methodology in a broader perspective. Our methodology provided a way to give preference to the classifier trained on the data annotated by a group involving the victims of hate speech, if this information is available, in order to "give voice" to the targeted group through the computational model. Furthermore, given its transparency, this classifier shows potential for providing an explicit explanation of its decisions, being able to track them back to the specific (highly cohesive) groups of people who annotated the training data.

Abstract

Negli ultimi anni, in concomitanza con l'incremento dell'uso di linguaggio abusivo online, la comunità scientifica ha mostrato grande interesse verso l'analisi del linguaggio d'odio, anche per via del considerevole impatto sociale di questo fenomeno sul benessere collettivo. Gli utenti impegnati in discussioni su argomenti controversi sui social media mostrano una tendenza più forte ad esprimere le proprie opinioni, e questo talvolta conduce a una più ampia diffusione di odio rivolto contro comunità vulnerabili come immigrati e migranti. Gli sforzi per regolare l'hate speech contro le minoranze sono estremamente complessi e falliscono nel momento in cui non considerano le conseguenze del linguaggio d'odio sulla salute fisica e mentale delle vittime. È importante quindi comprendere la prospettiva delle vittime di odio online e violenze allo scopo di avere un quadro completo della situazione. È necessaria l'attenzione di ricercatori e *policy maker* per proteggere i gruppi sociali svantaggiati.

Si osserva che in casi di crimini d'odio offline, quando i membri di una comunità vulnerabile sono le vittime, essi hanno l'opportunità di esprimere le proprie opinioni sull'incidente in questione. In questi casi, l'identità delle persone coinvolte negli episodi di odio è solitamente nota. Considerando invece uno scenario di identificazione automatica di linguaggio d'odio, tipicamente basata su un processo di annotazione, nella maggior parte dei casi le piattaforme di crowd-sourcing non rendono disponibile alcuna informazione sugli annotatori, come etnia, background socio-culturale, etc. Questo pone un limite all'efficacia dei sistemi di hate speech detection. In effetti, le espressioni di linguaggio abusivo e di odio verso comunità vulnerabili contengono spesso parole legate a stereotipi che sollecitano sentimenti positivi o negativi, e differenti reazioni in persone con background differente. A causa della mancanza di informazioni sugli annotatori, si osserva un alto livello di polarizzazione dei giudizi degli annotatori a proposito degli stessi messaggi potenzialmente abusivi, che può portare a un accordo tra annotatori basso e a una qualità inferiore dei dataset usati per l'apprendimento di modelli di hate speech detection.

Questa tesi si propone di indagare il fenomeno del linguaggio abusivo e le sue varie forme sui social media con particolare attenzione alla modellazione delle prospettive dei singoli annotatori. Si intende indagare la divergenza di opinioni tra i diversi gruppi di annotatori, comprese le vittime del linguaggio abusivo e dei discorsi di odio. Idealmente, il coinvolgimento delle vittime dell'odio, come i migranti, nel processo di annotazione aiuta a capire le loro opinioni sugli episodi di odio online. La motivazione principale del lavoro si basa sul presupposto che un gruppo di annotatori può essere diviso in sottogruppi basati su alcune caratteristiche personali come il background culturale, il comportamento sociale comune e altri fattori simili. Le vittime sono quelle che più soffrono emotivamente e psicologicamente nell'abuso online. Tuttavia, le informazioni informazioni a grana fine sul background culturale, etnico o demografico degli annotatori di solito non sono disponibili o non sono un fattore primario quando si selezionano esperti o volontari come annotatori. Si propone quindi una metodologia per modellare automaticamente le diverse prospettive degli annotatori nei confronti di alcuni fenomeni altamente soggettivi, come il linguaggio offensivo e i discorsi d'odio. Nel metodo proposto in questa tesi, i modelli di apprendimento supervisionato sono addestrati per apprendere diversi punti di vista degli annotatori umani. Questo risultato è ottenuto dividendo gli annotatori in gruppi utilizzando un indice di polarizzazione (P-index), una metrica basata sulla divisione automatica degli

annotatori in gruppi. Tale divisione permette di modellare la polarizzazione delle annotazioni a livello di singolo messaggio, di creare gold standard separati per ogni gruppo ed addestrare e valutare rispettivi modelli supervisionati che codificano le diverse prospettive, su più lingue. Inoltre viene proposto un classificatore *ensemble* che considera tutte le prospettive in maniera inclusiva. Tale metodo mira a “dare voce” a diverse prospettive possibili a proposito di un certo fenomeno, senza forzare una preferenza.

Per studiare empiricamente questo approccio è stato sviluppato un dataset chiamato BREXIT, costruito in modo da preservare diverse prospettive e annotato da vittime di linguaggio d’odio. Il corpus è analizzato qualitativamente per studiare le prospettive di gruppi il cui background personale è noto. La metodologia è stata poi estesa per affrontare problemi come la gestione di un gran numero di annotatori, problemi di ricerca esaustiva con l’indice di polarizzazione e l’indisponibilità di annotatori al di fuori delle piattaforme di crowd-sourcing. È stato sviluppato un metodo quantitative per estrarre automaticamente le prospettive di gruppi di annotatori mediante tecniche di clustering (k-means) per dividere gli annotatori in più di due gruppi individuando il numero ottimale di partizioni. Il dataset originale è stato esteso allo scopo di misurare l’efficacia della metodologia proposta in questa tesi su una casistica più ampia. Il risultato del metodo applicato a questo dataset esteso fornisce la possibilità di orientare il classificatore a preferire la prospettiva codificata dalle annotazioni fornite da un gruppo che coinvolge le vittime di discorsi d’odio, se questa informazione è disponibile, al fine di “dare voce” a tale gruppo sociale attraverso il modello computazionale. Inoltre, a causa della sua natura trasparente, questo tipo di classificatore ha potenziale per fornire una spiegazione più esplicita delle proprie predizioni, avendo a disposizione l’informazione per risalire agli specifici gruppi omogenei di persone che hanno annotato i dati di addestramento.

Acknowledgments

Here, I would like to acknowledge important people, who have continuously been a source of courage and inspiration to accomplish this work and towards achieving the milestones throughout this doctoral period.

I consider myself extremely lucky to met people who guided me in my personal and professional growth as an early stage researcher. I consider it as one of the best days in my life when i heard the news of my PhD acceptance at Computer Science department of the University of Turin. The lucky part was the privilege of working with Prof. Viviana Patti and Prof. Valerio Basile, the best supervisors one can have as a PhD student. They always helped me and supported me through several matters, a foreigner faces in a new environment. I really enjoyed my time stay with in Torino. They also helped me out learning the real way how an early career researcher should take a step forward, including the identification of my PhD theme. Viviana supported me and helped me with administrative issues throughout my stay along with the research activities. I remember at the end of my first year, when i was a bit distracted and lost, Valerio steeped in and showed me how to proceed forward with my research. He helped me to organize my scattered ideas in a coherent research direction. I am thankful to both of them for the mistakes they have pointed out which have been a source of continuous and quick learning for me during these days. Here, I acknowledge that it could not have been possible without their support. I admire the way they obliged me offering their valuable time and suggestions whenever needed.

I'm really thankful to the reviewers of this thesis: Prfo. María Teresa Martín Valdivia from University of Jaén (Spain) and Dr. Felice Dell'Orletta from Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche (Italy). I am grateful for your valuable and encouraging comments and useful suggestions which made it possible for me to improve the quality of this thesis. Thank you very much to the members of the evaluation tribunal of this thesis: Prof. Fabio Massimo ZANZOTTO, Dr. Felice Dell'Orletta, and Dr. Marco Guerini for their feedback and suggestions for improvements in the future work of this research direction.

During my stay at the department of Computer Science, I met people who were extremely helpful and supportive. I would like to thank the members of my research group: Prof. Vincenzo Lombardo (such a humble personality), Prof. Cristina Bosco (simplicity with elegance), Prof. Rossana Damiano (she looks so younger and i wonder how?), Prof. Alessandro Mazzei (a person with friendly and polite nature), Prof. Ruggero Pensa (always so respectful to others and felt about him like a big brother) and Prof. Antonio Lieto (a quiet person but an excellent researcher and writer). They are the people who inspire younger people like I and other PhD students.

I would like to say a big "thank you" to Dr. Mirko Lai, such a nice and humble person and if you are facing any problem whether it is about programming, research, feeling lonely, having an issue with your laptop, you would always ask "Where is Mirko?". Not a single time during my stay there, I asked Mirko for help and he said no or told me that he is busy and will do it later. Always there when I needed his help. A special person I would always remember. I would like to thank Dr. Michael Fell with whom i worked during his stay as a post-doc researcher at the department. A very nice and helpful person. A special mention is surely deserved for some of my friends: Faisal Imran, Endang Wahyu Pamungkas add Komal Florio, the best fellow

PhDs one could ever desire. I would especially like to thank Alessandra Teresa Cignarella for helping me out when I required her assistance several times with the Italian language for administrative issues. I want to thank all the fellow PhDs with whom I started this journey and spent wonderful time and shared moments of joy (specially the time at Bertinoro summer school): Simone Pernice, Stefano Tedeschi, Mattia Cerrato, Daniele Perlo and Simona Frenda. I also met some junior PhD students including Narges Azizi Fard and Elisa di Nuovo. They are wonderful people one would love to meet with and spend time.

During my stay in beautiful Torino, I had the pleasure of meeting some great Pakistanis (mostly students) with whom I spent these years together. I would like to thank Zeeshan Saeed for his company and support. I will always remember my time with him during special occasions such as Eid and his invitations for dinner. I enjoyed the company of my friends: Muhammad Abeer Irfan and Zain Memon, the fellow PhDs from Pakistan. I spent some time with Muhammad Waqas Rabbani, a very nice person and together, we used to share the kitchen with Syed Umair, whose knowledge on Pakistani and Italian politics and his command on Urdu poetry and literature with a touch of humor has always impressed me.

To my parents, whose sacrifice in every aspect of life enabled me to execute as best as I could. They have all the way been behind me through their prayers to encourage me fulfilling the necessary requirements during this tedious process, making me more energetic and enthusiastic to successfully ending up this work in order to complete my PhD program. My inspiration has always been my father who, at one stage, started from zero and worked extremely hard to establish himself as a businessman. He provided me the opportunities to excel in education while he was working hard to build a better future for our family. My mother never went to school but she taught me how to live a life, how to behave, respect older, love younger, always respect and help others and contribute in the betterment of a society. I will always remember her love and prayers.

To my dearest wife for all her support and prayers for ending up this work in a graceful way. It would not have been possible for me to achieve these goals without her support and love. She has been very much supportive considering the fact that i had to stay thousands of miles away form her during this journey. To my brothers and sisters, whose moral and ethical support has never let me divert my attention to anything else that might lead me astray somewhere far away from where I am standing today. To my brother-in-law Dr. Masood ul Hassan, a person to look up to in need of help and guidance. The person who has always desired for me to perform exceptionally well in my personal and professional life. I would like to mention Muhammad Amjad, my brother-in-law, who was always there for moral and financial support.

Finally, I am also indebted to my dearest friend Dr. Muhammad Ayub from Comsats University for his wisdom, inspiration, encouragement, laughs and much needed information and support of different sorts and motivational speeches throughout this hard journey. He was always there to support when I needed even until the last moment when I defended my thesis at his house due to the availability of a peaceful environment with technical and moral support.

Thank you all,
Sohail Akhtar

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The technological advances in recent years encouraged users to spend more time on the Internet and connect with the others around the globe, engaging in public debates [Conover et al., 2011]. This provides an opportunity to share views, ideas, opinions, and information about common interests [Jurgens et al., 2019] on diverse topics including social issues and sociopolitical events resulting in the formation of ideologically homogeneous clusters [Himelboim et al., 2013]. The Internet has developed into a rich source of information for the younger generations like children, pre-teens and teens who join social networking sites, which allow them to make online friends, socialize with them, find existing acquaintances, search for new experiences, and expand their interests beyond what they can access within their local communities. The social media platforms, such as Facebook and Twitter, which are considered the most popular platforms used by people around the world, provide tools to communicate that considerably affected the communication patterns and the way people exchange information with the others around the world. The dissemination of information at a global level has enabled the organizations and institutions to accumulate a massive amount of data, which is useful for enterprise information systems and other scientific disciplines such as astronomy, bioinformatics, medical sciences, imaging technologies, remote sensing, history, and social and psychological disciplines. In most cases, the available data is in raw form and cannot be used for further analysis. The data need to be inspected, analyzed, and transformed with the goal of finding useful patterns and trends by selecting and modeling more useful formats which are easily accessible to the scientists and engineers which use powerful data mining tools to get the useful information from large datasets. For the information processing and extraction, the data mining techniques make use of statistics, artificial intelligence, machine learning, and pattern recognition disciplines. The storage of the raw data is as important as the extraction of useful information and pattern recognition for the intended use. The use of data storage technologies such as relational databases, data-warehouses, transactional databases, advanced database systems such as object-oriented, no-sql databases, time-based databases and multimedia databases have solved the problem of storing the huge and a wide variety of raw datasets.

People rely heavily on the social media platforms to learn from other people through social interactions [Ito et al., 2009]. The environment provided by social media platforms to communicate online is quite complicated due to its dynamics, to the characteristics of the shared contents, e.g., persistence, visibility, spreadability and searchability [Boyd, 2014], and to the unrestricted nature of the information exchange. This online environment is also exploited for the dissemination of aggressive and offensive content, in a dangerous and hostile manner, against people through messaging, comments, wall posts, or even through the verbal abuse. This sudden rise in the hate speech related incidents on social media is considered a major issue and has raised serious concerns about the safety and inclusion of people in online communications. The Pew Research Center has conducted an online survey in 2017 about the online harassment in the US and found that 40 percent of the US adults were personally harassed online and almost 18 percent were either sexually harassed or physically threatened [Duggan, 2017]. A survey by Hinduja and Ptachin [2020] revealed that around 36 percent of the college students in the US face cyberbullying during their lifetime. Similar surveys conducted world-wide in the past years show identical or even more alarming patterns.

Long before the advances in digital age, people have been victimized based on the hostile portrayals and stereotyping categories such as race, color, gender, sex, religion, and the other similar factors. However, today, the Internet allow the spread of hatred faster and at a global level. Since the identity of people is often hidden from rest of the world as they prefer to communicate anonymously, this may result in heated arguments and aggressive communications [Burnap and Williams, 2015] when the others tend to disagree on a topic of mutual interest due to the availability of freedom of expression and limited legislation against the massive online abuse. This can also result in enclosing the admission of offensive and insulting language. There have been legal efforts and established regulations by social media platforms and the governments to detect and stop the spread of aggressive behaviour; still the automatic detection systems these platforms rely on are not effective in detecting the aggressive behaviour in general, and hate speech (HS) in particular. This increase in the toxic phenomenon may have serious negative consequences for the victims being exposed to such hate, with an impact not only on the individual but also on society. One example is the role of social media platforms in sociopolitical events, which is a hot topic of analysis these days, and highlights some challenges to be faced related to the so called "dark side" of Internet's unregulated and open nature, which allows everyone to share everything, regardless of the authenticity of the information, and is a fertile environment for the spread of information leading to aggressive and offensive behaviours in general public.

In this thesis, we aim to conduct an in-depth study of the novel approaches to address the task of automatic abusive language detection (particularly hate speech) in social media against the particular targets such as immigrants, Muslims, Jews, women, colored people, and the other disadvantaged minorities, within a computational linguistic perspective. We propose to address the task by considering some important aspects of abusive language detection, i.e, the role of polarized opinions in the datasets and the need of involving the victims of abusive language in the decision process to rely, in addition, on the external information of a particular event or debate, which can help to detect the presence of abusive content in online conversations. We believe that the characteristics such as ethnicity, social background, culture etc. can influence the opinions of people and based on such characteristics, people can be grouped together.

The chapter will first introduce the motivation behind the research work. We will highlight the importance of considering the polarization of opinions expressed by online users and also, the impact of involving the victims of HS in the detection process to let their voice be heard, so that we can understand their perspectives on the abusive phenomenon. The chapter is organized as follow: Section 1.1 introduces the theoretical concepts about the abusive language and explains its role in social media. Next, in Subsection 1.1.1, we briefly introduce what a language is and its role in a society. Subsection 1.1.2 describes abusive language with several definitions and comparison with the other similar phenomena related to the online abuse. Subsection 1.1.3 describes the hate speech as an abusive phenomenon with several definitions and problems in identifying the HS in social media and the urgency of building robust models to automatically detect such phenomenon. In Section 1.2, we introduce some open challenges related to abusive language detection and HS that guided our research. First, in Subsection 1.2.1 we present a short introduction to the HS against immigrants which are often the victimized minorities in a society. Second, in Subsection 1.2.2 we explore the problems related to the bias in data annotations that are often dependant on the available datasets. Third, Subsection 1.2.3 presents the open challenges related to the opinion polarization in data annotations and why they are important to model a robust abusive language detection task. Finally, Sections 1.3, 1.4, and 1.5 describe research questions and objectives, contributions, and the structure of the whole manuscript respectively.

1.1 Abuse in Social Media

Social media platforms provide users the freedom of expression and a medium to exchange information and express diverse opinions on various domains including debates on public affairs and social values. Unfortunately, this has also resulted in the massive rise of online abusive content and uncensored online hate with the purpose of discriminating people especially targeting the vulnerable communities such as immigrants, Muslims, women and LGBT+ people [Duggan, 2017]. The minority groups are often targeted based on race, ethnicity and gender [Duggan, 2017, Waseem, 2016]. In this section, we will thoroughly go through the phenomenon of abusive language in social media platforms by providing a short overview of language in general, abusive language and the role of hate speech against people.

1.1.1 Language in a Society

Language, as a social phenomenon and a way to communicate and understand each other, plays an important role in everyone's life. Language is responsible for developing and sustaining relationships among the members of a society and between the societies [Fitch, 2006]. The words as a way to communicate serve an indented purpose, usually the verbal delivery of a message to other people for many different reasons. Human languages are also a source to express feelings and emotions, which may represent love, sympathy, joy, anger or hate. Language is a very important part of our society and plays an important role in the survival and progress of a society to fulfil the needs of people [Panicacci, 2019].

Language can bring people together, create harmony, boost mutual respect, allow to express individuality and define the differences in an understandable manner. The choice of words to covey intended messages is a crucial step in any kind of communication. The words chosen for a conversation must be polite and convey humbleness and should not express threatening or insulting behaviour [Mills and Kádár, 2011]. Unfortunately, language is often used as a weapon to insult and degrade people including minorities. The misuse of language can harm people, emotionally disturb them, and also act as a catalyst for adverse psychological effects which may span the entire life of a normal human. Lying is considered a very serious and arguably the most dangerous instance of a verbal abuse. Lying can seriously alter or destroy the foundations of a relationship among people [1]. Lying can cause misunderstandings, manipulating emotions, exaggerating something which is not there and withhold the information important for a relationship to function properly. Unlike lies, the inappropriate language by using abusive words pertaining both the sexual, emotional or physical sphere, can act in a more dramatic manner. This can have adverse effect on the identity of a person and can create a sense of insecurity and self-disgust. The actual physical threats against the victims of physical and emotional abuse may cause fear and insecurity. We have seen in last few years that the way we communicate is evolving rapidly. The face-to-face communication may enhance the quality of life. A recent research revealed that the online communication can not have similar effects as in-person communication and may be less satisfying [Lee et al., 2011]. It is also seen that trying to guilt someone or using the arrogant language may contribute towards the current divide in the societies. How abuse spreads in a society is also important to consider while tackling these issues within a society [Napolitano et al., 2018].

1.1.2 Abusive Language

In online discussions, the users have shown a stronger tendency to express their opinions [Wendling, 2015] on controversial topics, and this has resulted in the spread of online abuse and hate speech. The increase in social media usage has provided new opportunities to collect and analyze the rich data in the form of messages and conversations including content related to the online harassment and abuse. In recent years, hand in hand with the massive rise of online hatred [Zhang and Luo, 2018], the research community has shown great interest in the HS detection, also due to the considerable social impact of the phenomenon on the well-being of a society [Jurgens et al., 2019].

While we do not have a universal compliance on what “offensive contents” are, however, there is a general agreement that define it as any type of language that attacks a person or a group of people based on the personal characteristics linked to their social and demographic backgrounds [Schmidt and Wiegand, 2017]. Abusive language, which is expressed either in verbal or in textual form, is considered a derogatory language containing words or phrases which are abusive and dirty in the form of unnecessary jokes, vulgar or sexual conversation, or insults linked to the cursing of human beings [Chen et al., 2012]. There are many reasons behind the spread of such uncontrolled abusive communications: the social media platforms lack proper tools to identify and filter such hate on a large scale, there is a lack of mutual respect and empathy among online users [Turkle, 2015], and there is a lack of proper guidance for online users to behave and communicate in a friendly manner, and also unawareness with respect to the law or to the codes of conduits which try to regulate such unwanted abusive behaviours.

[1] https://poemachronicles.com/language/

Similarly, Hinduja and Ptachin [2020] defined offensive language as a communication referring to vulgar, pornographic, and hateful language. The vulgar language means showing coarse and rude behaviour including explicit and offensive terms related to the sex or bodily functions; pornographic language refers to the explicit sexual matter which may be responsible, outside the law, for sexual arousal and erotic satisfaction. When people from the ethnic minorities are exposed to these characteristics of hate, they face radicalization, psychological trauma [Gelber and McNamara, 2015], and in most extreme cases, self-harm and suicide (e.g., Saha et al. [2019] observed such effects in college students). Nobata et al. [2016] defined hate as a "Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity". The laws defined by some countries against abusive language consider certain expressions hateful and consequently illegal [Abbondante, 2017].

Being able to identify abusive language in online conversations is important for supporting content moderation activities and raising awareness about hurtful contents. Because of the large amount of abusive content online, it is not possible to manually identify and filter the offensive content. Word matching is a very simple technique used to identify the hateful and offensive text online, where a dictionary containing hateful or abusive words is used to match the keywords that are hateful, and then to filter out contents if there is a match between the keywords and the words from the available hate lexicon, but this is a very limited approach [Nobata et al., 2016], which, among other things, fails to account for implicitly abusive language [Wiegand et al., 2021]. The language used by online users to write messages is often very informal and hard to understand. It may contain slightly modified or reduced spellings and grammar when the text is short such as in Facebook and Twitter messages, and the use of figurative language is not infrequent. The short messages make it really hard for the tools to detect the abusive messages as often the users use abbreviations to post the messages. Such words often contain emotions, character repetition to put more emphasis on a word or expression, slang words, and also numbers instead of vowels [Hanafiah et al., 2017]. Therefore, the detection of abusive language on the social media is a very difficult and challenging task and needs automated tools and machine learning technologies to process and filter abusive textual content.

There are several other terms which often overlap and have similar meanings to abusive language. Here, we try to provide a comparison between these terms and abusive language, which is often considered a broader term. These terms are hate speech, hate, discrimination, cyberbullying, flaming, toxicity and profanity. Table 1.1 provides an overview of these terms and their comparison with abusive language.

1.1.3 Hate Speech as Abusive Phenomenon

Hate speech is a special type of abusive language whose detection on social media platforms is a rather difficult but important task. *Online hate speech* (or *cyber-hate*) may take different forms. The sudden rise in hate speech related incidents on social media is considered a major issue. By nature, HS can be active for longer periods of time, violating people's privacy and also seriously affecting one's personality and self confidence with a viral spread across the Internet [Schmidt and Wiegand, 2017]. Another concern is that the users, while posting hateful content against the others, not only feel comfortable with the spread of such offensive content, but also the Internet

Term	**Definition**	**Comparison with abusive language**
Hate Speech	Hate speech usually refers to disparaging individuals or groups because of their ethnicity, gender, race, color, sexual orientation, nationality, religion, or other similar characteristics — see for instance the U.S. constitution [Nockleby, 2000].	Abusive language is a broader term and HS is a special type of abusive language.
Hate	Hostile expressions without any indication of a reason behind the use of such language [Tarasova, 2016].	Abusive language is a general term and related to many concepts. Hate here refers to more specific type of abuse.
Discrimination	Process through which a difference is identified and then used as a basis for unfair treatment [Thompson, 2019].	Abusive language is way of discriminating people based on their race, color and with derogatory words.
Cyberbullying	Aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend himself or herself [Chen et al., 2012].	Abusive language has many targets and not like cyberbullying, mostly focusing on the individuals.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [Jurgens et al., 2019].	Abusive language is dependent on a particular context or content, whereas flaming is more centered towards a participant within a specific context of a discussion.
Toxicity	Toxic is defined as rude, disrespectful or unreasonable comments that are likely to make a person to leave a discussion [Morzhov, 2020].	Abusive language can have toxic comments but Not all toxic comments contain abuse towards a person.
Profanity	It refers to either offensive or obscene words or phrases Dictionary.	Abusive language is a broader term which encompasses profanity.

Table 1.1: A comparison of some hateful terms with abusive language.

community who understands and makes positive use of social media platforms is becoming more and more tolerant towards the display of online hateful manners.

Hate speech is an extremely complex notion. Judging whether a message contains hate speech is quite subjective in nature, depending on the nature of given phenomenon. The term ‘Hate Speech’ is most frequently used by the research community as it provides a broader perspective of what hate is and also encompasses various insulting terms created by the online users to convey hateful messages. Also, there are factors which might influence the presence of hateful content such as discourse context, domain of an uttered text and also in a context of the occurrence of other media objects (audio, video, images etc.). The targets of hate speech, events that might cause a burst of hate against a community, identity of the author and time and place of posting such content are also important factors while considering hate in online messages [Brown, 2017].

Hate Speech online typically concerns how various social media groups and communities develop a relationship while exploring these platforms. Hate speech is not always necessarily toned in an aggressive or offensive way. Rather, it is characterized by an explicit call to violent actions [Poletto et al., 2017] and often considered a reason for hatred towards the vulnerable communities [Izsak-Ndiaye, 2015]. This issue needs to be addressed at a global level to counter the growing hate. It has been observed that this increase in online hate may result in violent acts against people, bringing very serious social consequences [Mossie and Wang, 2019, O'Keeffe and Clarke-Pearson, 2011].

Hate speech as an offensive phenomenon directed at specific targets is a challenge despite technological advances. *Online hate speech* (or *cyber-hate*) takes different forms and its rapid growth raises concerns that it may be a catalyst for harmful behavior [Izsak-Ndiaye, 2015]. Fortuna and Nunes [2018] mentions several definitions of hate speech. Although, there are several definitions of what constitutes hate, still there is no formal agreeable consensus on HS definition [Ross et al., 2016]. Therefore, in order to develop systems that automatically detect hate, we need to clearly define what HS is, as it becomes easier for a system to detect hate and function properly [Ross et al., 2016]. Hate speech usually refers to disparaging individuals or groups because of their ethnicity, gender, race, color, sexual orientation, nationality, religion, or other similar characteristics — see for instance the U.S. constitution [Nockleby, 2000]. Spertus [1997] in his earlier work on hate speech termed it as a hostile message, abusive message or flames. Many authors recently employed the term cyberbullying to represent online hate [Xu et al., 2012, Hosseinmardi et al., 2015, Van Hee et al., 2015]. The more famous and common term 'Hate Speech' is used by many researchers [Warner and Hirschberg, 2012, Burnap and Williams, 2015, Gitari et al., 2015, Kwok and Wang, 2013]. Razavi et al. [2010] called hateful messages as offensive language. The work on hate speech detection by Sood et al. [2012] called it insults, profanity, posts that intend malicious purposes. Therefore, it is difficult to develop automatic systems that determine whether a message contains any fragments of hate speech. This has also, unfortunately, made it difficult for the state-of-the-art HS models to effectively combine and compare the performance of detection systems developed for different types of online abusive content (e.g., hate speech, aggressiveness, misogyny etc.) in social media datasets.

Several social media platforms, such as Twitter[2], typically implement their own definition of hate speech. For instance, the definition of HS in case of Twitter is:

> *"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national Twitter origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease" [Twitter, 2017].*

Similarly, Facebook [3] which is another popular social media platform implemented its own definition of hate speech:

> *"Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability*

[2] https://www.twitter.com

[3] https://www.facebook.com

> *or disease is not allowed. We do, however, allow clear attempts at humor or satire that might otherwise be considered a possible threat or attack. This includes content that many people may find to be in bad taste (ex: jokes, stand-up comedy, popular song lyrics, etc.)".*

YouTube [4], a video streaming platform, defined hate speech as:

> *"We encourage free speech and try to defend your right to express unpopular points of view, but we don't permit hate speech. Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their ethnicity".*

A code of conduct was signed between the EU and popular social media companies such as Facebook, Twitter, YouTube and Microsoft Commission [2016]. They defined a common definition which states that is:

> *"All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnicity"* [5].

If we analyze all the definitions described above, we can find some common dimensions:

Targets: By reading any definition of hate speech, it can be realized that HS always has specific targets. These targets can be individuals or group of people belonging to *vulnerable categories* defined based on certain attributes: people from a certain community, race, religion, ethnic origin, country, demographic background or other factors.

Violence or Hate: We can also observe that in all the definitions, with slight variations, the hate speech always *incite* hatred and violence towards certain communities including minorities.

Attack or Diminish: It is also stated by many definitions that the purpose of hate speech is to attach or diminish the targets by using extreme and abusive words and derogatory terms.

Humor: Additionally, in some definitions as the one by Facebook, there are textual messages which are either slightly offensive or expressed in a humorous way. Such contents are allowed. Let us observe that this exception of humourous contents as allowed contents makes it really hard for the systems to identify abusive contents. The border between what is allowed and what

[4] https://www.youtube.com

[5] https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code

is not allowed is, indeed, very blurry in such cases, and also when humor is present we can identify posts where abuse is implicitly expressed, which are not harmless.

In order to understand how HS behaves in online conversations, we can implement several definitions of HS in an automatic HS detection task. Consider some examples taken from a hate speech dataset downloaded from a social media platform.[6]

> *#Brexit Facts about Pakistanis 1) Everyone hates them 2) Pakis only hate Hindus 3) Pakis dogs of Arabians*

> *Put a loving face of a raping murdering savage refugee terrorist up. https://t.co/raMdb5wJBK*

> *HillaryClinton a liar and a criminal who supports globalism and potential Syrian illegal immigration terrorist https://t.co/p2WqbP9AST*

> *thats great they can always go and fuck goats with there muslim terrorist friends https://t.co/sE4QQKKAJH*

Because abusive language and HS are generally subjective in nature, there might be highly polarizing topics or events involved in abusive language datasets. Therefore, we need novel approaches to model the conflicting perspectives and opinions in data annotations coming from people with different personal and demographic backgrounds. These conflicting perspectives might raise issues concerning the quality of annotation itself and might also impact the gold standard data to train Natural Language Processing (NLP) models created for training data to detect abusive content. The annotators might also show different sensitivity levels against the particular forms of hate, which result in low inter-annotators agreements. The online platforms used for abusive language annotation do not provide any background information on the annotators and the personal opinions and views of the victims of online abuse are often ignored while annotating these datasets for HS detection tasks against these victims.

However, when we consider the case of HS datasets, there are also further task specific relevant issues that need to be considered when we reason about different perspectives and disagreements among the annotators on the presence vs. absence of online hate. This is typically true in toxic online environments, where the discussion often turns into abusive expressions of hate against various targeted communities and vulnerable groups.

1.2 Open Challenges in Abusive Language Detection

In this section, we introduce several open challenges in abusive language and hate speech detection, that guided this thesis path, which covered the following themes.

[6] The examples in this work are included to illustrate the severity of hate speech problem. They are taken from actual web data and in no way reflect the opinion of the author of this work.

1.2.1 Hate Speech against Minorities

Since 2015, because of the political conflicts in Libya, Iraq, Afghanistan, and Syria, thousands of families are migrating and an estimated more than one million asylum seekers traveled to Europe alone (European Union, 2015). When the world (especially America and Europe) witnessed the refugee crisis at its peak, we saw an immense increase in hate speech against refugees, immigrants and migrants. This alarming situation has specifically concerned the United Nations [Fuchs, 2019]. As International law does not properly define hate speech or xenophobia, human rights treaties globally devised some policies and rules to protect refugees from the offline and online hateful speech. These policies often contradict with protection of the freedom of expression which is also protected by international law.

With the start of refugee crisis in Europe, the far-right political parties which were not very popular in the political spectrum, took it as an opportunity to criticize the migration policy. They successfully introduced their political complains by populating the common fears to xenophobia, Islamophobia, and the other aspects of migration within the general public [Vrielink, 2014]. The use of hate, discrimination against immigrants has also resulted in the spread of misinformation, which produced a hostile environment against the migrants and turned out a success for their electoral campaigns [Mudde and Kaltwasser, 2018]. This increase in hate speech and racism was mentioned in the annual report of the European body in 2016 as "racist insults have become increasingly common and xenophobic hate speech has reached at unprecedented levels" (ECRI, 2017). Similar strategies are now becoming popular in other parts of the world and the hate speech against migrants and refuges is becoming increasingly challenging and important to consider to understand the severity of the issues.

There is an urgency to re-define the laws against abusive language and speech and regulate the spread of these phenomena. The efforts to regulate the policies of hate speech against minorities are extremely challenging and fail due to the fact that the consequences of hate speech on the victims' physical and mental well-being and how it can diminish the victim's social standing are not considered. Furthermore, the immediate harm caused by the hate crimes are ignored and often assumed that such harm will not affect the victims with the possibility to retaliate against the harm inflicted, implying the possibility of the status of immigrants in the society as any other group, ignoring the fact that there exist historical and contextual disparities and unequal power structures in a society [Waldron, 2006, Benesch et al., 2016].

It has been observed that in real life hate incidents, when the members of any vulnerable community such as immigrants, are targeted, the victims have an opportunity to express their views and opinions on hate incidents. In such cases, the background of people involved in these crimes including the people responsible for these crimes is often available. However, we do not often see the same scenario in online hate speech especially when the datasets are annotated for hate speech detection and classification tasks by crowdsourced workers. A recent study shows that a boost in online hate over the years in specific geographic locations can result in an increase in offline hate crimes and physical violence in same areas [Relia et al., 2019].

The governments and policy makers are facing the challenges of regulating hate speech against minorities and general public. We believe one of the main reason in failing to come to an agreement between the stakeholders is not understanding point of views of the victims. The victims suffers physically and psychologically from the effects of hate as it can diminish their esteem, self-respect and confidence levels as a member of the society. We need to involve them not only in policy making but also in the process of detecting online hate. Our research work will facilitate the process of identifying hate speech against the victims by involving them in data annotation process that will help to understand their perspective and how these minorities, especially immigrants and migrants, feel when they are victimized and targeted online and offline.

1.2.2 Bias in Social Media Datasets

Due to an extensive use of Internet and the dissemination of information at a global level, the content from heterogeneous sources are easily circulated with the help of various platforms. An unrestricted and unlimited access to these contents is provided to people around the globe. One can argue that this free flow of information is beneficial for the mankind to gain knowledge on the topics of global importance, however, in recent years, it has also encountered the problem of disinformation, i.e., "false, inaccurate, or biased information propagated with the intention of causing harm or for profit purposes" [Zhang and Luo, 2018]. These developments have raised concerns at the center of public and academic attention; the issue of credibility of information, which involves the quality and veracity of the disseminated information [Metzger, 2007, Jasanoff and Simmet, 2017]. The policy makers, scientist, and activists are arguing to address the challenges related to the credibility of information by identifying and countering the so-called fake news, hoaxes, rumours and political and personal bias [Venturini, 2019]. However, the problem of mitigating bias and differentiating between true and false information or exposing the ideological bias is a difficult task to tackle [Graves, 2016]. We need computational solutions with the adequate training and evaluation solutions that can capture the specifics of bias in the data.

The abusive language and HS detection tasks in natural language processing require large datasets. The widespread approach for creating new datasets is through crowdsourcing platforms, where the labelled examples are generated by workers [Zaidan and Callison-Burch, 2011, Richardson et al., 2013, Trischler et al., 2017]. Commonly, a small group of workers are asked to produce quality annotations by labeling the dataset examples with specific guidelines. Because of the availability of only a few workers for the annotation process to annotate the dataset examples, data diversity and the ability of models to generalize can be affected. If the language patterns adopted by an annotator during the annotation are consistently correlating with the labels, a machine learning model can pick up on those, leading to an over-estimation of the model performance with biased evaluation.

One of the challenges faced by researchers in developing the annotated corpora is the level of inter-annotator agreement (IAA) — the extent to which different annotators "agree to something" when annotating the examples of a corpus. The term often referred to as annotator bias is considered, among the others, as the differences between the individual preferences of various annotators on a given topic [Artstein and Poesio, 2008]. Some common approaches to control bias include the development of comprehensive annotation guidelines, detailed and explicit manuals, and extensive training. Nevertheless, the differences between the individual preferences of various annotators still exist and need to be tackled.

In this thesis, we will tackle the problem of annotator bias by providing new qualitative and quantitative methods and by performing the experiments on abusive language datasets to evaluate the effectiveness of employed methodology.

1.2.3 Polarized Opinions

The opinions expressed by users about a social phenomenon in an online discussion might differ from the opinions of other people. The background knowledge and the prior experiences of users may shape their current understanding on the concepts and beliefs related to a topic and influence their perspectives in giving opinions on the topic of mutual interest. The HS and abusive language detection are formulated as text classification tasks. As a machine learning technique, text classification is considered as a fundamental task in natural language processing with broader applications such as sentiment analysis, topic labeling, spam detection, and abuse detection. The task assigns pre-defined categories or labels by structuring, organizing, and categorizing textual data from many online and offline sources such as documents, medical files etc., and all over the web [7]. Similarly in a broader perspective, the HS or abusive language detection task requires a textual instance or a phrase as an input, analyzing its content, then automatically assigning the relevant tags to the text, such as hate vs. not hate.

The issue needs considerable attention to protect the disadvantaged social groups. While some countries define some expressions of hate speech as illegal in their laws and regulations [Abbondante, 2017], a culturally shared definition of what constitutes hate speech is still under debate. The highly polarizing nature of the topics involved raises concerns about the quality of annotations these systems rely on, because not all the annotators are equally sensitive to different kinds of hate speech.

We assume that the deviating opinions on a dataset given by different groups of annotators are a source of valuable information and should not be considered as noise in the gold standard data. This information can help to create better quality data to train machine learning models for the prediction of highly subjective phenomenon such as cyber-hate. Such varying responses expressed by human judges can impact the level of agreements between the human annotators resulting in polarized judgements when they asked to manually annotate the datasets of a particular domain. The topics which create polarization among the online users are often controversial. Hate speech is also a highly controversial topic. The idea that the disagreement among the annotators is not noise but rather resourceful knowledge which can help to further study the phenomenon to create better training sets is studied by Soberón et al. [2013].

In this study, we model the polarized opinions expressed in data annotations of abusive language corpora by devising new methods. All the open challenges listed above are explored in the form of research questions presented in the next section.

1.3 Research Questions and Objectives

The ambiguous nature of polarized opinions in data annotation is one of the challenges faced by a general abusive language detection task. On the one hand, the polarization expressed by annotators could help an abusive classifier to spot the opinions in abusive content. On the other hand,

[7] https://monkeylearn.com/machine-learning/

such information can cause confusion for an abusive language classifier at run-time. The remaining challenges are related to bias in current social media datasets where the abusive content stems from the conversations between people belonging to different social and cultural communities. Therefore, developing an abusive language detection model needs to consider the issues related to opinion polarization and racial and cultural bias in the data. Considering the impact of these challenges on abuse detection, the main objective of this thesis is to investigate the possibility and challenges to build a robust model to detect abusive language in social media that considers the opinions of victims and tackle the polarization during the detection process. To tackle these challenges, in this thesis, we propose four main research questions, with related sub-questions. The research questions with the quantitative and qualitative approaches are discussed below:

RQ1: How to measure the opinion polarization in data annotation of abusive language corpora?

- How the availability (or absence) of information on the personal or socio-demographic background of annotators impacts the measurement of the level of polarization among the annotators, when their judgements reflect different opinions, as in the case of abusive language and HS corpora?
- Can we automatically partition the annotators into homogeneous groups based on the polarization of their judgments which reflect different perspectives?

RQ2: How high level of polarization in data annotations can influence the training datasets?

- Can we manually explore the datasets with respect to the measured level of polarization?
- Does this exploration of the individual instances of subjective content help to understand the topics and issues involved with polarizing nature?

RQ3: How to build a robust model which facilitates the modeling of polarized opinions for detecting abusive language across different topical focuses and targets?

- Can we improve the classification performance of a machine learning model by introducing training sets with such polarized opinions?
- Is it possible to effectively represent different perspectives expressed by annotators in polarized opinions in an inclusive model?

RQ4: How to mine annotator's perspectives in abusive language corpora?

- Can we identify the conflicting perspectives expressed by annotators in abusive language corpora with quantitative (feature-based) methods?
- Can we cluster the annotators into groups by identifying the individual and shared perspectives with clustering techniques?
- What are the best strategies to visualize the identified perspectives to understand the controversial topics and events?

1.4 Contributions

Automatic abusive language detection is organized as a text classification task. We develop the NLP models to differentiate between hateful and non-hateful, sexist and non-sexist, racist and non-racist, homophobic and non-homophobic, aggressive and non-aggressive, offensive and non-offensive, and stereotype and non-stereotype content in a binary classification scenario. However, It is not an easy to recognize and eliminate the abusive content and behaviour in social media conversations. Common approaches to tackle the problems related to online abuse are detecting abusive language automatically by developing robust models to detect abusive language across different domains and languages.

This section provides details on our contributions related to the abusive language and HS detection tasks with a focus on opinion polarization and mitigating the bias in data annotation by involving the members of vulnerable communities as the victims of hate and abuse, the goals we desire to achieve by performing the tasks on available datasets, and the methodology applied to perform the desired tasks. Finally, we give the details on the tasks evaluation within the context of present research.

1. We present an overview of the current approaches to deal with online abusive language and HS detection in a multilingual settings. We did not find many research studies on robust models which deal with opinion polarization and bias in the datasets that can affect automatic online abuse and HS detection, despite the introduction of several benchmark datasets in recent studies. Overall, we conclude that developing a robust model in a context where the bias and polarized opinions are modeled is challenging. We came across several issues which add up to the difficulties of this task, and can be considered as the main objective for future research.

2. The polarized opinions and annotator bias in annotated datasets could become a problem in abusive language detection task across several languages. The measurement of such differentiating opinions could be valuable information to handle the data inconsistencies or it can cause confusion for a classifier during the learning process affecting the classification performance. Therefore, modeling polarized opinions at instance level is important for building a robust model to detect abusive language detection. To this end, we conduct several investigations:

- We develop novel methods which include a polarization index to measure the level of polarization in opinions expressed by annotators in the individual instances of abusive phenomena when the annotated data is crowdsourced.
- We collected several publicly available datasets of abusive language in social media, covering phenomena such as hate speech, sexism and offensive language.
- In order to measure polarization, we need to divide the annotators into homogeneous groups. The polarization index is a useful tool that aided in the division of annotators into groups based on their judgements reflected in data annotations.
- We develop and experiment with supervised learning models for automatic abusive language and HS detection. We manipulate the training sets by removing the most polarized

instances from a set and duplicating the less polarized instances. Such models are trained on the benchmarked abusive language corpora to predict abusive contents from the Twitter datasets.

- We amplify the role of polarized opinions in abusive language detection task, by incorporating the knowledge stemming from the conflicting opinions, to improve the performance of abusive language detection.

3. Abusive phenomena are subjective in nature. This means that the quality of data annotation on these datasets is either based on or influenced by personal feelings, tastes, or personalized ideas. We need to develop models which can consider these properties of data annotation in the learning process. We conduct the following tasks to achieve this objective:

- We partition the annotators into groups with the help of P-index. The task implies that the datasets we employed are based on crowdsourcing and we do not have any information on the annotator's background.

- We create separate gold standards for developed annotator groups. The rationale behind this approach is that the data annotated by human judges reflect different opinions and we need to model these opinions separately before they are averaged out in gold standard creation. We also propose an ensemble classifier which at run-time considers all the learned perspectives in an inclusive fashion.

- We experiment with state-of-the-art transformer based BERT model and train two separate classifiers on the gold standards crated to represent opinions of different groups of people and perform classification tasks to measure the performance of perspective-aware supervised models in multilingual settings.

4. The corpora annotated by workers from crowdsourcing platforms, in general, do not have any information on the background of annotators. In order to validate the P-index and expand our analysis on the polarization of opinions in data annotation, we develop a novel resource, a multi-perspective dataset, to understand how the different perspectives are reflected in the annotations. To achieve this objective, we performed the following analysis:

- We develop a novel multi-perspective abusive language Twitter dataset on the Brexit referendum. We clean and filter the dataset to represent Islamophobic, xenophobic and hate speech contents against immigrants and Muslims.

- The dataset is divided into four sub-categories which are hate speech, aggressiveness, offensiveness and stereotype and manually labelled by human annotators and the complete information on the background of annotators is available.

- As a novel approach, we involve migrants as the victims of abusive language who volunteered to annotate the datasets. This helped us to capture their opinions. It is an important step because we believe that the benchmarked datasets available for the abusive language and HS detection fail to model and understand the feelings and emotions expressed by the victims of hate speech.

- We perform binary classification on our datasets and then evaluate the performance of different classifiers. We also employ an ensemble approach which combines all the perspective-aware classifiers into a single inclusive model.

5. We explore aforementioned contributions into a specific kind of abusive language analysis. We focus more on the annotated datasets and design experiments to mine annotators perspectives expressed in data annotations who belong to homogeneous groups. The work is more focused on identifying and understanding individual and shared perspectives which influence the judgements of annotators. We also investigate to understand how a specific background can influence the decision making of annotators belonging to same group. To achieve this, the following steps are taken:

- We develop quantitative feature-based approaches to mine annotators' perspectives expressed in the annotations of abusive language corpora.
- We add extra datasets available online to further expand our analysis.
- We implement clustering techniques to cluster the annotators into groups based on the identified perspectives by measuring the feature-based agreements.
- We present a comprehensive qualitative analysis of these datasets and visualize them using word clouds which provide us a deep and thorough understanding of individual and shred perspectives.

1.5 Thesis Structure

The rest of this chapter provides a brief overview of each chapter included in this thesis. The core of the manuscript consists of the chapters that comprise most relevant research articles submitted, accepted for publication and published during the duration of the PhD research period. It includes a paper submitted to an international journal, two published conference papers, one accepted workshop paper, and also further work which was still not included in the submitted research papers.

Chapter 2 Literature Review

This chapter contains a brief review of the research studies on abusive language and hate speech detection in social media platforms, specifically focused in a multilingual environment. We present an extensive study to obtain the current open problems with respect to the technological and sociological aspects of these tasks. This chapter introduces some state-of-the-art approaches developed to tackle the difficult tasks of abusive language and HS detection. In addition, we also provide related studies in the use of opinion polarization for different NLP related tasks.

Chapter 3 Identifying Polarized Opinions in Abusive Language Detection

This chapter presents the research work published at the 18th International Conference of the Italian Association for Artificial Intelligence (AIxIA) in 2019 [Akhtar et al., 2019]. However, some of the contents and experiments presented in this chapter are not yet published/will be published. In this chapter, we deeply investigate the role of polarized opinions in abusive language and HS detection tasks. We develop novel methods to measure the level of polarization in data annotations expressed by annotators from crowdsourcing platforms. We also develop supervised learning models to study the impact of most polarized opinions in the automatic classification of abusive content. When the annotators annotate a dataset, they may differ in opinions about a particular topic such as hate speech based on certain factors. Although, most of this information is lost during the creation of gold standard data with majority voting, still, If the annotators differ extensively in their opinions (high polarization), this can also affect the learning process of an algorithm and may cause confusion at the time of topic classification. Furthermore, some annotators in a group might not have either extensive knowledge of the domain in question or experience in data annotation so the quality of the training set might be compromised. If we can identify those annotators, then removing these annotators from the group may improve the quality of underlying data and also provide a boost in the classification performance. We design several experiments in which we apply the methodology discussed above to study its impact on the classification performance.

Chapter 4 Modeling Annotators' Perspectives by Developing Group-based Models and by an Inclusive Classifier

This chapter consists of the research work published at The 8th AAAI Conference on Human Computation and Crowdsourcing (HCOMP) in 2020 [Akhtar et al., 2020]. This chapter is mainly focused on the investigation of building robust perspective-aware models to detect abusive language and HS in social media data. We characterise the available datasets by diving the annotators into groups based on the opinions expressed by annotators in data annotation. We also present a supervised learning architecture with state-of-the-art BERT transformer based model which detects the abusive language and HS for separate groups of annotators. Finally, we provide a qualitative exploration of the datasets used in this research study to understand how topics and events with polarizing nature and high subjectivity can influence the opinions of annotators.

Chapter 5 Case Study: The Development of a Multi-perspective Corpus of Abusive Language on Brexit

This chapter presents the research work accepted with revisions in Information Processing and Management Journal and archived in [Akhtar et al., 2021]. In this chapter, we develop a corpus of abusive language which is annotated manually by six annotators for different abusive language categories. The dataset is multi-perspective in nature. The targets of online hate are often vulnerable groups, such as immigrants, migrants and Muslims. To understand their point of views, a group of migrants annotated the dataset. We called it multi-prospective because the data annotations represent perspectives coming from people with different social and demographic backgrounds. The factors behind such hate are often ethnicity, race, gender and discrimination among the people based on these physical characteristics. We focus on investigating the objective of building a robust model to detect hate speech representing these perspectives from different groups of annotators with different cultures and demography. Finally, we also provide a deeper qualitative analysis the dataset.

Chapter 6 Mining Annotator Perspectives in Hate Speech Corpora with Clustering Technique

This chapter describe the work published at the Natural Language for Artificial Intelligence workshop co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (NL4AI at AIxIA 2021) [Fell et al., 2021][8]. We mainly focus on mining annotators' perspectives from HS corpora. We develop a feature-based quantitative method which can automatically identify individual and shared perspectives expressed by annotators that belong to a similar group. We used clustering techniques to partition the annotators based on feature agreement between the annotators within a group. We also introduce the concepts of individual perspectives and shared perspectives within the context of data annotations for HS corpora. The identified individual and shared perspectives are then visualized with the help of word clouds. Finally, we provide a comprehensive qualitative analysis with many examples from each dataset in English and Italian languages.

Chapter 7 Conclusion and Future Work

This chapter outlines the conclusion of this thesis. It also includes a list of publication derived from the thesis, and describes some possible directions for future work.

[8] The first author of the paper is responsible for devising and designing the methodology presented in this chapter.

Chapter 2

Literature Review

This chapter gives a comprehensive review of the current state-of-the-art in abusive language and HS detection. The chapter throws light on recent machine learning methodologies to detect HS and abusive language online, with information on the possible application domains that will benefit from its integration and development. The corpora developed in various research studies to detect HS and abusive language online are briefly described. A high-level overview of polarization in opinions is provided with details on bias in social media data and its impact on data annotation to develop training sets.

The chapter is divided into the following sections: Section 2.1 introduces the HS and abusive language detection tasks in general. In Section 2.2, we give an overview of the modeling approaches applied to abusive language and HS detection. In Section 2.3, we give a short overview of computational linguistic approaches applied to text classification tasks. In Section 2.4, we give an overview of the resources and benchmarks required for abusive language and HS detection. Meanwhile in Section 2.5, we present an overview of the issues related to aggregation and agreement and we explain different strategies applied to data annotations, the common agreement measures between the raters and problems with them. In Section 2.6, we discuss the opinion polarization in social media datasets and in data annotations and also the strategies used to model these polarized opinions in training datasets to boost text classification performances. Finally, in Section 2.7, we provide a summary of the findings.

2.1 Abusive Language and Hate Speech Detection

Recent years have seen a massive increase in the use of social media around the world. Due to the arrival of new social media platforms, the number of social media users increased globally. Users on a social media platform are free to publish content they like or the information that comply o their personal or political affiliation. Since the identity of online users is often secret, the opinions are expressed without any fear of being caught or held accountable. Also, the majority of online users are not familiar with laws that protect online users against the online frauds, hate speech and foul language [Izsak-Ndiaye, 2015]. This resulted in the spread of hate against people. Although these platforms implement actions to moderate the contents, which involves teams of human moderators, it is not always easy to manually identify content which are aggressive, threatening, abusive or hateful. The authors in Raisi and Huang [2016] studied social and political events that might be responsible for hate in general public.

A simple click by a user can post anything across the globe and this information can reach millions in no time. The receiver of a message, if the message or content adhere to his personal beliefs and thoughts, may forward or re-post it without bothering to verify the authenticity of this information and thus the information becomes viral in very short period of time. So someone's freedom of expression can seriously affect others psychological or emotional state of mind and personality. It has become very hard to maintain a complex balance between free speech and the defense of human dignity, as these platforms are the grooming grounds for discourses that are harmful to certain groups of people. The challenges of tackling online hate speech has been rapidly recognized as a serious problem by the scientific and political community with the adaptation of digital social science methodologies and computational approaches.

The hate speech detection task has some similarity with the sentiment analysis and opinion mining task, where the goal is to automatically classify the opinions expressed by online users into predefined categories reflecting the negative or positive polarity of the user's sentiment, which is also highly subjective. The work in Benesch et al. [2016] focused not only on detecting HS but also finding the counter measures based on certain social or political events that actually triggered HS among the general public. Most of the classifiers proposed in the literature for sentiment analysis are based on the supervised machine learning approaches, which requires large corpora annotated by humans [Bosco et al., 2013]. The data is used to create gold standards and benchmarks to train machine learning models for specific classification tasks and compare the results of different systems. The data for HS detection is searched, filtered and then downloaded from various social media platforms by using special keywords related to abusive language and HS. The downloaded data is then cleaned and converted into a corpus for annotation process.

The technologies being developed for HS detection mainly employ supervised machine learning approaches in Natural Language Processing (NLP). Training such models require manually annotated data by humans, either by crowdsourcing paid workers or by domain experts, for training and benchmarking purposes. The NLP being a a sub-field of artificial intelligence, helps to understand human language to automatically perform various tasks such as machine learning, machine translation, summarizing, question answering, ticket classification, and spell checking. The NLP techniques are critical to online businesses and social media platforms as it can be used to analyze huge volumes of textual data, including social media comments, online reviews, news reports, abusive content detection, and many more. Human language is quite complex in nature. It is ambiguous, disorganized, and diverse which makes it difficult for the machine learning algorithms to interpret and understand it. Even it is not possible for humans to understand and make sense of the language without proper knowledge. Similarly, for machines to understand natural language, the language must be transformed into something that can be interpreted by a machine. In an NLP system, syntactical and semantic analysis are important elements in understanding the grammatical structure of online text in order to identify how words are related to each other in a given context. But, it is not a straightforward process to transform text into the language that machines can easily understand and process.

For any abusive language detection system to automatically classify online abusive content, we need to extract, collect and label the data from social media platforms [Richardson et al., 2013]. The collected dataset is either in the form of Twitter tweets, Facebook conversations, or Reddit messages. The data is in raw form and can not be processed for a detection task. We need to clean the dataset by performing several pre-processing steps to prepare it to train the machine learning models. Data pre-processing is a data mining technique which helps us to transform the raw data into a machine understandable format. The data we gather from social media is often incomplete, inconsistent, and/or lack certain trends and behaviours with many errors [Singh and Bhatia, 2011]. By pre-processing the data properly, we can reduce the noise in text which can improve the performance of a detection system and also speed up the whole process, resulting in real-time abusive language detection. The basic pre-processing of textual data involves several steps which include:

- Cleaning the text extracted from online media.
- Removing white spaces within the data.
- Expanding abbreviation if present in the raw data
- Removing non-alphabetic signs.
- The process of stemming and lemmatization i.e, reducing the words to their stems.
- Removing stop-words and non-language characters.

The pre-processing steps commonly used in text classification to prepare the raw data for further process are shown in Figure 2.1. It is often a good idea to label the data first before pre-processing so that the important information might not be lost from the data and the annotators can easily understand the messages within a proper context.
Once the dataset is cleaned, annotated and ready for further processing, we develop gold standard training data based on the majority annotations and give the input data to a model for training purposes and then detecting the abusive content. Figure 2.2 explains the architecture of a simple approach for an abusive language detection system.

2.2 Models for Abusive Language and Hate Speech Detection

In order to perform abusive language or HS detection, we need machine learning algorithms which are trained on the gold standard training datasets. The gold standard data are created by aggregating the individual annotations of all the annotators who annotate the dataste and then a baseline is defined and after the classification experiments, the evaluation metrics are compared with baseline results to evaluate the performance of a machine learning model.

The common approach used by the researchers for abusive language detection is building a machine leaning model that can automatically classify a text as either abusive or not abusive or no hate or hate speech. For data annotation process in a general purpose detection task, the researchers use either experts or amateur annotators to annotate the corpora either downloaded from social media or publicly available. Fortuna and Nunes [2018] in their recent survey on HS literature provided a deep and thorough analysis on the issues and challenges faced by the

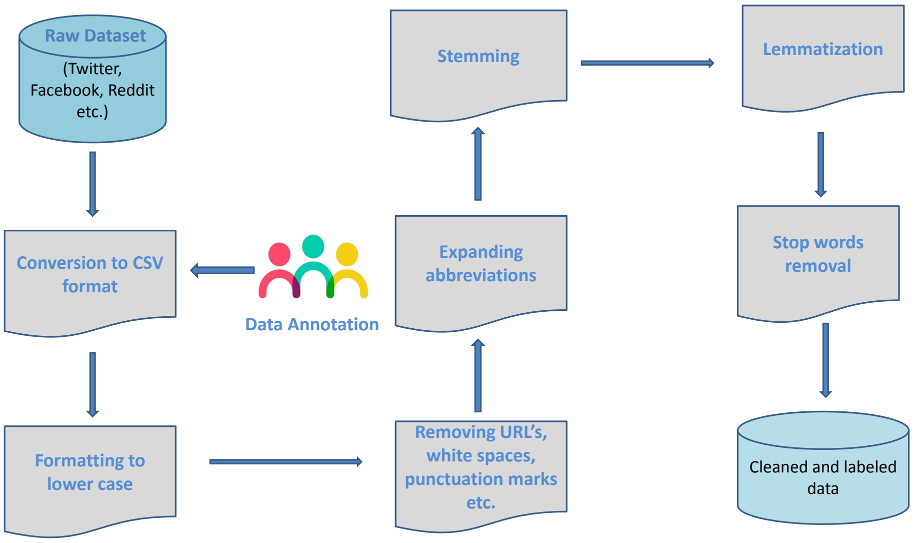


Figure 2.1: The pre-processing steps to clean a social media dataset.

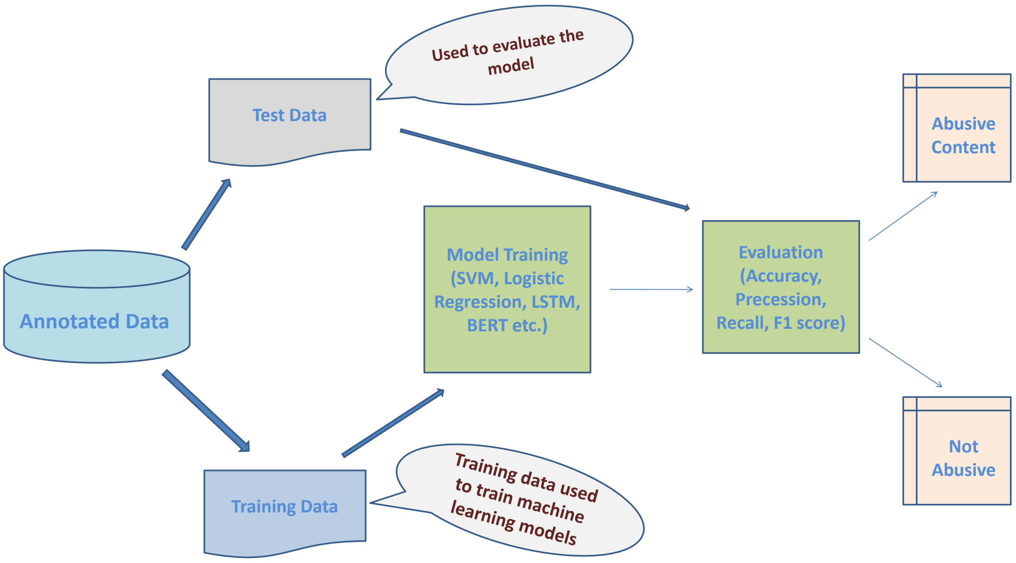

Used to evaluate the model
Test Data
Abusive Content
Model Training (SVM, Logistic Regression, LSTM, BERT etc.)
Evaluation (Accuracy, Precession, Recall, F1 score)
Annotated Data
Not Abusive
Training Data
Training data used to train machine learning models

Figure 2.2: The architecture of a simple approach for abusive language or hate speech detection task.

researcher community in automatic HS detection. They also addressed the issue of the availability of high quality datasets for benchmarking and training the models for HS detection. The majority of such computational approaches include ‘Deep Learning’, but also ‘Support Vector Machines’, ‘Random Forest’, ‘Logistic Regression’ and ‘Decision Trees’ [Fortuna and Nunes, 2018, Schmidt and Wiegand, 2017]. In these systems, different types of machine learning approaches are employed. Among the common ML approaches employed by the researchers, they are either supervised, semi-supervised, and or unsupervised approaches. A recent survey on such approaches revealed that most of the studies adopted supervised methods (73%). The analysis showed that these models perform well and achieve high accuracy, and there is no substantial evidence that one approach can be used in favor of another as each approach depends on the data, the underlying task, and the context in which a dataset is available (e.g., availability and quality of training samples). These factors can play a role in a decision to chose one category over the other. An as example, Chen et al. [2012] employed an unsupervised method and lexical resources and syntactic features to achieve very high accuracy up to 98%. Similarly, several other studies were based on supervised and semi-supervised approaches achieving better performances [Waseem and Hovy, 2016, Akhtar et al., 2019, 2020]. However, it is evident from the analysis that the supervised learning approaches are more common and favourable among the NLP researchers. The reason might be the availability of the benchmarking datasets and machine learning/deep learning platforms that promote supervised approaches.

Hate speech detection is often approached with similar techniques to sentiment analysis however, online hate can be characterized by incitement to hate and to violent acts [Sanguinetti et al., 2018], rather than just a display of emotions. As a consequence of inter-annotator agreement issues, the benchmarks based on the datasets created with traditional methods may be inadequate, leading to unstable results. In particular, when the inter-annotator agreement are computed for the subdivisions of an annotator set, and on the level of polarization of the human judgments. Researchers who recently started tackling hate speech detection from a natural language processing perspective are designing operational frameworks for HS, annotating corpora with several semantic frameworks, and automatic classifiers based on supervised machine learning models [Fortuna and Nunes, 2018, Schmidt and Wiegand, 2017, Poletto et al., 2019].

Several campaigns were held to evaluate targeted hate in a multilingual perspective [Basile et al., 2019, Fersini et al., 2018b, Bosco et al., 2018, Fersini et al., 2018a, Sanguinetti et al., 2020]. This has encouraged the development of hate speech detection systems in languages other than English, such as Spanish and Italian, the comparative study of different abusive phenomena with different targets, such as sexism and xenophobia [Plaza-del Arco et al., 2020], and the comparative study of the same abusive phenomenon across languages [Pamungkas et al., 2020a, Lazzardi et al., 2021]. The work by Warner and Hirschberg [2012] used the term hate speech and focused on collecting HS messages against Jews from various social media sites and classifying HS based on stereotypical words commonly used in antisemitic manner. Benesch et al. [2016] developed a system which not only detects HS on social media text but also provides counter-narratives against such hate arising from certain political events. While there exists legislation governing hate speech, often social media platforms implement their own regulations. The increase in the number of social media users also provided new opportunities to discuss events and topics and then collect and analyze rich data on abusive language including online harassment and hate against women and communities often targeted based on their ethnicity and gender [Duggan, 2017, Waseem, 2016].

Literature on hate speech detection in NLP has been recently surveyed by Fortuna and Nunes [2018]. Scholars addressing the task utilized various surface level features such as bags of unigrams and ngrams [Chen et al., 2012, Xu et al., 2012], syntax, lexical semantics, and combinations thereof [Warner and Hirschberg, 2012, Sood et al., 2012, Van Hee et al., 2015, Waseem and Hovy, 2016, Nobata et al., 2016]. An interesting by-product of such initiatives is the creation of publicly available gold standard datasets, annotated by experts or by crowdsourcing. A large number of such models employ surface-level features [Burnap and Williams, 2015] such as Bag-of-Words (BOW) [Burnap and Williams, 2016, Greevy and Smeaton, 2004]. The BOW approaches provide good prediction performance in the form of high recall but there is the issue of high false positive because of the wrong classification of abusive words into hate speech [Kwok and Wang, 2013, Burnap and Williams, 2015]. Another common feature based approach used by authors is the application of ngrams in HS related tasks [Davidson et al., 2017, Greevy and Smeaton, 2004]. These are further divided into character based ngrams and token based ngrams and character ngrams are found more effective in classification tasks [Mehdad and Tetreault, 2016]. The TF-IDF (term frequency-inverse document frequency) based approach focus on the importance and frequency of words in a corpus [Dinakar et al., 2011]. Syntactic features play an important role in the identification of targets and hate speech by exploiting the noun and verb relationships [Bhowmick et al., 2008a].

Recently, deep learning approaches also gained importance in sentiment analysis and other related tasks [Zook, 2012]. Deep learning methods are prominently used for hate speech detection, for instance in Mikolov et al. [2013] and in Pennington et al. [2014], and among the participating systems to the SemEval 2019 Task 5: Multilingual Hate Speech Detection [Basile et al., 2019]. The neural language models have also gained popularity. These models have been effectively applied to many NLP related tasks showing substantial improvements in the performances Peters et al. [2018]. Some of these NLP tasks include sentence level inference [Brown et al., 1992], name entity recognition, and question answering [Rajpurkar et al., 2016].

Pre-trained language models has gained significant importance recently. In particular, state-of-the-art is represented by deep learning models based on Transformer networks pre-trained on large amounts of unlabelled data from the web and wikipedia data and fine-tuned on the task-specific annotated corpora. In order to detect hate speech, natural language processing techniques are developed which use machine learning models and the gold standard created might be inadequate and of low quality causing poor classification performances. Some of these pre-training based language models involve either feature-based approaches in which they only use pre-training as the extra features, and depend on the task-specific architectures such as ElMo [Peters et al., 2018]. OpenAIGPT, proposed in Radford et al. [2018], uses a multi-layered transformer-based architecture with a left-to-right approach and fine-tuning, and it is less dependent on the task-specific architectures. Howard and Ruder [2018] proposed the ULMFiT model for text classification tasks, achieving state-of-the-art performance on several benchmarks. These models are unidirectional while processing the tokens, meaning that they either work from left-to-right or from right-to-left, thus limiting the power of pre-trained language representations.

BERT is a state-of-the-art achievement in pre-trained language transformation models which used bi-directional approach instead of one side approaches as in the previous similar models [Devlin et al., 2018]. Many authors [Zhang and Luo, 2018] recently used BERT for many NLP related tasks. BERT [Devlin et al., 2018] is one of the best known Transformer-based models employing a bi-directional approach that also achieved state-of-the-art performance in text classification [Yu et al., 2019]. BERT trains bidirectional language representations from unlabelled text and it considers both left and right contexts in a layered architecture [Munikar et al., 2019]. Unlike other language models, BERT can be fine-tuned with just one extra output layer for various downstream tasks without depending on the task-specific modifications in the model architecture.

2.3 Computational Linguistic Approaches

Computational linguistics is the scientific and engineering discipline concerned with the synthesis and comprehension of written and spoken language from a computational perspective. Language from a computational linguistic perspective helps human being in understanding different linguistic patterns a language represent and provides insight into thinking and intelligent behaviour. Computational linguistics is used in building computational artifacts to produce human language in the form of instant machine translation, speech recognition, text-to-speech synthesis, interactive voice response, search engines optimization, text editing and language instruction materials. The aim of any computational linguistic approach – which has both theoretical and applied elements – is to improve the relationship and understanding between computers and basic language structure such as semantics and grammar. With the help of such approaches, the computers can build artifacts to produce and process a human language. Computers equipped with computational linguistic capabilities facilitate our interaction with machines and software to make the textual resources readily available in multiple languages from the internet. Massive amounts of written and spoken language resources are required to build such systems in both structured and unstructured formats. As language is considered a natural and most versatile means of human communication, when the computers are made linguistically competent, we can build systems for our interaction with machines and software of different kinds to meet our needs by analyzing the vast and versatile textual and other resources available on the Internet.

Some important goals of computational linguistics which interest researchers include [1]:

- Language translations between different written and spoken languages.
- Retrieving topic specific textual data from different online sources.
- Context base text or spoken language analysis, such as sentiment analysis.
- Question answering tasks, including once requiring inference with a descriptive or discursive answer.
- Text summarizing task.

[1] https://searchenterpriseai.techtarget.com/definition/computational-linguistics-CL

- Dialogue agents build for several business tasks such as making a purchase, planning a trip or answering customer's questions about the products and services.
- Developing chatbots capable of passing the Turing Test.

The task of abusive language or HS detection, within research community, is considered a as an unstructured text analysis problem. In order to extract insights and patterns from such unstructured text, we need comprehensive computational linguistic approaches to cope up with the challenges such as the context-dependent interpretation of natural language. To handle the ambiguous and variable unstructured data, text mining technologies have the capabilities to handle such problems [Irfan et al., 2019]. Natural Language Processing and computational linguistic approaches are the main pillars of text mining as with the help of these approaches, computational systems employ a number of tasks to make human languages tractable and understood by the machine [Hirschberg and Manning, 2015]. NLP researchers need rich data resources to perform desired tasks and such datasets are available in social networks. After downloading the unstructured data, we need to mine and put it into a practical use. Several researchers working on HS and abusive language detection already developed such datasets and these datasets are publicly available.

2.4 Resources and Benchmarks

In any detection task, the first and most important step is the data collection. The datasets for abusive language and hate speech are often collected from several social media platforms (Facebook, Youtube, Twitter, Reddit etc.). Here, we will discuss the available datasets from two recent survey and from other related resources on abusive language and HS corpora. The authors in both papers comprehensively explained all the available resources in a multilingual setting and in a multi-dimensional scenario.

Poletto et al. [2021] provided a comparative analysis of different strategies and methods used to design and build resources for HS detection with five main dimensions. The first dimension is the "type" of annotated corpora which represent the textual instances of the data from various resources and labelled according to one or more dimensions. They also include the "lexica", which are resources based on the words or phrases having common semantic meaning. Out of all the resources described in their survey, 56 of them are HS corpora, 8 are lexica and 4 resources are composed of both a corpus and one or more lexica. Among all the corpora, 11 are the benchmark datasets released and experimented in shared tasks. The "Topical focus" being the second dimension, refers to the specific topics in connection with the abusive phenomena addressed, also dependent on the target of addressed hateful content with a number of overlapping concepts. Another dimension is the "source" of a dataset. They found that Twitter is the most exploited source as it provided the flexibly of using relatively reduced length of texts and easy access to the availability of data: 32 resources contain textual instance in the form of tweets downloaded from Twitter. Other online resources for data collection include Facebook, Instagram, Reddit, Youtube, user's comments on the news articles, white-supremacist forum Stormfront, and the posts from Wikipedia articles. They also observed different "annotation schemes" as another dimension of HS corpora. Some corpora were annotated in a binary classification scenario, some as a non-binary scheme with more than two labels and some of the corpora featured multi-level annotation, with fine-grained schemes accounting for different phenomena. The fifth and most important dimension is the "language" of HS corpora. English language is considered as a primary language among the researchers world wide. They found that 37 out of 64 HS resources are English corpora or lexica. The remaining resources were in several languages spoken worldwide including Indian researchers with an effort to create resources with the predominance of Hindi–English code–mixed data which could be explained by the large spread of mixed forms and Hindi words written in Latin script in non-formal online communication among the Indians.

Jahan and Oussalah [2021] in a recent survey found 69 HS datasets in 21 different languages. The findings are similar to the work by Poletto et al. [2021]. The authors noticed that most of the collected datasets are from social media platforms and are manually annotated based on the task requirements. Several annotations have been carried out with experts, the native speakers, volunteers, or through crowdsourcing from anonymous users. English is the most dominant language among the research community representing 26 datasets as expected. However there are other languages as well including Arabic, German, Hind-English, Indonesian and Italian being represented in a total of 6, 3, 4, 4 and 5 open datasets respectively. Most of the described dataset in this study are available on GitHub source repositories. Therefore, almost all the datasets are publicly available with the exception of some datasets from Twitter having only Twitter Id's of the instances which should be used to retrieve the full tweet messages. In addition to the above details, the authors also provided the statistics of dataset sizes and the ratio of offensive content in each of the datasets. Around 41% of the listed datasets are relatively small in size (only (0-5)k posts). Another 14% have (5-10)k sentences. Therefore, most of the datasets (55%) can be referred as very small, indicating the fact that how difficult and challenging it is to acquire large-scale labeled data for hate speech detection.

Here, we will only list the most important datasets which are widely used by the research community. We will also provide a separate listing of the benchmark datasets employed for various shared tasks in different languages.

Table 2.1 provides an overview of the most widely used HS datasets which are publicly available. The content of the table are taken from the surveys by Poletto et al. [2021], and Jahan and Oussalah [2021] and other similar resources.

We will also list the available benchmarked datasets used in several shared campaigns. Table 2.2 provides an overview of the benchmarked datasets in different languages. The contents of the table are taken from the surveys by Poletto et al. [2021], and Jahan and Oussalah [2021] and other similar resources.

2.5 Issues with Annotation, Aggregation and Agreement

2.5.1 Data Annotation

Data annotation is the process of labelling textual corpora with pre-defined categories or labels. The process is important to crate gold standard data to train machine learning models which are then evaluated for an underlying text classification task. HS corpora are highly diversified and may contain several pre-defined categories for annotation task. There are varieties of hate

Dataset	**Language**	**Type and Source**	**Author Reference**	**Available (Github)**
Hate Speech & Offensive Language	English	Twitter approx. 25k dataset used for HS detection.	[Davidson et al., 2017]	Yes
An Italian Twitter Corpus	Italian	Twitter datasets for HS against immigrants and Roma	[Sanguinetti et al., 2018]	Yes
Twitter Hate Speech	English	Deep learning for hate speech detection.	[Badjatiya et al., 2017]	Yes
Hate Speech Dataset	English	10,568 manually labelled HS data extracted from Stormfront, a white supremacist forum	[de Gibert et al., 2018]	Yes
HateXplain	Multilingual	Multilingual multi-aspect hate speech analysis dataset	[Aluru et al., 2021]	Yes
MLMA Hate Speech	Multilingual	Multilingual multi-aspect hate speech analysis dataset	[Ousidhoum et al., 2019]	Yes
Transformers	English	Transformers based pre-trained models to perform information extraction, question answering, summarizing, translation, text generation, sentiment analysis.	[Wolf et al., 2019]	Yes
Homophobia Dataset	Italian	1859 tweets in Italian annotated as "homophobic/not homophobic" by 5 trained volunteers	[Akhtar et al., 2019, 2020]	No
Hate Speech	Hindi-English	Tweets in Hindi-English code-mixed variety, annotated as 'HS/ normal speech' by two annotators	[Bohra et al., 2018]	Yes
Hate Speech Dataset	English	27,330 tweets annotated with crowdsourcing as 'hateful [personal attack/ no]/ not hateful'	[ElSherief et al., 2018]	Yes
Abusive Language Dataset	English	Three corpora of 2 million comments in English from the news websites Yahoo!News and Yahoo!Finance.	[Nobata et al., 2016]	Yes
Hate Speech Dataset	Italian	4000 tweets in Italian, to which three different schemes are applied with crowdsourcing platform	[Poletto et al., 2019]	Yes
Hate Speech Dataset	English	6909 tweets annotated with experts and crowdsourcing as 'sexist/racist/neither'	[Waseem, 2016]	Yes
Hate Speech Dataset	English	16,907 tweets annotated with experts and crowdsourcing as 'sexist/racist/both/neither'	[Waseem and Hovy, 2016]	Yes

Table 2.1: Publicly available datasets for abusive language and HS detection.

Dataset	Language	Shared Event	Task and Focus	Size
AMI	English, Spanish	IberEval 2018	HS and Misogyny	8115
AMI	English, Italian	EVALITA 2018	HS and Misogyny	10,000
HASOC	English, German, Hindi	FIRE 2019	HS, Offensiveness	17,657
HaSpeeDe	Italian	EVALITA 2018	HS and Racism, generic	8000
HatEval	English, Spanish	SemEval 2019	HS and Misogyny, Racism	19,600
HSD	Vietnamese	VLSP 2019	HS, Offensiveness	25,431
–	, German	GermEval 2018	Offensiveness	8541
task 6	Polish	PolEval 2019	HS and Cyberbullying, generic	11,041
TRAC-1	English, Hindi	TRAC 2018	Aggressiveness	15,000
OffensEval	English	SemEval 2019	Offensiveness	14,100

Table 2.2: Different benchmarked datasets used for shared tasks in different languages.

speech annotation categories in the original datasets (e.g., hate, offensive, race, gender, sexism, misogyny, toxicity, group. target, political, etc.) [Jahan and Oussalah, 2021]. For each annotated dataset, usually the annotation framework, predefined labels used, and the number and type of annotators involved are considered. In most cases, there are three types of labeling strategies used for the annotation process. In first case, the annotations are based on the binary classification with two mutually exclusive classes with a typical yes/no labeling (e.g., hate versus non-hate, sexism versus non-sexism, etc.) [Akhtar et al., 2019, 2020]. The second strategy refers to a multi-labeling scenario in which there are more than two mutually exclusive or non-exclusive labels, such as strong hate, weak hate, no hate [Del Vigna et al., 2017] or more than two labels with multiple classes (e.g., racism, sexism, both, none etc.) [Waseem and Hovy, 2016]. The third strategy refers to a multi-level annotation scheme targeting different phenomena for a certain type of hate speech, its severity, and the target group, [Jahan and Oussalah, 2021]. This scheme is the most complex of all the annotation schemes involving a number of different traits and variability. For example, Basile et al. [2019] employed a three-layer binary annotation for HS, aggressiveness, and also specifying nature of the target (individual or group).

Since the manual annotation of a public dataset is a crucial step in the creation of language resources that are used for training the predictive models of a language, the controversial texts might cause performance issues for Natural Language Processing (NLP) approaches that rely on such supervised machine learning techniques. In NLP, one typically relies on manual human annotation in order to create reference data to train models. Most abusive language detection corpora are composed of data collected from social media platforms [Poletto et al., 2021], such as Twitter and Facebook. Most of them are collected by querying social media APIs with lists of keywords. Then, the data are annotated by human judges either relying on the crowdsourcing platforms or on the experts (often judges with knowledge of the subject). The annotated datasets are then utilized to detect the opinions expressed by online users for pre-defined categories such as presence vs. absence of a specific phenomenon (e.g., offensive behaviour, hate speech against immigrants, cyberbullying, and so on).

When the annotation process relies on crowdsourcing, in most cases, the available platforms do not provide any background information (culture, ethnicity, social background etc.) on the workers annotating the datasets. This is a limitation in general, and especially in case of HS detection. Indeed, the HS against different communities often contains stereotypical words which might arouse positive or negative sentiments or different reactions from the annotators with different backgrounds [Sheerman-Chase et al., 2011, Waseem, 2016]. Due to the constant unavailability of this information about the annotators, there are high chances of observing polarization among the annotators judgements on the same potentially abusive messages, which might result in low inter-annotator agreement. The problem of the quality of gold standard data when dealing with subjective phenomena have been investigated recently, e.g., by Basile et al. [2018b], where the manual annotation of subjective phenomena is found to be tainted by serious issues in terms of inter-annotator agreement. As a consequence, the benchmarks based on datasets created with traditional methods are found to be inadequate and leading to unstable results.

Ideally, involving the victims and targets of hate speech in the annotation process would help us to understand their views about online hate incidents. However, the fine-grained information about the cultural, ethnic, or demographic background of the annotators is usually not available, or it is not a primary factor when selecting experts or volunteers as annotators. Therefore it is important to investigate the automatic modeling of different perspectives that annotators may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. Therefore, we need supervised machine learning models which are trained to learn different points of views of human annotators on the same data in order to subsequently take them into account at prediction time.

2.5.2 Inter-rater Agreement

In order to provide measures for agreement between the annotators to evaluate the quality of an annotation process in computational linguistics, the most common measures are Cohen's Kappa (between two annotators) [Cohen, 1960], Fleiss' Kappa and Krippendorff's Alpha [Artstein and Poesio, 2008, Fleiss and Cohen, 1973] (amongst multiple annotators) adopted by the NLP researchers, as they help to assess the quality of data annotations. However, such agreement measures are not always free of shortcomings [Artstein and Poesio, 2008]. In particular, the issues with these commonly used agreement measures for datasets annotated by workers from crowdsourcing platforms has been recently investigated by Checco et al. [2017], who highlighted these shortcomings of such inter-annotator agreement measures in the crowdsourcing scenario, and proposed an improved measure to solve these problems with alternatively developed methods.

Similarly, Hovy et al. [2013] noted that how the reliability of crowd contributors may be inconsistent compared to the traditional expert annotation scenario, and proposed the MACE method to create gold standard datasets accounting for the annotators' reliability. The quality of gold standard datasets can be tested empirically as in study by Basile et al. [2018b], where the authors compared experiments on the agreements between different systems, expert annotators, and the results of crowdsourcing annotations. Soberón et al. [2013] proposed a method to leverage the disagreement of annotators as a source of knowledge rather than treating it as noise in the data. We believe that the disagreement can be leveraged too, however with a focus on its interpretation as a measure of the divergence of opinion between different annotator groups.

Recent studies on inter-annotator agreements provide us an insight into the methodology and effectiveness of the annotation process. Bhowmick et al. [2008b] used Kappa coefficients to measure the quality and reliability of effective human annotations and the resulting corpora by classifying single items into more than one category. Gold standard datasets that are used for training models in NLP are traditionally created with manual annotation, whose quality is assessed by metrics of inter-rater agreement (such as Fleiss' Kappa [Fleiss and Cohen, 1973]). Checco et al. [2017] introduced new agreement metrics that aim to account for the polarization of annotator opinions.

2.6 Polarized Opinions in Social Media

People use social media platforms extensively to communicate their ideas in complex ways. However, these communications often turn into heated arguments resulting in abusive and hateful conversations. Hate speech is difficult to define, but its presence is evident in social media platforms. Most HS detection corpora are composed of social media data collected from social networks by querying social media API's with lists of keywords. In online discussions, different groups of people may share a common belief (e.g., political left or right). The reason is that social media users may have different social and demographic background and people belonging to same community and with a common demographic background may express similar opinions on a topic of social and global interest. Because of the presence of segregated communities in online social media interaction, recent years have seen increased interest by research community in such social interactions [Marozzo and Bessi, 2017, Bessi et al., 2014, Conover et al., 2011, Nevin et al., 2017], where people having common interests interact with each other with the exposure to different viewpoints from other people. This phenomenon is commonly called the echo chamber effect which means that the online interactions among people are conducted in a polarized pattern [Prasetya and Murata, 2020].

Consequently, we need deeper studies into the nature of such polarized communities to model the opinions coming from people with different backgrounds. Recently, such polarization patterns has been deeply studied as such patterns are often associated with significant political events such as Brexit phenomenon and the surprise win of the US elections by Donald Trump in the 2016. It is also known to cause significant harms to the process of discussion and democracy [Del Vicario et al., 2016]. Such polarized discussion can limit one's viewpoints, reinforce personal biases, and may create and foster environments where hoaxes and misinformation thrives.

2.6.1 Polarization in Annotations

Controversy in social media texts stem from events, topics or social issues that generate different responses from online users [Popescu and Pennacchiotti, 2010]. High controversiality can impact the manual annotation of such phenomena in terms of agreement between human judges because it can lead to polarized judgments. Controversiality is not a new concept in the study of social media. Usually, the controversial topics are identified and user responses or opinions are detected on those topics or issues [Beelen et al., 2017, Basile et al., 2017]. The focus is on the words or texts that are controversial about a particular topic or news items in Sentiment Analysis [Beelen et al., 2017]. Some of the controversial topics include climate change, abortion, and vaccination among the others [Basile et al., 2017]. People from different communities and backgrounds react differently to controversial topics and sometimes, the discussions also make some topics controversial because of the presence of bias against a certain community [Beelen et al., 2017, Basile et al., 2017]. The discussion on controversial topics often results in the spread of online hate [Popescu and Pennacchiotti, 2010]. There are approaches that try to measure the level of controversy by analyzing user opinions or responses on controversial topics. The focus of such studies is the controversy of topics or issues and they ignore the level of polarization in the opinions of annotators when they annotate these datasets [Beelen et al., 2017]. A recent literature survey on hate speech detection [Fortuna and Nunes, 2018] addresses many issues faced by researchers, including the scarcity of high quality datasets available as benchmarks for the hate speech detection tasks. Social media users discuss events and topics, and some of these topics are controversial [Basile et al., 2017].

When the annotation process relies on crowdsourcing, in most cases, the available platforms do not provide any background information (culture, ethnicity, social background etc.) about the workers annotating the datasets. However, even if this information is available, the cultural background of the annotators is usually considered as a secondary aspect or only a single culture is preferred for annotation [Sheerman-Chase et al., 2011]. The work by Sap et al. [2019] highlighted the problems of racial bias in the existing approaches for online abusive language detection tasks and provided an empirical characterization of such bias prevalent in social media platforms. They emphasized that there is strong relationship between AAE markers (e.g, “n*ggas”,“ass”, "f*ck") and HS annotations and models trained on such annotated data are highly likely to pick up and replicate this bias in the data. Such problems with the datasets may raise ecological and methodological issues especially when studying phenomena that target specific communities such as immigrants. This a limit in general, and especially in the case of HS detection. Indeed, HS against different communities often contains stereotypical words which might arouse positive or negative sentiments or different reactions in annotators with different backgrounds [Warner and Hirschberg, 2012, Waseem, 2016]. Due to the constant unavailability of information about annotators, there are high chances of observing polarization among the annotators judgements about the same potentially abusive messages, which might result in low inter-annotator agreement.

2.6.2 Modeling Polarized Opinions

A recent study shows that a boost in online hate over the years in specific geographic locations can result in an increase in offline hate crimes and physical violence in the same areas [Relia et al., 2019]. There is a divergence of opinions between the victims and people responsible for hate crimes. However, the fine-grained information about the cultural, ethnic, or demographic background of the annotators is usually not available, or it is not a primary factor when selecting expert or volunteer annotators. We need methodologies that can automatically model the different perspectives that annotators may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. In such methods, supervised machine learning models can be trained to learn different points of view of the human annotators on the same data, in order to subsequently take them into account at prediction time.

We need to discuss the open challenges and issues faced by the researchers in the development of detection models in order to improve the robustness and accuracy of HS detection models to efficiently categorize the instances of harmful effects of hateful behaviors. More specifically, We need a deeper analysis for the development of novel datasets with a natural division of the annotators into groups with the aim of improving the automatic detection of online hate speech by taking into account the single opinions of annotators and how they diverge on certain topics. The working hypothesis behind the development of such datasets should be that the difference in opinions expressed by the groups of annotators is a valuable source of information rather than a noise factor in the creation of a gold standard dataset. By processing such information, we can create better quality data to train hate speech models for the prediction of highly subjective phenomenon such as cyber-hate. In particular, the focus of such studies could be on inter-annotator agreement computed for subdivisions of the annotator set, and on the level of polarization of the annotated texts. It is often possible to observe the level of sensitivity by examining the data annotated by the targets of hate speech. Such analysis can help us to understand that how online hate can impact the real life hate incidents by analyzing the annotated data. The development of perspective-aware models can help us to understand their point of view better and then counter the spread of hate speech more accurately. We can analyze the opinions of victims at macro-level and by comparative analysis, we can differentiate them from the opinions of people who annotate HS related data but hardly experienced any hate in their life.

2.6.3 Modeling Polarized Opinions with Clustering Approaches

Clustering or cluster analysis techniques are either supervised, unsupervised and semi-supervised learning problems used to analyze the data to find promising patterns in data mining and knowledge discovery [Lin et al., 2006]. Clustering automatically discovers natural groupings such as customer segmentation based on their usage behavior, textual analysis and the grouping of similar documents etc. by finding similarities and dissimilarities among the data elements Shafeeq [2012]. Many clustering algorithms are available and there is no single best technique for different kinds of cluster analysis [Singh and Bhatia, 2011]. Researchers have experimented with different clustering techniques depending on the requirements of an underlying task.

Most common domains in which clustering techniques are applied include measuring documents similarity and customer segmentation. K-means is a heuristic clustering method widely used by the research community and it minimizes the sum of the square of distances from all the available samples from clustering centers to find a minimum of k clustering based on an objective function [Kumar and Ramaswami, 2011]. K-means requires a fixed number of clusters to start with and often uses cosine similarity or euclidean distance to measure the distances between the-data points and cluster centers. K-means has been subdivided into three clustering techniques: K-means [Al-Anazi et al., 2016], K-means fast [Elkan, 2003], and K-mediods [Kaufmann and Rousseeuw, 1987].

Luo et al. [2009] enhanced the performance of K-means clustering by a conceptual implementation of document neighbours and links in the clustering domain by using pairwise similarity function based on a certain threshold value and comparing the documents similarity with a proposed value. Lin [1991] proposed the KL-Divergence; a frequently used measure for finding word-sense disambiguation. It represents a more useful way of finding similar words and their collocation for similarity measure. Kansal et al. [2018], Qiuru et al. [2012], Ezenkwu et al. [2015] implemented K-means clustering for customer segmentation for various business domains. Researchers also employed k-means clustering for recommender systems such as movie recommendation [Yadav et al., 2021] and customer recommendation for a restaurant based on based on psycho-graphic and demographic factors in mobile environment [Katarya and Verma, 2015].

Most of the current research on clustering techniques mainly focus on customer segmentation and the development of product and service recommender systems. Unfortunately, We did not find any studies related to the annotators grouping based on similar opinions in data annotation with the help of clustering techniques especially in the context when the data annotations are sparse and the datasets are annotated by a large number of annotators.

2.7 Summary

This chapter presented an extensive review of recent literature on abusive language and hate speech study in a multilingual setting. This literature review is an important initial step to understand the background of current research work required to grasp the importance of the research carried out during the course of this PhD. We observed that the research area in this direction is relatively new and still has a lot to offer by addressing the challenges faced by researchers in recent years. We found several recent studies which developed abusive language and HS detecting models against minorities and general pubic but the main challenges faced are related to the bias and opinion polarization either in the datasets, in data annotation, or in the developed models. In order to develop comprehensive NLP models, we need good quality resources to provide better results. In this way, user-generated contents represent a big challenge for the NLP community to handle. The several studies presented in this review on abusive language and HS detection so far achieved a performance by employing the syntactic, lexical and semantic levels of natural language processing. Most of the approaches addressed the task as a binary classification with mutually exclusive labels. Current work on the automatic detection of various forms of hate speech (HS) typically employs supervised learning, requiring manually annotated data. From a natural language processing perspective, hate speech detection is often approached by the research community with similar techniques to sentiment analysis i.e., the task of identifying the opinions expressed in subjective utterances, from product and service reviews to the comments in political events.

At this moment, the biggest effort concerns the development of HS corpora labeled by human judges with known background to model the polarized opinions expressed by these annotators. From the abusive language detection perspective, the presence of polarization in opinions of annotators affects the performance of the task. As we pointed out, state-of-the-art systems generally have good results when dealing with regular content, but without the identification of perspectives expressed in the annotations. Therefore, robust HS detection systems should be build on the concept of perspective-aware system modeling, which can identify and model polarized options expressed in the annotations by the humans belonging to a different social, cultural or demographic background.

For abusive language and hate speech detection, another issue is mainly related to the availability of abusive language resources. A supervised learning approach typically relies on manual human annotation in order to create reference data to train models. The annotation is done either by experts or paid contributors on crowdsourcing platforms. In supervised learning, during the process of annotating data, the cultural background of annotators is usually ignored or we do not have any information on the background of annotators. Also, for HS corpora annotated by crowdsourced workers, platforms providing these annotating capabilities do not provide any personal information on the workers who annotate a given dataset. We believe that such information is crucial in mitigating annotator bias and modeling polarization in data annotations. We believe that its relevant and important to conduct studies to investigate these factors to build robust models for detecting abusive language which also reflect the perspectives of minorities which are often the victims of hate speech and online abuse.

Chapter 3

Identifying Polarized Opinions in Hate Speech Detection

Most of the current studies on automatic abusive language detection, and in particular hate speech detection, are based on the majority voting methods. This means that the opinions emerging from different annotators, regardless of their background, are merged to create a gold standard which is basically a single majority based aggregated perspective. Then, the models trained on such datasets always reflect majority opinions ignoring the opinions of individuals which we believe are important to consider in hate speech detection. Such cases are especially true when the targets of hate speech or abuse are minorities or vulnerable communities and the voice of people belonging to theses communities is often ignored while developing the polices and making important decision to counter the online hate. We need to create awareness among the research community to develop robust systems that identify online hate speech across multiple languages by studying the opinions of people with different backgrounds and having different viewpoints. We also need to consider the fact that most popular social media platforms are multilingual. This means that such platforms foster the users to communicate in different languages and the users may belong to different parts of the world with a different ethnic or cultural background and may also belong to a social group comprising people having identical ideological mindset. We need to design systems that consider these important factors.

In this chapter, we start our work on measuring the opinion polarization in hate speech datasets by answering the research questions asked in Chapter 1. We focus on hate speech, and in general on abusive phenomena in online verbal communications, for several reasons. First, hateful discourse online is growing at a worrying rate [Zhang and Luo, 2018], and it is linked to an increase of violence and hatred towards the vulnerable communities, with strong negative social impact [Izsak-Ndiaye, 2015, Mossie and Wang, 2019, O'Keeffe and Clarke-Pearson, 2011]. We propose different approaches to leverage the fine-grained knowledge expressed by individual annotators when they annotate a dataset. This may help us to refine the quality of training sets for hate speech detection. We measure the degree of polarization of the annotations at the message level and divide the annotators into groups that maximize such measure, under the hypothesis that annotators will be grouped by having similar personal characteristics (ethnicity, social background, culture etc.). The annotated datasets are then exploited to detect the opinions expressed by online users for pre-defined categories such as presence vs. absence of a specific phenomena (e.g., sexist and racist behaviours, hate speech against immigrants, homophobia, and so on).

The chapter is organized as follows. Section 3.1 explains the background, motivation and the objectives to conduct the research study presented in this chapter. We describe the models employed in the experiments, including a novel measure of polarization of opinions among the annotators belonging to different groups in Section 3.2, with a pilot study to validate the index. Then, Section 3.3 lists all the datasets that are available to perform the experiments. In Section 3.4, we present the result of an experimental evaluation on several datasets of hate speech in social media (described in Section 3.3). We present a discussion on the qualitative analysis of the datasets in Section 3.5, and finally summarizing the important finding of this research study in Section 3.6.

The majority of the findings in this chapter including the methodology, datasets, and the results were published in 2019 in Akhtar et al. [2020].

3.1 Motivation and Objectives

In Chapter 1, we discussed in detail that hate speech, which is also called *Online hate speech* (or *cyber-hate*), is a form of abusive language directed at specific targets and quite subjective in nature. The subjective nature of a phenomenon refers to the thoughts processing conscious experiences, perspectives, feelings, beliefs, and the desires of people related to a particular topic of interest. We can also address subjective utterances as some relevant information, situation, physical thing or an idea considered true only from the perspective of one subject or different subjects sharing similar thoughts [Gonzalez Rey, 2019]. It is also important to note that emphasizing and recognising something as a social issue also depends on how many people in a society perceive and consider the issue as a major social phenomenon targeting a community within the society and needs considerable attention [Lauer, 1976].

Modeling a subjective phenomenon, such as hate speech, requires considerable effort and resources. We need comprehensive datasets annotated by people with different viewpoints on a certain topic. We also personal information on the annotators to identify how the cultural or demographic background can influence the opinions of a group of people and cause polarization in the messages. The crowdsourcing platforms do not provide such information and without this information, it is challenging to identify the opinions of individuals and link them to their ethnicity or common background. We also assume that in case of people having different cultures and demographics, we can expect very low agreement values among the annotators because they possess diverging opinions on the topics of mutual interest. This also raises concerns about the quality of data annotations due to different sensitivity levels of annotators to different kinds of abusive language.

In Chapter 1, we mentioned several general research questions bu there we will tackle more specific related research questions concerning the studies in this chapter. The research questions that we will work with in this chapter are listed below:

RQ1: *How can we measure the level of polarization among the annotators at instance level when their judgements reflect different opinions in an experimental setup when no personal or* *demographic background is available?*

RQ2: *Can a measure of polarization for individual instances of subjective content help us manually explore the datasets and understand the topics and issues with polarizing nature?*

In order to answer these questions, We develop novel quantitative approaches. The first method is based on a metric of annotators' reliability, and its application to remove the annotators deemed the least reliable within a set of annotators. The second method introduces a measure of polarization at the instance level to manipulate the training sets and reduce the impact of the most polarized, often controversial text that may hinder the learning process of a classifier. We also divide the annotators into groups by measuring the level of polarization in opinions expressed by the annotators in individual instances of hate speech. We test these approaches on three datasets, in English and Italian, annotated by the experts and workers hired on a crowdsourcing platform and show how our approaches improve the prediction performance of a supervised classifier. Moreover, the proposed polarization measure helps towards the manual exploration of the individual instances of tweets in our datasets. A pre-requisite for this methodology is that we need pre-aggregated data which contains annotations by individual annotators. Also, the annotated data is mostly crowdsourced, where no background information on the annotators in available. Our work is based on the assumption that fine-grained knowledge expressed by individual annotators is a useful resource to look into the opinions of individual annotators before their subjectivity is averaged out with the creation of gold standard data.

There are two main contributions of this chapter:

1) *To improve the quality of hate speech detection corpora, and consequently models trained on them, by considering the impact of different opinions of annotators and how they differ on individual messages.*

2) *A mean to manually explore the data and understand the topics and issues with polarizing nature.*

In the next section, we will discuss the detailed methodology adopted to attain the above mentioned objectives.

3.2 Method

In order to answer the research questions presented in this chapter, we designed a series of experiments to evaluate several strategies for hate speech detection with a focus on measuring the level of polarization in opinions expressed by the annotators belonging to heterogeneous groups. We believe that each annotator from a group may perceive hate speech differently and based on such perception, will annotate an instance of a dataset differently from the other annotators either in the same group or from a different group. We also analyze the inter-annotator agreements in a setting where the annotators do not form one homogeneous group. We split the annotators in two groups with the highest divergence of opinions by performing an exhaustive search, and define a quantitative index of opinion polarization based on such split. Finally, we use such measure to automatically manipulate the training set of a supervised algorithm for hate speech detection.

3.2.1 Polarization Index

We proposed a novel method aiming at creating higher quality benchmarks for supervised learning of subjective phenomena by introducing a new index that measured how polarized a message is, when annotated by two different groups. Our approach exploits the fine granularity of single annotations, e.g., resulting from crowdsourcing.

We aimed at understanding the role of factors like ethnicity and social background of the annotators and how it is reflected in their judgment. In a sense, we tested a *homophily* hypothesis [Mcpherson et al., 2001] with respect to opinions, and on a larger (even global) scale: just as homophily in social groups strongly shape their social network, we postulated that the common background of some annotators shapes their opinions as well, leading to polarized judgments on certain kinds of messages. While polarization of opinions stems from the high subjectivity of some phenomena (e.g., hate speech), it differs from the inter-annotator agreement, as the latter is influenced by the factors such as text comprehension and interpretation, e.g., of ironic content. Our goal was instead to capture the influence of personal background of the annotators at a macro-level. Note that high polarization does not necessarily equates to low agreement: we considered the set of judgments on an utterance to be highly polarized if different groups show high agreement on different opinions. On the contrary, if the agreement is low overall, including among members of the same group, then there is no polarization, according to our definition.

We measured the level of polarization in a message given a set of annotations provided by two groups of annotators. Given a set of messages N and a set of annotators M, $g_{i,j}$ denotes the annotation of an annotator j on the message i. For each message $i \in N$, we can split the set of its annotations $G_i = \{g_{i,1}, ...g_{i,m}\}$ into k subsets G_i^1, ..., G_i^k. As a measure of agreement of the annotations on a single message, we use the normalized χ^2 statistics, that is, a test of independence of the distribution of the annotations against a uniform distribution. The rationale for this choice is that we consider a uniform distribution of annotations as total disagreement. For instance, if three out of six annotators decide for a label in a binary classification setting, and the other three assign the other label, the distribution (3,3) is uniform, and therefore the disagreement is maximum. Normalizing the χ^2 by dividing the statistic by the number of annotation, we obtain a value between 0 (total disagreement) and 1 (perfect agreement):

$$a(G_i) = 1 - \frac{\chi^2(G_i)}{|M|} \tag{3.1}$$

We compute the *polarization* index (P-index) of a message i as:

$$P(i) = \frac{1}{k} \sum_{1 \leq w \leq k} a(G_i^w)(1 - a(G_i)) \tag{3.2}$$

P is a number between 0 and 1, where 0 indicates no polarization and 1 indicates maximum polarization. It is designed to take a higher value when at the same time, the agreement between the members of same group is high and the agreement between the members of different groups is low. To give a few examples with $k = 2$:

Example 1:

If $G_i^1 = \{1, 1, 0\}$ and $G_i^2 = \{1, 1, 1\}$, then $a(G_i^1) \approx 0.11$ (low intra-group agreement), $a(G_i^2) = 1$ (high intra-group agreement), $a(G_i) \approx 0.44$, thus $P(i) \approx 0.31$.
Example 2:
If instead $G_i^1 = \{0, 0, 0\}$ and $G_i^2 = \{1, 1, 1\}$, that is, each group is in total agreement but on different labels, then $a(G_i^1) = 1$, $a(G_i^2) = 1$, $a(G_i) = 0$, thus $P(i) = 1$.

3.2.2 Pilot Study

In order to validate the metric, we created a small, manually annotated dataset of English tweets on the Brexit phenomenon, called the BREXIT dataset, gathered from the corpus developed by Lai et al. [2017], where around 5 million tweets were collected by querying Twitter with the hashtag *#Brexit* between the June 22nd and 30th, 2016. This dataset was initially annotated and used for stance detection. We filtered the dataset to only retrieve the tweets containing keywords related to immigrants and Muslims that reflect our work on HS detection. The keywords used are selected based on a study by Miller et al. [2016]. The keywords are shown in Table 3.1:

Immigration	Immigration, migration, immigrant, migrant, foreign, foreigners
Islamophobia	terrorism, terrorist, Muslim, Islam, jihad, Quran
Xenophobia	illegals, deport, anti-immigrant, rapefugee, rapeugee, Paki, Pakis, nigger

Table 3.1: The keywords used to filter the dataset from the original corpus.

We manually labelled 119 randomly selected tweets by following the scheme and guidelines proposed in Poletto et al. [2017], Sanguinetti et al. [2018] with 4 dimensions: *hate speech*, *aggressiveness*, *offensiveness*, and *stereotype*. The next experiments involved data annotations. We involved migrants as the victims of hate speech to annotate randomly chosen dataset for four categories that include hate speech, aggressiveness, offensiveness and stereotype. Since the categories are subjective in nature, it would be interesting to see results of the annotation experiments. We decided to involve women in annotation process so one of the annotators in each group is a female to mitigate the gender-bias. We asked three volunteers with specific demographic features, i.e. first- or second-generation migrants and students from the developing countries to Europe and the UK, of Muslim background, to annotate the dataset. The other three volunteers were researchers with western background with experience in linguistic annotation. The two groups annotated exactly the same data with the same guidelines 5.2.3. The final data set is therefore annotated by six people divided into two groups, which we refered to as *Target* (T) and *Control* (C).

For the experiments in this study to validate the P-index, we only focused on the main class, i.e., hate speech. For a comparative analysis, we also highlighted the details of the agreements for the other categories. We measured the inter-annotator agreements between all the annotators by using Fleiss' Kappa obtaining a value of 0.35. We hypothesized that the high subjectivity of the task is one of the reasons for low Kappa value. Interestingly, the agreement on hate speech classification is higher than the other labels included in the schema: aggressiveness (0.21), offensiveness (0.30), and stereotype (0.20). Since the groups are formed by people having different ethnic background and culture, we expected a higher level of polarization than what we could measure by splitting the groups randomly. In the presence of a given split of the annotator groups, in addition to the overall agreement (*inter-group* agreement), we can also calculate the *intra-group* agreements for each group. On the BREXIT data, we computed an intra-group agreement of 0.54 for both groups. We first measured group-wise agreements between group members. The group one is named as the Control group whereas the group two constitutes the immigrant's category and named as the Target group. For the Control group, we got following results: As the results show that for group one, the overall agreement rate for all the categories is low whereas the highest agreement was measured for hate speech with a value of 0.54 whereas, the lowest agreement rate was for stereotype (0.16).

The results for the Target group show almost similar patterns. Again, the overall percentage of the agreement is low but for HS, the agreement is better than the other categories with a kappa value of 0.54 and stereotype is again the category with the lowest agreement rate having a value of 0.30 but still better than the Control group (0.16). The aggressiveness and offensiveness categories have different results. The interesting comparison was the results of stereotype category as the Target group has better agreements for stereotype category than the Control group. By computing pairwise agreements, we induced a network of fine-grained agreements between the annotators. The topology of such network provided an insight on the relationships between the opinions of single annotators and their groups. On the BREXIT dataset, the pairwise agreement between the couples of annotators from the same group is rather high, between 0.52 and 0.56 in the Control group and between 0.46 and 0.60 in the Target group. However, the pairwise agreements between the pairs where the two annotators are from different groups drops significantly, between 0.16 and 0.36 with a median of 0.24 and a standard deviation of 0.06. The values of group-wise agreements and overall agreements of the two groups for the BREXIT dataset categories are shown in Tables 3.2 and 3.3 respectively.

Data	HS	Aggr.	Offen.	Str.type
Group 1	0.54	0.26	0.38	0.16
Group 2	0.54	0.24	0.39	0.39

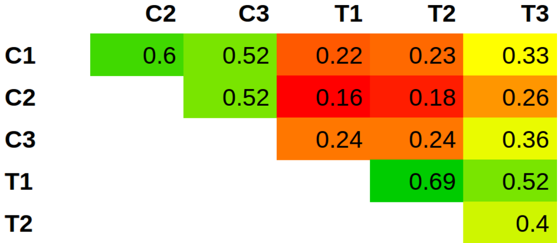
Table 3.2: Group-wise agreements for all BREXIT categories.

HS	Aggr.	Offen.	Str.type
0.35	0.21	0.30	0.20

Table 3.3: Overall agreements between annotators groups (BREXIT categories).

The mean P-index for the original split is 0.18, while the average mean P-index for the 9 other possible splits is 0.09. This result indicates that the P-index successfully picked up the divergence of opinions coming from different communities and ethnic backgrounds.

The pairwise agreements measured on the BREXIT dataset for HS category is shown in Figure 3.1. It is clear from the picture that the two groups of annotators show a much higher intra-



	C2	C3	T1	T2	T3
C1	0.6	0.52	0.22	0.23	0.33
C2		0.52	0.16	0.18	0.26
C3			0.24	0.24	0.36
T1				0.69	0.52
T2					0.4

Figure 3.1: Th pairwise agreements on the BREXIT dataset annotation, between the target group (T) and the control group (C).

groups agreements (top-left and bottom-right area of the figure) than their inter-group agreement (top-right area).

3.2.3 Annotator Reliability Index

We proposed an approach to measure the reliability of each annotator based on its pairwise agreements with other annotators in the same and in different groups, that is, computing the average of pairwise agreements of each annotator with all the other annotators.

Formally, given a set of annotators $x_1, ...x_N$, and a pairwise agreement function $0 \leq a(x_i, x_j) \leq 1$, the reliability of the generic annotator x_i is computed as:

$$R(x_i)\frac{1}{N}\sum_{1\leq j\leq N; i\neq j} a(x_i, x_j) \tag{3.3}$$

Simply for a single-class binary classification, the pairwise agreement function can be as straightforward as the ratio of common annotations over the total number of instances (percent agreement). We first measured the pairwise agreements between two annotators and then the average of pairwise agreements to find the most reliable annotator for each of the categories in both groups. For example; in order to measure the average of annotator A, we combined the average values of the pair AB and the averages of pair AC and then divided it by two to get the average of annotator A.

For the Control group, the values of the average of pairwise agreements are shown in Table 3.4. It is apparent from the Table 3.4 that the average agreement rate (aa(A)) for the annotator A is much better for the categories HS (0.55), offensiveness (0.39) and stereotype (0.23) when compared to the average agreement rates for the other annotators for these categories. The annotator B has a better average agreement rate (aa(B)) for offensiveness category and an equal rate for the HS (0.56) to the annotator A. Whereas, the annotator C has the lowest average agreement rates for all four categories. For the offensiveness, the annotators A and B show better results than the annotator C. For stereotype category, the annotator A the highest agreement rate (0.23) among all the annotators.

Similarly, the average of pairwise agreements for the Target group are listed in Table 3.5. The results of the pairwise agreements for the Target group are quite interesting when compared

Data	HS	Aggr.	Offen.	Stereotype
aa(A)	0.56	0.28	0.39	0.23
aa(B)	0.56	0.37	0.37	0.13
aa(C)	0.52	0.23	0.35	0.11

Table 3.4: The average of the pairwise agreements between Control group.

to the Control group. Again, the HS is the category having highest agreement rates (0.60) for the annotator A when compared to the annotator B (0.54) and the annotator C (0.46). For the aggressiveness, apparently the annotator A has lowest average agreement (0.20) when we compare it with the annotator B (0.23) and C (0.27). A similar pattern is observed with offensiveness category as the annotator A has slightly lower average agreement (0.37) in compassion with the annotators B (0.39) and C (40). Interestingly, a similar pattern is also observed for the stereotype where the annotator C has slightly better average agreement (0.42) when compared to the annotator B (0.40) but if we see the agreement rate of A, it is very low (0.17) among all the annotators.

Data	HS	Aggr.	Offen.	Stereotype
aa(A)	0.60	0.20	0.37	0.17
aa(B)	0.54	0.23	0.39	0.40
aa(C)	0.46	0.27	0.40	0.42

Table 3.5: The average of the pairwise agreements between annotators for Target group.

Figure 3.2 shows the results of the calculations in a more comparative and informative way for all the categories and for the individual annotators in both groups.

As we can observe in the figure that for the Control group, in the hate speech category, the annotators A and B have similar agreement values and they are equally reliable but annotator C is less reliable when compared to the other two annotators. For aggressiveness, the annotators B and C are more reliable when compared to the annotator A. A different pattern is observed for the offensiveness category in which the annotator C is less reliable than the annotators A and B. Similarly, for stereotype, the annotator C is less reliable than the other two annotators.

For the Target group and HS category, the annotator A is the most reliable with a kappa value of 0.60. The annotator C is less reliable than the other two annotators with a value of 0.46. For offensiveness, the annotators B and C have similar values hence are most reliable in this group. The annotator A is least reliable for the offensiveness category. For aggressiveness, again the annotator C is the most reliable in the whole group and the annotator A being the least reliable for this category.

The most interesting results are for stereotype category. The annotators B and C have almost similar results and deviate considerably from the annotator A, who has a kappa value of 0.17. We can assume that since stereotype is the most difficult category to detect, it is be possible that the annotators B and C didn't follow the guidelines properly for this category that's why their

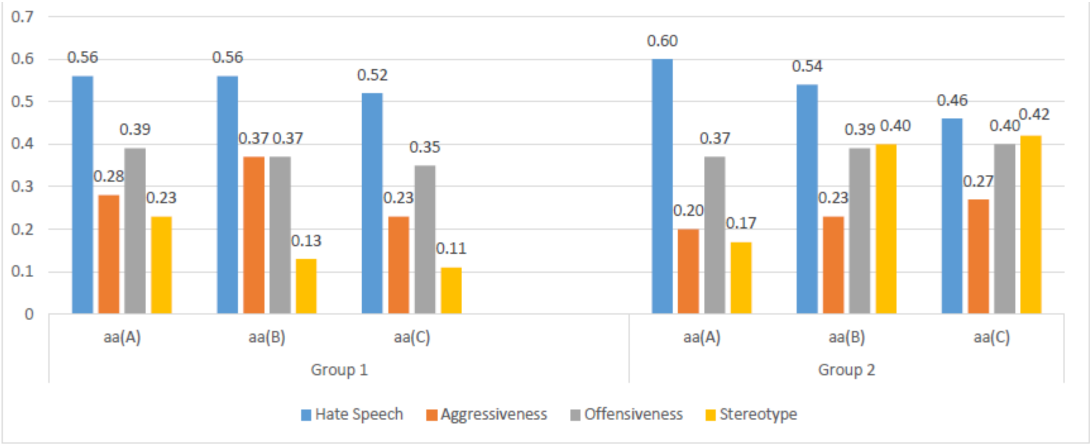


Figure 3.2: The average of pairwise agreements on the BREXIT dataset between two groups for all the categories.

results are very different from rest of the annotators and also a very high kappa agreement score.

3.2.4 Enhancing the Training Data

The polarization index introduced in Section 3.2.1 and the annotator reliability index introduced in Section 3.2.3 provide useful information on the annotation of highly subjective messages. First, we propose to employ this metric to improve the gold standard data which can impact the classification performance. In a supervised learning fashion, a training set is needed, made by manually annotated instances of the text paired with the judgments of a set of annotators. Supposing that complete information about the annotation is available, i.e., not only the aggregated values but each single annotation, then we can compute the P-index of each instance in the dataset. It is important to note that even when the complete annotation is available, in general, we do not have background information about the annotators. However, based on the result of the pilot study presented in Section 3.2.2, we assume that it is reasonable to split the annotators in two groups in a way that maximizes the total polarization.

We compute the P-index for each instance in the training set and then replicate the instances based on its value. The intuition is that if the P-index of an instance is low, a classifier can learn more than if the instance is more polarizing. Therefore, we replicate the instances in the training set a number of times inversely proportional to their P-index. Instances with a P-index of 1 are removed from the training set. In order to verify that our approach works, we experimented with different strategies. First, we only remove instances with a maximum P-index value and do not replicate the rest of the instances. Alternatively, we do not remove the tweets with a maximum value of P but only replicate the instances. Finally, we combine the two approaches.

The methods introduced can only modify the training set in a supervised learning setting, and are fully automated, provided the fine-grained data on the annotation of the training set.

3.3 Datasets

In order to test the methodology introduced in Section 3.2, we gathered a dataset of hate speech in social media. The corpus is borrowed from a previous study by Waseem [2016] on HS detection in English language. The original dataset was composed of 6,909 tweets annotated with racism and sexism. The dataset is available on a Github repository[1], where only the Twitter IDs and the labels are provided. Querying Twitter to retrieve the messages by using the IDs resulted in the collection of a smaller dataset consisting of 6,361 tweets, due to the perishability of the data on the online platform.

Experts (feminist and anti-racism activists) annotated the data. These experts were allowed to skip any instances that they were unsure of. The annotations from experts were aggregated into a single label. Non-experts were hired via a crowdsourcing platform[2] and they worked on the same tweets annotated by experts, following the guidelines developed by Waseem and Hovy [2016]. Each tweet was annotated by at least four annotators. The total number of annotators was not disclosed for privacy reasons. The gold labels are computed by majority vote, and ties are broken by giving preference to the judgment of expert annotators. For current work, we treat all annotators (experts and non-experts) equally.

We also employ an additional set of tweets in Italian, to test the application of our method in a multi-lingual perspective. The Italian dataset comprises 1,859 tweets on topics related to the LGBT community. The Homophobia dataset was annotated by the volunteers.

Table 3.6 summarizes the size and their label distribution of the datasets employed in this work.

Dataset	Language	Positive class	Negative class	Total
Sexism	Train			5,088
	Test			1,273
	Total	810	5,551	6,361
Racism	Train			5,088
	Test			1,273
	Total	100	6,261	6,361
Homophobia	Train			1,487
	Test			372
	Total	224	1,635	1,859

Table 3.6: The datasets sizes and the label distributions.

3.3.1 Sexism

The dataset from Waseem [2016] contains tweets annotated according to four categories: *sexism*, *racism*, *both*, and *neither*, in a multi-label fashion. We isolated the sexism and racism classes to

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

[2] https://www.figure-eight.com/

focus on them individually with two binary classification tasks. In other words, we converted the labels *sexism* and *both* to *sexist*, and the labels *racism* and *neither* to *non-sexist*. In the resulting Sexism dataset, 810 tweets out of 6,361 (12.7%) are marked as *sexist*.

The overall agreement (Fleiss' Kappa) among the four annotators in Sexism dataset is 0.58, indicating a moderate agreement. Following the methodology of the pilot study conducted on the BREXIT data (see Section 3.2.2), we compute the P-index of all the tweets in the dataset for all possible splits of four annotators, and select the combination that maximizes the average P-index, in order to create two annotator groups. We measured the intra-group agreement for two groups, resulting in 0.53 and 0.64 respectively.

Figure 3.3 shows the examples from Sexism dataset with their P-index values and the labels. Notice that the two examples with P-index = 1 are polarized in opposite directions with each group having different annotations for a single tweet.

[redacted] *Because she's not a "feminazi" and is really cool. I have a lot of friends with different views.*
(P-index=1), labels(1 1 0 0)

i just googled to find out if i was a basic bitch. buzzfeed says i am not. i remain suspicious.
(P-index=1), labels (0 0 1 1)

[redacted] *I'm sure you give good ones. Too bad you're probably ugly as dirt like most FemiNazi cunts.*
(P-index=0), labels (1 1 1 1)

Figure 3.3: The examples from Sexism dataset with their P-index values.

3.3.2 Racism

We extracted a binary labeled Racism dataset from the data of Waseem [2016] following the same procedure we applied to derive the Sexism dataset (Section 3.3.1). The annotation scheme remains the same as with the original dataset explained in Section 3.3. The only difference is the mapping of the original labels: *racism* and *both* are mapped to *racist*, while *sexism* and *neither* are mapped to *non-racist*. In the resulting Racism dataset, 100 tweets out of 6,361 (1.57%) are marked *racist*. The overall agreement (Fleiss' Kappa) between all annotators in the Racism dataset is 0.23, indicating relatively high disagreement between the annotators. We divide the annotators into two groups by selecting the split that maximizes the average P-index, and measure an intra-group agreement of 0.22 and 0.25. Figure 3.4 shows the examples from Racism dataset with their computed P-index values and also the labels. Each tweet in the first two examples is oppositely polarized, with each group having different annotation for the tweet.

Headed out #coon #hunting with some friends on the back of the #farm. This is his first time. Kinda... https://t.co/vDofdeemhY.
(P-index=1), labels (1 1   0 0)

jumping in the #BlameOneNotAll tag.I expect to find all kinds of bigoted fucktards telling me how I'm the problem http://t.co/MwgmqJXPiR
(P-index=1), labels(0 0   1 1)

Why do #Blacks #Coon on television or the movies? http://t.co/je4HUxhMEt If they don't...they won't work they won't make money...
(P-index=0), labels(1 1   1 1)

Figure 3.4: The examples from Racism dataset with their P-index values.

3.3.3 Homophobia

We exploited a dataset from the ACCEPT project[3] on the monitoring of homophobic hate online. The data consist of tweets selected with a number of LGBT+-related keywords and annotated by five volunteers contacted by the largest Italian LGBT+ non-for-profit organization (Arcigay)[4] selected along different demographic dimensions such as age, education and personal view on LGBT stances. The original dataset is labeled in a multi-class fashion according to four categories: *homophobic*, *not homophobic*, *doubtful* or *neutral*. We map *not-homophobic*, *doubtful* and *neutral* to *not homophobic* and leave the label *homophobic* unchanged, to restrict the problem definition to a binary classification task.

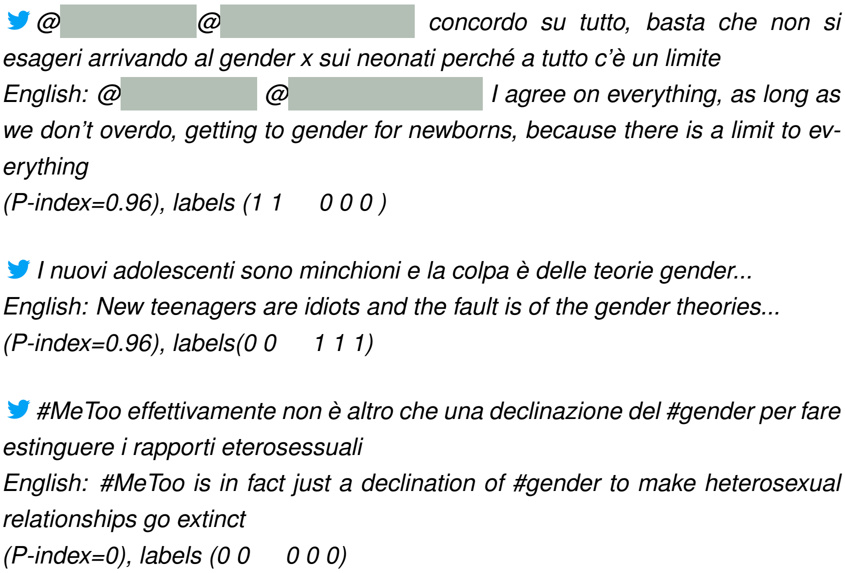
The agreement between the five annotators (Fleiss' Kappa) is 0.35 (moderately low). Similarly to the Sexism and Racism datasets, we split the annotators into two groups, by computing the average P-index for all the possible combinations of 3+2 groups, and selecting the split that maximizes the average P-index. The intra-group agreement for the two groups is 0.40 and 0.39. Figure 3.5 shows the examples from Homophobia data set with their English translations and their computed P-index values and also the labels. The tweets with high P-index are oppositely polarized, i.e., one group detected HS whereas the other group did not.

3.4 Experiments and Results

We evaluated the methods introduced in Section 3.2 with cross-validation experiments on the datasets described in Section 3.3. At each fold, we randomly split the dataset into a training set (80%) and a test set (20%). We refer to the "positive" and "negative" classes as a generalization over the actual labels, which are different (but comparable) for each dataset. All the datasets are highly unbalanced. We did not balance the data artificially, in order to obtain realistic results.

We employed a straightforward supervised learning approach, keeping the test set fixed and

[3] http://accept.arcigay.it/
[4] https://www.arcigay.it/en/



@[redacted] @[redacted] concordo su tutto, basta che non si esageri arrivando al gender x sui neonati perché a tutto c'è un limite
English: @[redacted] @[redacted] I agree on everything, as long as we don't overdo, getting to gender for newborns, because there is a limit to everything
(P-index=0.96), labels (1 1   0 0 0)

I nuovi adolescenti sono minchioni e la colpa è delle teorie gender...
English: New teenagers are idiots and the fault is of the gender theories...
(P-index=0.96), labels(0 0   1 1 1)

#MeToo effettivamente non è altro che una declinazione del #gender per fare estinguere i rapporti eterosessuali
English: #MeToo is in fact just a declination of #gender to make heterosexual relationships go extinct
(P-index=0), labels (0 0   0 0 0)

Figure 3.5: The examples from Homophobia dataset with their P-index values.

only modifying the training set prior to giving it as an input to the classifier. We employed a basic classifier, to focus on the impact of the modified training sets rather than the effect of hyper-parameters of more sophisticated models. The classifier is based on a Support Vector Machine model (SVM) with Bag of Word features and TF-IDF weighting. The reason behind using SVM is that it is well known in classification performance boost when compared to newer algorithms like neural networks. The two main advantages are: higher speed and better performance with a limited number of samples. Specifically, we implemented the classifier with the *Scikit-learn* Python library with default parameters, and the *TfIdfVectorizer* function. The only parameter we optimize for different datasets is the number of features (unigrams) in the vectorized representations of the tweets.

The performance is measured in terms of overall Accuracy, Precision, Recall and F1-score on the positive class, averaged over five folds. The baseline results are given by the classifier trained on the original, unmodified training sets.

3.4.1 Experiment 1: Training Data Manipulation with P-index

In the first experiment, we train the classifier on a training set modified according to the polarization of its textual instances. We compute the P-index for all the tweets in the training set, and replicate them according to their value. The first modification to the training set consists in the removal of instances with the maximum P-index value (*P-max filter*). For the Sexism and Racism datasets, the maximum P-index value is 1, whereas for the Homophobia dataset, the maximum P-index is 0.96[5]. The second modification consists in the replication of instances

[5] This difference is due to having five annotators in total, therefore uneven group sizes.

(*replication*) based on the following scheme: for the Sexism and Racism datasets, the tweets with $0 \leq P < 0.375$ are replicated one time (two instances in total). The reason behind choosing these numbers is that since the lower values show that the tweet is less polarized, hence we replicate it a number of times more than the tweet with a higher P-index value. For the Homophobia dataset, the tweets with $0.3552 \leq P < 0.5328$ are replicated once (two instances in total), tweets with $0.32 \leq P < 0.3552$ are replicated twice (three instances in total), tweets with $0.0528 \leq P < 0.32$ are replicated twice (four instances in total) and tweets with $0 \leq P < 0.0528$ are replicated to a total of five copies.[6] Finally, we combine both modifications (*P-max filter+replication*). These changes concern the training set only, while the test set is unchanged.

The results are presented in Tables 3.7, 3.8 and 3.9. The performance of the classification generally improves over the baseline on all three datasets. On the Sexism dataset, the performance boost is caused by a higher recall. The recall on the Racism and Homophobia datasets with baseline result is substantially low, due to the datasets being highly skewed. However, both precision and recall improve on these datasets. Interestingly, the recall improves in every experiment, including when some training data is removed (P-max filter). This indicates that indeed highly polarizing instances tend to generate confusion for the classifier.

Classifier	Accuracy	Precision	Recall	F1
SVM	95.11	**87.60**	71.60	78.74
SVM+P-max filter	95.13	86.40	73.01	79.11
SVM+replication	**95.27**	87.01	73.40	79.67
SVM+P-max filter+replication	**95.27**	86.60	**74.01**	**79.83**

Table 3.7: The prediction results on the Sexism dataset (1700 features used).

Classifier	Accuracy	Precision	Recall	F1
SVM	98.55	55.40	11.01	18.40
SVM+P-max filter	98.58	59.01	12.01	19.88
SVM+replication	**98.61**	**70.01**	19.60	29.49
SVM+P-max filter+replication	**98.61**	69.80	**19.80**	**29.74**

Table 3.8: The prediction results on the Racism dataset (1700 features used).

3.4.2 Experiment 2: The Annotators' Reliability

In this experiment, we computed the reliability of each annotator based on their pairwise agreements with all other annotators, as described in Section 3.2.3. This gives us a ranking of all the annotators based on their reliability score. According to such ranking, we remove the least reliable annotator from the set, and recompute the gold standard. The Sexism and Racism datasets

[6] The threshold values come from the observation of actual P-index values in the data.

Classifier	Accuracy	Precision	Recall	F1
SVM	**88.81**	61.01	11.40	19.02
SVM+P-max filter	**88.81**	**63.60**	13.60	22.30
SVM+replication	86.55	50.40	18.40	26.83
SVM+P-max filter+replication	87.63	47.90	**26.20**	**33.67**

Table 3.9: The prediction results on the Homophobia dataset (3500 features used).

had originally four annotations, therefore removing one annotator is straightforward, since it does not introduce ties. For the Homophobia dataset, moving from five to four annotators can produce ties. We resolved them by assigning weights to the annotations based on the reliability of their author.

The results of the experiment, presented in Tables 3.10, 3.11, and 3.12 in the rows marked by *+*Relab*, show that our method based on annotator reliability improves the performance on the classification of homophobic tweets, mainly due to a better recall. Recall is also improving on the Sexism dataset, while this method is not found beneficial on racism detection. The result suggests that employing a measure of annotator reliability can be beneficial for hate speech detection, but the classification performance is highly dependant on the nature of the data.

Classifier	Acc.	F1	Pre	Rec
SVM	**94.32**	**76.70**	**80.76**	73.10
SVM+Relab	93.70	76.20	73.72	**79.03**

Table 3.10: The prediction results on the Sexism dataset.

Classifier	Acc.	F1	Pre	Rec
SVM	**98.35**	**40.10**	**48.02**	**36.83**
SVM+Relab	97.96	29.60	32.50	27.60

Table 3.11: The prediction results on the Racism dataset.

3.5 Manual Evaluation

It is worth noting that the measure of polarization introduced in this chapter is useful to support manual exploration of the data, besides providing a tool for supervised text classification.

By ranking the instances of a dataset by P-index, the most polarizing tweets emerge naturally at the top of the list. It is also important to note that in order to measure P-index values, we need pre-aggregated annotations showing annotation values of individual annotators. The

Classifier	Acc.	F1	Pre	Rec
SVM	**86.39**	33.00	**40.41**	28.17
SVM+Relab	84.94	**35.20**	36.50	**34.14**

Table 3.12: The prediction results on the Homophobia dataset.

most polarizing tweets can be analysed in order to understand the multiple facets of the phenomenon under consideration and extract the most subjective controversial topics and keywords in a dataset. This divergence of opinions at message level can help us to understand how the social and personal background of annotators can impact the annotations and in-turn the creation of gold standard data. Consider the following tweet from the Racism dataset:

> *[redacted] [redacted] all he does is attack black men. He hates himself and he doesn't even know it #coon*
> *(P-index=1)*

In the above example, all the members of Group 1, have marked the message as racist and hateful while the members of Group 2 marked it as not conveying hate. This also shows the level of subjectivity of this kind of annotation task. Strong lexical expressions such as *attack* may have been perceived by one group as indicators of racism. The hashtag *#coon* is highly controversial too. In fact, there are other messages in the Racism dataset containing such hashtag and characterized by the maximum value of P-index, such as the following, where the same groups expressed opposite opinions with respect to the first example:

> *RT:[redacted] Can we re-retire the word #Coon*
> *(P-index=1)*

It is interesting to note that after ranking the Racism tweets by P-index values, we only found a few tweets marked as hateful by all the annotators. This shows how racism is a highly subjective phenomenon, and the sensibility of different groups of people plays an important role in its annotation.

We also believe that the sensitivity level of the Group 1, because they are activists, is high and that is the reasons why some of the messages in our dataset which were not racist were marked as racism by first group. The example in figure two explains this. We observed that most of the tweets with P-index equal to 1 are either racist remarks or typical words degrading certain minorities. We can also notice that the annotators having similar ethnic and social background have similar opinions on racism related messages.

Similar patterns are observed in the other datasets. The level of polarization of a message also seems to correlate with the simultaneous presence of different controversial issues, such as in the following example from the Homophobia dataset:

> *Silvana De Mari a Otto e Mezzo: 'dittatura delle minoranze, omofobia e islamofobia psicoalleati di un futuro totali... https://t.co/Wogb7Rj7sV*

> *English: Silvana De Mari at Otto e Mezzo: 'minority dictatorship, homophobia and islamophobia psycho-allies of a total future ... https://t.co/Wogb7Rj7sV (P-index=0.96)*

The above example contains only homophobia but also racism and Islamophobia, therefore different groups perceive it differently based on their inner moral values and cultural background. The tweet is homophobic, racist and Islamophobic at the same time. These words are used in a derogatory way.

In the Sexism dataset, the vast majority of the tweets with $P = 1$ contain race-related remarks along with misogyny, as in the following example:

> *@[redacted] uh... did you watch the video? one of the women talked about how it's assumed she's angry because she's latina.*

Similarly, we found several instances of sexism among the most polarizing tweets in the Racism set. Humour (albeit black) also seems to play a role in generating confusion and polarization among the annotators, as we found several instances of (often inappropriate) jokes at the top of the racism P-index ranking, e.g.:

> *Another #Arab car #terror attack in #Jerusalem #Israel. Will #Obama call it random traffic infringement? http://t.co/XrxajfBXKF*

Finally, by manually inspecting the Homophobia dataset ranked by P-index, we found that the most polarized tweets mention a restricted number of topics (gender theories and their education in school, family values) very consistently, while such topics are otherwise distributed in the corpus equally among other topics such as news, law, gossip, politics and homophobia. In fact, the relative frequency of the word *gender*[7] is about seven times higher among the tweets with $P = 0.96$ than those with $P = 0$. Tweets about homophobia are generally not controversial or polarized, with the relative frequency of the word *homophobic* (and its variations) being almost three times higher among the tweets with $P = 0$ than those with $P = 0.96$.

3.6 Summary

In this chapter, we explored the task of hate speech detection in a multilingual setting. We experimented with the traditional SVM based models by presenting novel methods that leverage different opinions emerging from the heterogeneous groups of annotators who annotated the available datasets. We designed two different experiments: in the first case, we trained the models on manually enhanced data by employing the developed polarization index. We removed the most polarized tweets from the training data and duplicated the tweets having lower P-index values. In the second case, we leveraged the annotation given by individual annotators to compute their reliability, and removed the least reliable annotators to improve the quality of the gold standard data and recomputed the gold standard after removing the annotators. We aimed at improving the automatic classification of highly subjective phenomenon such as hate speech.

[7] In Italian, the English word *gender* is used as a borrowing only to refer to the modern gender theories.

We tested our approach in a cross-validation experiment on social media datasets containing sexism and racism in English language and homophobia in Italian language.

The impact of our first method on the classification of sexist messages is reflected in a higher recall, at the cost of lower precision. This indicates that ignoring the disagreement between annotators is likely to generate greater confusion on borderline sexist messages, that in turn produces a higher number of false negatives. The results on the Racism data show a different pattern, where the precision in particular improves by a large margin by applying agreement-based training set manipulation. This suggests that injecting knowledge about polarization in the model helps us in disambiguation of potential false positives. A similar pattern is observed on the Homophobia dataset, but with the P-max filter providing a higher precision boost, as opposed to the replication strategy, which was giving the best precision performance on the Racism dataset.

In contrast, the results for second experiment show that by employing the annotator's reliability, we see improvements in the classification of homophobic and sexists tweets, mainly in the form of a better recall. Surprisingly, we did not observe any improvements in racism detection.

We can deduce from the evaluation of our results that the gold standard data based on general agreements between different annotators can not be considered as a reasonable choice when we have to tackle highly subjective phenomena, and we can utilize, to our benefit, the information extracted from the disagreement and the polarization of opinions. Finally, the manual exploration of datasets by measuring the level of polarization shows that the P-index is an effective tool that can be employed to explore and understand the data. We can also use the P-index to rank the individual instances to identify the messages that are more likely to generate confusion and polarization amongst the annotators.

One of the challenges we encountered during this study is that the information about the background of annotators is often not available in the datasets which are publicly available. Moreover, the set of annotations could be sparse, e.g., in a crowdsourcing context. Therefore, we will tackle these challenges in the next chapters by further expanding our study.

Finally, let us emphasize that this first work on polarization of annotators' opinions is rooted in the somewhat strong assumption that there exists a latent background divide in the annotator population. Even stronger is the assumption that the number of groups is fixed. Although the experimental results confirm the existence of the polarization phenomenon, it will be interesting to investigate how the method can be refined by relaxing the division constraints and aiming for a more flexible, perhaps clustering-based procedure.

In the next chapter, we will explore HS detection with a focus on perspective-aware systems. We will develop group-wise classifiers to measure the classification performance pertinent to an individual group of annotators. We will also explore the idea of inclusive classification to study its effects on the classicaiton of hateful messages with the available datasets.

Chapter 4

Modeling Annotators’ Perspectives by Developing Group-based Models and by an Inclusive Classifier

Perspective identification is an important task in Natural Language Processing (NLP) and can be considered a sub-task in the area of opinion identification and extraction from social media datasets. For example, the discussion about whether it was right for America to participate in the Vietnam war may be one topic but it may reveal multiple perspectives based on the knowledge of individuals on the global event. An expert on this phenomenon can easily identify different perspectives on the issue. Automatic identification of perspectives on domestic and global issues can be very useful for political and security analysts.

In NLP, the task of *stance detection* [Mohammad et al., 2016, Biber and Finegan, 1988] aims at identifying point of views, judgments or opinions on a given topic of interest. The social and political issues on which individuals tend to express their opinions are usually controversial in nature, causing polarization among people [AlDayel and Magdy, 2021]. This point of view can be in favor of or against a specific target of interest on social media platform such as Twitter. Stance detection is often considered a sub-problem of sentiment analysis aiming to extract stance detection of a person against a target(an entity, concept, event, idea, opinion, claim, topic, etc.). The social and political issues on which individuals tend to express their opinions are usually controversial in nature, causing differentiation or polarization among people [AlDayel and Magdy, 2021]. Political events, such as elections and referendums, generate heavily polarized opinions from the public and are used for stance detection studies. Beigman Klebanov et al. [2010] worked on perspective identification in public stance on controversial topics such as abortion. The authors describe perspectives as “a particular way of thinking about something, especially one that is influenced by one’s beliefs or experiences,” stressing the manifestation of one’s broader perspective in some specific issue. They employ term frequencies as features to train models that use term absence/presence and found them superior for opinion classification. An important example of perspectives identification is Israeli-Palestinian conflict on which online users from some communities and famous media houses tend to give opposite perspectives based on victim personalization to highlight the Palestinian perspective when covering the same conflict while the majority of west frame the events in favor of the Israeli side [Elmasry et al., 2013].

Chapter 3 centered at hate speech detection with a focus on the opinion polarization in data annotations. We also discussed how diverging opinions expressed by annotators with different background may affect the agreement measures among the annotators and can also influence the performance of machine learning algorithms for HS detection. We developed novel approaches to identify polarized opinions in data annotations. The first method, a polarization index, not only identified the polarized opinions in HS data but also helped us to divide the annotators into groups in a scenario where we do not have any background knowledge about the annotators. The second method measured the reliability of annotators within group and facilitated to remove the least reliable annotator within a group.

The work in the previous chapter 3 was based on the idea that different people perceive a phenomena differently (subjectivity). In order to make use of the diverging opinions expressed in the data annotations, we enhanced the training data by removing the most polarized instances and duplicating the instances with less polarized opinions. We also measured the reliability of annotators and removed the least reliable annotators within a group. The results suggested that these modifications improve the classification of hateful instances. Similarly, by removing the least reliable annotator slightly improves the classification performance of some datasets.

The experiments preformed in the previous chapter proved significant with good results and can be considered as a starting point in this research area. Still, we could not model the perspectives of individual or group-based annotators and lost useful information in the majority voting as we mainly focused on the opinion polarization and how it can cause confusion for the classifier. We did not consider the importance of the annotator groups. In this chapter, we investigate the annotator's grouping and how the individual annotations can be linked to the process of HS detection. We propose an interesting approach to model the perspectives of annotators in a systematic way. The fine-grained knowledge expressed by each individual annotators is rich in information and can be exploited to boost the performance of a system.

By isolating and describing the annotator perspectives in HS corpora will enable the training of classifiers that mimic the annotators perspective, i.e. train perspective-aware classifiers. The NLP community has long been aware that ignoring disagreement in NLP applications makes no sense. Wiebe et al. [2004] showed that the subjective utterances are collocations of multi-word expressions, rather than single lexemes, and these collocations can be used to identify the level of subjectivity in texts. The annotators might also show different sensitivity against particular forms of hate, which results in low inter-annotators agreements.

For example, in hate speech detection tasks, to the best of our knowledge, we did not find a study in which immigrants or migrants as the victims of online hate were involved in the annotation process except in our work explained in Section 3.2.2 of Chapter 3.

In order to understand the concept properly, let us consider an example from HS category of the BREXIT dataset introduced in Section 3.2.2 of Chapter 3. In this study, we developed a manually labeled abusive language dataset in which the background details on all annotators are available. We found that the average value of P-index for all the tweets of dataset in a random split of annotators was higher than the natural selection of the annotators in which one group was named as the Control (C) and the other group was named as the Target (T) which included migrants which are often the victims of online and offline abuse.

> *cuz hard-working Christian Eastern Europeans need to have solidarity with **Muslim rapists** &amp; welfare **leeches** https://t.co/Ue9usiYTLI#Brexit*
> *(P-index = 1), labels (0 0 0 1 1 1)*

The above tweet has a polarization value of 1 which means that the two groups totally disagree with each other on the contents of the tweet whether it contains hate or not. The Control group marked the tweet as not hateful whereas the Target group believes that the tweet contains hate against migrants, in particular Muslims. Since the annotators from second group are all first-or second-generation immigrants and Muslims and the annotators of the Control group are of European descendent, the Target group is more sensitive to the highlighted words in the tweet due to their demographic and cultural background and a direct attack on them in the tweet. This means both groups are showing a divergence of opinions due to their background and sensitivity level on the contents of this tweet. When we asked one of the annotators from the Target group about the tweet, he told us that these words are though stereotypes but are quite hateful and offensive for them and media is the main reason behind all this hate and they face such hate on regular basis in a foreign environment. We also believe that the Control group considered the messages as offensive but not hateful.

Ideally, involving the victims and targets of hate speech in the annotation process would help us to understand their views on online hate incidents. This can help us to understand their viewpoints in a realistic manner and then counter the spread of hate speech more accurately. We can analyze the opinions of the victims at macro-level and by comparative analysis, we can differentiate them from the opinions of people who annotate HS related data but hardly experienced any hate in their life. Indeed, HS against different communities often contains stereotypical words which might arouse positive or negative sentiments or different reactions in annotators with different backgrounds [Sheerman-Chase et al., 2011, Waseem, 2016].

However, the fine-grained information about the cultural, ethnic, or demographic background on the annotators is usually not available, or it is not a primary factor when selecting expert or volunteer annotators. We therefore propose a methodology to automatically model the different perspectives that annotators may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. In our method, supervised machine learning models are trained to learn different points of view of the human annotators on the same data, in order to subsequently take them into account at prediction time.

The chapter is organized as follows. Section 4.1 introduces the research questions we will investigate in this chapter. We describe the models employed in the experiments in Section 4.2. Then, Section 4.3 lists all the datasets that are available to perform the experiments. In Section 4.4, we present the results of an experimental evaluation on several datasets of hate speech in social media (described in Section 4.3). We present a discussion and qualitative analysis in Section 4.5, and finally summarizing the important finding of this research study in Section 4.6.

The majority of the findings in this chapter including the methodology, datasets, and the results were published in 2020 in Akhtar et al. [2020].

4.1 Research Questions

In this chapter, we will answer the following research questions.

1. Does an automatic partition of the annotators based on the polarization of their judgments reflect different perspectives on hate speech?

2. can we improve the classification performance of machine learning models by introducing training sets with such polarized perspectives?

3. Is it possible to effectively represent different perspectives expressed by annotators in polarized opinions in an inclusive model?

In order to test these research questions and to serve a multilingual perspective, we performed classification experiments on three different Twitter datasets in English and Italian, featuring different forms of hate speech: sexist, racist and homophobic content. For each dataset, we created separate gold standards, one for each group, and trained a state-of-the-art deep learning model on them, showing that the supervised models informed by different perspectives (perspective-aware models) on the target phenomenon outperform a baseline represented by models trained on fully aggregated data. With separate classification experiments, we get an opportunity to analyze the point of views of the annotators of a group at message level. This also helps us to track the changes in the evaluation scores by experimenting with different classifiers.

Finally, we implemented an ensemble approach that combines the single perspective-aware classifiers into an inclusive model that aims at accounting for every perspective at once. The results show that this strategy further improves the classification performance, especially with a significant boost in the recall of HS prediction.

Moreover, the polarization measure and the grouping of annotators help us to understand the topics and issues creating polarization among the annotators. This information can also help us to understand and develop better training sets for NLP systems.

4.2 Methodology

Our proposed method is based on the assumption that a group of annotators can be divided into groups based on some characteristics such as cultural background, common social behaviour and other similar factors. The idea is to investigate how these characteristics can influence the opinions of annotators expressed while annotating HS data. The method works in two steps, and it is applied to an annotated dataset for which the single, pre-aggregated annotations are known:

1. We divide the annotators into groups (two, in this iteration of the study) by using a numeric index measuring the polarization of the judgments.

2. Different gold standard datasets are compiled following the division of the annotators, and each used to train a different state-of-the-art model.

The original and group-based models are tested against the same test set for comparison. The steps of the method are detailed in the rest of this section.

4.2.1 Polarization Index

Most datasets for NLP tasks are either annotated by the experts or by crowdsourced workers, thus in general, the background of annotators is not known. However, we hypothesize that a group of annotators can be effectively divided according to characteristics linked to their background, by analyzing their annotations. In particular for this work, we again make use of the polarization index (P-index) introduced in Chapter 3 and its application, described in the same chapter, for dividing the annotators into groups based on the polarization of their judgments, which can induce higher quality data for supervised learning tasks for subjective phenomenon such as hate speech and abusive language. The method leverages the information at the single annotation level, measuring the level of polarization of all the annotations on each instance individually.

4.2.2 The Division of Annotators into Groups

After measuring the P-index for individual instances of a dataset, the next step of our method consists in automatically dividing the set of annotators into groups. we perform an exhaustive search between all the annotator combinations and groups are formed based on the highest divergence of opinions. This means that for each combination, we measure the average P-index value for the whole dataset and then chose the partition having the maximum average P-index value. It is assumed that this division can approximately divide the annotators based on some common characteristics such as, cultural and ethnic background, social behaviour etc. We believe that when the gold standard or labeled data is created based on the majority voting, the subjectivity is averaged out in the gold standard creation process. The personal characteristics mentioned above are often ignored while annotating the datasets and then creating the labelled data.

Once the annotator bi-partition is found that maximizes the average P-index, we create two new gold standard datasets, one for each group, by aggregating the annotations with a standard procedure of majority voting and then perform the training and classification tasks on the gold standard data and test data respectively.

We postulate that instead of having a classification task performed on one gold standard created by the majority voting based on the annotations from all the annotators, we can also perform classification tasks for the groups separately and then analyze the effects to monitor the HS detection performance. The architecture of this approach is explained in Figure 4.1.

4.2.3 Supervised Classification

We employ the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] as prediction framework for the binary classification task of hate speech detection. BERT is a modeling technique developed by Google AI for Natural Language Processing pre-training tasks, which minimizes the need of heavily engineered complex features required for specific tasks. Since BERT is a multi-layer architecture, the downstream specific tasks only need one additional fine-tuning layer at the top. BERT has been applied to a large number of NLP tasks and the source code is freely available.

The functionality of BERT is divided into two steps, namely pre-training and fine-tuning. In the pre-training step, the model is trained on different tasks on large unlabelled datasets.

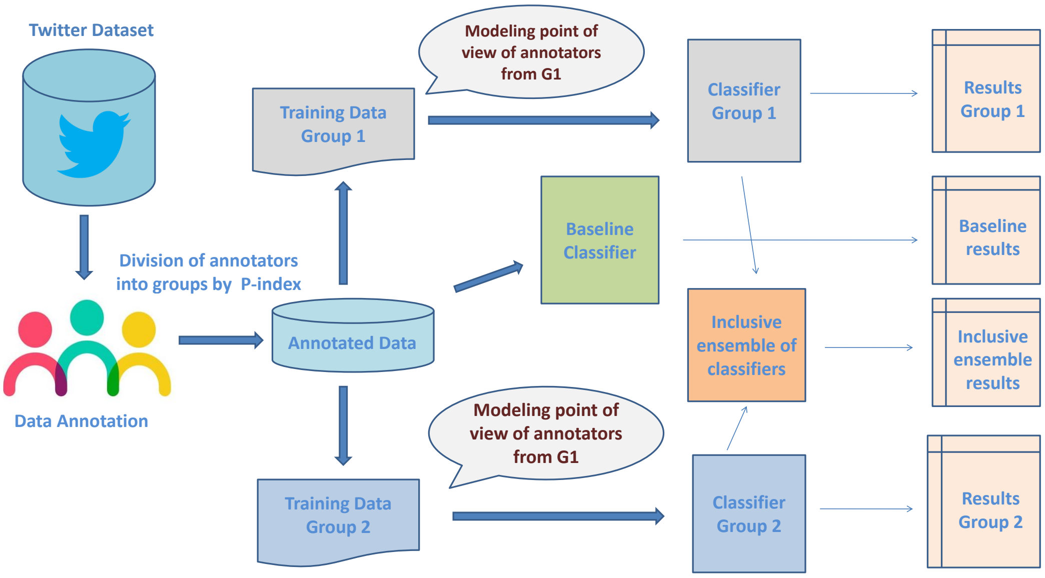

Twitter Dataset
Division of annotators into groups by P-index
Data Annotation
Annotated Data
Training Data Group 1
Modeling point of view of annotators from G1
Training Data Group 2
Modeling point of view of annotators from G1
Baseline Classifier
Classifier Group 1
Inclusive ensemble of classifiers
Classifier Group 2
Results Group 1
Baseline results
Inclusive ensemble results
Results Group 2

Figure 4.1: The architecture of the proposed approach for perspective-aware hate speech detection.

During the fine-tuning step, the pre-trained parameters are adjusted according to a specific task requirements and all the parameters are fine-tuned with labelled data from a downstream task. This architecture allows BERT to be applied to different NLP tasks while sharing the knowledge modeled from pre-training. It is important to note that there are different models for different NLP related tasks and for all the tasks, but same parameters are initialized for fine-tuning step for a downstream task. BERT framework is unique in a sense that same unified architecture is implemented across different tasks.

In the second step of our method, *we fine-tune BERT models to the group-based gold standard datasets obtained in the previous steps, in order to learn different points of view on the perception of the same phenomenon (HS) on the same data.* By contrast, the model trained on the original dataset encodes all the possible points of view of the annotators.

Many BERT pre-trained models are available for multiple and individual languages, and trained on text from different genres and domains [Nozza et al., 2020]. In this work, we use the uncased base English model provided by Google for English (*uncased_L-12_H-768_A-12*[1].). For Italian, we use AlBERTo [Polignano et al., 2019], a model for Italian pre-trained on Twitter data. AlBERTo has similar specifications to the BERT English base model, namely 12 Transformer blocks, a hidden size of 768, and 12 attention heads (L-12, H-768, A-12) and is available on Github [2].

[1] https://github.com/google-research/bert#pre-trained-models
[2] https://github.com/marcopoli/AlBERTo-it

4.3 Datasets

We test our methodology on three datasets. We employed same dataets explained in Section 3.3 of Chapter 3. Since the datasets are already explained in detail, here we will only briefly described them without adding too many details. In order to explore a multilingual perspective, the first two datasets are in English language whereas the third one is in Italian.

The first and second datasets in English language are taken from previous work by Waseem [2016], available online via its Github repository[3]. The original dataset contains 6,909 messages from Twitter annotated in a multi-label fashion with four labels: *sexism*, *racism*, *both*, and *neither*. We separated the corpus into two binary datasets, namely Sexism and Racism. Since only tweet IDs and labels are provided, due to data decay we were able to retrieve a smaller dataset containing 6,361 tweets.

The third dataset is in Italian, containing 1,859 tweets on topics related to LGBT community.

4.4 Evaluation

The datasets presented in Data Section 4.3 are employed to experiment with the method introduced in Method Section 4.2. For evaluation purposes, each dataset is divided into training and test sets. For all the datasets, the training set contains 80% of the dataset whereas, the remaining 20% constitutes the test set. For simplicity, the actual labels are referred as "positive" (presence of hate speech) and "negative" (absence of hate speech) class. All the datasets are highly unbalanced, with more negative instances than positives ones.

The technical details about the BERT are explained in Section 4.2.3. BERT is a more sophisticated model in which we require fine tuning by modifying pre-training parameters for the individual datasets. The information like the size of a dataset, language and the average size of a sentence and a word provide basis for the BERT fine-tuning process.

We fine-tuned the BERT model on the training sets, keeping the test sets fixed for each dataset, for fair comparison. We explored the hyper-parameters space with regard to sequence length, batch size and learning rate. After a preliminary study, we fixed the sequence length at 128 words. The batch size was set to 12 for English and 8 for Italian, also due to memory limitations. The learning rate is 1^{-5}. We repeated each experiment five times, in order to average out the variance induced by the random initialization of the network.

The performance of the models in the classification task is measured in terms of precision, recall and F1 score. The classification performance on the gold standard created by majority voting from the original datasets (before partition) are reported as baselines. We then test the performance of the two models trained on gold standard training sets created by only considering one group of annotators at a time (Group 1 and Group 2).

We also include the results obtained by a straightforward ensemble classifier which considers an instance positive if any of the Group 1 or Group 2 classifiers (or both) considers it positive. We call this ensemble “Inclusive”. The rationale behind this ensemble is that hate speech is a sparse and subjective phenomenon, where each personal background induces a perspective that lead to different perception of what constitutes hate. This classifier includes all these perspectives in its

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

decision process. The Inclusive classifier will naturally have a bias towards the positive class, by construction.

Classifier	Pre. (0)	Rec. (0)	F1 (0)	Prec. (1)	Rec (1)	F1 (1)
Baseline	.953 (.007)	**.972** (.007)	**.962** (.002)	**.812** (.034)	.711 (.044)	.756 (.015)
Group 1	.960 (.007)	.955 (.013)	.957 (.004)	.745 (.048)	.764 (.045)	.752 (.008)
Group 2	.984 (.003)	.940 (.007)	**.962** (.002)	.720 (.019)	.907 (.018)	**.802** (.008)
Inclusive	**.989** (.002)	.920 (.012)	.953 (.006)	.665 (.033)	**.939** (.009)	.778 (.020))

Table 4.1: Results of the prediction on the Sexism dataset. Averages of 5 runs with standard deviation in parenthesis (positive and negative classes).

Classifier	Macro Pre.	Macro Rec.	Macro F1
Baseline	**.882** (.015)	.842 (.019)	.859 (.008)
Group 1	.853 (.021)	.859 (.016)	.855 (.005)
Group 2	.852 (.009)	.924 (.007)	**.882** (.005)
Inclusive	.827 (.016)	**.929** (.004)	.865 (.013)

Table 4.2: Results of the prediction on the Sexism dataset. Averages of 5 runs with standard deviation in parenthesis (Macro Averages).

Classifier	Prec. (0)	Rec. (0)	F1 (0)	Prec. (1)	Rec. (1)	F1 (1)
Baseline	.979 (.002)	**.999** (.001)	**.989** (.001)	**.852** (.159)	.194 (.059)	.312 (.085)
Group 1	.985 (.004)	.993 (.005)	**.989** (.002)	.654 (.154)	.424 (.140)	.488 (.104)
Group 2	.984 (.005)	.988 (.011)	.986 (.003)	.571 (.175)	.412 (.198)	.419 (.076)
Inclusive	**.990** (.004)	.982 (.012)	.986 (.005)	.532 (.141)	**.612** (.136)	**.542** (.091)

Table 4.3: Results of the prediction on the Racism dataset. Averages of 5 runs with standard deviation in parenthesis (positive and negative classes).

Classifier	Macro Prec.	Macro Rec.	Macro F1
Baseline	**.916** (.080)	.596 (.030)	.650 (.043)
Group 1	.819 (.076)	.708 (.068)	.739 (.053)
Group 2	.778 (.086)	.700 (.094)	.702 (.037)
Inclusive	.761 (.070)	**.797** (.065)	**.764** (.048)

Table 4.4: Results of the prediction on the Racism dataset. Averages of 5 runs with standard deviation in parenthesis (Macro Averages).

Tables 4.1, 4.3, and 4.5 show the results of the performed experiments on the datasets for positive and negative classes. Tables 4.2, 4.4, and 4.6 show macro averages of the results.The

Classifier	Prec. (0)	Rec. (0)	F1 (0)	Prec. (1)	Rec. (1)	F1 (1)
Baseline	.898 (.006)	.941 (.040)	.919 (.017)	.415 (.146)	.231 (.079)	.273 (.038)
Group 1	.921 (.017)	.841 (.071)	.877 (.034)	.302 (.038)	.471 (.154)	.355 (.040)
Group 2	.895 (.003)	**.976** (.011)	**.934** (.005)	**.531** (.112)	.178 (.031)	.262 (.033))
Inclusive	**.924** (.017)	.831 (.067)	.873 (.033)	.302 (.039)	**.502** (.142)	**.367** (.035)

Table 4.5: Results of the prediction on the Homophobia dataset. Averages of 5 runs with standard deviation in parenthesis (positive and negative classes).

Classifier	Macro Prec.	Macro Rec.	Macro F1
Baseline	.657 (.072)	.586 (.021)	.596 (.014)
Group 1	.612 (.018)	.656 (.046)	.616 (.022)
Group 2	**.713** (.056)	.577 (.013)	.598 (.017)
Inclusive	.613 (.019)	**.667** (.042)	**.620** (.021)

Table 4.6: Results of the prediction on the Homophobia dataset. Averages of 5 runs with standard deviation in parenthesis (Macro Averages).

results report the arithmetic mean of the evaluation metrics across five runs, along with their standard deviation.

Generally, we see an overall improvement over the baseline on all the datasets. In terms of Marco-averaged F1 score, the classifiers trained on the datasets annotated by single groups almost always outperform their counterparts trained on the datasets annotated by all the annotators.

It is important to note that the improvement on the positive class is particularly important in this setting, since this binary classification task is actually a *detection* task. That is, it is more important to be able to individuate precisely hateful messages than non-hateful messages.

For the Sexism and Racism datasets, the overall improvement is mainly due to a better recall on the positive class. Precision drops, however less substantially, leading to better F1 scores, both on the positive class and macro-averaged. For the Homophobia dataset, group-based classifiers obtain an even greater improvement over the baseline, with higher precision, recall and F1 scores for both the positive and negative class.

The baseline results on the Racism and Homophobia datasets see substantially low recall values, which is expected given the highly skewed class distribution. Group-based classifiers largely correct this problem, although introducing some false positives (hence the lower precision on the positive class).

Finally, the results of the Inclusive ensemble classifier show that including multiple perspectives into the learning process is beneficial to the classification performance on all the datasets, however at the cost of lower precision.

4.5 Qualitative Analysis

In Section 5.5 of Chapter 3, we explained that the P-index tool employed to partition the annotators is also a valid tool to manually explore the datasets. We ranked the messages with polarization value and the messages with maximum polarization appear on the top of the list which help us to easily to observe the controversial messages.

We push the analysis further, by looking at the predictions of the group-based classifiers, with particularly focus on the cases where the two classifiers diverge in their prediction. Analysing one of the runs, we counted 82 classifier disagreements in the Sexism dataset (6.4%), 15 cases in the Racism dataset (1.1%), and 38 in the Homophobia dataset (10.3%). The disagreement is always fairly unbalanced, that is, one classifiers predicts the positive class and the other classifiers predicts the negative class in 81.7%, 60%, and 78.9% of the cases for the Sexism, Racism, and Homophobia datasets respectively.

From a more qualitative point of view, following manual exploration of the results of one run per dataset, we noticed how a considerable portion of the cases where the two classifiers are in disagreement are relatively hard to interpret without access to external knowledge, that is, they mention people and topical events. For instance, from the Sexism dataset:

> *Wadhwa thinks women only disagree vocally with him because they want "attention". #stopwadhwa2015*

The above tweet mentions the technology entrepreneur Vivek Wadhwa who was at the center of a controversy involving women in science.

Highly controversial topics also induce confusion between the classifiers, mirroring what we observed for the annotators and the polarity index, such as in this example from the Racism dataset:

> *WHO SAID GOON'S DON'T EXIST? TELL THAT TO THE BLK MEN WHO LOSE THEIR LIVES DAILY TO OTHER BLK MEN. #COON-LUMINATI [URL]*

The results of our manual analysis therefore confirm that the supervised models trained on gold standard data annotated by partitioned groups of annotators successfully pick up prototypical abstractions of their background, as well as the indecisiveness due to high level of controversy.

4.6 Summary

The chapter tackled the challenges of annotator's division into groups based on their annotation for multilingual Twitter abusive language datasets. The rationale behind the approach is that such partitions reflect some common characteristics (culture, demography, ethnicity, etc.) and can influence the annotators' perception on certain phenomena and shape their opinions on social media posts. Our polarization-based methodology groups the annotators based on their opinions and stance toward given a phenomenon, effectively acting as an empirically-driven substitute to the unavailability of information on the background of the annotators, e.g. in a crowdsourcing scenario. When getting separate groups for individual datasets, we designed experiments to model different perspectives expressed in the annotations and also analyze how these perspectives are represented in classification tasks. We employed a state-of-the-art BERT algorithm for classification. BERT is a more sophisticated model in which we require fine tuning by modifying pre-training parameters for the individual datasets.

The experimental results show overall improvements on three different social media datasets in English and Italian over the baseline by training multiple, group-based classifiers instead on a single, all-comprehensive one on the target phenomenon. We can draw the following observations from the general picture emerged from the classification experiments. The results suggest that the detection of sexist behavior is an easier task than the detection of racism on our data, and that the detection of homophobic content is the harder of three tasks. This is in line with the results of recent evaluation campaigns on misogyny detection [Anzovino et al., 2018] and hate speech detection on online media about immigrants [Basile et al., 2019], and likely due to the vocabulary of misogynistic hate being somewhat more restricted.

The ensemble approach provides an even stronger classification performance overall, especially for the positive class. We saw substantial improvements in recall and F1 score for all the dataset. We can observe that with all the datasets except for Racism and Homophobia, the precision values are low for positive class compared to baseline. This shows that the classifier is producing higher number of false positives and most probable reason is that the disagreement (polarization) between the annotators is generating greater confusion along the borderline messages.

We can also observe that the evaluation results are somewhat circular rather than linear. This reason we believe are that the phenomenon targeted are different in nature and the annotators approach them differently when they annotate these datasets. For some phenomenon, such as Sexism, The annotators are somewhat agreed on most of the tweets but for the others, such as Racism, the annotators differ more in their perspectives on the individual messages.

It is possible to get volunteers to annotate a dataset but again the annotation process takes considerable amount of time and it is hard to find the volunteers willing to annotate larger datasets. Also in special cases, like involving the victims of HS in the annotation process that was done for HS category of BREXIT dataset in Chapter 3, it is very hard to find educated people who could be trained for the annotation process. We also need well developed guidelines to produce higher quality annotated gold standards. Another issue is epistemological: our methodology and the subsequent empirical evaluation show that there is a great deal of information that is effectively wiped out by the aggregation step employed in the standard procedure to create benchmark datasets. Therefore, evaluating perspective-aware machine learning models on traditionally aggregated datasets is unfair. This consideration inspires us to promote the publication of datasets in pre-aggregated form, and to develop new paradigms of evaluation that take all the perspective due to different backgrounds into account.

Like in the previous chapter, where we applied the methodology to manually explore an annotated dataset, here, we qualitatively analyzed the prediction results from separate classifiers to understand whether a classifier trained on a gold standard training data from an annotators group can successfully pick up the polarization in opinions expressed by that particular group.

Apart from the raw performance metrics, one may wonder which classifier should be selected when multiple group-based models are available trained on the same data. One possibility is to give preference to the classifier trained on data annotated by a group involving the victims of hate speech, if this information is available, in order to “give voice” to the targeted group through the computational model. Another possibility is to implement an inclusive classification framework, such as the ensemble classifier proposed in this work. Such method aims to “give voice” to all the existing perspectives on a certain phenomenon equally. Furthermore, given its transparency, the latter classifier shows potential for providing an explicit explanation of its decisions, being able to track them back to the specific (highly cohesive) groups of people who annotated the training data.

In the next chapter, we will expand our perspective on the phenomenon of online abuse by analysing beyond the linguistic aspects and with the design of extensive experiments for the extended BREXIT dataset in which we involve migrants as the victims of online abuse to explore how the demographic and cultural aspects in a proper context can influence the detection and classification of abusive language.

Chapter 5

Case Study: The Development of a Multi-perspective Corpus of Abusive Language on Brexit

The growing number of economic migrants, refugees to Europe and the new waves of extremism and xenophobia has increased the importance of studying hate in international and European socio–political context. It is often observed that there is a thin line between the fundamental human rights to the freedom of expression and the increasing need for tolerance and mutual respect in a growing society which is racially, ethnically, and religiously diverse and multicultural in nature. Unfortunately, a major proportion of the online abuse is targeted towards the most vulnerable communities such as immigrants, LGBT, Muslims, Jews and women. The factors often responsible for such hate may include the demographic background and personal characteristics of the victims such as ethnicity, religion, race, sexual orientation and color [Nobata et al., 2016, Duggan, 2017]. This online growth of HS is often considered a reason for violent behaviour and hatred towards such communities [Izsak-Ndiaye, 2015].

The manual annotation of a public dataset is a crucial step in creation of language resources that are used for training predictive models of a language, the controversial text might cause performance issues for Natural Language Processing (NLP) approaches. Most abusive language detection corpora are composed of data collected from social media platforms [Poletto et al., 2021], such as Twitter and Facebook. The annotation process relies on crowdsourcing, in most cases, such platforms do not provide any background information (culture, ethnicity, social background etc.) on the workers annotating the datasets. Similarly, the victims of online hate speech are never involved in the annotation process. Therefore, it is hard to understand their viewpoints and how they feel and what they think of a particular communication or a message as hateful or portraying no hate.

In Chapter 3, we focused on HS detection in an experimental scenario where the data annotation has high polarization in messages. We developed novel approaches to tackle the high level of polarization. The work emphasized that different people may react to hateful messages differently and such difference in opinions are based on how they perceive a phenomenon. Traditionally, such difference of opinions is considered noise in the data but rather they contain useful information. The experiments preformed proved significant with good results. Still, we could not model the opinions of individual or group-based annotators. We did not consider the importance of the annotator groups.

Chapter 4 focused on the annotator's grouping and how the individual annotations can be linked to the process of HS detection. We proposed a new approach to model the perspectives of annotators in a systematic way. The fine-grained knowledge expressed by each individual annotators is rich in information and can be exploited to boost the performance of a system. Although we presented a pilot study in Chapter 3, in which we worked on BREXIT dataset but only few tweets were annotated to analyze the rich multi-perspective dataset and it was not possible to perform classification experiments on such a small dataset so there was a need to expand the annotation and the analysis of the BREXIT data.

The study in this chapter extends and completes the preliminary studies presented in previous chapters, by providing further analysis and additional experiments. In previous work, we initially experimented on three available datasets in English and Italian, featuring different topical focuses. In this chapter, we present a case-study to model polarized opinions coming from different communities under the hypothesis that similar characteristics (ethnicity, social background, culture etc.) can influence the perspectives of annotators on a certain phenomenon. We propose a novel resource, a multi-perspective English language dataset annotated according to different sub-categories relevant for characterising online abuse: hate speech, aggressiveness, offensiveness and stereotype. Unlike previous work, where the annotations were based on crowdsourcing, here, we involved the victims of the targeted communities in the annotation process, who volunteered to annotate the dataset, providing a natural selection of the annotator groups based on their personal characteristics and providing a deeper insight and better understanding of our hypothesis regarding the importance of capturing annotators' perspectives and developing perspective-aware models. By training state-of-the-art deep learning model on this novel resource, we show how our approach improves the prediction performance of a state-of-the-art supervised classifier. Moreover, we also perform an in-depth qualitative analysis on the novel dataset to identify and understand the relevant keywords, topics and events causing polarization among the annotators in expressed opinions. By processing such information, we aim at creating better quality data to train abusive language models for the prediction of highly subjective phenomenon such as cyber-hate. In particular, we focus on inter-annotator agreements computed for subdivisions of the annotator set, and on the level of polarization of annotated texts. We experimented with state-of-the-art transformer-based Neural model, such as BERT [Devlin et al., 2018], which is a pre-trained language model fine-tuned for each down-stream detection task.

The rest of the chapter is organized in the following sections. Section 5.1 presents a detailed description of the research questions that we will try to answer in this study and our contributions as a subsection. We introduce Section 5.2 which provides comprehensive details on the newly developed multi-perspective BREXIT dataset including the details on how the dataset was collected and filtered, data annotation process and the presentation of detailed statistics on the dataset. Section 5.3 explains the extrinsic evaluation of the BREXIT dataset. We present the results of our empirical evaluation on the dataset in Section 5.4. Then we present the qualitative analysis performed by manual exploration of the BREXIT dataset in Section 5.5, before we summarize our findings in Section 5.6.

5.1 Research Questions

Well crafted and properly documented research questions are an essential part of any research study. The questions which are feasible, concise, well focused and research-able provide us a clear directions for what we want to find and gives our work a clear focus and purpose. This section explains why the work conducted in this chapter is important to further investigate the research questions formulated in Chapter 1 and subsequently explored in Chapters 3 and 4. The questions are not entirely different from the broader questions explained in the previous chapters. The important difference to consider here is that in previous chapters and experiments we preformed, the data was entirely based on crowdsourcing platforms. Since working with the crowdsourcing data do not provide any personal information on the annotators hence we assumed that the polarization in opinions is depended on various factors (ethnicity,s social background, cultural information, demographics, social and political affiliations etc.) hence modeling their opinions required division into homogeneous groups by implementing innovative methods and evaluation techniques to validate the working of qualitative analysis on the annotator's segregation. In order to further expand our analysis of annotator's partitioning based on social and cultural profiles, we required a dataset annotated manually by diverse annotators who can be easily divided based on the known personal information. Such datasets are easier to investigate to find the patterns in data annotations to understand the divergence in perspectives expressed by these annotators and also explore how the ethnic or cultural background or even the topics that are admired by these annotators can influence their decision while annotating a datasets. In this chapter, we develop a multi-perspective dataset on abusive language and HS with a natural selection of the annotator groups who annotated the dataset with known background information so the questions are more focused on this dataset. Let us recap the questions again:

RQ1: *How can we measure the level of polarization among the annotators when their judgements on data annotation at message level reflect different perspectives in an experimental setup where the cultural or demographic backgrounds of the annotators is available?*

In order to answer the first question, we have to understand the methodology explained in previous chapters. In Chapter 3, we developed a novel approach, a polarization index (P-index) which measures the level of polarization (personalized perspectives) expressed by annotators in opinions to exploit the fine granularity of single annotations of highly subjective phenomena resulting from crowdsourcing platforms.

First, we showed that how traditional inter-annotator agreement metrics can provide new insight when applied in a setting, where the annotators do not form one homogeneous group. Then, we introduced a new index that measured the level of polarization of a message with respect to the annotations given by two different groups. The P-index aided in the division of annotators into groups by performing an exhaustive search and finding the partition having maximum polarization. We tested it on several available benchmark datasets. In this work, we do not need the P-index to aid the division of annotators into similar groups because we have a natural selection of the annotator groups based on their personal characteristics.

RQ2: *Can a measure of polarization for individual instances of subjective content help us* *exploring the datasets and understanding the topics and issues involved with polarizing nature?*

In order to answer this qualitative research question, we employ P-index with the newly developed multi-perspective abusive language dataset. With the help of P-index, the most polarized messages naturally emerge at the top of the list. Similarly, we can analyze the group-wise predictions to find the topics and keywords relevant to certain events which cause polarization among the annotators. Due to a natural selection, it will be much easier and more feasible to analyze the perspectives in different groups of annotators because we know who they are, their personal characteristics are known so we can deduce from the analysis why a group of annotators annotated an instance differently from the other group and which topics and keywords may influence the decisions of annotators because they are more sensitive to certain keywords due to their personal experiences in a multi-cultural environment.

RQ3: *Can we improve the classification performance of machine learning models by introducing training sets with different perspectives and is it possible to effectively represent these perspectives expressed by annotators in an inclusive model?*

To answer the question, we employed the methodology developed in Section 4.2 in which we create separate gold standards for each of the annotator groups and perform classification tasks to measure the performance of perspective-aware supervised learning models in which the model learns from the training sets created from different groups. We also employed an ensemble classifier, which was proposed in previous work, that considers all the learned perspectives in an inclusive fashion.

5.1.1 Contributions

The summary of our contributions for this work is given below:

1. We provided a brief review of the novel approaches developed in previous chapters which include a polarization index, the unique concept of the creation of group-based gold standards and subsequent classification experiments, and the details of an inclusive model to detect different forms of abusive language in social media with state-of-the-art NLP models. We employed the approaches to the datasets developed for the study conducted in this chapter.

2. We developed a multi-perspective abusive language dataset containing tweets downloaded from Twitter. The dataset is manually labelled and the complete information about the personal, cultural and demographic backgrounds of the annotators is available. For the first time, we involved migrants as the victims of abusive language to annotate the dataset. This helped us to capture the common perspectives that may shape their opinions. It is an important step in our study because we believe that the current datasets and annotation schemes available for abusive language detection fail to model and understand the feeling and emotions expressed by the victims of online and offline abuse.

3. We presented a comprehensive and in-depth qualitative analysis of our dataset which provides us a deep and thorough understanding of polarized instances and how they might affect the gold standard creation process and particularly, the annotation process. The qualitative analysis also helps us to explore and understand the topics, events and keywords related causing polarization among the annotators.

5.2 Abusive Language Dataset on Brexit

Brexit is a portmanteau of the words "British" and "exit" merged to refer to the exit of the great Britain from the European Union (EU). Social media platforms are highly influential in encouraging users to express opinions on certain world level events, such as the Brexit. There is an increased user participation to openly express their opinions and suggestions on such events. During the voting period of Brexit referendum, there was an exponential increase in the number of opinions expressed online, suggesting different reasons and expressing intent concerning the importance of this global event.

The abusive language and HS dataset for current research are gathered from Lai et al. [2019] in the context of a study concerning *stance detection* (referred as BREXIT dataset henceforth). The data was collected in the form of tweets for different time intervals from Twitter. The time frame for the collection of data on Brexit was before the voting, during the voting, and immediately after the voting period for referendum to measure the frequency of tweets and the stance of online users on the event.

The motivation to focus on the Brexit political debate is explained in the following. The process of creating an abusive language dataset with a step-by-step narrative requires the raw data to be rich enough for mining the contents which are more resourceful and provide better insights about a topical phenomenon, in our case abusive language. The research community should be able to consider the approaches for transforming the existing insulting environment of social media into a non-hate inclusive online society. Because of its global importance and strong impact on the European society, we decided to have a through review of the available BREXIT dataset, also stimulated by studies highlighting how the Brexit debate have been overdetermined by the racist and xenophobic attitudes [Miller et al., 2016]. A careful manual analysis of the tweets revealed interesting patterns from the discussions on the topics and events that are strongly linked with the abusive contents including racism. The majority of the tweets blamed immigrants and Muslims as the main reasons for Brexit and also using derogatory and abusive words against minorities. This motivated us to select the dataset for our research work on abusive language detection.

5.2.1 Data Collection and Filtering

Initially, around 5 million tweets were collected between the June 22^{nd} and 30^{th}, 2016 by using the hashtag #Brexit. After getting the dataset, we performed many pre-processing steps to filter and clean the dataset for further work. We divided the collected data into sub-corpora related to three categories: Immigration, Islamophobia, and Xenophobia. The details about different categories are given below. For each category, keywords have been selected based on a previous study by Miller et al. [2016]. The following list shows the keywords used to filter the dataset:

Immigration, migration, immigrant, migrant, foreign, foreigners, terrorism, terrorist, Muslim, Islam, jihad, Quran, illegals, deport, anti-immigrant, rapefugee, rapugee, paki, pakis, nigger.

Table 5.1 shows the frequency of the occurrence of these keywords in the whole dataset.

Immigration	**Islamophobia**	**Xenophobia**
immigration (43287)	Jihad (1974)	Illegals (823)
migration (50365)	Jihadi (1140)	deport (5003)
immigrant (23803)	terrorist (5161)	rapefugee (550)
migrant (48334)	terrorism (3765)	rapeugee (41)
foreign (23661)	muzzie (243)	anti-immigrant (1776)
foreigner (5793)	muzzies(42)	anti-immigration (1490)
foreigners (5148)	islam (30239)	paki (8117)
refugee (18519)	kuffar (46)	pakis (7692)
refugees (14416)	kaffir (9)	nigger (183)
	Quran (878)	
	muslim (26556)	

Table 5.1: The frequency of keywords in the dataset which were used to filter the data.

Immigration:

We filtered the BREXIT dataset by using different keywords to select the tweets related to immigrants, and to gather and analyze the opinions expressed by online users about the immigration and the role of immigrants in a society. The list of keywords for immigration corpus is shown in Table 5.1. We retrieved a total of 53,824 tweets by filtering the data with related keywords.

Xenophobia:

We further filtered the dataset by using keywords that are linked to Xenophobia. The idea was to get a subset of the BREXIT corpus to analyze whether the data also contain xenophobic contents, which refer to hate against minorities and less privileged communities in a society. We retrieved a total of 4585 tweets by using keywords that are mentioned in Table 5.1 with the frequency of occurrence of these keywords in the dataset.

Islamophobia:

Similar to the Immigration and Xenophobia corpora, we filtered the BREXIT corpus by using keywords specific to Islamophobia to get a subset of the data. The details about the keywords and their occurrence in the overall dataset are mentioned in Table 5.1. We retrieved a total of 17222 tweets by using keywords related to the Islamophobic content.

Discarded racism and racist keywords:

When reasoning on the relevant keywords to select our data, we also investigated the occurrence of two important keywords, *racist* and *racism*, in order to determine their use by tweeters of the BREXIT dataset in various contexts. The frequency of the occurrence of keyword ‘racist’ in the whole data set was 72,023, which shows a relatively high usage frequency. Similarly, the word ‘racism’ occurred 54,024 times which also shows a high usage of the keyword. However, a close manual analysis of the tweets where these keywords occur revealed that most of the tweets were targeting natives of the Britain as racist people or were linking racism to the reason why Britain is leaving the EU. Some of them highlighted that the racism in the UK is on rise, other tweets also included abusive language targeting people living in the UK and voting for the Brexit.

Therefore, we came to the conclusion that most of the tweets filtered using such keywords are not containing abusive language against immigrants, but are directed towards natives of the UK. Therefore, we decided not to further use *racism* and *racist* as keywords to select the data. We decided to report here about the process that led us to consider and discard these keywords, to bring to the light some further aspects on the nature of the original dataset.

Pre-processing:

After selecting the data as specified before, we applied pre-processing steps before the annotation experiments. We cleaned the dataset by removing duplicates including the retweets. Figure 5.1 explains the procedure which resulted in the filtered and processed datasets. Figure 5.2 shows a comparison of three sub-corpora with the number of retrieved tweets.

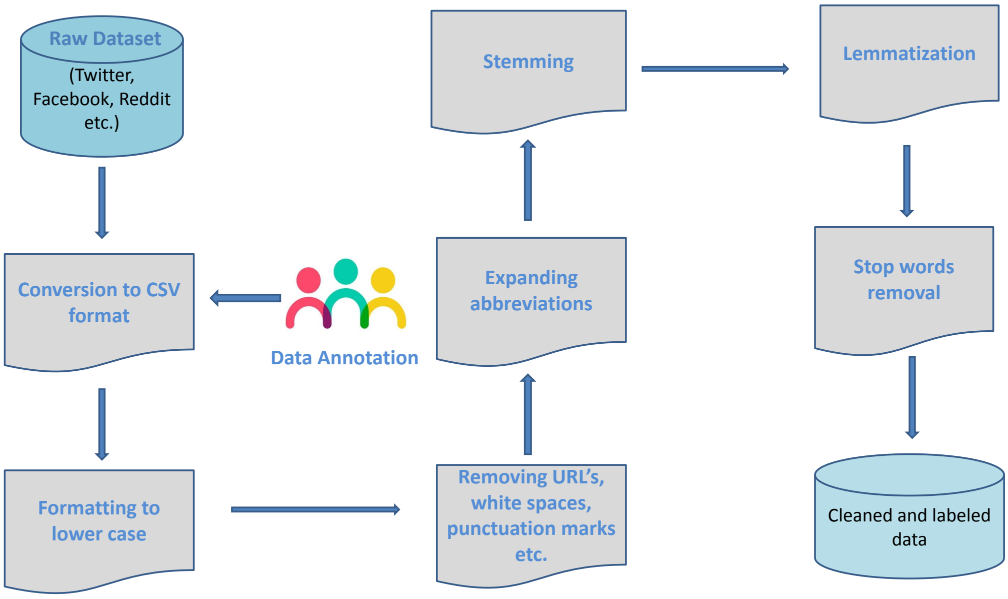


Figure 5.1: The pre-processing steps to clean the dataset.

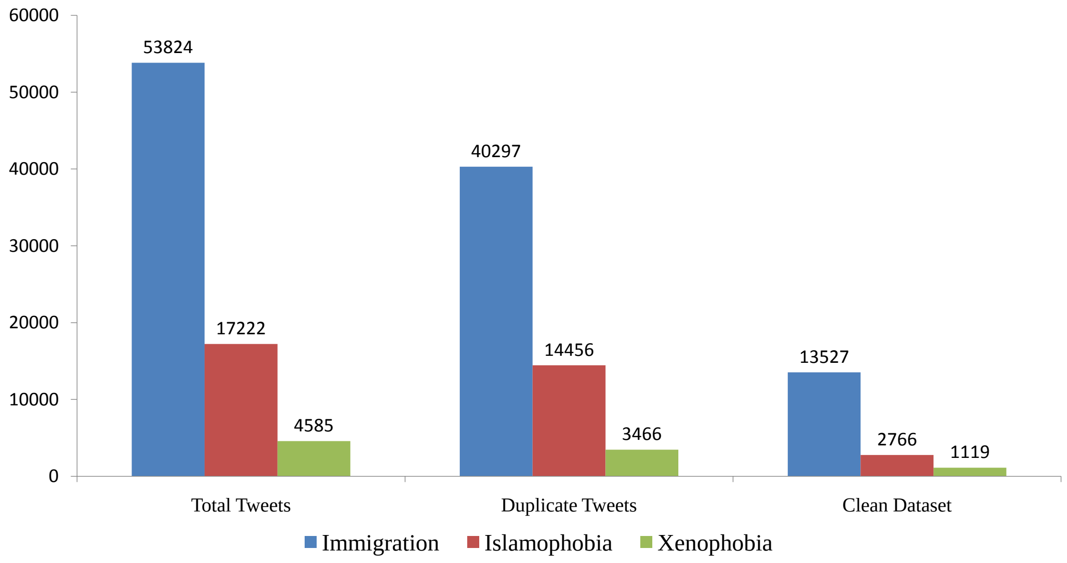


Figure 5.2: The details of the clean sub-sections of the dataset.

5.2.2 Data Annotation

The next step is the data annotation process. One of the focus of this study is to involve the victims of abusive language such as migrants and Muslims in the annotation process. To create an annotated corpus, we randomly selected 1120 tweets from the dataset. The corpus was annotated according to four categories that include *Hate Speech*, *Aggressiveness*, *Offensiveness*, and *Stereotype*, following the multi-layered annotation scheme suggested by Sanguinetti et al. [2018] to develop a corpus of hate speech against immigrants. A total of six annotators were selected for the annotation process and then divided into two groups. Since the categories are subjective in nature, we thought it would be interesting to see the results of annotation experiments and agreement measures between the two groups.

The annotation process for different types of abusive language is a rather difficult and vague process which usually results in low agreement scores as also acknowledged in Schmidt and Wiegand [2017]. Here, the annotators were briefed and trained so that they have a similar understanding of the abusive categories.

The scheme and guidelines proposed are described and referred in Poletto et al. [2017], Sanguinetti et al. [2018]. The authors used the scheme and guidelines to annotate an Italian language dataset on HS against Muslims and Roma. The guidelines were written in Italian language. We translated them into English and modified them according to our requirements to educate the annotators with minor modifications. All the annotators were volunteers with certain demographic and cultural background. The first group of the three volunteers were first-or second-generation immigrants and students from the developing countries to Europe and the UK, of Muslim background. The other group has three volunteers who were researchers with western background and having experience in linguistic annotation. The same dataset was annotated by two groups and also followed the same scheme of annotation and guidelines. We named the groups as *Target* which are the migrants and *Control* which are the locals. For further processing, we will also call the *Control* group as *Group 1* and the *Target* group as *Group 2*. In order to mitigate the possible gender bias, we also involved people with different gender in each group of annotators.

5.2.3 Annotation Guidelines

Our goal is to detect Abusive Language and HS on social media (Twitter). We came up with a set of annotation categories and guidelines that attempt to gather various categories in a single coherent framework. Such categories include, besides HS, aggressiveness, offensiveness, and stereotype. For our work, we have two targets of interest namely migrants and Muslims. Although these two categories can overlap, there exist different stereotypes and hateful discourses for each of them. For this reason, we preferred to keep them separate. The targets are already provided for annotators.

Hate Speech:
In order to annotate hate speech we need an operational definition of the phenomenon. In particular you should focus on the following aspects:

- The presence of the target of interest, meant as the group identified as one of the vulnerable categories we focus on, or as an individual, for its membership in that category (and not for its individual characteristics). For example, while insulting someone for his/her look, although certainly offensive, is not hate speech, insulting someone for his/her origin, skin color or religious beliefs is indeed hate speech.

- Whether the writer of a tweet has the intention to spread, incite, promote or justify hatred or violent acts towards the given target. This intention can either be explicitly expressed or remain implicit. It can also include attempts to threat or terrify the target, or to convince other people to adopt hostile attitudes that can result in violent, abusive or discriminatory acts against the targets.

- The Organization category has two labels: Individual and Group. The purpose is to distinguish whether the contents of a single tweet, to detect our categories, refer to an individual person or group of people with common attributes (or community).

Labels: No - Yes
Yes: is the label we chose when there is single or joint presence of above-mentioned elements in a tweet. HS against immigrants and Muslims may include:

- Insults, threats, denigrating, ridiculing or hateful expressions:

> 🐦 *#Brexit just in time. Fungus of Islam infecting non-Muslim brains across EU, cowardly officials excuse sex assaults https://t.co/ysCivWNvtU*

- Incitement to hatred, contempt, violence or violation of rights to individuals or groups due to their somatic traits (e.g. skin color), origin, cultural traits and language:

> *Let's start throwing out these black and brown slaves coming into the UK from around the world #brexit*

- Presumed association of origin/ethnicity with cognitive abilities, tendency to commit crimes, laziness, disability etc. or association with animals or entities considered inferior:

> *London should kick all Muslim Refugees out before they all kill them.#Trump2016 https://t.co/pmHE3mLc4F*

- Assuming that some ethnic groups are superior with respect to others or questioning the social position or credibility based on origin/ethnicity:

> *These fucking immigrants are useless as they have ignorant mindset, uneducated and can't contribute in the economy of the Country; just a burden.*

- References to certain backgrounds/ethnicity as a threat to people's security or welfare or as competitors in the distribution of government resources:

> *UK should have invaded Syria; removed Assad to avoid these 5,000 Syrian refugees from coming in; causing #Brexit. https://t.co/K5L8gVBD7w*

- Associations between Islamic faith and fundamentalism, terrorism, murder or a supposed plan of invasion or conquest of Europe:

> *"When the English vote #BREXIT, they should immediately eject all Muslim immigrants"*

> *#voteleave if your not a left wing bigot who wants the uk to become a state of islam and feminism #itscancer #Brexit*

> *The United Kingdom preferring not to be overrun by Immigrant animals is racist, you guys. #Brexit*

No: in case even just one of these conditions is not detected, HS is assumed not to occur.

Aggressiveness:
Labels: No - Yes
If user intends to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target. Conversely to offensiveness, it focuses more on the physical harm that can affect the target. It is important to note that a tweet that might incite violence among the targets is not aggressive from our point of view.
Yes: A message is considered aggressive if:

- It implies or legitimates discriminating attitudes or policies:

> *We Britain's don't need filthy immigrants*

- There is an allusion to a potential threat posed by the presence of the target, or its alleged outnumbering with respect to country's population:

> *Stop these immigrants otherwise they will outnumber us soon*

- There is a sense of dissatisfaction and frustration due to the (perceived) privileged treatment granted to the target group by the government:

> *Spending a lot of money and resources on these immigrants is just waste of everything.*

- Which also results in overt hostility (though expressed using measured tones) or if there is the reference to whether explicit or just implied to violent actions of any kind:

> *Let's start a new crusade against these invading immigrants.*

> *Let's clean Britain from all immigrants and Muslim invaders*

No: if a tweet is considered as hateful or offensive, it does not necessarily imply it is also aggressive:

> *SadiqKhan lets Vote #remain for more migration .... 330,000 and rising we don't have the infrastructure but who cares #brexit does*

Offensiveness:
Labels: No -Yes
When compare to aggressiveness, it rather focuses on the potentially harmful effect of the tweet content on a given target. Conversely to aggressiveness, it focuses more on the emotional or psychological harm that can affect the target.
Yes: a tweet is considered offensive in at least one of the following cases:

- The given target is associated with typical human flaws (laziness in particular) Or, in general, with negative characteristics:

> *Lad in Q at shop said he glad he voted #Brexit Foreigners lazy bastards. Not seen one at Polling Station.Too lazy to vote #howscaryisthat*

> *These pathetic ugly foreigners flooding Europe*

- The status of disadvantaged or discriminated minority is questioned or they are mocked or ridiculed with serious insults:

> 🐦 *Terrorist are coming disguised as immigrants. Throw them out*

> 🐦 *How ironic (sad) that so many in #Scotland would choose foreign rule ;servitude https://t.co/6P0RrpONO9 #Brexit #WilliamWallace*

- The targeted people are described, or just considered as unpleasant people, or just the kind of people you better have nothing to do with. May also include extreme insults.

  "To be honest; i don't like these immigrants"

  > 🐦 *This US immigrant woman escapee of banana islamized land can now hell out: GOD IS GREAT! #Brexit Dominoes will fall. In US also #MAGA #MAGA*

No: a tweet may be aggressive, or it may portray a stereotypical image of the given target group, but it does not mean this is also offensive:

> 🐦 *If migration increases, Economy weakens and prices go up. Don't say you haven't been warned... #EUref #Brexit*

Stereotype:
Labels: No - Yes
Stereotypes are (mostly untrue) beliefs about a given target. It is important to consider the presence of stereotype, as previous experiments show that hateful contents are often based on generic negative considerations about the target group. It is important to note that in our case, the stereotype refers to untrue beliefs about the given target that prevails in the society and it's not the opinion of an individual person about the target.
Yes: a tweet is considered as containing a stereotype in (at least) one of the following cases:

- The members of a given target are referred to as invaders, criminals and or having filthy habits:

  > 🐦 *A nation that cannot protect young girls from being raped by foreign invaders is no nation at all Remember Rotherham. #Brexit*

  > 🐦 *Nobody wants foreign beggars all over the streets. Nobody wants criminal muslims freely passing every border in Europe https://t.co/9Mf16E11K4*

  > 🐦 *Filthy immigrant food....#Brexit https://t.co/tCqOUN1t32*

No: tweets that disprove a stereotype, or debunk fake news, should not be labeled as containing stereotypes

> 🐦 *PM Blames Brexit On Immigration Concerns"not true it's about governing ourself #brexit https://t.co/cRwrZuv4bN https://t.co/bRJ6d1ZNf4*

IMPORTANT NOTE
Whenever a clearly hateful tweet does not actually refer to one of the targets selected in our corpus, HS is assumed not to occur. On the other hand, the remaining categories are expected to be annotated accordingly.

> *RT* [redacted]*: The idiots are out today. Desperate stuff. Everyone's a racist who's #VoteLeave because we want to #TakeControl*

Hate Speech: no; Aggressiveness: no; Offensiveness: yes; Stereotype: yes The following example shows the presence of all the categories in a single tweets:

> [redacted] [redacted] [redacted] *The history suggests that immigrants always created problems for a nation. They are filthy thieves. Let's wipe them out of Britain.*

Hate Speech: yes; Aggressiveness: yes; Offensiveness: yes; Stereotype: yes

5.2.4 Dataset Statistics

In this section, we give a statistical description and quantitative analysis of the final dataset with all sub-categories. We can see the distribution of labels for all categories in Table 5.2. It is clear from the distribution that the abusive language categories are highly unbalanced. This means that there are less instances of positive class and more of the negative class in our data. Although, for the Offensiveness and Stereotype, the ratio of positive class is relatively better in comparison with the Hate Speech and Aggressiveness categories.

Dataset	**Positive class**	**Negative class**	**Total**
Hate Speech	106	1,014	1,120
Aggressiveness	87	1,033	1,120
Offensiveness	206	914	1,120
Stereotype	151	969	1,120

Table 5.2: The BREXIT dataset with all sub-categories and the distribution of labels.

Table 5.3 shows the average P-index measured for all the possible splits by performing an exhaustive search for each of the BREXIT category. In order to validate the metric, we need an annotated dataset in which the personal and demographic backgrounds of the annotators are known and the annotators having similar background are grouped together. In order to measure the average, we first calculated the P-index values of individual sentences in a category and then we measured the average value of P-index for all the instances in each of the categories. If the average is higher, we have more polarization and divergence of opinions among the annotators for that category.

As shown in Table 5.3, the value is higher for HS, Aggressiveness and Offensive for the natural grouping of annotators which indicate how effectively the P-index can divide the annotators coming from different communities and backgrounds into groups based on the divergence of their opinions. This Table also shows the maximum and minimum values for all the other possible splits which are less than the values of natural grouping. It is interesting to see that for Stereotype, the average P-index value of another split is slightly higher than the natural split. One possible reason is that we found high pairwise agreements between one member of the Control group and all other members of the Target group and this might cause the average p-index value to go slightly higher than the natural grouping.

It is important to note that we conducted a similar study in our previous work in Section 3.2.2 of Chapter 3, but the study was only limited to the HS category and only a few tweets were available for the analysis. Here, we extended the abusive language categories and also the number of tweets for each category to further verify the validation process of the division of annotators into groups by using the P-index.

Category	**Natural**	**Max.(other splits)**	**Min.(other splits)**
Hate Speech	**0.14**	0.09	0.06
Aggressiveness	**0.10**	0.09	0.07
Offensiveness	**0.18**	0.14	0.11
Stereotype	0.14	**0.15**	0.12

Table 5.3: The BREXIT categories with natural grouping and maximum (other splits) and minimum (other splits) average P-index values.

We measured different types of agreements among the annotators from the two groups. Table 5.4 shows intra-group and overall agreements. We measured the agreements by using Fleiss' kappa coefficient. As seen in Table 5.4, the value of overall agreement for each category is low when compared to the agreements between the members of individual groups. Also, the agreements for Hate Speech and Offensiveness are relatively higher than the other categories. We hypothesize that if the level of subjectivity of a task is higher, the value of kappa is low and vice versa. This means that the Stereotype is the most subjective category in the dataset having very low kappa values because stereotypes are often more implicit in nature.

Agreements	**HS**	**Aggress.**	**Offens.**	**Stereotype**
Overall Agreement	0.35	0.30	0.36	0.29
Control Group	0.43	0.34	0.44	0.28
Target Group	0.58	0.37	0.49	0.33

Table 5.4: The group-wise and overall agreements for all the categories.

The computation of pairwise agreements between the members of individual groups and between the members of different groups provided us a fine-grained network of agreements between all the annotators. Such topology of the network indicates the relationships between the opinions of annotators within a group and also between the groups. Tables 5.5, 5.6, 5.7 and 5.8 show us the pairwise agreements in an explainable manner. We can observe that for each of the categories in the BREXIT dataset, the pairwise agreements between the members of same groups are rather high than the agreements between the members of different groups.

The value for HS category for the members of the Control group is between 0.41 and 0.44 and between 0.52 and 0.66 in the Target group. However, the value between the members of two groups is between 0.20 and 0.28 which is significantly low. Similarly, for Aggressiveness, the value between the pairs from same group is between 0.26 and 0.44 for the Control group and between 0.29 and 0.48 for the Target group which is high but the value between the pairs of different groups is in between 0.17 and 0.34 which is low than the high value between the pairs from same group. For Offensiveness, we see a similar pattern and the value is between 0.39 and 0.49 for the Control and between 0.30 and 0.36 for the Target group, a relatively higher value but between the pairs of different groups, the value is between 0.27 and 0.35 which is still lower than the other groups. For Stereotype, we saw a slightly different pattern in which the pairwise agreements between one member of the Control group and all members of the Target group are higher than all other pairwise agreements (three values in the top left row of Table 5.8). The pairwise agreements between the members of the Target group (between 0.30 and 0.36) were relatively higher than the Control group (between 0.23 and 0.37).

We can still deduce from the pairwise agreements that the two groups of annotators show much higher intra-group agreements (top-left and bottom-right area of the Tables) than their inter-group agreements (top-right area).

	C2	C3	T1	T2	T3
C1	0.41	0.45	0.22	0.20	0.28
C2		0.44	0.22	0.20	0.23
C3			0.28	0.25	0.27
T1				0.66	0.56
T2					0.52

Table 5.5: The pairwise agreements between the members of two groups for Hate Speech.

	C2	C3	T1	T2	T3
C1	0.44	0.33	0.34	0.27	0.27
C2		0.26	0.24	0.17	0.31
C3			0.27	0.25	0.24
T1				0.48	0.29
T2					0.31

Table 5.6: The pairwise agreements between the members of two groups for Aggressiveness.

This section provides a detailed description of the abusive language detection task. First, we continue with a brief description of the task, the goals we desire to achieve by performing the task, and then we explain the methodology applied to perform the task. Finally, we give details about the evaluation strategy.

	C2	C3	T1	T2	T3
C1	0.49	0.39	0.35	0.28	0.38
C2		0.40	0.34	0.28	0.32
C3			0.27	0.24	0.33
T1				0.55	0.49
T2					0.43

Table 5.7: The pairwise agreements between the members of two groups for Offensiveness.

	C2	C3	T1	T2	T3
C1	0.37	0.23	0.41	0.43	0.41
C2		0.25	0.32	0.29	0.19
C3			0.15	0.15	0.12
T1				0.36	0.30
T2					0.34

Table 5.8: The pairwise agreements between the members of two groups for Stereotype.

5.3 Extrinsic Validation of BREXIT Corpus

5.3.1 Task Description

The extrinsic validation of the BREXIT datasets is substantiated through automatic abusive language detection which is cast as a text classification task. The purpose of the task is to differentiate between hateful and non-hateful, aggressive and non-aggressive, offensive and non-offensive, and stereotype and non-stereotype content in a binary classification scenario, which can be featured by different topical focuses depending on the targets of hate. The main task is divided into many sub-tasks.

The sub-tasks are defined below:
1. Measurement of polarization index: The first sub-task is to measure the level of polarization in opinions given by annotators belonging to different groups (human judges) who annotated the multi-perspetive BREXIT dataset. It is important to note that the polarization index was explained in the previous chapters and briefly described in this chapter.
2. Creating annotator groups: If we do not have any information on the annotators background, we divide the annotators of that dataset into groups by performing an exhaustive search with the help of P-index. The split that maximizes the average polarization index is selected for the division of annotators. In principle, If we have any information on the personal and demographic backgrounds of all annotators, we do not need to find a split by performing an exhaustive search as in such a case, we have a natural selection of annotator groups. In this study, we developed a multi-perspective dataset with manual annotation process in which, we have complete informa-

tion on the annotators so we have a natural selection for grouping the annotators based on the available information.

3. Developing group-wise gold standards: We create separate gold standards for annotator groups. The idea is that each gold standard training data can represent individualized perspective expressed in the annotations by annotators belonging to that particular group. These perspective are modeled using machine learning algorithms for abusive language detection.

4. Classification tasks: Finally, we perform binary classification and evaluate the performance of group-wise classifiers. We also employ an ensemble approach which combines all the perspective-aware classifiers into a an inclusive model. The classification performance is evaluated in terms of accuracy, precision, recall and f1 score.

5.3.2 Measurement of Polarization Index for BREXIT Corpus

Many datasets for NLP tasks are annotated either by experts or crowd-sourced workers. Thus in general, the background of annotators is not known. However, we hypothesize that a group of annotators can be effectively divided according to the characteristics linked to their background, by analyzing their annotations. Here, we do not have crowdsourced datasets. Instead, we have a multi-perspective BREXIT dataset with known annotators' background. In particular, we make use of the polarization Index (P-index) introduced in Chapter 3 and its application, described in the same chapter.

The method leverages the information at the level of single annotations, measuring the level of polarization for all the annotations individually. We believe that the knowledge about a topic, the background, and or social circumstances may generate polarized opinions among different communities. It is important to note that there is a difference between the polarization of opinions and inter-annotator agreement as the latter might be influenced by factors such as the knowledge of a language, text comprehension and interpretation, e.g., of ironic content while the former stems from the level of subjectivity of some phenomena (e.g., hate speech is highly subjective). In our work, we capture the annotator background at a macro-level. It is interesting to note that when we have high polarization among the instances of dataset, it does not necessarily mean that there is low agreement between the annotator groups: according to our definition, we consider the judgements highly polarized when different groups have high agreement on different judgements. On the contrary, there is no polarization at all if we have an overall low agreement between different groups or among the members of same group.

Here we will give a few examples from the BREXIT datasets. To give a few examples with suppose for $k = 2$,

if $G_i^1 = \{1, 0, 0\}$ and $G_i^2 = \{1, 1, 1\}$, then $a(G_i^1) \approx 0.11$ (low intra-group agreement), $a(G_i^2) = 1$ (high intra-group agreement), $a(G_i) \approx 0.11$, thus $P(i) \approx 0.49$.

Alternatively, If we have,

$G_i^1 = \{0, 0, 0\}$ and $G_i^2 = \{1, 1, 1\}$, which means that each group is in total agreement but on different labels, then $a(G_i^1) = 1, a(G_i^2) = 1$, $a(G_i) = 0$, thus $P(i) = 1$.

5.3.3 Perspective-Aware System Modeling

In Section 4.2.2 of Chapter 4, we introduced a method to create group based gold standards to automatically model the perspectives coming from different annotators that may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. The natural selection of the annotator groups is beneficial in terms of creating such gold standards and developing models trained to learn different perspectives of human judges on similar data.

As we assumed that the instances annotated by each group represent the opinions of people belonging to that group. We create separate gold standards; one for each group based on majority voting method and train state-of-the-art classifiers on group-based training data. We believe that the models trained on a gold standard annotated by people with a common personal background can represent these personalized perspectives and also give us an insight how well the classifier performed with that particular group after the models are evaluated.

The ensemble classifier considers an instance positive if any of the group 1 or group 2 classifiers (or both) considers it positive. We call this ensemble "Inclusive". The rationale behind this ensemble is that hate speech is a sparse and subjective phenomenon, where each personal background induces a perspective that leads to different perceptions of what constitute hate. This classifier includes all these perspectives in its decision process.

The Inclusive classifier will naturally have a bias towards the positive class, by construction. Figure 5.3 explains the architecture of perspective-aware system to detect abusive language. The figure was introduced in the previous chapter but it is important to repeat it here as it elaborates the idea of group-wise classification tasks with respect to the multi-perspective BREXIT dataset with multiple categories.

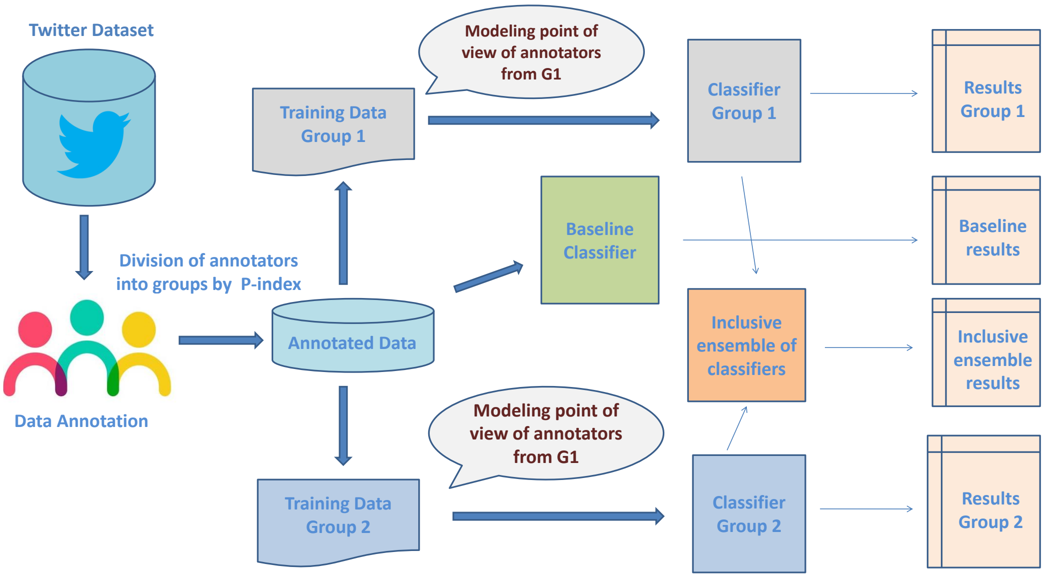

Twitter Dataset
Modeling point of view of annotators from G1
Training Data Group 1
Classifier Group 1
Results Group 1
Baseline Classifier
Baseline results
Division of annotators into groups by P-index
Annotated Data
Inclusive ensemble of classifiers
Inclusive ensemble results
Data Annotation
Modeling point of view of annotators from G1
Training Data Group 2
Classifier Group 2
Results Group 2

Figure 5.3: The architecture of proposed approach for perspective-aware abusive language detection for BREXIT corpus.

5.4 Results

We developed models representing the perspectives of individual groups of annotators by employing the section 5.3.3 of our methodology. The experiments performed in the previous chapters gave us an insight how well the classifier performed with that particular group after we evaluated the models.

We employed the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018], the prediction framework for binary classification. BERT was developed by google and it is considered state-of-the-art in all transformer-based models. BERT uses a bidirectional approach to classify textual data and achieve state of the art performance in many NLP tasks when compared to other available NLP models [Yu et al., 2019]. BERT is trained on bidirectional language representations from unlabelled text from different domains obtained from Internet and Wikipedia pages by considering left and right contexts in a multi-layer architecture [Munikar et al., 2019]. Unlike other language models, BERT can be fine-tuned with just one extra output layer for various downstream tasks without depending on task-specific modifications in the model architecture. We fine-tuned the BERT model on the training sets, keeping the test sets fixed for each dataset, for a fair comparison.

Let us now apply our methodology to the BREXIT dataset, where the personal and demographic background of the annotators are known. We employed the Section 5.3.3 of our methodology for the BREXIT categories. We created separate gold standards as with previous experiments; one for each group based on majority voting and then trained our classifiers on group-based training sets. For all the categories, the training set contains 85% of the data whereas, the remaining 15% constitutes the test set. The baseline results are given by the classifier trained on the original, unmodified training sets. Here, we again employed BERT for the binary classification task. We fine-tuned the BERT model on the training set, keeping the test set fixed for each category. We used the same uncased base English model provided by Google for English (*uncased_L-12_H-768_A-12*). We changed the hyper-parameters and fixed the sequence length at 128 words, batch size to 8 for all the categories. The learning rate was set to $2e^{-5}$. The prediction results for BREXIT categories are shown in Tables 5.9, 5.10, 5.11 and 5.12.

Classifier	Accuracy	Precision	Recall	F1
Baseline	0.845	0.658	0.742	0.684
BERT Control Group	**0.881**	0.573	0.518	0.514
BERT Target Group	0.845	**0.698**	**0.889**	**0.736**
Inclusive	0.839	0.694	0.886	0.730

Table 5.9: The prediction results on HS category.

For all categories, the classification performance generally improved over baseline results.

For HS and Stereotype, the classifiers trained on group-wise training sets and the inclusive classifier always outperformed their counterparts trained on the sets annotated by all the annotators. For Aggressiveness, the accuracy and micro-averaged precision is higher for the baseline but recall and f1-score are showing good improvements over the baseline results. For Offensive-

Classifier	Accuracy	Precision	Recall	F1
Baseline	**0.917**	**0.957**	0.632	0.686
BERT Control Group	0.899	0.759	0.644	0.679
BERT Target Group	0.911	0.798	0.697	**0.733**
Inclusive	0.899	0.748	**0.713**	0.729

Table 5.10: The prediction results on Aggressiveness category.

Classifier	Accuracy	Precision	Recall	F1
Baseline	0.893	**0.896**	0.670	0.720
BERT Control Group	0.893	0.809	0.748	0.773
BERT Target Group	**0.911**	0.839	0.806	**0.821**
Inclusive	0.887	0.782	**0.807**	0.793

Table 5.11: The prediction results on Offensiveness category.

ness, we have a higher recall at the expense of lower precision but very high f1-score.

It is also important to note that we see substantial improvements in the recall and f1-score for all the categories. It is also interesting to note that the Target group provides best results for all the datasets except Stereotype where the Control group performed better than the Target group. A possible reason might be, since Stereotype is often more implicit, it is recognised better by experts in linguistic annotation (group Control).

Finally, the results of the Inclusive classifier provide best recall for almost all the categories showing that including multiple perspectives into the learning process is beneficial for performance boost.

5.5 Qualitative Exploration of BREXIT dataset

In this section, we perform a deep qualitative analysis of the multi-perspective BREXIT dataset with all categories. The most polarizing tweets can be analysed with the help of P-index in order to understand the multiple facets of the phenomenon under consideration and extract the most subjective controversial topics and keywords related to certain events, such as immigration, in a dataset. Consider as an example the following tweet from HS category in the BREXIT dataset:

> 🐦 *Put a loving face of a raping murdering savage refugee terrorist up. https://t.co/rMdb5K (P-index=1), labels (1 1 1 0 0 0)*

In above example, all the members of the Target group marked the message as racist and hateful while the members of the Control group marked it as conveying no hate or racism. This also shows the level of subjectivity of such annotation task. Strong lexical expressions such as *raping or murdering* may have been perceived by one group as the indicators of extreme hate.

Classifier	Accuracy	Precision	Recall	F1
Baseline	0.869	0.611	0.536	0.541
BERT Control Group	**0.881**	**0.693**	0.522	0.514
BERT Target Group	0.792	0.623	0.709	0.640
Inclusive	0.792	0.631	**0.730**	**0.649**

Table 5.12: The prediction results on Stereotype category.

We also found that the hashtags *#illegals and #rapists* in some tweets are highly controversial too. In fact, there are other messages in the BREXIT categories containing such hashtags and are characterized by maximum P-index values. Interestingly, we did not find a single tweet in HS category, where the Control group expressed opposite opinions with respect to the first example i.e, they marked the message as hateful. The closest example is given below, where two members of the Control group marked it as hateful, whereas all members of the Target group marked it as not hateful:

> 🐦 *Boost UK exports, deport Anna Soubry, apparently #brexit https://t.co/mV0rA1gndQ*
> *(P-index=0.49), labels (0 0 0   1 1 0)*

It is interesting to note that after ranking HS instances by measuring the P-index values, we only found a few tweets, 13 to be precise (1.13%), on which all the annotators agreed that they contain hateful messages (p-index = 0). This shows how HS is a highly subjective phenomenon, and that the sensitivity level of people with different cultural backgrounds to a particular topic or event plays an important role in the annotation process.

Similar patterns are observed in other categories. The level of polarization of a message also seems to correlate with the simultaneous presence of different controversial issues, such as in the following example of a tweet annotated as aggressive:

> 🐦 *The two biggest evils in the world: Islam and lily white progressive idiots. https://t.co/RbKgvk7ibN*
> *(P-index=1), labels (1 1 1   0 0 0)*

According to the Target group, the above example not only contain hate speech but also aggression against Islam but, the Control group did not perceive it in a similar way. Therefore, different groups perceive it differently based on their inner moral values and the cultural background.

We also measured the number of instances having maximum polarization (P-index = 1) for all the BREXIT categories. For HS, we counted a total of 62 disagreements (5.5%), 12 cases in the Aggressiveness category (1.1%), 50 cases in the Offensiveness category (4.5%), and also 12 instances in case of Stereotype category (1.1%). This shows that the polarization is fairly unbalanced, that is, one group of the annotators marked an instance as positive whereas the other marked it as a negative class.

We observe that if we compare all the categories, HS and Offensiveness are more related phenomena. Since they are related, we were expecting a higher number of common tweets with maximum polarization (P-index = 1) in both categories but interestingly, we only found 13 common instances (maximum polarization) between them.

It is also interesting to observe that we did not find a single instance in HS category in which the Control group marked it as hate speech because all 62 cases were marked as hateful by the Target group. For Aggressiveness, out of 12 cases, the Control group marked only 3 as aggressive in nature whereas the Target group marked 9 as the aggressive content. Similarly, for Offensiveness category, there was only 1 tweet marked as offensive by the Control group and rest of the 49 tweets were marked as offensive by the Target group. Finally for Stereotype, 5 instances were marked as contacting stereotype by the Control group and 7 of them are marked as contacting stereotype by the Target group. The analysis is shown in Figure 5.4:

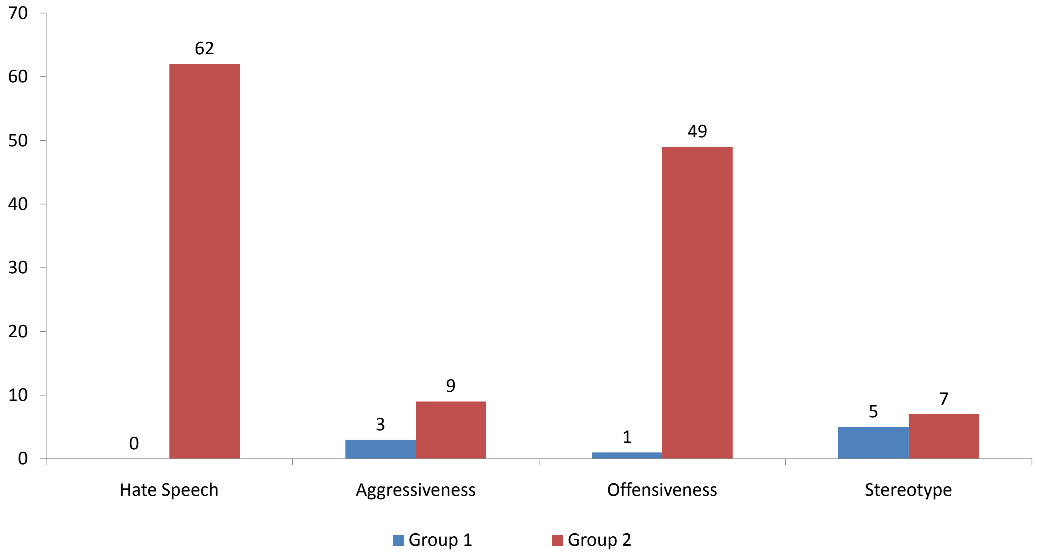


Figure 5.4: A comparison of the cases marked as hateful by two groups for all categories.

We push this kind of analysis further, by looking at the predictions of group-based classifiers, with particular focus on the cases where the two classifiers diverge in their predictions. Analysing the predictions, we counted 40 classifier disagreements in HS (23.8%), 10 cases in the Aggressiveness (5.9%), 15 in the Offensiveness category (8.9%), and 39 in the Stereotype category (23.2%).

The disagreement in all the categories is always fairly unbalanced, that is, one classifier predicts the positive class and the other classifier predicts the negative class in 97.5%, 60%, 60%, and 95.1% of the cases for the HS, Aggressiveness, Offensiveness, and Stereotype categories respectively. Figure 5.5 shows a comparison of different values.

The numbers from the group-wise classifiers already show how polarized the two groups are in classification results but, we also wanted to explore the tweets from the test sets to find the topics and keywords responsible for causing high polarization.

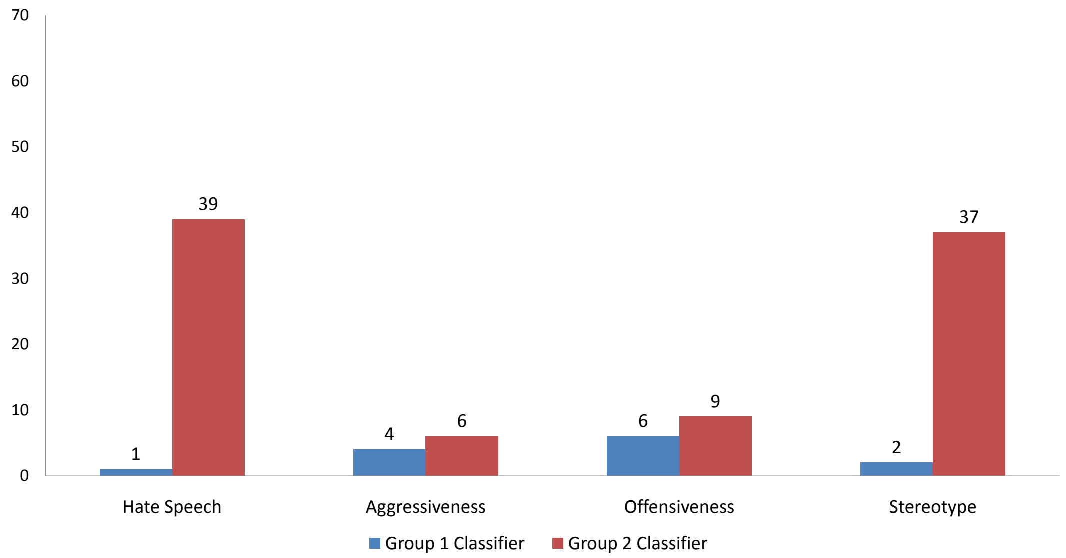


Figure 5.5: A comparison of the cases marked as opposite by group-wise classifiers for all categories.

In HS data from the test set, we found that the keywords in most polarized tweets mention a restricted number of highly related topics (Muslim invasion, Muslims rule over the world, illegals, stealing jobs, deporting or banning immigrants, terrorism, radical Islam, and rapists etc.) very consistently, while such topics are otherwise distributed in the corpus equally among other topics such as immigration control, negative effects of immigration, racism, economy, Brexit voting, working environment, politics and Islam.

Interestingly for Aggressiveness category, the few tweets that were polarized mentioned topics related to Islam and Muslims. Most of the tweets contained derogatory words such as radical Islam, global genocide by Muslims, disgraceful and evil religion, biggest idiots, Muslim insurgency and similar other related words. The few polarized tweets from the Offensiveness test set contained similar keywords that we found in the HS data with the addition of some extra keywords like Muslim caliphate, dumb nigger (referred to Obama), radicalized and savaged refugees etc.

The polarized tweets from the Stereotype test set, like HS, consistently mentioned similar topics which are highly related (Muslim invasion, illegals, global invasion, massive migration, deporting immigrants out of the UK, defeating Muslim nations, terrorist attacks, rapists, and derogatory words for Pakistanis). There were also additional topics related to war over economy and world rule, Brexit due to racism and similar topics mostly related to immigration.

It is interesting to note that we also found similar statistics regarding the keywords and topics in the gold standard data of both groups for all BREXIT categories. All these keywords and topics are strongly linked to the cultural and demographic background of people in the

Target group which derived their perception and influenced the stance on these messages. This means that the group-wise classifiers successfully picked up the keywords and topics causing polarization in the annotator groups.

From a qualitative point of view, following the manual exploration of the results of predictions per category, we also noticed how a considerable portion of the cases where the two classifiers are in disagreement are relatively hard to interpret without access to external knowledge, that is, they mention people, their faith, nationality, migration, and topical events. For instance, from the HS category:

> *I am sure HRC and Obama Bin Lying can put a good spin on it! Looking 4 underwater route 4 illegals maybe? https://t.co/z8fJGWy8n9*

Mentioning the former US president Barak Obama, who was at the center of a controversy involving the import of immigrants. Highly controversial topics also induce confusion between the classifiers, mirroring what we observed for the annotators and the polarity index, such as in this example from the Stereotype category:

> *@[redacted] London is the Muslim Brotherhood's Head Office..I wonder when they'll be given their walking papers. #Brexit @[redacted]*

The results of our manual analysis therefore confirmed that the supervised models trained on gold standard data annotated by partitioned groups of annotators successfully pick up prototypical abstractions of their background, as well as the indecisiveness due to high level of controversy.

5.6 Summary

In this chapter, we presented a case-study in which we developed a multi-prospective BREXIT corpus with four annotation categories. These categories include HS, Aggressiveness, Offensiveness and Stereotype. we performed experiments on the newly developed resource. We preformed an exploratory analysis in relation to several personal characteristics seemed interesting for our analysis. Our results show interesting patterns, encouraging us to further pursue the development of the proposed methodology, implement the more recent state-of-the-art deep learning techniques and compare the performances of different dataset categories.

In the novel approach, we experimented with abusive language corpus by involving the migrants as the victims of online hate who annotated the dataset under the hypothesis that some common characteristics (cultural, demographic, ethnic, etc.) can influence their perception on certain phenomena and also shape their opinions on social media posts. Interestingly, our results indicate a higher correlation between the data annotation process and the personal background of the annotators. Such signals may help us to shed some light on the abusive language phenomenon and how it affected the immigrants of a society when they are the targets of online and offline hate. The study is an extension of our work presented in previous chapters in which we developed novel methods to measure the level of polarization in HS messages and divide the annotators into groups based on the polarization of their judgments also effectively acting as an empirically-driven substitute to the unavailability of information on the background of the annotators, e.g., in a crowd-sourcing scenario.

The experimental results suggested that the methodology improves the classification performance on the novel BREXIT corpus, by training multiple, group-based classifiers instead on a single, all-comprehensive one. The results are improving for all the categories over the baseline. Moreover, the addition of an "inclusive" ensemble classifier in the experiments further boosted the performance by outperforming the baseline models with a particular increase in the recall on abusive messages.

By observing the raw matrix, we can see that prediction are results for HS category are the best among the others. The improvements can be seen in accuracy, precision, recall and f1 score. For Aggressiveness, the improvements can be seen only in recall and f1 score. For Offensiveness, precision drops but other factors are showing considerable improvements. For Stereotype, we see best results as with HS category in which everything is improving at a considerably.

We also preformed a manual analysis of the BREXIT corpus by measuring the P-index and exploring the most polarized tweets by ranking the instances filtering the tweets with highest divergence of opinions for further qualitative analysis, to better understand the controversial topics and issues, and create more compact and better guidelines to improve the annotation quality and hence solve the inconsistencies in the gold standard data. Our results suggest that the migrants group was more sensitive towards the words usually linked to Islamophobia and xenophobia such as biggest evil, evil Islam, idiots, savage refugees, murders, rapists etc. It seemed that such words influenced the decisions of migrants group while annotating abusive language datasets. We can deduce that the knowledge coming from the disagreement and the polarization of opinions is indeed highly informative and valuable in abusive language detection. The results also helped us to understand how the victims of online and offline abuse feel when they are targeted for the crimes they never committed.

In future work, we also plan to explore multi-dimensional approach towards the background of annotators, including native language, other demographic factors, and how they interplay with the measured polarization of their annotations in a group. We plan to apply the methodology presented in this paper to other abusive language phenomena such as cyberbullying, radicalization, and extremism. We are also interested to test the method on sentiment analysis tasks applied to specific domains such as political debates.

The finding of this research (which are accepted with revisions in Elsevier Information Processing and Management Journal and archived in Akhtar et al. [2021]) would certainly lead to even more interesting insights if we increase the training data and also involve the immigrants in the annotation process. We also plan to expand the methodology and test it on other domains such as user reviews.

In the next chapter, we will focus on the polarization as an indicative of the conflicting perspectives held by different annotator groups. We will propose new quantitative methods to model this phenomenon. Moreover, we will introduce a method to automatically identify shared perspectives stemming from a common background. To test our methodology, we will experiment on several corpora in English and Italian including some additional datasets, manually annotated according to their hate speech content, validating prior knowledge about the groups of annotators, when available, and discovering characteristic traits among annotators with unknown background. We will focus only on the identification of conflicting perspectives and will not perform any classification experiments.

Chapter 6

Mining Annotator Perspectives in Hate Speech Corpora with Clustering Technique

The majority of people are not familiar with the laws that protect online users against online frauds, hate speech and foul language. The perspectives expressed by users on a social phenomenon in an online discussion might differ from the perspectives of other users. The background knowledge and prior experiences of users may shape their current understanding about concepts or beliefs related to a topic and influence their perspectives in giving opinions on the topic of mutual interest. Also on social media platforms, individuals get a chance to explore and express their point of views on various aspects of emerging topics, social and political debates and get instant feedback and may find others who may have similar stances on the topics of mutual interest.

Traditionally, the disagreement in annotation is treated mostly as noise. Now, more and more often it is considered as a source of valuable information. In our work, We mainly investigated a particular form of disagreement which occur when the annotated datasets are based on the topics which are subjective and controversial in nature. These controversial topics can induce a certain degree of polarization among the annotators' judgments which are based on the perspective the annotators rely on while make judgements on such textual data. It can be argued that the conflicting perspectives held by annotators of different groups are main reason for this polarization of opinions.

In Chapter 3, we introduced a quantitative measure of the polarization of the annotation induced by the controversiality of the messages. In Chapter 4, we leveraged the polarization in the annotation to create multiple perspective-encoding classifiers, boosting the classification performance in the process. We extended our work in Chapter 5 by developing a novel resource, multi-perspective BREXIT corpus on abusive language annotated by six annotators with the information on the background of annotators. Although the novel method developed and explained in Section 3.2.1 is a useful tool to aid the divison of annotators into groups based on some common characteristics but it only works when there are less number of annotators to group. A careful study of the available literature on clustering analysis revealed that there are no significant studies on the use of clustering techniques to cluster the annotators for measuring perspectives in abusive language detection.

In this chapter, we also aim at further exploring the perspectives expressed by the annotators in HS corpora, and particularly at providing a methodology to quantitatively study the emerging groups of annotators holding different, and sometimes conflicting, perspectives. The research in this chapter is different from the work explored in previous chapters in the context of using feature-based agreement methods to cluster the annotators into groups by using state-of-the-art clustering algorithm. With the method (p-index) employed in previous chapters to aid the division of annotators into groups, it was not possible to methodologically divide a very large number of annotators into common groups. Furthermore, we required full annotation matrix in order to implement the methodology. The clustering techniques might be the best solution to handle the limitations faced by the P-index method. One of the goal of this work is to identify precisely defined perspectives, described in terms of increased sensitivity towards textual content expressing attitudes such as xenophobia, Islamophobia, and homophobia. More specifically, in this work, we deal with shared perspectives, that is, the set of factors that cause a certain annotation by a group of human annotators (each holding an individual perspective). By analyzing the annotation with computational methods, we aim at:

i) Distinguishing groups of annotators holding different shared perspectives.

ii) Identifying the nature of the shared perspectives, providing a human-readable description.

The chapter is divided into the following sections: Section 6.1 introduces our motivation and some background information that helped us to conduct the research study presented in this chapter. The research questions are explained in Section 6.2. Then, Section 6.3 lists all the datasets that were available to perform the quantitative analysis. Section 6.4 highlights concepts important to understand for this research study. We describe the methodology in Section 6.5. In Section 6.6, we present the result of an experimental evaluation on several datasets of hate speech in social media (described in Section 6.3), including a discussion and qualitative analysis on the datasets with example tweets, and finally summarizing the important finding of this research study in Section 6.7.

The majority of the findings in this chapter are published in 2021 in Fell et al. [2021].

6.1 Background and Motivation

Mostly, the research related to the identification of perspectives mainly focuses on the perspective of the author of the messages. The literature is typically concerned with subjective phenomena in natural language, such as abusive language, where an abundance of expressions of emotions, opinions and sentiments is found. Subjective language is considered a catalyst for multiple perspectives [Wiebe et al., 2004, Riloff and Wiebe, 2003] and varying opinions at sentence level [Yu and Hatzivassiloglou, 2003]. Political discourse analysis is an important research area and many researchers worked in identifying different perspectives on political topics including election campaigns as a qualitative analysis task [Pan et al., 1999]. Lin et al. [2006] automatically identified perspectives at the document and sentence level with high accuracy by developing statistical models and learning algorithms on articles about the Israeli-Palestinian conflict.

Most modern approaches to Natural Language Processing tasks rely on supervised Machine Learning. This is true, among other tasks, for text classification tasks such as abusive language and hate speech detection [Zampieri et al., 2020, Basile et al., 2019]. However, while bias in datasets has been investigated [Wiegand et al., 2019, Razo and Kübler, 2020], the bias in the annotation of the datasets used for training hate speech models is relatively less studied. Highly controversial topics, such as hate speech, are a rich source to identify and analyze conflicting perspectives in online environments. When social media users express different opinions on topics or social issues, the text depicts high level of controversy due to varying perspectives [Popescu and Pennacchiotti, 2010]. When such phenomena are manually annotated by human judges, high controversy is bound to have an impact on such annotations, in terms of agreement between the human judges.

Hate speech is a highly *subjective* phenomenon. While no phenomenon is neither totally subjective nor totally objective, the position of hate speech on a hypothetical *inter-subjectivity* spectrum [Maul et al., 2019] is far from the center, as its judgment is influenced by factors such as socioeconomic background, ethnicity, gender, and so on [Warner and Hirschberg, 2012]. Moreover, the hatred is typically directed towards targets carrying specific socio-economic, cultural, or demographic traits, which are likely aligned to the factors influencing the judgments of hateful messages by human annotators. Indeed, messages containing hateful content are often *controversial*, that is, they reference events, people, and issues that prompt very different reactions depending on the recipient of the message [Popescu and Pennacchiotti, 2010].

6.2 Research Questions

In this chapter, we will try to answer the following research questions.

1. Can we cluster the annotators into groups by using feature-based agreements with clustering techniques on the datasets with very large number of annotators and without a full annotation matrix (empty annotations)?

2. Can clustering annotators into groups by employing feature-based agreement methods help us to identify the individual and shared perspectives held by annotators stemming from a common background (known or unknown)?

3. What are the best ways to visualize the identified perspectives to understand the controversial topics and events?

In order to test these research questions and to serve a multilingual perspective, we devised quantitative methods and experimented on different Twitter datasets in English and Italian, featuring different forms of hate speech.

First, we provide a formal definition of *perspective* in the context of the annotation of NLP datasets, hinging on the difference between *label agreement* and the novel concept of *feature agreement* We then empirically demonstrate the emergence of perspectives in real datasets of hate speech, computed with a straightforward yet effective procedure, and shown in the form of word clouds and selected examples.

6.3 Datasets and Annotation Process

The experiments described in this chapter are conducted on several hate speech corpora. These corpora all consist of Twitter messages (tweets); they are published in various research studies on hate speech. These datasts were already explained in the previous chapters. We will just give a short introduction to them. However, we employed new datasets in this chapter, the addition of new datasets helped us to further expand our analysis. This section provides details on the datasets, the annotation process with scheme and guidelines, and the information on the annotators. In this work, we focus on hate speech corpora. Each corpus is labeled by several annotators and exhibits a certain degree of disagreement, as expected by this kind of data. We measure the agreement in terms of the Krippendorff Alpha reliability [Krippendorff, 1970]. Table 6.1 gives an overview of the different datasets used, presented in rest of this section.

6.3.1 Hate Speech Dataset on Brexit

The first dataset that we will employ for our analysis is hate speech dataset on the Brexit phenomenon, which is developed as a multi-perspective dataset to automatically detect abusive language on social media with the intention to model annotator perspectives and polarized opinions. The details of the dataset are already presented in Chapter 5. It is important to note that we only use the hate speech category of the BREXIT dataset.

6.3.2 Sexism

The Sexism dataset is derived from the hate speech corpus by Waseem [2016] and is available on a Github repository[1] and explained in Section 3.3. The original dataset has four annotation categories: *sexism*, *racism*, *both*, and *neither*.

6.3.3 Racism

The Racism dataset was complied by following the same procedure, annotation scheme and guidelines applied to the Sexism dataset. The labels *racism* and *both* were mapped to *racist*, while the labels *sexism* and *neither* were mapped to *non-racist*. The Sexism and Racism datasets are therefore significantly overlapping.

6.3.4 Homophobia

The Homophobia dataset, in Italian language, was developed to detect and monitor homophobic and hateful contents on social media against the LGBT+ community. We converted the Homophobia dataset in a binary labelled dataset, changing the labels *not-homophobic*, *doubtful* and *neutral* to *not homophobic* and leaving the label *homophobic* intact to identify the homophobic content in the dataset.

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

6.3.5 Hate Speech Dataset in Italian

The hate speech dataset in Italian language [Florio et al., 2019] (HS Italian) consists of 3,200 tweets collected from TWITA [Basile et al., 2018a] in 2017, partially overlapping with the Italian Hate Speech Corpus Sanguinetti et al. [2018]. The collection was filtered by the authors of the original dataset with a list of handcrafted keywords related to migrants and ethnic and religious minorities in Italy. The tweets were annotated on the Figure Eight platform[2]. A minimum of three annotators annotated the whole corpus, subsequently aggregated by the crowd-sourcing platform to create a gold standard dataset. We requested and obtained the dataset from the authors.

6.3.6 Hate Speech Dataset in English

Davidson et al. [2017] developed a hate speech dataset to perform automatic hate speech detection as a multi-classification task. The authors gathered around 85.4 million tweets from a total of 33,458 Twitter users. A hate speech lexicon containing hateful words and phrases was used to query the tweets. This lexicon was compiled by Hatebase.org and the hateful words in the lexicon were identified by internet users. The authors randomly selected about 27,000 tweets from the dataset by using the keywords from the hate speech lexicon. CrowdFlower (now Appen) workers were hired to manually annotate the tweets. The annotation scheme comprises the labels *hate speech*, *offensive but not hate speech*, and *neither offensive nor hate speech*. The authors developed detailed guidelines with their own definitions of different hate speech terms including the context in which the words were used. Each tweet in the dataset was annotated by three or more annotators. The Davidson dataset is only available for download in an aggregated gold standard form[3]. As we needed pre-aggregated data for our work, we only found aggregated gold standard data on author's GitHub repository. Therefore, we requested the authors to provide us with pre-aggregated data and we are grateful to them for providing us the required format of the dataset.

Dataset	*Lang.*	*Domain*	*Items*	*Annotators*	*Ant./Item*	*Classes*	*Cod/Lab*	*Pos. label %*	*Lab.Agr. α*
BREXIT	en	Twitter	1,120	6	6.0	2	3.0	12.9	0.35
Racism	en	Twitter	6,361	4	4.0	2	2.0	3.6	0.23
Sexism	en	Twitter	6,361	4	4.0	2	2.0	15.0	0.58
Homophobia	it	Twitter	1,859	5	5.0	2	2.5	15.8	0.35
HS Italian	it	Twitter	3,199	14	4.0	2	2.0	26.6	0.23
Davidson	en	Twitter	27,341	111	3.0	3	1.0	71.2/10.0	0.58

Table 6.1: Key statistics of the datasets used for the experiments. Ant./Item: annotations per item, Cod/Lab: coder to label ratio. Classes: possible labels (negative: no hate speech, positive: hate speech). The positive labels in the Davidson dataset can be either *offensive* or *hate speech*.

[2] https://www.figure-eight.com/, now Appen.

[3] https://github.com/t-davidson/hate-speech-and-offensive-language

6.4 Defining Perspectives

In this section, we will explain some important concepts related to our work and are important to understand before we elaborate our methodology.

Individual Perspectives Given a (possibly infinite) list of items on a topic T, an annotator A judges these items, according to their opinion on each of them. We call this labeling the *individual perspective of annotator A on topic T*. By modelling annotator perspectives as vectors, we can compare them quantitatively.

In order to identify perspectives in annotations, we require items to have disagreeing annotations. This is only possible in the case where annotations have not previously been aggregated into a single label, what in Campagner et al. [2021] has been called *diamond standard*. This is in contrast with the usual *gold standard* paradigm where multiple annotations are harmonized into one gold label, often implemented by majority voting. Under the paradigm of annotator perspectives that we have introduced above, the reduction of multiple labels (annotator opinions) into a gold label by majority vote is equivalent to taking the *majority perspective*.

Limitations In practice, we have two restrictions: first, we do not have infinite items that exhaustively describe a topic, and second, the items, however carefully selected or constructed they may be, are not always concerned with only the topic T. For the sake of this work, we assume that an annotator's value judgements on a finite number of items on topic T defines his perspective on T. Furthermore, the topic in this work is limited to different forms of *hate speech*. So, the perspectives we uncover in this work can be paraphrased as *what is annotator A's perspective on hate speech?*

Shared Perspectives While each annotator takes his own perspective, we are more interested in finding perspectives which are shared amongst annotators. We call perspectives p_A, p_B shared based on their similarity. We employ clustering to find clusters of annotators that share perspectives.

While shared perspectives arise from an agreement of annotators on item labels (label agreement), we also aim to understand how shared perspectives are linguistically defined. To this end, we analyze the importance that different annotators give to different linguistics features, i.e. which words annotators in a shared perspective agree to be important (feature agreement).

6.5 Methodology

The ideas behind the method of this work which is based on the feature agreement between the annotations of a given datasets was devised and destined by our colleague and the first author of the paper submitted on this research work. The methodology was then implemented on given datasets by the author with the help of co-authors in a quantitative scenario. In order to mine a perspective in a given dataset, we postulate a two-step procedure.

In order to mine a perspective in a given dataset, we postulate a two-step procedure. First, we **detect** the perspectives. To this end, we measure how much annotators agree on labels, the *label* *agreement*. Second, we aim to **understand** what the perspective means. To this end, we measure to what extent annotators agree on the importance of linguistic features of the items, the *feature agreement*. For the sake of a simpler explainable solution, in this work we only use *unigrams* as linguistic features. However, the method is agnostic to the type of feature extracted from the messages, and it could therefore be used in conjunction with other, more refined features. For instance, annotators holding the perspective that the word "fag" is especially hateful, tend to always label items containing this word as *hate speech*, i.e., they exhibit a high feature agreement on this unigram.

We finally perform analysis on annotators that form the same clusters by label agreement and feature agreement. For these annotators we know that they agree in individual perspective (label agreement) and they also agree on the importance of linguistic features of the items (feature agreement). This enables us to (start to) explain their shared perspective.

Formally, consider a given topic T. We postulate that annotators and their individual perspectives influence how they annotate different items related to T. This is particularly relevant to annotation tasks that exhibit a high degree of *subjectivity* as here the influence of the perspective on the ratings may be higher. According to a common definition, a judgment is considered *subjective* when it is mainly "based on, or influenced by, personal feelings, tastes, or opinions."; we usually contrast this concept with that of *objective*, a term that characterizes judgments that, ideally, are not influenced by personal feelings or idiosyncrasies and which, on a practical level, the vast majority of people would see and label in the same way. For instance, in hate speech detection, different annotators have been shown to diverge highly in their ratings and are polarized [Basile, 2020]. Furthermore, the offensiveness of words depends on the context in which the words are uttered [Pamungkas et al., 2020b]. For example, consider the difference between the use of the word "nigga" in a Rap song, where it is considered as lowly offensive, as opposed to using such words in a political discourse, where it is understood as highly offensive. We assume that annotators implicitly or explicitly take perspectives on topics, and we model this as described in the following.

6.5.1 Label Agreement

We measure the pairwise label agreement in terms of inter-annotator agreement. We use Krippendorff's alpha reliability [Krippendorff, 1970] and cluster the annotators based on the label agreement. Since we have very large number of annotators for some datasets and also the data annotations are sparse, with the issue of missing entries for some tweets. We applied Krippendorff's alpha because it can handle various sample sizes, categories, and the numbers of raters, and can be applied to any measurement level (i.e. (nominal, ordinal, interval, ratio).

Note that Krippendorff's alpha is also defined in the case of incomplete annotation, i.e., where not all annotators covered all the instances. This is a typical scenario in crowdsourcing, but could happen with other annotation procedures as well.

6.5.2 Feature Agreement

We want to measure whether annotators agree on the importance of linguistic features of the textual items. In this paper, we use a simple *bag of words* (BOW) model, i.e., texts are modelled as bags of words and the features are unigram counts. Feature agreement between annotators i, j arises when i and j give similar *importance* to features. As a measure of agreement of the annotations on a single message, we use the normalized χ^2 statistics, that is, a test of independence of the distribution of the annotations against a uniform distribution. We measure the importance of each feature to an annotator by computing the chi-square (χ^2) statistics between the feature distribution and the label distribution in the annotator, following a univariate feature selection approach. This measures how the annotator's label depends on the presence of a word in an item. For instance, the presence of the unigram *bitch* often coincides with the label *hate speech*, while this is not the case for the word *sunny*. The χ^2 statistics captures this; it is much higher for *bitch* than it is for *sunny*. When annotators tend to agree on the importance of words, they exhibit an overall high feature agreement.

Since the χ^2 statistics requires a dense label matrix, if an annotator has not labelled an item, we insert the negative label (i.e., *not hate speech*). Truly unimportant words then correctly get low importance, while truly important words get assigned a somewhat diminished importance.

6.5.3 Label Feature Agreement

We consider two different ways of clustering the annotators: by label agreement and by feature agreement. These two clustering techniques sometimes differ, for instance, when two annotators agree on the item labels (label agreement), but do not agree on the importance of words related to those labels (feature agreement). Since our goal is to find shared perspectives and describe them via feature agreement, we analyze all annotators that cluster in the exact same way in both labels and features. The effect of our label feature overlap procedure on one of the datasets (the Davidson dataset, introduced below) is illustrated in Figure 6.1.

6.5.4 Cluster Analysis

Given the clusters of annotators we obtain, we analyze certain cluster properties statistics, how the clusters differ and which words are important to each cluster. Specifically, we perform the following analyses:

- Quantitative cluster description: the number of annotators in the cluster, the positive label rate %, the label agreement α, the number of features, and the feature agreement β. We compare the cluster numbers also with the numbers for all annotators disregard of their cluster affiliation.

- Qualitative cluster description: we inspect the most characteristic unigrams for the clusters, i.e. the words with the highest *relative importance* R to the cluster. We illustrate the most important words for the clusters. Furthermore, we look into the most important shared words for all annotators disregard of their cluster affiliation.

- We inspect examples that are polarized between the clusters, i.e., they are annotated with disagreement between the clusters. These examples often carry *important words* as vocabulary.

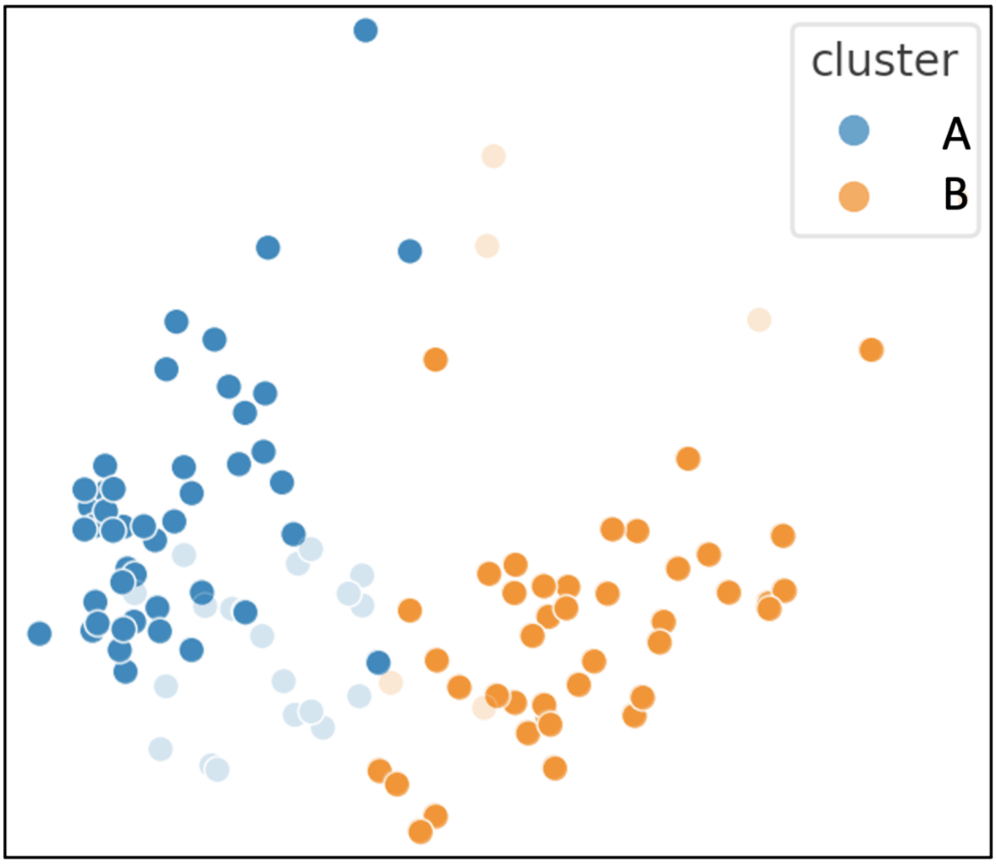


Figure 6.1: Clustering of the annotators of the Davidson dataset into two shared perspectives A and B, according to α agreement. After label feature overlap, only the annotators with the same cluster affiliation under β agreement (solid color) remain, while the ones that switch clusters (transparent color) get removed.

6.6 Results and Evaluation

We performed analysis on all the datasets that we introduced in the previous section. Table 6.1 gives an overview over their key metrics. An important factor for our experiments is what prior knowledge we have about the annotators that annotated the datasets. Where such background information is given, we can confirm or reject our findings by comparing our empirically found annotator clusters (shared perspectives) with groupings of human annotators. As stated in the dataset section before, we have the following information on the dataset annotators. On the BREXIT dataset: the personal details of all annotators such as cultural and demographic background and ethnicity are known. On the Sexism'/'Racism dataset: the annotators were a mix of experts (feminists and anti-racism activists) and crowd-sourced workers. The annotations from experts were aggregated into a single label by majority voting and, due to privacy concerns, no information was provided about the experts. On all other datasets: no background information

Dataset		*Cluster A*	all raters	*Cluster B*
BREXIT	Raters	3	6	2
	Positive labels %	**20.5**	12.9	**5.8**
	Label agreement α	**0.58**	0.35	**0.44**
	Features		266	
	Feature agreement β	**0.86**	0.7	**0.86**
Racism	Raters	3	4	1
	Positive labels %	4.2	3.6	**1.7**
	Label agreement α	0.25	0.23	-
	Features		996	
	Feature agreement β	0.96	0.91	-
Sexism	Raters	3	4	1
	Positive labels %	16.4	15.0	**11.1**
	Label agreement α	0.63	0.58	-
	Features		996	
	Feature agreement β	1.0	0.99	-
Homophobia	Raters	2	5	2
	Positive labels %	**9.3**	15.8	**22.1**
	Label agreement α	0.38	0.35	0.39
	Features		361	
	Feature agreement β	0.67	0.67	**0.77**
HS Italian	Raters	7	14	7
	Positive labels %	30.9	26.6	22.3
	Label agreement α	**0.03**	0.23	**0.42**
	Features		623	
	Feature agreement β	0.34	0.35	**0.48**
HS English	Raters	45	111	41
	Positive labels (offensive) %	77.2	71.2	65.6
	Positive labels (hate speech) %	**4.1**	10.0	**15.4**
	Label agreement α	0.64	0.58	0.60
	Features		2366	
	Feature agreement β	0.22	0.29	**0.46**

Table 6.2: Quantitative cluster statistics for different datasets.

on the annotators is available.

Note that the HS Italian and Davidson datasets are sparsely annotated, as annotators have only labeled a fraction of the instances. This is in opposition to all other datasets which have a dense annotation matrix.

Table 6.2 gives an overview over the quantitative statistics and differences between the clusters. As for the qualitative analysis, we provide word clouds of the important words for each analyzed corpus as well as a few examples where important words are shown in context. For all clustering experiments we used the experimental setup as described below.

6.6.1 Experimental Setup

- Pre-processing: we removed URLs and Twitter handles (username) from the tweets, tokenized them using the NLTK[4] Tweet Tokenizer and lemmatized them using spaCy[5].
- BOW features: we created the BOW feature space with the scikit-learn[6] CountVectorizer, where we set the minimum document frequency to 10. This number was decided based on the fact that some tweets occured duplicate or near-duplicate, because of the dialog structure of Twitter, users will cite each other. Such tweets are contained in several of the datasets, hence we alleviate the problem by setting a rather high minimum document frequency of 10. Furthermore, we counted each word once per document ("binary") and extracted solely unigrams.
- Clustering: we used the KMeans algorithm with different numbers k of clusters, we settled to $k = 2$ which appeared most reasonable based on the inspection of 2D-PCA embeddings of the datasets. This parameter choice makes us conform with the polarization paradigm, i.e. we analyze two conflicting/polarized perspectives.
- Important words: for each cluster, the top 20 words with highest relative importance are considered. For important shared words, the top 20 words not important to a single cluster are considered. The polarized examples are extracted by computing the polarization index method introduced in Chapter 3.

6.6.2 Perspectives in the BREXIT Corpus

As shown in Table 6.1, the BREXIT dataset contains 1,120 tweets, which are rated by each of the 6 annotators by one of 2 possible labels ("yes, this is hate speech" vs. "no, this is not hate speech"). The inter-annotator agreement is measured as $\alpha = 0.35$. The positive label rate is 12.9%. We extract 598 features from the corpus and obtain $\beta = 0.68$ as feature agreement between all annotators.

We observe in the BREXIT dataset the pairwise label agreement α and feature agreement β as depicted in Figure 6.2. The label agreement induces the clustering $A = \{3, 4, 5\}$ and $B = \{0, 1, 2\}$. The feature agreement, however, gives us the different result $A = \{0, 3, 4, 5\}$ and $B = \{1, 2\}$, i.e. annotator 0 belongs to a different cluster, depending on α or β agreement. Since annotator 0 is closer to annotator 1 in label agreement than to annotator 5 ($\alpha(0, 1) = 0.41 > 0.25 = \alpha(0, 5)$), but the opposite is the case for the feature agreement ($\beta(0, 1) = 0.66 < 0.81 = \beta(0, 5)$). As annotator 0 cannot be definitely attributed to one cluster, annotator 0 is excluded from further analysis, according to the label feature overlap procedure, as described in Section 6.5.3. After label feature overlap, we obtain the clusters $A = \{3, 4, 5\}$ and $B = \{1, 2\}$. Since we know the annotator backgrounds, we know that A corresponds to the migrants with

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

[5] https://spacy.io

[6] https://scikit-learn.org

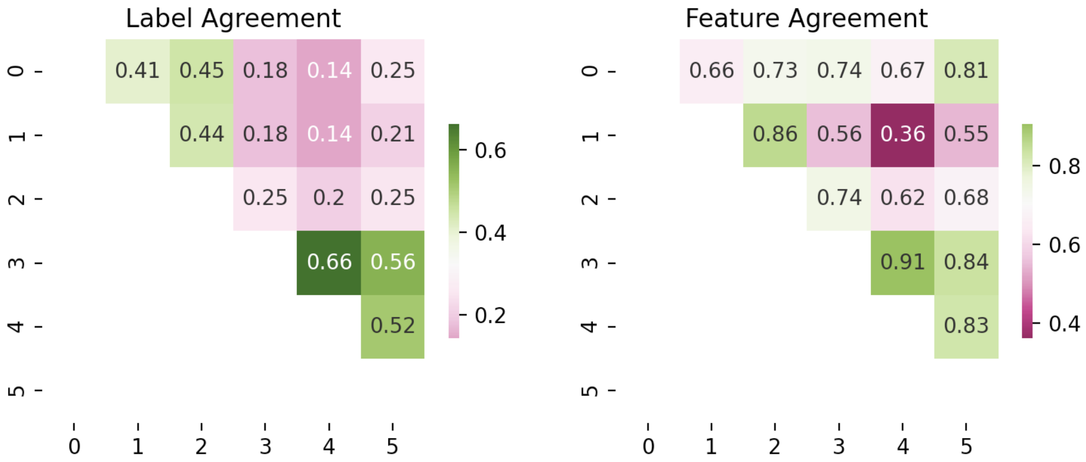


Figure 6.2: BREXIT dataset, discovered label agreement and feature agreement.

Muslim background (*target group*) and that B corresponds to the non-migrants (*control group*). This result effectively validates our clustering methodology based on label and feature agreement to extract perspectives empirically.

6.6.2.1 Quantitative Cluster Description

We find the following differences between the clusters:

- The positive label rate is much higher in A (20.5%) as compared to positive labels in B (5.8%), indicating the annotators in A behave much more sensitive in this task (all annotators 12.9%).
- The label agreement is higher in cluster A ($\alpha_A = 0.58$) as compared to $\alpha_B = 0.44$, indicating that cluster A holds more coherent opinions as B. Both values are much higher than the average, meaning that the groups hold polarized opinions.
- The feature agreement is higher in both clusters ($\beta_A = 0.86 = \beta_B$) compared to the dataset feature agreement ($\beta = 0.7$), indicating polarization of the feature agreements of the clusters as well.

6.6.2.2 Qualitative Cluster Description

Certain words are highly correlated with the positive label in both groups, and some words are specific to the annotator clusters. Figure 6.3 depicts on the left side the words most characteristic for cluster A, in the middle the words that indicate the positive label (“yes, this is hate speech”) for both clusters, and on the right side we show the most characteristic words for cluster B.[7] As we can see, the shared vocabulary (see Figure 6.3, middle) contains words such as “islam”, “kill”

[7] word clouds generated with www.wortwolken.com

and hashtags related to US president Donald Trump (#maga, #trump2016). When inspecting the corpus, we find examples such as the following that exemplify the use of the words; matched words are **bold**. And indeed, in this example, both annotator groups give the positive label.

> *RT @[redacted]: The U.K. Must ban **Islam** and close all mosques! URL*

> *London should kick all Muslim Refugees out before they all **kill** them. **#Trump2016** URL*

Other shared words are “deport”, “paki” (meaning pakistani, a person from Pakistan) that is instantiated by the following example, which both groups gave the positive label.

> *gud news for **paki** bastards-India is now a member of MTCR.Now v hope predator drones to fuck pakistanis’ asshole haha https://t.co/jEHR5iBZ18*

> *#Brexit @[redacted] BETTER @[redacted] DEPORT **ISLAM** PACK YOUR PRAYER RUGS! [redacted] @[redacted] @[redacted] @[redacted] https://t.co/N76eNri0Ry*

From these shared vocabulary examples, we can see that since “islam” is one of the important words, the hate speech in this corpus appears to be at least partially islamophobia. An inspection of the important words for the potential target group of islamophobia, cluster A (see Figure 6.3, left side), supports this claim. We find a specific and distinctive vocabulary related to muslims, invasion, terrorists. The following examples illustrate the important words for cluster A. The examples got the positive label in A and the negative label (“no, this is not hate speech”) in cluster B. Interestingly, while “islam” is a shared top word, we found “radical islam” to be typical for cluster A.

> *FYI world, the ppl of GB supporting #Brexit know if they don’t control their own immigration/borders **radical** Islam will end their lives.*

Stealing jobs, a well-known negative assumption towards foreigners, is also amongst the examples that are important for cluster A:

> *Bloody foreigners coming here & taking our **jobs** though! #Brexit URL*

Note how positive words such as “love” and “thanks” are used in a satirical way in the examples below (labeled positive by B and negative by A):

> *"You can **love** a jihadi as much as you like, but if you ain’t a **muslim** he’s still going to kill you. #brexit for border control."*

> *"**Thanks** to **Merkel**, Germany destined to become a **muslim** country ruled by Sharia Law #Brexit was wise to leave URL"*

> *#Brexit England Voted Against a Strange Union of Billionaires and **Muslim jihadis** / Mujahideen mullahs / ISI Imams https://t.co/l159Tb35FT*

On the other hand, the vocabulary for cluster B (see Figure 6.3, right side) appears to be mostly noise. We conclude from this finding, that, except the shared vocabulary, there is no cluster-specific vocabulary that makes annotators in B label items as positive.

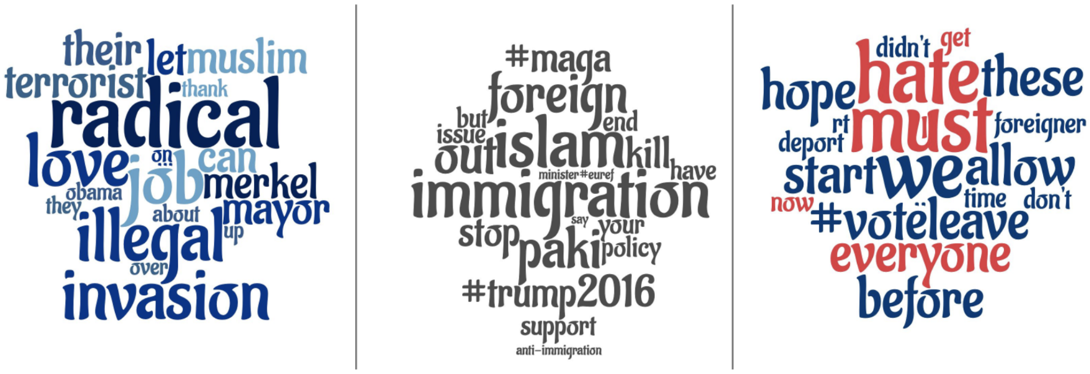


Figure 6.3: Important words in the Brexit dataset. Cluster A words (left), shared words (middle) and cluster B words (right).

Identified Perspectives Overall, we find two polarized groups, both by label and feature agreement. Cluster A - the target group - is much more likely to give the positive label and this group of annotators consistently bases their opinion on a specific and distinct vocabulary which can be described as Islamophobic. Given the background information we have on all annotators, we identify cluster A as the *Muslim perspective* on the topic, highly sensitive to Islamophobic content. In opposition, for cluster B we did not find a positive identity, those annotators form more a counter position to the migrant group, therefore we describe them as control group or *non-muslim perspective*.

Here are some more polarized examples from the BREXIT dataset showing the important words from word clouds as bold.

*You can love a **jihadi** as much as you like, but if you ain't a **muslim** he's still going to **kill** you. #brexit for border control.*

*illuminati own EU, provide for, and apply, **muslim** insurgency, as mean to global genocide; so **muslims** for EU, for now https://t.co/oeShKeINVB*

*My wife just saw a cashier at Clerkenwell [redacted] be racially abused: "fucking **pakis** are the worst". Apparently **racism's** ok now. #Brexit*

*NOTE: George Soros &amp; Angela **Merkel** encouraged the **Muslim invasion** of EU &amp; your countries-&gt; WELCOME TO THE EU https://t.co/QHggsEmp2c #BREXIT*

*Muslims demand London's **Muslim mayor** declare independence from UK in the name of voter **jihad**. #Brexit #KilroyRising https://t.co/pht4oRc10c*

*RT [redacted]: The fact that **Islam** is so mad about #brexit should be an eye opener to Brits on how bad you stumped their **invasion**.*

*As an 18 yo Briton the saddest thing is knowing my children will NOT have their Father blown up in the name of **Islam** #brexit [redacted]*

6.6.3 Perspectives in the Sexism Dataset

In the Sexism dataset we obtained the same clustering according to α and β agreements: $A = \{1, 2, 3\}$ and $B = \{0\}$. Note how the detected clustering coincides with a natural grouping, as annotator 0 is the expert annotator.

Quantitatively, we found the positive label rate to be lower in cluster B (11.1%) as compared to cluster A (16.4%), indicating the annotators in B behave less sensitive in this task. See Table 6.2 for an overview of all cluster statistics.

Qualitatively, we find that words such as “feminazi” and “bitch” are important shared words, and both clusters label the following example positive:

> *@[redacted] @[redacted] @[redacted] @[redacted]2 we'd be ranting **feminazi bitches** during the wrong time of the month URL*

> *RT @[redacted]: Stay strong Missionary Man. Don't let these **feminazi** fascists tell you where you can and cant piss @[redacted]_ @[redacted] @A...*

> *@[redacted] @[redacted] Evolution doesn't have a brain to be sexist with but it did make **men** & women different & unequal biologically.*

> *@[redacted] Time for the Feminazi's to put their money where their mouth is and fight against female oppression in the middle east.. #feminism*

We noticed that the annotators in cluster B tended to pay more attention to the context of tweets as compared to the annotators in A. The following two tweets contain the typical indicator words, but they are indirect quotations (labeled as positive by A and negative by B):

> *RT @[redacted]: .@[redacted] As a mom of daughters, I am asking if you would call them '**sluts**, **sexretary** or FemiNAZI'? **Rush** does & you pay f...*

> *@[redacted] @[redacted] "5 year olds are **sexual**, the feminazi are the ones with problems"- tweeted one of his "friends" *shudder**

Annotators in cluster B are more sensitive to degrading talk about women. We suspect that B also has more background information, associated with hashtags, not conveyed in the tweets (labeled positive by B and negative by A):

> *These 2 delusional, narcissistic hostesses on **#MKR** make me want to vomit my own dinner up.*

On the other hand, it is unclear why the following example is tagged as positive. Probably, the hashtag is an indicator as they know the context better.

> *Katie, Nikki and Pete Evans are everything that is wrong with society. **#MKR***

Here are some more polarized examples from Sexism dataset showing the important words from word clouds as bold.

> *@[redacted] @[redacted] I was blocked by Alexa because i'm pro dick and breast-plates... POW! #Equality* ***#Sexism #Feminazi*** *#Dick #GamerGate*

> *@[redacted] the anti-abuse thing is my main focus, always. but in my spare time? his bullshit *is* abusive towards* ***feminism****.*

> *esr is a* ***sexist*** *bag of* ***dick****s hiding behind ideas of a false meritocracy to defend his position of shitting on* ***women****.*

> *Okay, they've brought it on themselves and I don't like them either, but everyone needs to stop calling* ***Katie*** *fat.* ***#mkr #MKR2015***

> ***Nikki*** *&* ***Katie*** *is what happens when parents continually tell their kids, "You're a winner in my eyes." & avoid constructive feedback* ***#mkr***

> *Ugh these #KillerBlondes are killer boring, and what's with spreading their dna all over the table? Unhygienic -and- gross!* ***#MKR2015 #MKR***

6.6.3.1 Identified Perspectives

The expert (annotator 0) annotates differently. The Amateurs sometimes give a positive label (yes) to a tweet when the tweet is about a quote, in comparison with the expert who is more critical and gives a negative (No) label. This indicate that an expert annotator is more familiar with the hashtags and sometimes can give positive label when there is no visible sexist wordsn in the tweet.

In this dataset, we found different perspectives being held by people with different levels of expertise. Cluster B, coinciding with the expert annotation, seem to base their judgment on a broader context, taking into account clues like specific persons or hashtags mentioned in the messages, while cluster A, coinciding with the crowd-sourced amateurs, performed a more shallow analysis. so experts focus more on specific targets like personalities as models etc. Amateurs have a more generic view on the topic, maybe superficial, as they fail to see that there is a quote, not a statement of the tweet author.

6.6.4 Perspectives in the Racism Dataset

In the Racism dataset we obtained the same clustering according to α and β agreements: $A = \{1, 2, 3\}$ and $B = \{0\}$. Note how the detected clustering coincides with a natural grouping, as annotator 0 is the expert annotator.

Quantitatively, we found the positive label rate is much lower in cluster B (1.7%) as compared to cluster A (4.2%), indicating the annotators in A behave much more sensitive in this task. See Table 6.2 for an overview of all cluster statistics.

Qualitatively, we find that in the shared vocabulary words such as "coon" are highly indicative for the positive label in both clusters:

> *Man wtf is this damn* ***#Coon*** *ass nigga #BillyBlanks doing on that damn commercial? ? #CooneryAtItsFinest*

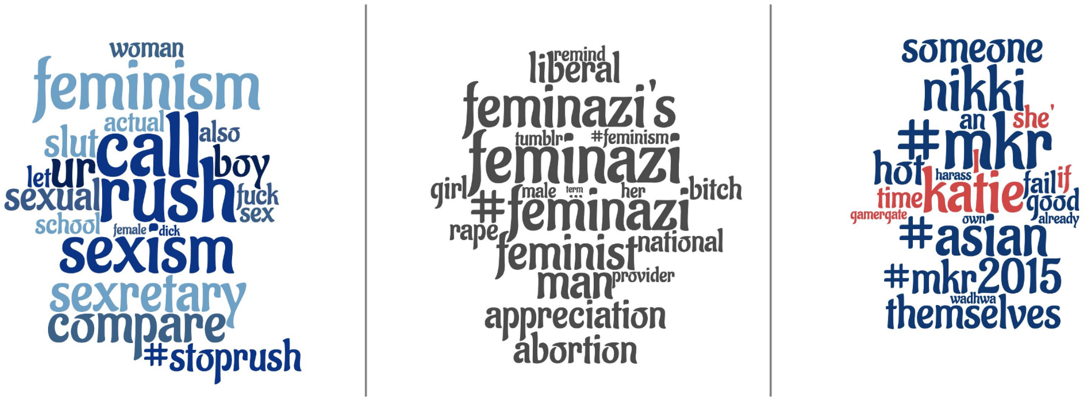


Figure 6.4: Important words in the Sexism dataset. Cluster A words (left), shared words (middle) and cluster B words (right).

> *@[redacted] quit cryin ya fuckin **coon** the dumshit shoukdnt fucked with tha cops yard ape got had @[redacted] @[redacted] @[redacted] **#coon***

> *RT @[redacted]: Another **#Arab** car #terror attack in #Jerusalem #Israel. Will #Obama call it random traffic infringement? http://t.co/Xrx*

The examples below are indicative for important words in cluster A, such as the hashtag #tcot (top conservatives on Twitter) adn white:

> *#Libya Accuses #US of Trying to Put #Muslim #brotherhood in Power http://t.co/4QHFzAh6ku #news #media #tlot #p2 **#tcot** #terror #ISIS #arab*

> ***color** me surprised when i look up dude defending piracy on linkedin and he's some young **white** dude claiming to be a CTO.*

> *RT @thetrudz: And **White** people who bring up queerness or being a survivor as a way to SILENCE **WoC**? Because no **WoC** are queer? None survivors. . .*

While we did not find a clear pattern for polarized examples in cluster A, we found examples related to "asian" and "israel", for instance the examples below show important words in cluster B (labeled negative by A, positive by B).

> *Tokyo Hot n1049 Endless Sex **Drive** - URL #dailyxLover #jav **#asian** URL*

> *baby you can **drive** my car http://t.co/CyhEMc8aVy **#asian** #juicyasian #sex #nsfw #adult #xxx*

> *"RT @[redacted]: **#ISRAEL** FOREVER BLOG/ **Terror attack** in **#Jerusalem** injures seven, #terrorist shot - URL - #ARAB"*

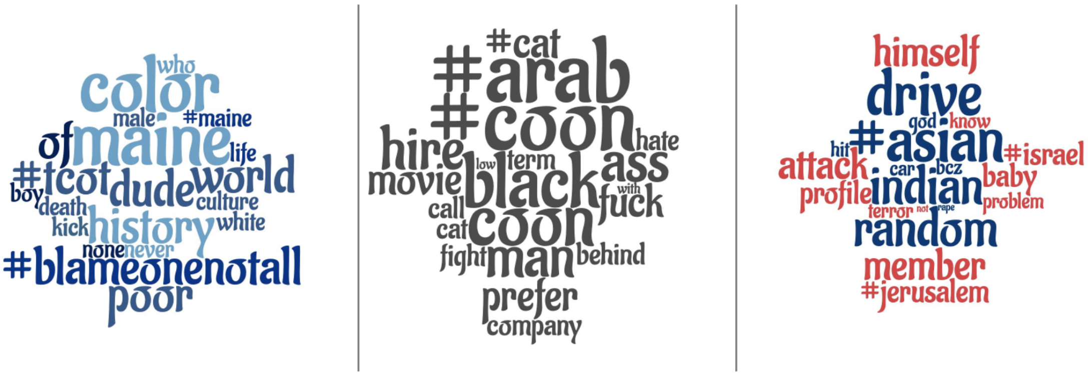


Figure 6.5: Important words in the Racism dataset. Cluster A words (left), shared words (middle) and cluster B words (right).

6.6.4.1 Identified Perspectives

We found one defined perspective (cluster B). This group appears especially sensitive towards degrading comments against asians as well as jews.
Other examples from Racism corpus are given below.

http://t.co/ZxbZV39jru: Rio, Maine ***Coon****, #takes a bath http://t.co/eo02meFyRy* ***#coon #maine***

RT @[redacted]: Rosie dismissed Lauren's claims, defended exploitative Eve Ensler, called Lauren "bully" after valid points, defended ***racist*** *. . .*

@[redacted]_ @[redacted] @[redacted] man ***fuck*** *these feminazi s Kube. We don't need a* ***black*** *history month.* ***Blacks*** *don't care why solo them out ...*

reformed ***white*** *male kotaku commenter bionicle building vaping virgin atheist antisjw&feminazi libertarian naruto watching gamergater here*

@[redacted] when you find him let me know so I can hunt him ***#coon***

WHO SAID GOON'S DON'T EXIST? TELL THAT TO THE ***BLACK*** *MEN WHO LOSE THEIR LIVES DAILY TO OTHER* ***BLACK*** *MEN.* ***#COON****-LUMINATI http://t.co/6Ai8lZXUBA*

6.6.5 Perspectives in the Homophobia Dataset

We observe in the Homophobia dataset the agreements α, β as depicted in Figure 6.6. The label agreement induces the clustering $A = \{2, 3\}$ and $B = \{0, 1, 4\}$. The feature agreement, however, gives us the different result $A = \{2, 3, 4\}$ and $B = \{0, 1\}$, i.e. rater 4 switches cluster affiliation. Since rater 4 is closer to rater 0 in label agreement than to rater 3 ($\alpha(4, 0) =$

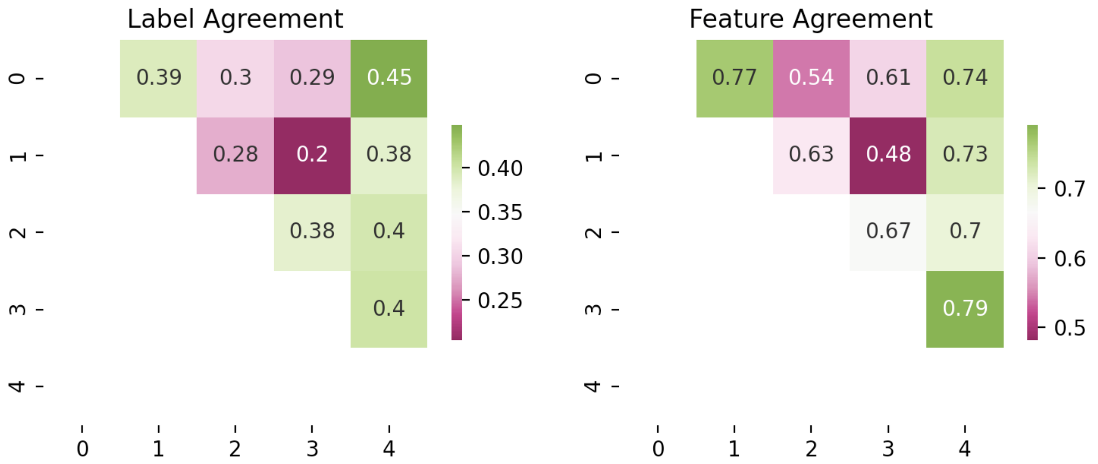


Figure 6.6: Label agreement and feature agreement in the homophobia dataset.

$0.45 > 0.4 = \alpha(4,3)$), but the opposite is the case for the feature agreement ($\beta(4,0) = 0.74 < 0.79 = \beta(4,3)$). As rater 4 cannot be definitely attributed to one cluster, we ignore rater 4 in following the cluster analysis, as described in Section 6.5.3. According to label feature overlap (see Section 6.5.3) we obtain the clusters $A = \{2,3\}$ and $B = \{0,1\}$.

Quantitatively, we found that the positive label rate is much higher in B (22.1%) as compared to cluster A (9.3%), indicating the annotators in B are much more sensitive in this task. Furthermore, we found the feature agreement elevated in cluster B ($\beta_B = 0.77$), compared to $\beta_A = 0.67$, indicating that the important words in cluster B is more clearly defined than that in cluster A.

Qualitatively, we found the terms “teoria gender” (gender theory) and “propaganda” are important in the shared vocabulary, see these examples:

> *Le lezioni del pupazzo transgender: l’ultima follia nella* ***propaganda*** *lgbt URL*
> *(English: The lessons of the transgender puppet: the latest madness in lgbt propaganda URL)*

> *#18settembre I governi passati “illuminati” dalle* ***teorie gender*** *hanno spinto per matrimoni gay, uteri in affitto, d. . . URL*
> *(English: #18september Past governments "enlightened" by gender theories have pushed for gay marriages, rented wombs, d. . . URL)*

The following examples contain important words of cluster A:

> *Il Primato Nazionale: Le lezioni del* ***pupazzo transgender****: l’ultima follia nella propaganda lgbt. https://t.co/ii7VMSVwp6*

For cluster B, we find examples with important words, such as **chiesa** meaning the church and **adulterio** which means adultery. The theme appears to be insults against the church:

*@[redacted] @[redacted] @[redacted] @[redacted] La **Chiesa** è un enorme gay pride.*
*(English: @[redacted] @[redacted] @[redacted] @[redacted] The **Church** is a huge gay pride.)*

*L'AGENDA PER LA **CHIESA**: **NORMALITÀ DI ADULTERIO**, OMOSESSUALITÀ, SESSO LIBERO E NOZZE GAY. URL di @[redacted]*
*(English: THE AGENDA FOR THE **CHURCH**: **NORMALITY OF ADULTERY**, HOMOSEXUALITY, FREE SEX AND GAY WEDDING. by @[redacted])*

Other examples from Homophobia corpus are given below.

@[redacted] Anche se fosse temo sarebbe comunque lesbica
(English: @[redacted] Even if it were I'm afraid it would still be a lesbian)

@[redacted] ma pensa che questa tua ostentazione della omosessualità sia di orgoglio ai gay . Quelli seri ? Ma... https://t.co/VYIYXQVSpL
(English: @[redacted] but think that your ostentation of homosexuality is a pride for gays. The serious ones? But... https://t.co/VYIYXQVSpL)

@[redacted] E voglio Vilde lesbica io, magari teniamocela femmina solo per quello....magari se la fa con femmina PChris...
(English: @[redacted] And I want Vilde lesbian, maybe let's keep the female just for that .... maybe she does it with female PChris ...)

Grande Silvana De Mari che cita letteralmente #MarioMieli con frasi del libro "Elementi di Critica Omosessuale" , g... https://t.co/Zt0bcd3Epq
(English: Grande Silvana De Mari who literally quotes #MarioMieli with phrases from the book "Elements of Homosexual Criticism", g... https://t.co/Zt0bcd3Epq)

Putroppo le nuove parole del papa non smettono di condannare l'omofobia, il sesso fuori dal matrimonio e il semplic... https://t.co/bJtp0hVP0A
(English: Unfortunately, the pope's new words do not stop condemning homophobia, sex out of wedlock and simplicity...) https://t.co/bJtp0hVP0A

@[redacted] Quando una notizia sui commenti razzisti e omofobi dei fascio-leghisti? Non è difficile trovarli
(English: @[redacted] When news about the racist and homophobic comments of the fascists? It is not difficult to find them)

Godimento neolibertino e gender, il nuovo ordine erotico turboglobalista https://t.co/jaFExlkECA di @[redacted] Que... https://t.co/rBUj2fyKCf
(English: Neolibertine and gender enjoyment, the new turboglobalist erotic order https://t.co/jaFExlkECA by @[redacted] Que... https://t.co/rBUj2fyKCf)

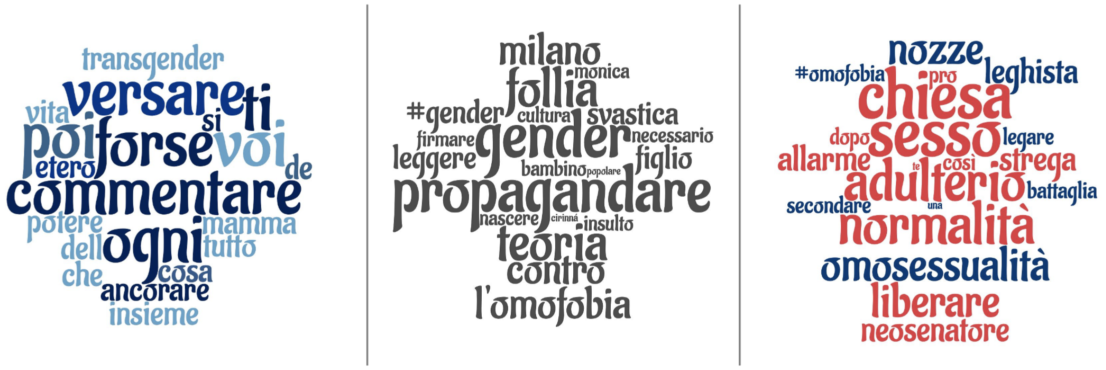


Figure 6.7: Important words in the Homophobia dataset. Cluster A words (left), shared words (middle), and cluster B words (right).

6.6.5.1 Identified Perspectives

In this dataset, we found a perspective in cluster B. These annotators showed an especially high sensitivity with respect to anti-church tweets.

6.6.6 Perspectives in the HS Italian Dataset

In this dataset, we found large differences between the number of items annotated by the different annotators. To avoid biasing our model to this factor, we only analyze annotators with a high rating count[8]. When clustering using both α and β agreement, we obtained the same clustering into two clusters A, B of each 7 annotators.

Quantitatively, we found an anomaly here, as the label agreement in cluster A is almost zero ($\alpha_A = 0.03$), whilst in cluster B it is rather high ($\alpha_B = 0.42$). This already indicates that A is a *cluster of outliers*. Furthermore, the feature agreement is higher in cluster B ($\beta_B = 0.48$) as compared to A ($\beta_A = 0.34$). The latter appears to be due to noise only.

Qualitatively, we found that degrading talk about immigrants get positive labels from both clusters. For example:

> *Tu **immigrato** spacciatore e che porti degrado vuoi venire in Italia? Vieni pure, ti aspetto qui sulla costa. #nuota **#dallavostraparte***
> *(English: You immigrant drug dealer and who brings degradation do you want to come to Italy? Come on, I'll wait for you here on the coast. #wheel #from your side)*

> *Anche io voglio essere un profugo! e cosi' chiedo asilo politico!hahahahaha **#dallavostraparte***
> *(English: I also want to be a refugee! and so I ask for political asylum! hahahahaha #dallavostraparte)*

[8] this means for this dataset at least 800 ratings per annotator

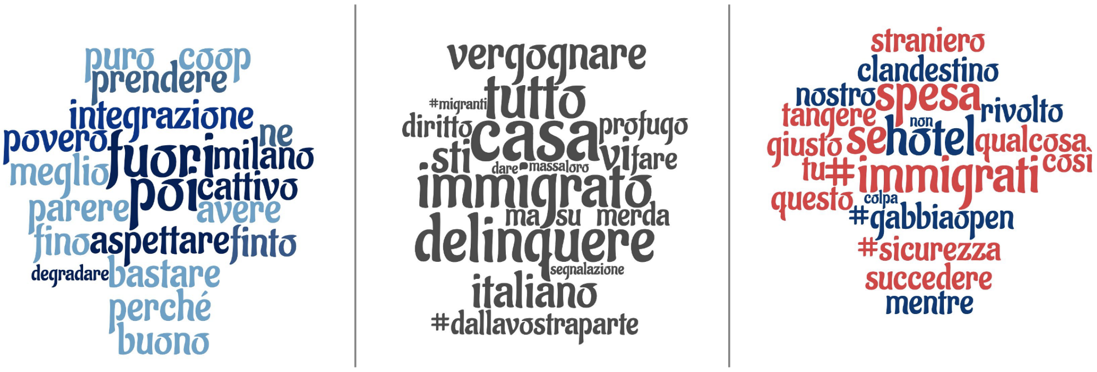


Figure 6.8: Important words in the HS Italian dataset. Cluster A words (left), shared words (middle) and cluster B words (right).

For cluster B, we found examples with complains about immigrants driving up public costs by living in "hotels" as well as concerns about "sicurezza"(security) being diminished in the country after immigration:

> *E gli immigrati a 35 al giorno negli **hotel**. È una vergogna!!! 1!uno!*
> *(English: And 35 immigrants a day in hotels. It's a shame !!! 1! One!)*

> *Altro che #alberghi x **#immigrati** e **spese** folli x loro #mantenimento! intanto Italia #collassa! #spesapubblica **#sicurezza** @*[redacted]
> *(English: Other than #hotels x #immigrants and crazy expenses x their #maintaining! meanwhile Italy #collassa! #spesapubblica #safety @*[redacted]*)*

6.6.6.1 Identified Perspectives

In this dataset, we found a defined perspective in cluster B. The annotators tend to label a large spectrum of content - from critical, conservative, nationalistic, to openly hateful tweets, all as hate. We conclude, annotators in cluster B are very sensitive towards xenophobia.

Some more examples from Italian HS corpus are given below.

> *@*[redacted] *povera disgraziata, pur d far voce, specula sulla massa d migranti, che invadono l'Italia, l guerra senza armi, con vittime*
> *(English: @*[redacted] *poor unfortunate, while making a voice, speculates on the mass of migrants, who invade Italy, the war without weapons, with victims)*

> *In un paese dove'immigrato clandestino ha piÃ¹ diritti di un cittadino ti chiedi se non valga la pena fare l'immigrato #siono #iovotono https://t.co/n8vZbopoOk*
> *(English: In a country where an illegal immigrant has more rights than a citizen, you wonder if it's not worth being an immigrant #siono #iovotono https://t.co/n8vZbopoOk)*

Quindi lo lasciamo fare a immigrati senza documenti ..? (Pare che siano molti Myrta...ma forse la tua stampa non lo dice) Žfrase ambigua https://t.co/PTeMaVexfQ
(English: So we leave it to undocumented immigrants ..? (Apparently there are many Myrta ... but maybe your press doesn't say it)Ž ambiguous phrase Ž https://t.co/PTeMaVexfQ)

Un quartiere tutto di stranieri. Ma vi pare normale? Intere zone colonizzate
(English: A neighborhood full of foreigners. But does it seem normal to you? Entire colonized areas)

#profughiuncazzo #stopinvasione https://t.co/SLrI7xvwg8

Con l'arrivo poi di centinaia di migliaia di presunti profughi i salari saranno ancora piu' bassi...se tutto va bene siamo rovinati.. https://t.co/lvqXOoyIB0
(English: With the arrival of hundreds of thousands of alleged refugees, wages will be even lower ... hopefully we are ruined .. https://t.co/lvqXOoyIB0)

Se volete accogliere fatelo gratis.Bloccare subito gli sbarchi e mandare via i presunti profughi!!! Ridiamo l'italia agli italiani. https://t.co/kONAMcrnmP
(English: If you want to welcome, do it for free. Immediately block the landings and send the presumed refugees away !!! Let's give Italy back to the Italians. https://t.co/kONAMcrnmP)

Noi italiani schedati e controllati in tutto degli immigrati non si sa nulla nome origini malattie vi sembra normale?
(English: We Italians registered and checked in all immigrants do not know anything name origins diseases seems normal to you?)

Ma sti migranti non si potrebbero mettere tutti su un'isola deserta? #GABBIAOPEN
(English: But couldn't these migrants all be put on a desert island? #CAGEOPEN)

a_meluzzi @[redacted] *assolutamente urgente fermare l'immigrazione selvaggia da africa mussulmana:non si entra di forza in casa altri*
(English: a_meluzzi @[redacted] *It is absolutely urgent to stop the wild immigration from Muslim Africa: you do not enter the house by force others)*

#GABBIAOPEN gli immigrati olgono il lavoro , questo Ã¨ assolutamente vero. Pensateci bene , pensateci
(English: #GABBIAOPEN immigrants take away work, this is absolutely true. Think about it, think about it)

Ma questi che vogliono tutti sti migranti c'hanno la merda in testa? #omnibusla7
(English: But do these migrants who want all these migrants have shit in their heads? #omnibusla7)

> *Coppia #gay accoglie #immigrato @matteosalvinimi quanto stai crepando in questo momento? #integrazione #SinistraItaliana #norazzismo*
> *(English: #Gay couple welcomes #immigrant @matteosalvinimi how much are you dying right now? #integration #SinistraItaliana #norazzism)*

> *#migranti #terremoto gli italiani al freddo e i clandestini accolti con tutti gli onori. governo di merda*
> *(English: #migrants #terquake Italians in the cold and illegal immigrants welcomed with all honors. government shit)*

6.6.7 Perspectives in the HS English dataset (Davidson)

In this dataset, we found large differences between the number of items annotated by the different annotators. To avoid biasing our model to this factor, we only analyze annotators with a high rating count[9]. We obtained different clusters according to α and β agreement. After computing the label feature overlap, we obtained cluster A of size 45 and cluster B of size 41. As seen in Figure 6.1, after the label feature overlap step, only the annotators with the same cluster affiliation under β agreement (solid color) remain, while the ones that switch clusters (transparent color) get removed. Note that, in contrast with all previous datasets, we have two kinds of positive labels in this dataset, one for "offensive language" content and one for "hate speech" (stronger label).

Quantitatively, we found cluster B to have a much higher hate speech label rate (15.4%) over cluster A (4.1%). The base rate is 10%. While both cluster have comparable positive label rates, this indicates that cluster B has a tendency to give the hate speech label when the offensive label would have been an option. as compared to cluster A. Further, feature agreement is much lower in cluster A ($\beta_A = 0.22$) as opposed to cluster B ($\beta_B = 0.46$), indicating the annotators in cluster B agree much more on their important words.

Qualitatively, we found some words are understood by both clusters as hateful:

> *@[redacted] shut up **nigger** whore! Hope u get raped by one of those **animals**. Might change your tune.*

This more moderate example is seen as offensive by both:

> *RT @[redacted]: You females overlook us geeks. When I take these glasses off, I'm no longer Clark Kent. I go superman in that **pussy**. - FA...*

As cluster A has a much lower positive label rate, it is hard to find examples where A was more critical than B. For cluster B we find several examples with the same keywords, centered around homophobic slurs such as "faggot" and variations. The following examples are each rated almost unanimously as "offensive" in cluster A, as opposed to cluster B, where they are rated as "hateful" (stronger label).

> *@[redacted] word is you use roids, stupid **hypocrite** lying **faggot**.*

[9] for this dataset, at least 500 annotations per annotator

*"@[redacted]: "@[redacted]: @[redacted] Is a **fag**" jackie jealous" Neeeee*

*"Warning: penny boards will make you a **faggot***

We find the following example with the keyword “terrorist” (an important word from A according to our method), but the labels are very similar in both clusters:

*@[redacted] your a dirty **terrorist** and your religion is a fucking joke, you go around screaming Allah akbar doing **terrorist** shit. Dirty faggot.*

Other example tweets for the Davidson dataset are shown below.

@[redacted] You heard me bitch but any way I'm back th texas so wtf u talking about bitch ass nigga

I'm tired of people saying I look like my brother &amp; calling me Deondre' like serious Succ My Ass fag asses

#VoteBlue2014 Yeah. CUZ 8 million people in faggot ass #newyork are #chickenshit JEWS&gt; FUCK THEM right? Fuck Bibi Netanyahu. RIGHT?

@[redacted] I love bad bitches that my fucking problem....

@[redacted] @[redacted] @[redacted] what's your problem? Don't u know Jews control niggers? Read a book or two.

RT @[redacted]: Niggas with face tats are the same ones that stole your animal crackers in elementary smh they just dont give a fuck

6.6.7.1 Identified Perspectives

For cluster A we mostly find that this group is less sensitive for the kinds of hateful language used in this dataset as they give much less positive labels. We did not find any characteristic vocabulary in the polarized examples.

In this dataset, we find a defined perspective for cluster B. Annotators in this group give harsher labels when homophobic slurs are present in a tweet, as compared to annotators in A. We conclude that the annotators in B are highly sensitive towards homophobic slurs.

6.7 Summary

In this chapter, we developed a novel approach to employ quantitative identification of human perspectives expressed in data annotation by the annotators with a common background. As a novelty, we measured feature agreements amongst the annotators along with the traditional measurement of label agreements. This helped us to identify the features (unigrams) on which the annotators agree within a given group of annotators with common background. The method not

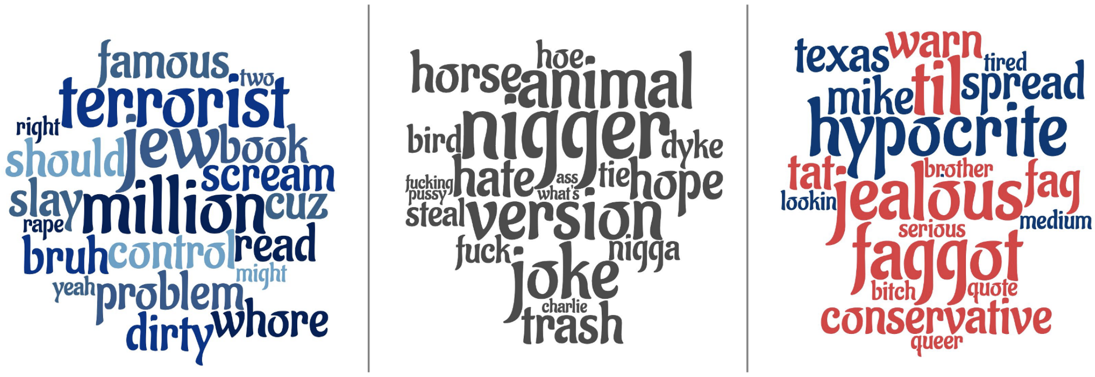


Figure 6.9: Important words in the Davidson dataset. Cluster A words (left), shared words (middle) and cluster B words (right).

only identified single perspectives specific to an annotator but also identified the shared perspectives stemming within group members due to a common cultural or demographic background. We employed clustering to find clusters of annotators that share perspectives. The clustering of annotators also helped us to remove the least significant annotators in a dataset. In order to test the method, we analyzed a number of annotated hate speech corpora in English and Italian languages, showing how the opinions of the annotators, reflected in their annotation, are far from uniformly distributed.

The results suggested that annotators do shape their opinions based on the held perspectives on a topic and these perspectives are often shared amongst the annotators who form a homogeneous group. In fact, this is more visible in the annotation of hate speech corpora because they tends to be polarized, and our methodology is able to highlight the groups of annotators sharing similar opinions which are also visible in the example tweets taken from the HS corpora.

For BREXIT corpus, where we have natural selection of the annotators with known common background, the identified perspectives are based on Islamophobic and xenophobic keywords and we can deduce that because the annotators were Muslims and migrants so they showed higher level of sensitivity for these topics. Similarly, for Sexism, the perspectives were identified in a broader context. For Racism, the identified perspectives were based on anti-Asian and antisemitism. For Homophobia, there was anti-christian sentiment. Annotators for HS dataset in Italian language were more sensitive towards xenophobic content. Finally, the HS dataset in English was more homophobic based and the groups showed divergence in individual and shared perspectives.

Furthermore, we the automated method, that we developed, supported the manual exploration of the perspectives emerging from a polarized annotation of hate speech, resulting in consistent patterns describing why certain groups of people are more or less keen on judging a message as hateful.

As future work, we plan to expand our analysis and investigate feature based agreement methodology with more refined linguistic features, to abstract away from individual words and therefore provide a more robust approach. Here we implemented K-means clustering algorithm and we plan to perform clustering with other state-of-the-art clustering techniques. The methodology developed in this chapter may also be employed to other NLP tasks traditionally considered less subjective, but recently found to contain informative disagreement [Uma et al., 2020], as well as non-linguistic tasks such as image labeling.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis we addressed several challenges on abusive language and HS detection in social media datasets. We focused on the opinion polarization and annotator bias in abusive language and HS corpora and experimented with state-of-the-art methods to improve the performance of downstream abusive language and HS detection tasks. We investigated how different opinions emerging from various communities on the topics of mutual and social interest cause polarization among the annotators, when they are asked to annotate abusive language and HS in a dataset. In this research work, we also investigated whether the machine learning methodologies can effectively leverage different opinions emerging from different groups of annotators to improve automatic classification of highly subjective phenomena such as abusive language and hate speech. In an attempt to take the advantage of polarized opinions in data annotations, we presented a novel method, a quantitative index, to measure the level at which the individual instances of a dataset are polarized. For characterizing a polarized utterance, majority of the detection tasks in this research work are addressed as binary classification tasks which mainly employ classical machine learning algorithms such as support vector machines and transformers based pre-learning methods such as BERT.

To evaluate the newly developed model, We tested our approach with several cross-validation experiments by collecting a set of Twitter corpora already employed in the previous research focusing on the abusive contents such sexism, racism and homophobia. The experimental results show a consistent improvement in the prediction performance due to the pre-processing induced by our method, even using simple models and features (bags of words). We also proposed to leverage the annotation given by individual annotators to compute their reliability, and remove the least reliable annotators to improve the quality of gold standard data and subsequently, the performance of machine learning models. We draw two observations from the general picture emerged from the experiments evaluated in Section 3.4. First, the results show that the detection of sexist behavior is an easier task than the detection of racism on our data, and that the detection of homophobic content is the harder when compared to the other tasks. This is in line with the results of recent evaluation campaigns on misogyny detection [Anzovino et al., 2018] and hate speech detection on online media against immigrants [Basile et al., 2019], and likely due to the vocabulary of misogynistic hate being somewhat more restricted. Our results suggest that the consensus-based methods to create gold standard data are not necessarily the best choice when dealing with highly subjective phenomena, and the knowledge coming from the disagreement and polarization of opinions is indeed highly informative. Finally, we show how the P-index can effectively be employed as a tool to manually explore the data, ranking the instances to identify messages that are more likely to generate confusion and polarization among the annotators. Our experimental setting demonstrates and confirms that our method is able to recognize the presence of polarization in the datasets annotated by annotators which may have different social or cultural background (Chapter 3).

Furthermore, we presented a method to divide the annotators of a dataset into groups based on their annotation, under the hypothesis that such partitions reflect characteristics such as cultural background, common social behaviour and similar other factors. The P-index appeared to be a useful tool in dividing the annotators into homogeneous groups based on the similar preferences in data annotations within the context of crowdsourced datasets where we do not have any background information on the annotators. We also validated the division of annotators into groups by designing a Pilot Study 3.2.2 in which a small set of English tweets on abusive language are manually annotated by the annotators belonging to two different groups with known annotator background. The validation experiments suggested that the P-index can successfully picks up the divergence of opinions in the data annotations and hence aid to divide the annotators into groups at a macro-level.

To extend our research further, we created separate gold standards for each group of the annotators in Section 4.2.2. The idea is to investigate how these characteristics can influence the opinions of annotators expressed while annotating abusive language data. The method aimed at modeling the different perspectives of the annotators toward complex, subjective phenomenon. In turn, we proved that this methodology is able to improve the classification performance on several benchmarks of online hate speech, by training multiple, group-based classifiers instead on a single, all-comprehensive one. To test the method, we experimented with three different social media datasets in English and Italian. The results show us an improvement over the baseline across all the datasets. Moreover, the implementation of an “inclusive” ensemble classifier further boost the classification performance, in particular by strongly increasing the recall on hateful messages (Chapter 4).

By expanding our research in the area of annotator polarization, as a novel case study, we developed a multi-perspective abusive language dataset on the Brexit debate and involved migrants as the victims of abusive language to annotate the datasets under the hypothesis that some common characteristics (cultural, demographic, ethnic, etc.) can influence the annotators’ perception on certain phenomena and shape their opinions on social media posts (Chapter 5). Our polarization-based methodology groups the annotators based on their opinions and stance toward a given phenomenon. In our previous work, we developed a novel method to divide the annotators into groups based on the polarization of their judgments also effectively acting as an empirically-driven substitute to the unavailability of information on the background of the annotators, e.g., in a crowd-sourcing scenario. Here, we have a natural selection of annotator groups and do not need to methodologically divide the annotators. The results on BREXIT dataset in Section 5.4 by employing group-based classification methodology for perspective-aware system modeling explained in Section 5.3.3 show us an improvement over the baseline across all the categories. Moreover, the implementation of an "inclusive" ensemble classifier further succeeded to boost the classification performance by outperforming the baseline models, in particular by strongly increasing the recall on abusive messages. We preformed deep manual exploration of BREXIT dataset with all underlying categories. By ranking the instance with P-index values, the most polarizing tweets can be filtered for further qualitative analysis, to better understand the controversial topics and issues, and create more compact and better guidelines to improve the annotation quality and hence solve the inconsistencies in the gold standard data. The qualitative analysis of the BREXIT dataset in Section 5.5 provided us deeper insights into the nature of data and how annotators belonging to different cultures and communities perceive subjective phenomena and how these perspectives can influence their opinions and the results suggest that these opinions are strongly linked with the background of annotators.

Apart from the raw performance metrics, one may wonder which classifier should be selected when multiple group-based models are available trained on the same data. One possibility is to give preference to the classifier trained on data annotated by a group involving the victims of hate speech, in order to "give voice" to the targeted group through the computational model. Another possibility is to implement an inclusive classification framework, such as the ensemble classifier proposed in this work. Such methods aim to "give voice" to all the existing perspectives on a certain phenomenon equally. Furthermore, given its transparency, the latter classifier shows potential for providing an explicit explanation of its decisions, being able to track them back to the specific (highly cohesive) groups of people who annotated the training data.

In order to identify perspectives in hate speech corpora in a quantitative way, we analyzed a number of annotated hate speech corpora, showing how the opinions of the annotators, reflected in their annotation, are far from uniformly distributed. In fact, the annotation of hate speech tends to be polarized, and our methodology is able to highlight the groups of annotators sharing similar opinions. Table 7.1 summarizes all identified perspectives, defined as increased sensitivity towards certain types of textual content by specific groups of people (Chapter 6). Furthermore, we introduced an automated method to support the manual exploration of the perspectives emerging from a polarized annotation of hate speech, resulting in consistent patterns describing why certain groups of people are more or less keen on judging a message as hateful. We also provided a visual exploration of the identified perspectives in annotator clusters with the help of word clouds.

Dataset	*Sensitivity A*	*Sensitivity B*
Brexit	-	islamophobia, xenophobia
Sexism	-	broader context
Racism	-	anti-asian, antisemitism
Homophobia	-	anti-christian
HS Italian	-	xenophobia
Davidson	-	homophobia

Table 7.1: The perspectives identified in the HS corpora defined by groups of people that show increased sensitivity towards certain phenomena.

The results described in this thesis have attempted to answer the research questions and related sub-questions introduced in Chapter 1:

1: ***How to measure the opinion polarization in data annotation of abusive language corpora?***

We developed a novel approach in Section 3.2.1, a quantitative index to measure the level of polarization in opinions expressed by annotators while annotating a dataset. The index, called polarization index (P-index) can aid in the division the annotators into groups. With the help of index, we can measure different opinions expressed by the annotators in the individual instances of the abusive phenomena when the annotated data is crowdsourced and in general, we do not have any information about the background of annotators. To the best of our knowledge, the P-index is the first quantitative method to broadly exploitthe polarized opinions and utilize the information coming from the polarized utterances for characterizing hateful contents. We evaluated the robustness of our model over Twitter corpora. The results show that the models based on training data exploited with the help of P-index performs quite well over the baseline across the experiments carried out.

2: ***How high level of polarization in data annotations can influence the training datasets?***

The novel resource developed and explained in Section 3.2.1 is a useful resource to manually explore abusive language corpora which are annotated by two different groups of annotators. It is important to note that we might have the information on the background annotators such as in BREXIT dataset explained in Section 5.2 and we may not have any background information such as for the datasets annotated by crowdsourcing workers explained in Section 3.3. In both cases, the P-index is a useful tool to explore these datasets. By ranking the instances of abusive dataset with the P-index, the most polarized instances naturally emerge at the top of the list and we can identify those messages that are more likely to generate confusion and polarization among the annotators.

3: ***How to build a robust model which facilitates the modeling of polarized opinions for detecting abusive language across different topical focuses and targets?***

To answer this question, we collected several publicly available social media corpora of hateful language, covering phenomena such as hate speech, sexism, racism, homophobia, and offensive language. We then created separate gold standards for each of the groups and perform classification tasks to measure the performance of perspective-aware supervised models in a multilingual setting explained in section 4.2.2. We also proposed an ensemble classifier that considers all the learned perspectives in an inclusive fashion.

4: ***How to mine annotator's perspectives in abusive language corpora?***

To answer this question, we introduced an automated method to support the manual exploration of the perspectives emerging from a polarized annotations on several datasets of HS in multiple languages. We developed feature-based (ngrams) methods to measure the pairwise agreements and perform clustering with state-of-the-art K-means algorithm to cluster the annotators into groups based on similarity in features rather than the labels. This helped us to understand how sensitive the annotator groups are to particular hateful messages that might be linked to their ethnicity, religion, culture or demographic background. Furthermore, hate speech, resulting in consistent patterns describing why certain groups of people are more or less keen on judging a message as hateful (see Chapter 6).

To summarize, we demonstrated the usefulness of perspective-aware models and the information extracted from the opinions coming from people with different backgrounds and from different communities for dealing with the presence of abusive content in social media. The findings of this research were presented in different publications that are introduced in the following section.

7.2 Research Contributions

Below, we present the publications derived from this research by grouping them according to the theme of each publication.

1. Abusive language detection with a focus on opinion polarization and annotator bias:
We describe our approach for addressing the abusive language detection by developing a novel index to utilize the important information expressed by annotators in the form of single annotations which are often considered as noise in the training datasets. The findings were published in a conference paper:

- **Akhtar, S.**; Basile, V.; and Patti, V. 2019. A new measure of polarization in the annotation of hate speech. In Alviano, M.; Greco, G.; and Scarcello, F., eds., AI*IA 2019 – Advances in Artificial Intelligence, 588–603. Cham: Springer International Publishing [Akhtar et al., 2019].

2. Modeling polarized opinions to improve abusive language and HS detection with perspective-aware modeling based on the group-wise classifiers:
We developed a novel method to model polarized opinions expressed by annotators in the annotation task and experimented with datasets in both English and Italian languages. The model is based on the idea of creating separate gold standards for each training set based on the annotations from individual groups of annotators. The work resulted in a conference publication:

- **Akhtar, S.**, Basile, V., & Patti, V. (2020). Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8(1), 151-154. [Akhtar et al., 2020]

3. The development of a multi-perspective abusive language dataset with multiple subcategories on the Brexit phenomenon in English langauge:
We developed a novel resource, a multi-perspective dataset on the Brexit phenomenon and called it BREXIT dataset to investigate how polarized opinions can influence the data annotation process and in turn the training models for abusive language detection. We involved migrants as the victims of hate speech to annotate the dataset. We subdivided the dataset into four abusive language categories to model polarized opinions expressed by annotators in the annotation task and also performed classification experiments with an extensive qualitative analysis of the BREXIT dataset. The research resulted in a journal paper acceptance with revisions:

- **Akhtar, S.**, Basile, V., & Patti, V. (2021). Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. Accepted with revisions in Elsevier Information Processing and Management Journal on 23-12-2021. [Akhtar et al., 2021]

4. Mining annotator's perspectives in hate speech corpora:
We developed a novel quantitative method to automatically identify individual and shared perspectives stemming from a common background. We tested our methodology on several corpora in English and Italian, manually annotated according to their hate speech content, validating prior knowledge about the groups of annotators, when available, and discovering characteristic and traits among the annotators with unknown background. The work resulted in a conference paper accepted in NL4AI (Natural Language for Artificial Intelligence) workshop at AIxIA conference (The 20th International Conference of the Italian Association for Artificial Intelligence) 2021. The first author of the accepted paper is responsible for devising and designing the methodology presented in this paper.

- M. Fell, **S. Akhtar**, and V. Basile. Mining annotator perspectives from hate speech corpora.In E. Cabrio, D. Croce, L. C. Passaro, and R. Sprugnoli, editors, Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021, volume 3015 of CEUR Workshop Proceedings. CEURWS. org, 2021. URL http://ceur-ws.org/Vol-3015/paper136.pdf. [Fell et al., 2021]

7.3 Future Work

Abusive language and HS detection is an interesting computational linguistic task within NLP context that is rapidly gaining exposure within the research community. There are state-of-the-art approaches that achieve best performances but hardly any of such approaches tackle the problem of polarization in opinions expressed by annotators which can also be termed as the bias in the datasets. Most of the available resources have crowd-based annotations where we do not any any information on the background of annotators. That means that there is still plenty to do for improving abusive language detection performances with respect to modeling perspectives emerging from a group of annotators belonging to different backgrounds. Following, we mention some areas that could be investigated for addressing this issue in further studies:

- Although the methods we introduced in our research for abusive language and HS detection boosts the hate speech classification performance, there are limitations which are important to consider. For the methodology to work, we need pre-aggregated data, e.g., full reports from the crowdsourcing platforms, which are often not available. Another issue is epistemological: our methodology and the subsequent empirical evaluation showed that there is a great deal of information that is effectively wiped out by the aggregation step employed in the standard procedure to create benchmark datasets. Therefore, evaluating perspective-aware machine learning models on traditionally aggregated datasets is unfair.

This is in line with recent work by Basile [2020] and the *Perspective Data Manifesto*[1], an initiative that promotes the publication of datasets in pre-aggregated form and to develop new paradigms of evaluation that take all the perspectives linked to different backgrounds into account.

- The future work also aims at exploring more dimensions of the background of annotators, including native language, demographic factors, and how they interplay with the measured polarization of their annotations in a group. Moreover, we believe that involving the victims of hate speech in the process of annotating the data in hate related detection tasks can provide new insights and improve the quality of the data. Unfortunately, such datasets or the information on the background of annotators in not available. We need to develop such datasets and involve the victims in the process to understand their feelings and emotions as they are then expressed in the annotations.

- Another issue often encountered when dealing with highly subjective datasets is that, with no background information of annotators in publicly distributed datasets, the set of annotations could be sparse, e.g., in a crowdsourcing context. Also,we limited the number of annotator groups to two. However, this is more a practical limit than a theoretical one, therefore we plan to investigate the effect of dividing the annotators into more than two groups, and how to find an optimal number of partitions. Therefore, we need methods to effectively cluster larger annotator sets based on their annotation. In this direction, unsupervised clustering of the annotators based on their annotations with standard methods may be a solution, also to the issue of the unavailability of background information on the annotators in general. we clustered large number of annotators with K-means, we need to test and compare the results with other available techniques (e.g., with agglomerative and spectral clustering) and more than two groups in order to make diverging opinions emerge on polarizing topics.

- Our initial work on the polarization of annotators' opinions is rooted in somewhat strong assumption that there exists a latent background divide in the annotator population. Even stronger is the assumption that the number of groups is fixed. Although the experimental results confirm the existence of the polarization phenomenon, it will be interesting to investigate how the clustering methods can be refined by relaxing the division constraints and aiming for a more flexible procedure.

- We also plan to apply the methodology presented in this thesis to other abusive language phenomena such as cyberbullying, radicalization, and extremism. We are also interested to test the method on sentiment analysis tasks applied to specific domains such as political debates.

- As future work, we plan to test our methods with deeper and more refined linguistic features, to abstract away from individual words and therefore provide a more robust analysis. We also plan on investigating other NLP tasks traditionally considered less subjective, but

[1] https://pdai.info/

recently found to contain informative disagreement [Uma et al., 2020], as well as the non-linguistic tasks such as image labeling.

Bibliography

F. Abbondante. Il ruolo dei social network nella lotta all'hate speech: un'analisi comparata fra l'esperienza statunitense e quella europea. *Informatica e diritto*, 26(1-2):41–68, 2017.

S. Akhtar, V. Basile, and V. Patti. A new measure of polarization in the annotation of hate speech. In M. Alviano, G. Greco, and F. Scarcello, editors, *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603, Cham, 2019. Springer International Publishing. ISBN 978-3-030-35166-3.

S. Akhtar, V. Basile, and V. Patti. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154, 10 2020. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7473.

S. Akhtar, V. Basile, and V. Patti. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021. URL https://arxiv.org/abs/2106.15896.

S. Al-Anazi, H. AlMahmoud, and I. Al-Turaiki. Finding similar documents using different clustering techniques. *Procedia Computer Science*, 82:28–34, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.04.005. URL https://www.sciencedirect.com/science/article/pii/S1877050916300199. 4th Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia.

A. AlDayel and W. Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102597. URL https://www.sciencedirect.com/science/article/pii/S0306457321000960.

S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. *A Deep Dive into Multilingual Hate Speech Classification*, pages 423–439. ECML/PKDD, 02 2021. ISBN 978-3-030-67669-8. doi: 10.1007/978-3-030-67670-4_26.

M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91947-8.

R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://doi.org/10.1162/coli.07-034-R2.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188, 2017. URL http://arxiv.org/abs/1706.00188.

A. Basile, T. Caselli, and M. Nissim. Predicting controversial news using facebook reactions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.*, 2017.

V. Basile. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In G. Vizzari, M. Palmonari, and A. Orlandini, editors, *Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020), Anywhere, November 27th, 2020*, volume 2776 of *CEUR Workshop Proceedings*, pages 31–40. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2776/paper-4.pdf.

V. Basile, M. Lai, and M. Sanguinetti. Long-term social media data collection at the university of turin. In *CLiC-it*, 2018a.

V. Basile, N. Novielli, D. Croce, F. Barbieri, M. Nissim, and V. Patti. Sentiment polarity classification at evalita: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*, pages 1–1, 11 2018b. doi: 10.1109/TAFFC.2018.2884015.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://www.aclweb.org/anthology/S19-2007.

K. Beelen, E. Kanoulas, and B. van de Velde. Detecting controversies in online news media. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1069–1072, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5022-8. doi: 10.1145/3077136.3080723. URL http://doi.acm.org/10.1145/3077136.3080723.

B. Beigman Klebanov, E. Beigman, and D. Diermeier. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 253–257, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-2047.

S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, and L. Wright. Counterspeech on twitter: A field study. In *A report for Public Safety Canada under the Kanishka Project*, 2016.

A. Bessi, M. Coletto, G. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10, 08 2014. doi: 10.1371/journal.pone.0118093.

P. K. Bhowmick, A. Basu, and P. Mitra. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65, Manchester, UK, Aug. 2008a. Coling 2008 Organizing Committee. URL https://aclanthology.org/W08-1209.

T. Bhowmick, A. Robinson, A. Gruver, A. MacEachren, and E. Lengerich. Distributed usability evaluation of the pennsylvania cancer atlas. *International journal of health geographics*, 7: 36, 08 2008b. doi: 10.1186/1476-072X-7-36.

D. Biber and E. Finegan. Adverbial stance types in english. *Discourse Processes*, 11:1–34, 1988.

A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava. 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, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1105. URL https://aclanthology.org/W18-1105.

C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63, 2013.

C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, and M. Tesconi. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

D. Boyd. *It's Complicated: The Social Lives of Networked Teens*. Yale University Press, USA, 2014. ISBN 0300166311.

A. Brown. What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36, 08 2017. doi: 10.1007/s10982-017-9297-1.

P. Brown, V. Dellapietra, P. Souza, J. Lai, and R. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479, 01 1992.

P. Burnap and M. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5, 12 2016. doi: 10.1140/epjds/s13688-016-0072-6.

P. Burnap and M. L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2): 223–242, 2015. doi: https://doi.org/10.1002/poi3.85. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85.

A. Campagner, D. Ciucci, C.-M. Svensson, M. T. Figge, and F. Cabitza. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790, 2021.

A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, and G. Demartini. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada.*, pages 11–20. AAAI Press, 2017.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 71–80, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-0-7695-4848-7. doi: 10.1109/SocialCom-PASSAT.2012.55. URL http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.55.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 04 1960. doi: 10.1177/001316446002000104.

E. Commission. Code of conduct on countering illegal hate speech online, 2016.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, pages 89–96, 01 2011. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14126/13975.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL https://aclanthology.org/W18-5102.

M. Del Vicario, A. Scala, G. Caldarelli, H. Stanley, and W. Quattrociocchi. Modeling confirmation bias and polarization. *Scientific Reports*, 7, 06 2016. doi: 10.1038/srep40391.

F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017*, pages 86–95, 2017.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 11–17, 2011. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14209.

M. Duggan. Online harassment 2017. Technical report, Pew Research Center., 2017.

C. Elkan. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 147–153. AAAI Press, 2003. ISBN 1577351894.

M. Elmasry, A. Shamy, P. Manning, A. Mills, and P. Auter. Al-jazeera and al-arabiya framing of the israel-palestine conflict during war and calm periods. *International Communication Gazette*, 75(8):750–768, Dec. 2013. ISSN 1748-0485. doi: 10.1177/1748048513482545.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257, 2018. URL http://arxiv.org/abs/1804.04257.

C. Ezenkwu, S. Ozuomba, and C. Kalu. Application of k-means algorithm for efficient customer segmentation: A strategy for targeted customer services. *International Journal of Advanced Research in Artificial Intelligence(IJARAI)*, 4, 10 2015. doi: 10.14569/IJARAI.2015.041007.

M. Fell, S. Akhtar, and V. Basile. Mining annotator perspectives from hate speech corpora. In E. Cabrio, D. Croce, L. C. Passaro, and R. Sprugnoli, editors, *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021*, volume 3015 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-3015/paper136.pdf.

E. Fersini, D. Nozza, and P. Rosso. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018a.

E. Fersini, P. Rosso, and M. Anzovino. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, 2018b.

K. Fitch. Diana boxer, applying sociolinguistics: Domains and face-to-face interaction. *Kristine Fitch Munoz*, 35, 01 2006. doi: 10.1017/S0047404506230057.

J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973. doi: 10.1177/001316447303300309. URL https://doi.org/10.1177/001316447303300309.

K. Florio, V. Basile, M. Lai, and V. Patti. Leveraging hate speech detection to investigate immigration-related phenomena in italy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7. IEEE, 2019.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51:1–30, 07 2018. doi: 10.1145/3232676.

G. Fuchs. Countering hate speech against refugees and migrants: An evaluation of international human rights treaties and soft law instruments. *Relaciones Internacionales*, 92:103, 12 2019.

K. Gelber and L. McNamara. Evidencing the harms of hate speech. *Social Identities*, pages 1–18, 12 2015. doi: 10.1080/13504630.2015.1128810.

N. D. Gitari, Z. Zuping, Z. Zhang, H. Damien, and J. Long. A lexicon-based approach for hate speech detection. In *MUE 2015*, 2015.

F. Gonzalez Rey. Subjectivity in debate: Some reconstructed philosophical premises to advance its discussion in psychology. *Journal for the Theory of Social Behaviour*, 49(2):212–234, 2019. doi: https://doi.org/10.1111/jtsb.12200. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsb.12200.

L. Graves. Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, pages 1–19, 06 2016. doi: 10.1080/1461670X.2016.1196602.

E. Greevy and A. Smeaton. Classifying racist texts using a support vector machine. *Greevy, Edel and Smeaton, Alan F. (2004) Classifying racist texts using a support vector machine. In: SIGIR 2004 - the 27th Annual International ACM SIGIR Conference, 25-29 July 2004, Sheffield, UK.*, 01 2004. doi: 10.1145/1008992.1009074.

N. Hanafiah, A. Kevin, C. Sutanto, Fiona, Y. Arifin, and J. Hartanto. Text normalization algorithm on twitter in complaint category. *Procedia Computer Science*, 116:20–26, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.10.004. URL https://www.sciencedirect.com/science/article/pii/S1877050917320410. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

I. Himelboim, M. Smith, and B. Shneiderman. Tweeting apart: Applying network analysis to detect selective exposure clusters in twitter. *Communication Methods and Measures*, 7:169–197, 07 2013. doi: 10.1080/19312458.2013.813922.

S. Hinduja and W. Ptachin, J. Cyberbullying identification, prevention, and response. cyberbullying research center (cyberbullying.org), 2020. URL https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2018.pdf.

J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349:261 – 266, 2015.

H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909, 2015. URL http://arxiv.org/abs/1503.03909.

D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1132.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://www.aclweb.org/anthology/P18-1031.

R. Irfan, Z. Rehman, A. Abro, C. Chira, and W. Anwar. Ontology learning in text mining for handling big data in healthcare systems. *J. Medical Imaging Health Informatics*, 9(4):649–661, 2019. doi: 10.1166/jmihi.2019.2681. URL https://doi.org/10.1166/jmihi.2019.2681.

M. Ito, H. Horst, M. Bittanti, B. Herr-Stephenson, P. Lange, C. Pascoe, L. Robinson, S. Baumer, R. Cody, D. Mahendran, K. Martínez, and D. Perkel. *Living And Learning With New Media: Summary Of Findings From The Digital Youth Project*. The MIT Press, 06 2009. ISBN 9780262258937. doi: 10.7551/mitpress/8519.001.0001. URL https://doi.org/10.7551/mitpress/8519.001.0001.

R. Izsak-Ndiaye. Report of the special rapporteur on minority issues, rita izsak : comprehensive study of the human rights situation of roma worldwide, with a particular focus on the phenomenon of anti-gypsyism. Technical report, UN, Geneva :. 2015-05-11, 5 2015. URL http://digitallibrary.un.org/record/797194. Submitted pursuant to Human Rights Council resolution 26/4.

M. S. Jahan and M. Oussalah. A systematic review of hate speech automatic detection using natural language processing. *ArXiv*, abs/2106.00742, 2021.

S. Jasanoff and H. Simmet. No funeral bells: Public reason in a ‘post-truth’ age. *Social Studies of Science*, 47:751–770, 10 2017. doi: 10.1177/0306312717731936.

D. Jurgens, E. Chandrasekharan, and L. Hemphill. 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. Association for Computational Linguistics (ACL), 2019.

T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury. Customer segmentation using k-means clustering. In *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*, pages 135–139. IEEE, 2018.

R. Katarya and O. Verma. Restaurant recommender system based on psychographic and demographic factors in mobile environment. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 907–912, 10 2015. doi: 10.1109/ICGCIoT.2015.7380592.

L. Kaufmann and P. Rousseeuw. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 01 1987.

K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30:61 – 70, 1970.

S. Kumar and K. Ramaswami. Efficient cluster validation with k-family clusters on quality assessment. *European Journal of Scientific Research*, 53:25–36, 05 2011.

I. Kwok and Y. Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

M. Lai, M. Tambuscio, V. Patti, G. Ruffo, and P. Rosso. Extracting graph topological information and users' opinion. In G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 112–118, Cham, 2017. Springer International Publishing.

M. Lai, M. Tambuscio, V. Patti, G. Ruffo, and P. Rosso. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738, 09 2019. doi: 10.1016/j.datak.2019.101738.

R. H. Lauer. Defining social problems: Public and professional perspectives. *Social Problems*, 24(1):122–130, 1976. ISSN 00377791, 15338533. URL http://www.jstor.org/stable/800329.

S. Lazzardi, V. Patti, and P. Rosso. Categorizing misogynistic behaviours in italian, english and spanish tweets. *Proces. del Leng. Natural*, 66:65–76, 2021. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6323.

P. S. N. Lee, L. Leung, V.-H. Lo, C. Xiong, and T. Wu. Internet communication versus face-to-face interaction in quality of life. *Social Indicators Research*, 100:375–389, 2011.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116, New York City, June 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-2915.

C. Luo, Y. Li, and S. M. Chung. Text document clustering based on neighbors. *Data & Knowledge Engineering*, 68(11):1271–1288, 2009. ISSN 0169-023X. doi: https://doi.org/10.1016/j.datak.2009.06.007. URL https://www.sciencedirect.com/science/article/pii/S0169023X09000974. Including Special Section: Conference on Privacy in Statistical Databases (PSD 2008) – Six selected and extended papers on Database Privacy.

F. Marozzo and A. Bessi. Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, 8, 11 2017. doi: 10.1007/s13278-017-0479-5.

A. Maul, L. Mari, and M. Wilson. Intersubjectivity of measurement across the sciences. *Measurement*, 131:764–770, 2019.

M. Mcpherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–, 01 2001. doi: 10.3410/f.725356294.793504070.

Y. Mehdad and J. Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, Sept. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3638. URL https://www.aclweb.org/anthology/W16-3638.

M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *J. Assoc. Inf. Sci. Technol.*, 58:2078–2091, 2007.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.

C. Miller, F. Arcostanzo, J. Smith, A. Krasodomski-Jones, S. Wiedlitzka, R. Jamali, and J. Dale. From brussels to brexit: Islamophobia, xenophobia, racism and reports of hateful incidents on twitter. *DEMOS. Available at www. demos. co. uk/wpcontent/uploads/2016/07/From-Brussels-to-Brexit_-Islamophobia-Xenophobia-Racism-and-Reports-of-Hateful-Incidents-on-Twitter-Research-Prepared-for-Channel-4-Dispatches-% E2*, 80, 2016.

S. Mills and D. Z. Kádár. *Politeness and culture*, page 21–44. Cambridge University Press, 2011. doi: 10.1017/CBO9780511977886.004.

S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL https://aclanthology.org/S16-1003.

S. Morzhov. Avoiding unintended bias in toxicity classification with neural networks. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 314–320, 2020. doi: 10.23919/FRUCT48808.2020.9087368.

Z. Mossie and J.-H. Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57:102087, 07 2019. doi: 10.1016/j.ipm.2019.102087.

C. Mudde and C. R. Kaltwasser. Studying populism in comparative perspective: Reflections on the contemporary and future research agenda. *Comparative Political Studies*, 51(13):1667–1693, 2018. doi: 10.1177/0010414018789490. URL https://doi.org/10.1177/0010414018789490.

M. Munikar, S. Shakya, and A. Shrestha. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE, 2019.

F. Napolitano, L. Gualdieri, G. Santagati, and I. Angelillo. Violence experience among immigrants and refugees: A cross-sectional study in italy. *BioMed Research International*, 2018: 1–8, 09 2018. doi: 10.1155/2018/7949483.

S. Nevin, R. Gleasure, P. O'Reilly, J. Feller, S. Li, and J. Cristoforo. Large crowds or large investments? how social identity influences the commitment of the crowd. In I. Ramos, V. Tuunainen, and H. Krcmar, editors, *25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, June 5-10, 2017*, 2017. URL http://aisel.aisnet.org/ecis2017_rip/30.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883062. URL https://doi.org/10.1145/2872427.2883062.

T. Nockleby, J. Hate speech. In *Encyclopedia of the American Constitution*, page 3:1277–1279. Macmillan, 2000.

D. Nozza, F. Bianchi, and D. Hovy. What the [mask]? making sense of language-specific BERT models. *CoRR*, abs/2003.02912, 2020. URL https://arxiv.org/abs/2003.02912.

G. S. O'Keeffe and K. Clarke-Pearson. The impact of social media on children, adolescents, and families. *Pediatrics*, 127:800 – 804, 2011.

N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung. Multilingual and multi-aspect hate speech analysis. 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 4675–4684, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1474. URL https://aclanthology.org/D19-1474.

E. W. Pamungkas, V. Basile, and V. Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Inf. Process. Manag.*, 57(6):102360, 2020a. doi: 10.1016/j.ipm.2020.102360. URL https://doi.org/10.1016/j.ipm.2020.102360.

E. W. Pamungkas, V. Basile, and V. Patti. Do you really want to hurt me? predicting abusive swearing in social media. In *The 12th Language Resources and Evaluation Conference*, pages 6237–6246. European Language Resources Association, 2020b.

Z. Pan, C.-C. Lee, J. Man, and C. So. One event, three stories. *Gazette*, 61:99–112, 04 1999. doi: 10.1177/0016549299061002001.

A. Panicacci. Does expressing emotions in the local language help migrants acculturate? international journal of language and culture. *International Journal of Language and Culture*, 6(2): 279–304, 2019. doi: https://doi.org/10.1075/ijolc.17013.pan. URL https://www.jbe-platform.com/content/journals/10.1075/ijolc.17013.pan.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Trans. Internet Techn.*, 20(2):12:1–12:19, 2020. doi: 10.1145/3369869. URL https://doi.org/10.1145/3369869.

F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco. Hate speech annotation: Analysis of an italian twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

F. Poletto, V. Basile, C. Bosco, V. Patti, and M. Stranisci. Annotating hate speech: Three schemes at comparison. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–8. CEUR-WS, 2019.

F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523, 2021. doi: 10.1007/s10579-020-09502-8. URL https://doi.org/10.1007/s10579-020-09502-8.

M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, and V. Basile. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481 of *CEUR Workshop Proceedings*, Bari, Italy, 2019. CEUR-WS.org.

A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1873–1876, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871751. URL http://doi.acm.org/10.1145/1871437.1871751.

H. Prasetya and T. Murata. A model of opinion and propagation structure polarization in social media. *Computational Social Networks*, 7, 01 2020. doi: 10.1186/s40649-019-0076-z.

C. Qiuru, L. Ye, H. Xi, L. Yijun, and Z. Guangping. Telecom customer segmentation based on cluster analysis. In *2012 International Conference on Computer Science and Information Processing (CSIP)*, pages 1179–1182, 08 2012. ISBN 978-1-4673-1410-7. doi: 10.1109/CSIP.2012.6309069.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018. URL https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

E. Raisi and B. Huang. Cyberbullying identification using participant-vocabulary consistency, 2016.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://www.aclweb.org/anthology/D16-1264.

A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin. Offensive language detection using multi-level classification. In *Canadian Conference on AI*, 2010.

D. Razo and S. Kübler. Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.9. URL https://www.aclweb.org/anthology/2020.alw-1.9.

K. Relia, Z. Li, S. H. Cook, and R. Chunara. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities. *CoRR*, abs/1902.00119, 2019. URL http://arxiv.org/abs/1902.00119.

E. Richardson, F. Escalettes, I. Fotheringham, J. Wallace, and M. Watson. Meta4: A web application for sharing and annotating metagenomic gene predictions using web services. *Frontiers in genetics*, 4:168, 09 2013. doi: 10.3389/fgene.2013.00168.

E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, 2003. URL https://www.aclweb.org/anthology/W03-1014.

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, and T. Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, 2016. URL https://arxiv.org/pdf/1701.08118.pdf.

K. Saha, E. Chandrasekharan, and M. De Choudhury. *Prevalence and Psychological Effects of Hateful Speech in Online College Communities*, page 255–264. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450362023. URL https://doi.org/10.1145/3292522.3326032.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and M. Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association, 2018. URL http://aclweb.org/anthology/L18-1443.

M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2765/paper162.pdf.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://www.aclweb.org/anthology/P19-1163.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL https://www.aclweb.org/anthology/W17-1101.

A. Shafeeq. Dynamic clustering of data with modified k-means algorithm. In *Proceedings of the 2012 conference on information and computer networks*, 02 2012. doi: 10.13140/2.1.4972.3840.

T. Sheerman-Chase, E.-J. Ong, and R. Bowden. Cultural factors in the regression of non-verbal communication perception. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1249, 11 2011. doi: 10.1109/ICCVW.2011.6130393.

R. Singh and M. P. S. Bhatia. Data clustering with modified k-means algorithm. In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pages 717–721, 06 2011. ISBN 978-1-4577-0588-5. doi: 10.1109/ICRTIT.2011.5972376.

G. Soberón, L. Aroyo, C. Welty, O. Inel, H. Lin, and M. Overmeen. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, pages 45–58, Aachen, Germany, Germany, 2013. CEUR-WS.org. URL http://dl.acm.org/citation.cfm?id=2874376.2874381.

S. Sood, J. Antin, and E. Churchill. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1481–1490, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208610. URL http://doi.acm.org/10.1145/2207676.2208610.

E. Spertus. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, 1997.

K. Tarasova. Development of socio-emotional competence in primary school children. *Procedia - Social and Behavioral Sciences*, 233:128–132, 10 2016. doi: 10.1016/j.sbspro.2016.10.166.

S. Thompson. Hate speech and self-restraint. *Ethical Theory and Moral Practice*, 22, 06 2019. doi: 10.1007/s10677-019-10004-y.

A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623.

S. Turkle. *Reclaiming Conversation: The Power of Talk in a Digital Age*. Penguin Press, 2015. ISBN 9781101980460. URL https://books.google.it/books?id=V813rgEACAAJ.

Twitter. Twitter rules, 2017. Retrieved from https://help.twitter.com/it/rules-and-policies/twitter-rules.

A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. A case for soft-loss functions. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, pages 173–177, 2020. URL https://ojs.aaai.org/index.php/HCOMP/article/view/7478.

C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Detection and fine-grained classification of cyberbullying events. In *Proceedings of* *the 10th Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar, Bulgaria, 09 2015.

T. Venturini. From fake to junk news, the data politics of online virality. In D. Bigo, E. Isin, and E. Ruppert, editors, *Data Politics: Worlds, Subjects, Rights*, pages 123–144. Routledge, 03 2019. doi: 10.4324/9781315167305-7. URL https://doi.org/10.4324%2F9781315167305-7.

J. Vrielink. 'islamophobia' and the law: Belgian hate speech legislation and the wilful destruction of the koran. *International Journal of Discrimination and the Law*, 14(1):54–65, 2014. doi: 10.1177/1358229113500418. URL https://doi.org/10.1177/1358229113500418.

J. Waldron. The core of the case against judicial review, 115 yale l.j. *The Yale Law Journal: https://digitalcommons.law.yale.edu/ylj/vol115/iss6/3*, 03 2006.

W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390374.2390377.

Z. Waseem. 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, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL https://www.aclweb.org/anthology/W16-5618.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://www.aclweb.org/anthology/N16-2013.

M. Wendling. The year that angry won the internet. In *Retrieved from http://www.bbc.com/news/blogs-trending-35111707.*, 2015.

J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30:277–308, 09 2004. doi: 10.1162/0891201041850885.

M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL https://www.aclweb.org/anthology/N19-1060.

M. Wiegand, J. Ruppenhofer, and E. Eder. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.48. URL https://aclanthology.org/2021.naacl-main.48.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6. URL http://dl.acm.org/citation.cfm?id=2382029.2382139.

V. Yadav, R. Shukla, A. Tripathi, A. Maurya, et al. A new approach for movie recommender system using k-means clustering and pca. *Journal of Scientific and Industrial Research (JSIR)*, 80(02):159–165, 2021.

H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP*, 2003.

S. Yu, S. Jindian, and D. Luo. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, PP:176600–176612, 11 2019. doi: 10.1109/ACCESS.2019.2953990.

O. F. Zaidan and C. Callison-Burch. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-2007.

M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and c. Çöltekin. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*, 2020.

Z. Zhang and L. Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, Accepted, 10 2018. doi: 10.3233/SW-180338.

M. Zook. Mapping racist tweets in response to president obama's re-election. In *In floating sheep*, 2012.

Appendix

List of publications within the course of the PhD

[2019]

- A. **Akhtar, S.**; Basile, V.; and Patti, V. 2019. A new measure of polarization in the annotation of hate speech. In Alviano, M.; Greco, G.; and Scarcello, F., eds., AI*IA 2019 – Advances in Artificial Intelligence, 588–603. Cham: Springer International Publishing.

[2020]

- A. **Akhtar, S.**, Basile, V., & Patti, V. (2020). Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8(1), 151-154.

[2021]

- A. **Akhtar, S.**, Basile, V., & Patti, V. (2021). Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. Accepted with revisions in Elsevier Information Processing and Management Journal on 23-12-2021.

- M. Fell, **S. Akhtar**, and V. Basile. Mining annotator perspectives from hate speech corpora.In E. Cabrio, D. Croce, L. C. Passaro, and R. Sprugnoli, editors, Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021, volume 3015 of CEUR Workshop Proceedings. CEURWS. org, 2021. URL http://ceur-ws.org/Vol-3015/paper136.pdf.