Darkness Cannot Drive Out Darkness: Investigating the Impact of Bias in NLP Models on Hate Speech Detection Models

Fatma Elsafoury

Thesis submitted in partial fulfilment of the requirements
of the University of the West of Scotland
for the award of Doctorate of Philosophy

June 2023

Abstract

Hate speech on social media could have severe negative effects. This is why it is crucial to develop tools for automated hate speech detection. These tools should provide a safer environment for individuals, especially from marginalised groups, to express themselves online. However, recent research shows that current hate speech detection models falsely flag content written by members of marginalised communities, as hateful. Similarly, recent research indicates that there are biases in natural language processing (NLP) models. Yet, the impact of these biases on the task of hate speech detection has been understudied.

I identify three research problems: 1) the lack of studying the impact of bias in NLP models on the performance and explainability of hate speech detection models; 2) the lack of studying the impact of the imbalanced representation of hateful content on the bias in NLP models; and 3) the lack of studying the impact of bias in NLP models on the fairness of hate speech detection models. Investigating and understanding the impact of bias in NLP on hate speech detection models will help the NLP community develop more reliable, effective, unbiased, and fair hate speech detection models.

In this thesis, I first critically review the literature on hate speech and bias in NLP models. Then, I address my research problems by investigating the intersection of bias in NLP and hate speech detection models from three perspectives: 1) The explainability perspective, where I address the first research problem and investigate the impact of bias in NLP models on their performance of hate speech detection and whether the bias in NLP models explains their performance on hate speech detection. I run a series of experiments to investigate pre-training bias in 3 contextual word embeddings and 5 static word embeddings and test that impact of these models on five hate speech related datasets; 2) the offensive stereotyping bias perspective, where I address the second research problem and investigate the impact of imbalanced representations and co-occurrences of hateful content with marginalised identity groups on the bias of NLP models. I prospose two metrics to measure the offensive stereotyping bias in static and cotnextual word embeddings. I used 15 static word embeddings and 3 contextual word embeddings. Then I invetigate the impact of the measured bias on the downstream task of hate speech detection on 6 hate speech related datasets; and 3) the fairness perspective where I address the third research problem and investigate the impact of

3 sources of bias in NLP models on the fairness of the task of hate speech detection. I also investigate the impact of removing the different sources of bias on the fairness of hate speech detection.

The findings of this thesis provide evidence that the bias in NLP models has an impact on hate speech detection models from all three perspectives. This means that we need to mitigate the bias in NLP models so that we can ensure the reliability of hate speech detection models. On the other hand, I argue that the limitations and criticisms of the currently used methods to measure and mitigate bias in NLP models are direct results of failing to incorporate relevant literature from the social sciences.

Table of contents

List of tables

List of figures

Declaration

The work leading to this PhD Thesis has been conducted in the School of Physics, Engineering, and Computing at the University of The West of Scotland. The supervisors of this PhD are Keshav Dahal and Zeeshan Perves.

With the exception of chapters 1, 2, and 3 which contain introductory material and a literature review, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Acknowledgement

This thesis is the product of 4 years of work. In those 4 years I met many inspiring researchers who contributed to my work and my growth as a researcher in different ways. I'd like to acknowledge their positive impact that definitely led to produce this thesis in its current shape.

I'd like first to thank my main supervisor Keshav Dahl for organising my PhD defence and for his feedback on the early versions of my defence presentation. I'd also like to thank my internal examiner Qi Wang and the external examiner Walid Magdy for serving on my examination committee and for their valuable feedback on my thesis.

In 2019, I started my PhD on hate speech detection at the university of the West of Scotland where there is no research group or researchers who work on NLP. I understood that for me to progress my work and to be able to publish, I first need to be part of the NLP community in the UK and to find NLP researchers who are willing to collaborate with me. I contacted many researchers. A few replied and many have not. Among those who replied was Steven R. Wilson. Back then, Steven was a Post-doctoral researcher at the SMASH research group in Edinburgh university. Steven ended up co-authoring 3 of my published work. His contribution was not limited to reading and editing my writing but to have important discussions that led to improving my work.

Around the same time in 2019, I also realized that my supervisor at the time would not be able to help me with my research. So I contacted Stamos Katsigianis who was a post-doctoral researcher working with my supervisor at the time. I asked him to read one of my early papers and his feedback led to huge improvements in the paper and he became one of my main co-authors and de-facto my main supervisor. In the later stages of my PhD, Stamos read my thesis and gave me valuable feedback. More importantly, when I was not feeling good about my research, Stamos would reach out and encourage me to try again.

Summer 2022, I did an internship at IBM research in New York and I got the chance to meet many interesting NLP researchers and to have interesting discussions. Among them were my manager (Radu Florian), team lead (Bishwaranjan Bhattacharjee), Ioana Baldini and Katy Gero.

In November 2022, I gave a talk at the Interaction Lab at Harriot-Watt university where I met Verena Rieser, Gavin Abercrombie, and other researchers. Verena was such a positive example of a professor who showed interest in me as a research and what I do. Verena and Gavin asked me very important questions about the limitations of my work and how my PhD would benefit the NLP community. These questions stayed with me and I kept thinking how to answer them which resulted in my Conclusion chapter that was published as a paper at the BigPicture workshop at EMNLP 2023. I then asked them to read one of my papers and Verena gave me very useful feedback on my paper. Gavin not only gave me feedback but he also collaborated with me on the paper.

In December 2022, I gave a talk at the SMASH research group at Edinburgh university and got positive feedback on my work and interesting questions from Walid Magdy, Björn Ross, Eddie Ungless, and Wendy Zheng.

I met many inspiring researchers through the Women_in_NLP talk series: Helia Hashemi, Abhilasha Ravichander, Abeer Aldayel, Alexandra Olteanu, Marian Antoniak, Sabine Weber, Jasmin Bastings, Vered Shwartz, Khyathi Chandu, and Rachael Tatman. I'd like to thank all of them for accepting my invitation and for giving their interesting talks.

I'd also like to thank my current colleagues at the Data, Algorithmic Systems and Ethics research group at the Weizenbaum institute for their support during the tough period before my PhD defence: Milagros Miceli, David Hartman, Tianling Yang, Lena Pohlmann, Seyi Olojo, Laurenz Sachenbacher, and Camilla Salim Wagner.

I'm lucky to have my friends whose support has no end: Soha Yassin, Heba Fekry, and Patrizia Di Campli San Vito. Big thanks should go to my family in Cairo, Hünxe, Köln, Tiblisi, and Kerala (Yes, I'm lucky to have family members all over the world!).

In the dark times when I was burned out, depressed and stressed out, I got huge help from the counselling service at the university of the West of Scotland where therapists heard me out and gave me access to resources that helped me navigating those dark times. I particularly want to thank Hilary Groom.

Finally, I'd like to thank Michel Steuwer, for his eternal support, constant encouragement, and for believing in me and in my right to have a fair PhD examination with an expert in my research area as an external examiner. He supported me in my fight for that right when almost everyone else told me to go along with whatever my supervisor, at the time, decides. He is one of a few people who gave me very valuable feedback on my thesis and most of my writings that helped making them better. Without you Michel, I'd have given up.

Chapter 1

Introduction

Social media has provided many opportunities for connection and communication worldwide. It provided a space for millions of people to share their thoughts, experiences, and opinions, as well as spread misinformation and hateful and abusive content. With the concerning increase in the scale of online hate speech in 2021, as shown in Commission et al. [51], the concern about hate crime increased as well. Research has indicated that there is a strong positive correlation between online hate speech and offline hate crimes following significant events like elections, terrorist attacks, or court cases, as shown in Hanes and Machin [97], Williams and Burnap [285] or without such events happening, as shown in Williams et al. [286]. For example, in the Christchurch terror attack in 2019, 51 members of the Muslim community in New Zealand were killed in two mosques, with plans to target a third one. The attacker, Brenton Tarrant, an extreme right-wing terrorist, live-streamed the shooting on Facebook, indicating that he was taking his online hate dialogue into an offline action , as shown in Williams et al. [286].

As a result, many countries have passed legislation to stop online hate speech. For example, in January 2018, the German Network Enforcement Act (NetzDG) imposed a legal obligation on social media platforms with more than two million users to remove hateful content within 24 hours or risk a fine of up to 50 million euros. This example has been followed by over 20 countries worldwide, as shown in Mchangama et al. [150]. These legislative measures pressured social media platforms like Facebook, Twitter, and YouTube to implement algorithms to detect and remove hateful content. For example, Facebook removed a total of 78.6 million posts in 2020 for violating community standards on hate speech. Similarly, Twitter removed 14900 tweets and challenged 4.5 million tweets between March and July 2020 for spreading misinformation , as shown in Mchangama et al. [150].

However, there are claims that the policies and algorithms used to moderate content on social media platforms are vague, conflicting, and non-transparent, with negative consequences

for freedom of expression. This led to the silencing of the communities that the legislation aimed to protect, as shown in Mchangama et al. [150]. It urged a coalition of more than 70 social and racial justice organizations to write a letter to Facebook to ask them to fix their racially biased moderation system , as shown in Levin [137]. For example, Sap et al. [224] demonstrates that tweets written in African American English and tweets by self-identified African Americans are two times more likely to be labeled as toxic. Similarly, Dias Oliva et al. [66] demonstrates that Facebook's restriction of certain words without taking into consideration the context in which they are being used. This restriction led to the censoring of some comments by the LGBTQ community, who proclaimed some of these restricted words as self-expression. This problem also exists on other social media platforms like Twitter and YouTube, as shown in Mchangama et al. [150]. These biases and limitations on freedom of speech are the results of the moderation process that most social media platforms implement. The moderation process is a hybrid of machine learning (ML) models and human reviewers. First, the ML models find a post with potentially harmful content, and then this post is reviewed by human reviewers who make the final decision if the post is harmful or not, as shown in Jiang et al. [112]. Additionally, Facebook and Twitter rely on users to report harmful content. There are different types of ML models. Some ML models are trained to understand images and videos. Other ML models are trained to understand text, such as natural language processing (NLP) models. There is evidence that different NLP models contain different social biases like racial biases, as shown in Garg et al. [86], Manzini et al. [147], Sweeney and Najafian [254], gender biases, as shown in Bolukbasi et al. [27], Chaloner and Maldonado [41], Garg et al. [86], personality stereotypes, as shown in Agarwal et al. [4], and offensive stereotyping bias, as demonstrated in chapter 5. There is also evidence that human reviewers sometimes bring their biases into the process, as shown in Jiang et al. [112]. This thesis focuses only on NLP models and does not examine human reviewers.

1.1 Research Problems

The impact of bias in NLP models on NLP tasks like hate speech detection is still understudied, even though bias in NLP models has been an active research direction in the last few years, as shown in Caliskan et al. [38], Dev and Phillips [64], Nangia et al. [172]. In this thesis, I identify the following three research problems:

1.1.1 The lack of studying the impact of bias in NLP models on the performance and explainability of hate speech detection models

The first research problem is the lack of understanding of how the bias in NLP models impacts the performance of hate speech detection, and whether the bias in NLP models explains the performance of hate speech detection models. Prior research on bias and hate speech detection models focused mainly on the impact of bias on the fairness of hate speech detection models, not their performance.

In this thesis, I fill this research gap by investigating the explainability of some of the best performing hate speech detection models. As well as investigating whether different types of bias in the most commonly used NLP models explain the performance of these models on the task of hate speech detection.

1.1.2 The lack of studying the impact of imbalanced representations on bias in NLP models

The second research problem is that the impact of imbalanced representation and the co-occurrence of hateful expressions with marginalised identity groups has not been studied in NLP models, e.g., word embeddings, as well as their indirect impact on the performance of hate speech detection models. Prior research has focused only on the impact of imbalanced representations of marginalised identity groups in hate speech datasets on hate speech detection models, as shown in Dixon et al. [69].

In this thesis, I fill this research gap by investigating the bias resulting from imbalanced representations in NLP models and how it impacts the performance of hate speech detection models.

1.1.3 The lack of studying the impact of bias in NLP on the fairness of hate speech detection models

The third research problem is the impact of the bias in NLP models on the fairness of hate speech detection models. This has been studied in the literature. However, there are some significant limitations in the conducted research. For example, Goldfarb-Tarrant et al. [89] uses only one metric to measure social bias, and one fairness metric, which makes their findings inconclusive since different bias and fairness metrics tend to give different results, as demonstrated in Badilla et al. [14], Elsafoury et al. [75]. Similarly, in Steed et al. [241], the authors use only one bias metric and bleached template sentences to measure bias in contextual word embeddings. The problem with bleached sentences is that they do not

provide a real context, and hence their results to measure bias are unreliable, as indicated in May et al. [149].

In this thesis, I overcome these limitations by investigating the impact of different sources of bias on the fairness of hate speech detection models. I use different metrics to measure bias in NLP models, as well as multiple fairness metrics. I investigate the impact of removing the different types of bias on the fairness of the hate speech detection models.

1.2 Research Contributions

In this thesis, the research goal is to investigate the impact of bias in NLP models on hate speech detection models by addressing the identified research problems. Understanding the impact of bias in NLP models on hate speech detection models is crucial to ensuring their effectiveness and fairness. Since hate speech detection models that utilize biased NLP models, e.g., word embeddings, may learn to associate marginalised groups with extremism and hate. Consequently, they may lead to blocking them or flagging their content as inappropriate instead of providing a protective environment for marginalised people to express themselves. To address the identified research problems, I start by surveying the literature on the two aspects of the conducted research: hate speech and bias and fairness in NLP models. Then, I investigate the intersection between hate speech and bias in NLP to address the three research problems from three perspectives. The first perspective is the explainability perspective, where I investigate whether the biases in NLP models can explain the performance of some hate speech detection models. The second perspective is the offensive stereotyping bias perspective, where I investigate how hate speech makes the NLP models form associations between profanity and marginalised groups. Finally, the last perspective is the fairness perspective, where I investigate how the social bias in NLP models impacts the fairness of the task of hate speech detection. Each of the research contributions is explained in detail below.

In this thesis, I study bias and marginalisation from a Western perspective where the marginalized identities are groups like women, non-white ethnicities, and Muslims. My contribution is also limited to studying English language models and using English hate speech datasets.

1.2.1 Survey: Hate Speech

To precisely understand hate speech, I rigorously survey the literature on hate speech, its definitions, and types. I also review, the literature on the different hate speech datasets and

how they were collected and annotated, as well as the most common machine learning models and feature selection techniques used in hate speech detection. This survey on hate speech and hate speech detection aims to answer the following research questions:

1. What is hate speech, and what are the different forms of hate speech in literature?
2. What are the most commonly used datasets and models used to detect hate speech?
3. What are the limitations and challenges of the current research on hate speech detection?

The literature review on the automated detection of hate speech sheds light on challenges and limitations of the current literature regarding: (a) data collection; (b) feature selection; (c) model selection and training; (d) evaluation metrics; and (e) bias and fairness. The focus of this thesis is to address the last challenge by investigating the impact of bias in NLP models on hate speech detection models. This survey is published in the IEEE-ACCESS journal in 2021.

1.2.2 Survey: Bias and Fairness in NLP

Since the focus of this thesis is to investigate the impact of bias in NLP models and hate speech detection models, the second research contribution is a survey of the literature on bias and fairness in NLP models. More specifically, I review the literature on the definition of bias and fairness in ML and NLP models and the proposed methods to measure and mitigate them. I also review the literature in the social sciences and NLP on the sources of bias. This survey aims to answer the following research questions:

1. What are bias and fairness, and how do we measure and mitigate them in NLP models?
2. What are the origins of bias from a social science perspective? How do they relate to the sources of bias from an NLP perspective?
3. What are the limitations of studying bias and fairness in NLP? How can we mitigate these limitations?

Based on the literature review on bias and fairness from the perspective of social science in addition to the NLP perspective, I argue that, in fact, the sources of bias found in the NLP pipeline are rooted in those uncovered in the social sciences. I also discuss how the lack of inclusion of social sciences in attempts at mitigating bias in NLP models has resulted in problematic quantitative measures of bias and superficial mitigation techniques. Finally,

I propose recommendations to the NLP community to mitigate biases in NLP models by incorporating the social sciences.

After having thoroughly surveyed the relevant literature, I next investigate the interaction and intersection between hate speech and bias in NLP from three perspectives to address the three research problems.

1.2.3 The Explainability Perspective

For this perspective, I aim to address the first research problem and understand the performance of hate speech detection models and whether bias in NLP models explains their performance on the task of hate speech detection. I investigate this relationship between different types of bias in NLP models and their performance on hate speech detection. I examine the impact of two types of bias: 1) the bias resulting from pre-training NLP models, and 2) the bias resulting from biased pre-training datasets. In this work, I aim to answer the following research questions:

1. How does bias resulting from pre-training NLP models explain their performance on the task of hate speech detection?
    - What is BERT's performance on different hate-speech-related datasets?
    - What is the role that attention weights play in BERT's performance?
    - What does BERT learn during fine-tuning?
    - Does pre-training bias explain the performance of contextual word embeddings on the task of speech detection?
2. How do biased pre-training datasets impact the performance of NLP models on the task of hate speech detection?
    - What is the performance of the different word embeddings on offenses' categorization?
    - What is the performance of the different word embeddings on the task of hate speech detection?
    - Can we use certain static word embeddings to detect certain offensive categories within hate-speech-related datasets?
    - Do biased pre-training datasets explain the performance of static word embeddings on hate speech detection?

3. What is the impact of social bias in NLP models on their performance on the task of hate speech detection?

    - Does social bias explain the performance of contextual word embeddings on the task of hate speech detection?
    - Are social-media-based word embeddings more socially biased than informational-based word embeddings?
    - Does social bias explain the performance of static word embeddings on the task of hate speech detection?

The results in chapter 4 show that pre-training language models results in syntactic bias that improves their performance on hate speech detection tasks. Similarly, the results show that pre-training some word embeddings on biased datasets improved their performance on different tasks related to hate speech detection. This improved performance suggests that the different biases explain the performance of these models on the task of hate speech detection. They also suggest that hate speech detection models might be making the right decisions for the wrong reasons. For example, associating hateful content with marginalised groups could lead to flagging the mere existence of marginalised identities as hateful. On the other hand, the results show no strong evidence that the social bias in NLP models, whether static or contextual word embeddings, explains the performance of hate speech detection models. However, this could be because of the limitations of the metrics proposed in the literature to measure social bias. So, this finding remains inconclusive. Part of the work in this chapter is published at the SIGIR conference in 2021 and the SocialNLP workshop in 2022.

1.2.4 The Offensive Stereotyping Bias Perspective

For this research perspective, I aim to address the second research problem and understand how the hateful content against marginalised groups on social media and other platforms that are used to train NLP models is being encoded by those models to form offensive stereotyping bias. I introduce the novel systematic offensive stereotyping (SOS) bias. I formally define it, propose a method to measure it, and validate it. Finally, I study how it impacts the performance of hate speech detection models. In this work, I am interested in answering the following research questions:

1. How can we measure SOS bias? How to validate it?

    - How to measure SOS bias in static and contextual word embeddings?

- What are the SOS bias scores of common pre-trained static and contextual word embeddings?
- Does SOS bias in the word embeddings differ from social biases?

2. How strongly does SOS bias correlate with external measures of online extremism and hate?
3. Does the SOS bias in the word embeddings explain the performance of these word embeddings on the task of hate speech detection?

The findings of this work demonstrate that word embeddings, both static and contextual, are SOS-biased. The SOS bias is significantly higher against marginalised groups in static word embeddings. Even though there is no strong evidence that the SOS bias explains the performance of the word embeddings on the task of hate speech detection, the existence of the SOS bias might have an impact on the hate speech detection models in ways that we have not explored or understood yet. Part of the work in this chapter is published at the COLLING conference in 2022.

1.2.5 The Fairness Perspective

For this research perspective, I aim to address the third research problem and understand how the different sources of bias in NLP models, language models, impact the fairness of the downstream task of hate speech detection. I first investigate three of the four sources of bias and their impact on the fairness of the NLP task of hate speech detection. Then, I investigate the impact of removing these biases on the fairness of hate speech detection models. I aim to find out the most important sources of bias and what debiasing techniques to use to ensure that our hate speech detection models are as fair as possible. To this end, this work aims to answer the following research questions:

1. What is the impact of the different sources of bias on the fairness of the downstream task of hate speech detection?
2. What is the impact of removing the different sources of bias on the fairness of the downstream task of hate speech detection?
3. Which debiasing technique to use to ensure the fairness of the task of hate speech detection?
4. How to have fairer text classification models?

The findings of this chapter show that the examined types of bias have an impact on the fairness of the models on the task of hate speech detection. The results indicate that the bias in the fine-tuning datasets used in the downstream task has a stronger impact on the models' fairness than the bias in the pre-trained language models. This means that the bias in the current hate speech datasets and the bias in the most commonly used language models have a negative impact on the fairness of hate speech detection models. Hence, researchers should pay attention to these biases and aim to mitigate them before implementing unfair hate speech detection models.

1.2.6 Publications

The work conducted for this thesis led to the following peer-reviewed publications:

1. ***Fatma Elsafoury***. *2022"Darkness cannot drive out darkness: Investigating Bias in Hate Speech and Abuse Detection Models".* ***In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 31–43, Dublin, Ireland. Association for Computational Linguistics 2022.***

2. ***Fatma Elsafoury****, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. 2021 "When the timeline meets the pipeline: A survey on automated cyberbullying detection".* ***In IEEE Access, vol. 9, pp. 103541-103563, 2021, doi: 10.1109/ACCESS.2021.3098979.***

3. ***Fatma Elsafoury****, Stamos Katsigiannis, Steve Wilson, and Naeem Ramzan. 2022 "Does BERT Pay Attention To Cyberbullying?".* ***In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021). Association for Computing Machinery, New York, NY, USA, 1900–1904. https://doi.org/10.1145/3404835.3463029.***

4. ***Fatma Elsafoury****, Steve Wilson, and Naeem Ramzan. 2022 "A Comparative Study on Word Embeddings in Social NLP Tasks".* ***In Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media, pages 55–64, Seattle, Washington. Association for Computational Linguistics 2022.***

5. ***Fatma Elsafoury****, Steve Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022 "SOS: Systematic Offensive Stereotyping Bias in Word Embeddings".* ***In Proceedings of the 29th International Conference on Computational Linguistics, pages 1263–1274, Gyeongju, Republic of Korea. International Committee on Computational Linguistics 2022.***

6. ***Fatma Elsafoury***. *2023 "Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection".* ***In Proceedings of the Big Picture Workshop, pages 53–65, Singapore. Association for Computational Linguistics.***

The rest of the work that went into my thesis is currently under-submission or pre-print that are ready to be submitted:

1. ***Fatma Elsafoury****, Gavin Abercrombie. 2023 "On the Origins of Bias in NLP through the Lens of the Jim Code?".* ***ArXiv preprint https://arxiv.org/abs/2305.09281, 2023.***
2. ***Fatma Elsafoury***. *2023 "Systematic Offensive Stereotyping (SOS) Bias in Language Models".* ***Under-submission at LREC-COLING 2024***
3. ***Fatma Elsafoury****, Stamos Katsigiannis. 2023 "On Bias and Fairness in NLP: How to have a fairer text classification?".* ***ArXiv preprint arXiv:2305.12829, 2023.***

1.3 Thesis structure

- **Chapter 2: Hate Speech (Survey)**
  In this chapter, I make my first research contribution and review the literature on hate speech and hate speech detection models.
- More specifically, on the definition of hate speech and the different machine learning models, features, and evaluation metrics used in the literature to detect hate speech.
- I also discuss the limitations of the reviewed literature, provide suggestions to overcome these limitations, and propose directions for future research in the field of hate speech detection.
- **Chapter 3: Bias and Fairness in NLP (Survey)**
  In this chapter, I achieve my second research contribution and provide another literature review on bias and fairness in NLP models.
- This chapter is an attempt to incorporate the literature in both the fields of social science and NLP. I start by reviewing the literature on critical race theory and
- critical race and digital studies to understand how years of oppression resulted in bias and discrimination in NLP models.
- Then I move on to review the literature on bias from the NLP perspective to review the different methods in the literature used to measure bias and fairness

- in NLP models and to remove the bias from NLP models. Finally, I provide a discussion on the limitations of those methods and recommendations to overcome them.

- **Chapter 4: The Explainability Perspective**
  In this chapter, I aim to achieve my third research contribution. I investigate the impact of bias in NLP on the performance of hate speech detection models by investigating how that bias might explain the performance of hate speech detection models. I inspect the impact of two sources of bias: Pre-training bias and Biased pre-training datasets. I provide the background on which I build the work, a detailed description of the experiments, and an extensive analysis of the results and how they answer the research questions.

- **Chapter 5: The Offensive Stereotyping Bias Perspective**
  In this chapter, I aim to achieve my fourth research contribution. I investigate how hateful content leads language models to form offensive stereotyping between marginalised groups and profanity. To this end, I introduce a computational measure of *systematic offensive stereotyping* (SOS) bias and examine its existence in pre-trained word embeddings. I provide the background on which I build the work, a detailed description of the experiments, and an extensive analysis of the results and how they answer the research questions.

- **Chapter 6: The Fairness Perspective**
  In this chapter, I aim to achieve my fifth research contribution. I investigate different sources of bias and their impact on the models' fairness in the downstream task of hate speech detection. I aim to overcome the limitations of previous research by using different metrics to measure representation (intrinsic) bias and models' fairness. Moreover, I investigate the effectiveness of various debiasing methods for removing different sources of bias, as well as their impact on the models' fairness (extrinsic bias). I provide practical guidelines to ensure the fairness of the downstream task of text classification. I provide the background on which I build the work, a detailed description of the experiments, and an extensive analysis of the results and how they answer the research questions.

- **Chapter 7: Conclusion and Discussion**
  In this chapter, I summarize the work, findings, contributions, and limitations of each chapter. I also provide a discussion of how the findings of this thesis can

- benefit the fields of hate speech detection and bias and fairness in NLP models. Finally, I discuss possible future research directions.

Chapter 2

Survey: Hate Speech

2.1 Introduction

The internet has become an important development tool for young people. It provides a great source of information and a tool for communication. In recent studies, children and young people categorized their Internet activities into three groups: (a) Content-based activities, such as school work, playing games, watching video clips, reading the news, or downloading music; (b) Communication-based activities such as instant messaging, email, chatting or Skype; and (c) Conduct peer participation activities such as blogging, post photos or file-sharing websites, as shown in Omar et al. [182]. Despite all the benefits, the Internet could be an environment for bullying. In their research, Haddon and Livingstone, as shown in Haddon and Livingstone [95] showed that 17% of the children, who are interviewed between the age of 9 and 14 in the UK, are exposed to sexual content compared to 24% of children from the EU. The study also indicated that the children experienced bad language in the form of insults or swearing, aggressive communication, or harassment. Moreover, social media platforms provide a fruitful environment for hate speech in the forms of threats, harassment, and exploiting potential victims ,as shown in Chan et al. [42]. The Pew research center reported in 2017 that 40% of social media users have experienced some form of hate speech, as shown in Duggan [70]. Another study that included university students found that among 200 university students, 91% experienced cyberbullying, 55.5% of them on Instagram, and 38% on Facebook, as shown in Abaido [1]. The negative effect extends to social media moderators who get impacted by reviewing and removing the hateful content online as explained in Sarah Roberta's book *Behind The Screens*, as shown in Roberts [215].

Hate speech and cyberbullying experiences can have serious consequences for the victims, including depression, anxiety, low self-esteem, and self-harm, and may even lead in extreme cases to suicide, as shown in Sticca et al. [243]. Consequently, having tools for detecting and

Paper	Year	Systematic review	Definition section	Types section	Data section	Data annotation section	Features section	Preprocessing section	Models section	Evaluation metrics section	Replication experiments section	Extended experiments section
[178]	2015	✓										
[293]	2016			✓	✓							
[96]	2016		✓	✓					✓	✓		✓
[222]	2017	✓	✓		✓		✓	✓	✓			
[258]	2017											
[145]	2018		✓	✓	✓				✓			
[220]	2018	✓	✓		✓		✓			✓		✓
[255]	2018	✓	✓		✓	✓	✓		✓			
[54]	2018		✓		✓						✓	✓
[166]	2018		✓	✓	✓		✓		✓			
[83]	2018	✓	✓	✓	✓		✓		✓			
[8]	2019	✓	✓				✓		✓	✓		
[76]	2019				✓		✓				✓	✓
[274]	2020	✓	✓	✓	✓	✓						
[199]	2020	✓	✓	✓	✓	✓						
[158]	2021	✓	✓	✓	✓	✓		✓				
This thesis	2024	✓	✓	✓	✓	✓	✓	✓	✓	✓		

Table 2.1 Discussed sections in published literature review papers on the automated detection of hate speech

preventing hate speech is crucial for reducing the negative effects. Studying hate speech is rooted in Psychology, Education, Behavioral Science (BS), and Information Technology (IT). On the IT front, the automated detection of hate speech can help in the automated removal of flagged content, post, or communication, in the automated blocking of the perpetrators, and in reaching out to help the victims. Over the last decade, the body of literature on automated detection of hate speech has been growing, especially concerning detecting hate speech from social media networks like Twitter, as shown in Bosque and Villareal [30], Chatzakou et al. [43], Raisi and Huang [210], Raisi and Huang [211], Waseem and Hovy [282], Zhang et al. [297], Zhao et al. [300], Instagram, as shown in Cheng et al. [49], Hosseinmardi et al. [105], Kao et al. [119], Raisi and Huang [210], Raisi and Huang [211] and YouTube, as shown in Dadvar et al. [55], Dinakar et al. [67], Kumar et al. [131]. This body of research has been working towards automated hate speech detection using either rule-based models, as shown in Bosque and Villareal [30], Dinakar et al. [67], Reynolds et al. [214], conventional machine learning models, as shown in Agrawal and Awekar [5], Cheng et al. [49], Dinakar et al. [67], Kumar et al. [131], or deep learning models, as shown in Agrawal and Awekar [5], Mikolov et al. [154], Zhang et al. [295, 297].

The last decade brought significant advances in the fields of machine learning and natural language processing, which have been successfully applied in domains related to hate speech

detection, such as rumor detection, as shown in Bondielli and Marcelloni [28], sentiment analysis, as shown in Feldman [80], and fake news detection, as shown in Shu et al. [232]. Consequently, it is extremely useful to review the available literature on automated hate speech detection, in light of these recent advances. There have been various attempts to review that body of literature. An overview of the published literature review papers between 2009 and 2021 regarding automated hate speech detection is provided in Table 2.1. The works shown in Table 2.1 cover the following aspects of the examined problem: systematic review or how the literature is collected, as shown in Dadvar and Eckert [54], Salawu et al. [222], Tahmasbi and Fuchsberger [255]; hate speech definition, as shown in Al-Garadi et al. [8], Dadvar and Eckert [54], Mladenović et al. [158], Nadali et al. [166]; hate speech types, as shown in Haidar et al. [96], Mahlangu et al. [145], Zainudin et al. [293]; datasets, as shown in Dadvar and Eckert [54], Tarwani et al. [258], Zainudin et al. [293]; feature selection, as shown in Al-Garadi et al. [8], Emmery et al. [76], Salawu et al. [222]; model selection, as shown in Haidar et al. [96], Salawu et al. [222], Tahmasbi and Fuchsberger [255]; and evaluation metrics, as shown in Al-Garadi et al. [8], Haidar et al. [96], Rosa et al. [220]. However, only a few are comprehensive, as shown in Haidar et al. [96], Rosa et al. [220]. There are some important aspects that are rarely covered in the literature, like data annotation, as shown in Tahmasbi and Fuchsberger [255] and data preprocessing, as shown in Salawu et al. [222]. In addition, some review papers replicate experiments from their reviewed literature [76, 255], while others design their own experiments to fill in gaps in the literature [96, 220].

However, none of the reviews from Table 2.1 organize the reviewed literature around the steps of the text classification pipeline. The text classification pipeline (Fig. 2.1) is a series of ordered steps that constitute the machine learning workflow, consisting of data collection (data sourcing and data annotation), data pre-processing, feature selection, model training, and model evaluation, as shown in Raschka [212]. Organizing the literature review around the text classification pipeline would help to aggregate the different methods and approaches used to accomplish each step in the pipeline, giving the reader the opportunity to learn and compare these different approaches and methods. Taking this into consideration, in this work, I organize the reviewed literature around the steps of the text classification pipeline employed by each reviewed work.

In this chapter, I present my first research contribution and review the collected body of literature on automated hate speech detection, starting with explaining the search strategy for selecting literature works (Section 2.2) and then reviewing the different definitions of hate speech in the literature and the different types of hate speech (Section 2.3.1). Then, I review the different methods used in the literature for each step in the text classification

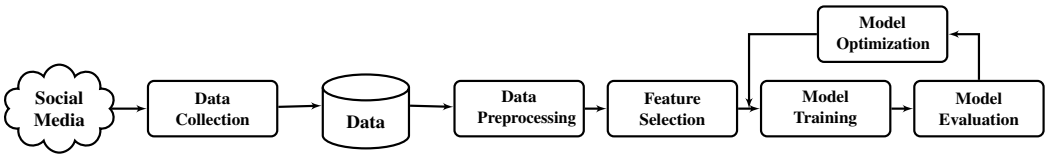


Fig. 2.1 Text Classification Pipeline

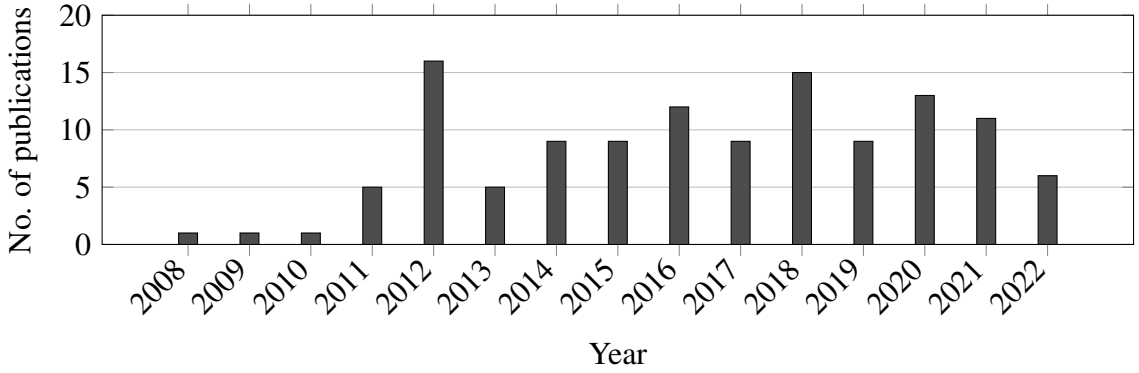


Fig. 2.2 The number of papers on automated detection ofHate speechthat I review, grouped by the year of publication, from 2008 to 2020.

pipeline: data collection (Section 2.4.1), pre-processing (Section 2.4.2), feature selection (Section 2.4.3), model training (Section 2.4.4) and model evaluation (Section 2.4.5). Then, I provide a critical analysis of the current challenges and limitations in the literature on hate speech detection (Section 2.5).

2.2 Search strategy and study selection

The papers reviewed in this chapter are selected by following a systematic literature review method to make sure that as many relevant papers as possible are covered. To achieve this, I first look at how other literature reviews selected their papers. Among the literature review papers in Table 2.1, the collection methods used in, as shown in Salawu et al. [222] and Tahmasbi and Fuchsberger [255] generated the highest number of relevant papers, which is 43. They used the search keywords “cyberbullying” and “detection” to search through Google Scholar, IEEE Xplore, Science Direct, ACM Digital Library and Wiley online databases. Following their method, I locate some key studies in the field of automated hate speech detection. To ensure that as many relevant and new papers as possible are covered, I review the papers that cited those key studies and especially those published after 2016. This process led to 122 papers related to computational methods for hate speech detection. Figure 2.2 shows the number of the reviewed papers grouped by publication year from 2008 to 2020.

Definition	Used in
Cyberbullying is a form of cyber-aggression that is defined as an intentional harmful act to another person that takes place through online means and is characterized by an imbalance of power between the individuals involved and repetition of the act [126, 190, 238]	[7, 47, 207, 270]
Cyberbullying is an individual's intentional and repeated harmful act to others through harmful posts or messages through various digital technologies [22]	[165, 189, 200, 219, 219]
The use of electronic forms of communication to abuse, threat, or harass another person [125]	[31, 85, 194, 210]
When the Internet, cell phones, or other devices are used to send or post text or images intended to hurt or embarrass another person [67]	[165, 189, 200, 219, 219]
Willful and repeated harm inflicted through the medium of electronic text [101]	[21, 122, 173, 214]
Online harassment includes being called offensive names, purposefully embarrassed, stalked, sexually harassed, physical threat in a sustained manner [71]	[289]
Any fierce, purposeful activity directed by people or gatherings, utilizing on the web channels over and again against a victim who does not can [231]	[17, 44]
Hate speech is defined as targeting individuals or groups based on their characteristics (targeting characteristics); demonstrating a clear intention to incite harm, or to promote hatred; it may or may not use offensive or profane words [282]	[282, 297]

Table 2.2 The most common cyberbullying definitions used in the reviewed literature

2.3 Hate Speech and Cyberbullying

2.3.1 Cyberbullying

Definition

The lack of a globally accepted definition of cyberbullying is one of the main issues detected in the reviewed literature on automated cyberbullying detection. For example, although some reviewed works claim to detect cyberbullying in their title, they detect child grooming, as shown in Potha and Maragoudakis [200], Romsaiyud et al. [218] or detect the participants in the act, like the bullies, victims, and bystanders, rather than the actual incident of cyberbullying, as shown in Chelmis et al. [45], Cheng et al. [48]. Out of the

106 reviewed papers, 65 papers defined cyberbullying. There are eight main definitions that most of the papers used, as shown in Table 2.2. However, despite these definitions being close in meaning, as most of them describe cyberbullying as "*one form or another of insulting, spread using mobile or internet technology*", the lack of a clear definition leads to difficulties in comparing and evaluating different works. For example, in, as shown in Belsey [22], Dinakar et al. [67], Kowalski et al. [125], cyberbullying is described as online aggression, bullying using new communication technologies, online harassment, or hate speech. This is problematic as each of these tasks is different, making it significantly difficult to replicate the studies and to compare the models' results and generalisability. Some studies consider cyberbullying as a subtype of cyber-aggression, as shown in Patchin and Hinduja [190], while others consider cyberbullying as a different task from cyber-aggression, as shown in Smith [237]. Mladenović et al. provided a detailed survey on the diversity of the definitions of cyberbullying, cyber-aggression, trolling, and cyber-grooming, as shown in Mladenović et al. [158]. Another issue is that some studies do not differentiate between bullying and cyberbullying apart from the usage of electronic means. As a consequence, they require the following three characteristics of bullying to be evident in cyberbullying cases: harmful, repetitive, and with power imbalance between the bully and the victim. These characteristics are sometimes hard to satisfy in the online space. For example, someone may send a bullying message to someone during an online conversation only once, which does not satisfy repetition. However, some studies claim that the fact that an online post makes permanent harm satisfies the repetition requirement, as shown in Tahmasbi and Fuchsberger [255]. In addition, in the case of the Twitter platform, Tian and Xin argue that negative messages on Twitter tend to be retweeted more often, which also satisfies the repetition requirement, as shown in Tian [261].

Cyberbullying Types

According to Mahlangu et al. [145], there are 12 types of cyberbullying. These types are described in Table 2.3. Most of the reviewed literature does not specify which type of cyberbullying they are detecting. Nevertheless, online harassment is the most common type of cyberbullying in the literature [7, 47, 118, 146, 165, 176]. There are subtypes of harassment mentioned in the reviewed literature like Aggression, as shown in Kao et al. [119], Nazar et al. [173] and Toxicity, as shown in Wulczyn et al. [289].

Type	Description
Flaming	Starting a fight online.
Harassment	Sending insulting messages frequently.
Cyberstalking	Sending intimidating messages to the victim, which causes fear.
Masquerade	The bully pretends to be someone else.
Trolling	Posting controversial comments to upset other members on the online platform.
Denigration	Negative gossip about another person.
Outing	Posting personal information about someone in public forums.
Exclusion	When a social group deliberately excludes someone.
Catfishing	Creating a fake profile using someone else's information.
Dissing	Posting information about someone to hurt them or defame them.
Trickery	Tricking someone to share their secrets or personal information.
Fraping	Using someone else's online account to post inappropriate content and tricking others into believing that the account owner posted them.

Table 2.3 Types of cyberbullying in the literature [233]

2.3.2 Hate Speech

In the last few years, research on hate speech detection has been increasing, as shown in Agrawal and Awekar [5], Arango et al. [12], Krasnowska-Kieras and Wróblewska [127], Kumar et al. [131], Waseem and Hovy [282], Zhang et al. [297]. In a survey chapter on the automated detection of hate speech in text, Fortuna et al. studied the definition of hate speech in the literature in relation to four dimensions: physical violence encouragement, targets, attack language, and humorous hate speech, as shown in Fortuna and Nunes [83]. From these four dimensions, the authors proposed a new definition for hate speech, i.e. "*Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used*".

In the NLP community, it is unclear what the difference in definition between hate speech and cyberbullying is. This lack of clarity can cause generalisability problems with the developed models, as each of the cyberbullying detection and hate speech detection tasks require different features. However, there are also some similarities between the two tasks. The main similarity is the abusive language, while the main difference is the target of the abusive language. In cyberbullying, the abusive language is targeted at specific individuals, while hate speech is targeted at groups of people who share specific characteristics, as shown in Fortuna and Nunes [83]. Examples of types of hate speech and their targeted groups are summarized in Table 2.4.

Categories	Example of possible targets
Race	Black people, white people
Behavior	Insecure people, sensitive people
Physical	Obese people, beautiful people
Sexual orientation	Gay people, straight people
Class	Ghetto people, rich people
Gender	Pregnant people, women
Ethnicity	Chinese people, Indian people
Disability	Bipolar people, people with mental disabilities
Religion	Religious people, Muslims, Jews, Atheists
Other	Drunk people, shallow people

Table 2.4 Types of hate speech and their targets in the literature Silva et al. [233]

The main focus of this chapter is to review the literature on hate speech detection. However, due to the similarities between cyberbullying and hate speech, I opt to include some of the cyberbullying datasets and features used in the literature in addition to the hate speech datasets and features. Consequently, the term hate speech will hereby cover both hate speech and cyberbullying in this chapter and the next chapters.

2.4 Text Classification pipeline

This section provides a thorough literature review on automated hate speech detection, organized by the steps in the text classification pipeline, as shown in Fig. 2.1.

2.4.1 Data collection

In the reviewed literature, the used datasets originated from various social media platforms. In this section, I provide an overview of the different datasets used in the literature, including the annotation processes followed, the ratio between positive and negative samples, and the sampling strategies used.

Data sources

The datasets used in the reviewed literature originated from twelve different sources, including seven social media platforms (Twitter, Instagram, FormSpring, Ask.FM, MySpace, YouTube, Vine, and Reddit), an online collaborative platform (Wikipedia Talk Pages), and a news website (Yahoo News). All of these platforms have experienced incidents of hate speech and are thus used for the creation of datasets for hate speech detection. Examples of offensive

comments from these data sources can be found in Table 2.5. In addition, details about all the datasets, including their source, the number of positive and negative samples, the proportion of positive vs. negative samples, their focus (e.g., cyberbullying, hate speech, cyber-aggression, etc.), their availability, and related references, are provided in Table 2.6.

- **Twitter** is one of the most famous social media platforms where hate speech takes place, as shown in Tian [261]. In the reviewed hate speech literature, there are 13 datasets collected from Twitter with different sizes, collection methods, and annotation methods. The tweets in the datasets are collected using the public Twitter API[1]. Some studies used hateful hashtags and profane words, like feminazi, immigrant, nigger, Islam, terrorism, and bully to filter the tweets, as shown in Bosque and Villareal [30], Chatzakou et al. [43], Raisi and Huang [210], Raisi and Huang [211], Waseem and Hovy [282], Zhang et al. [297], Zhao et al. [300]. Other studies used publicly available datasets, like for example, as shown in Nahar et al. [168] and, as shown in Xu et al. [290], who used the 2011 TREC Microblog Track corpus[2]. In 2019, the multilingual detection of hate speech against immigrants and women in Twitter (hateEval) dataset is released in two languages, English and Spanish. The dataset was used in SemEval 2019 Task 5, as shown in Basile et al. [19].

- **Instagram**[3] is a social media platform where people share photos and videos, and others can comment on them. This opens the door for hate speech, as people can either post offensive pictures or write insulting comments. In the reviewed literature, I find four Instagram datasets, as shown in Hosseinmardi et al. [105], Kao et al. [119], Raisi and Huang [210], Raisi and Huang [211]. The data is crawled from Instagram by first filtering images and videos using hate speech, harassment, and abusive words. Then, collecting those media sessions where offensive comments are made.

- **FormSpring.ME**[4] is a social media platform that allows its users to ask other users anything and start a conversation between them. Sometimes the questions or the answers are abusive. In the reviewed literature, there are three FormSpring.ME datasets, as shown in Agrawal and Awekar [5], Reynolds et al. [214], Rosa et al. [219]. Two datasets are made available as part of the Kaggle competition website[5] and are used by, as shown in Agrawal and Awekar [5], Rosa et al. [219, 219]. The third dataset is

[1]https://developer.twitter.com/en/docs
[2]https://trec.nist.gov/data/microblog2011.html
[3]https://www.instagram.com/
[4]https://domain.me/formspring-me/
[5]2https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullyingdetection

Comment	Source	Label
my boyfriend showed this song to me I love it Me tooo Is she having a seizure. Omg u have a corgi I am training for the Olympics and I am Russian You want some rapes... LOL	YouTube	hate speech (Aggression)
RT @BeepsS: @senna1 @BeepsS: I'm not sexist but f**k if you're a woman and you can't Cook get your s**t together.	Twitter	hate speech (Sexist)
@freemedialive F**k #Islam. Mohammed was a pedophile, murderer, bigot, sexist, rapist, slave trader, caravan robber, and liar. : racism	Twitter	hate speech (Racist)
You f**k your dad.	Kaggle-insults	hate speech (Insult)
f**k off you little a**hole. If you want to talk to me as a human start showing some fear the way humans act around other humans, because if you continue your beligerant campaign, i will cross another boundary and begin off-site recruitmehnt. I can escalate till I am rhetorically nuclear with the whole goddamed mob of you if that is where you think you will find what you want. You had better start expressing some interest in the concerns presented to you or your credibility as either a document or a community will be about that of a pile of shit.	Wikipedia talk pages	hate speech (Aggression)
You are not worth the effort. You are arguing like Viriditas and Pename now. 24 hours really means 24 and a half hours. Four reverts in more than 24 hours is violating the "spirit" of the three revert in 24 hour rule - as interpreted by you. "So tough." Who needs rules? Just make it up as you go along and do what you want - call it "discretion". You violate the rules by blocking me then claim I violated the "spirit of the rule." Your violation is "debatable" like the Occupied Territories are "disputed." You are just abusing your authority to push your petty authoritarian agenda that obviously reflects your personal insecurities. You think you can threaten and bully me. "And I guess you won't be reverting so quickly in future, will you now¿' What a weasel. Please go ahead and contribute your petty complaints to ban me so I don't bother wasting my time on a project populated by immature arrogant twerps, fascist Zionist bigots, Islamophobe hate-mongers, bunch of lamea** bigots and losers. Why waste my time?	Wikipedia talk pages	hate speech (Attack)

Table 2.5 Examples of hate speech comments on social media

Source	Dataset	Total samples	Positive samples	Negative samples	Focus	Available	Papers
Twitter	Twitter-DS 1	12,705	391 (3%)	12314 (97%)	Hate speech		[30]
	Twitter-DS 2	16,014	5,355 (33%)	11,559 (72%)	Hate speech		[282]
	Twitter-DS 3	14,742	3,370 (23%)	11423 (77%)	Racism	✓	[281]
	Twitter-DS 4	1,762	685 (38.8%)	1,078 (61.18%)	Hate speech		[300]
	Twitter-DS 5	4,865	93 (3%)	4,700 (98%)	Hate speech		[234]
	Twitter-DS 6	296,308	-	-	Hate speech		[210, 211]
	Twitter-DS 7	7,321	2,102 (28.7%)	5,219 (71.2%)	Hate speech		[57, 290]
	Twitter-DS 8	9,484	-	-	Cyber-aggression	✓	[43]
	Twitter-DS 9	16,000	5,074 (31.6%)	10,926 (68.28%)	Hate speech		[5]
	Twitter-DS 10	14,194	1,753 (12.3%)	12,441 (87.7%)	Hate speech		[107]
	Twitter-DS 11	2,435	414 (17%)	2,021 (83%)	Hate speech	✓	[297]
	Twitter-DS 12	10,041	850 (10%)	9,191 (90%)	Hate speech	✓	[127]
	SemEval-DS	12722	5313 (42%)	7410 (58%)	Hate speech	✓	[19]
ASK.FM	Ask-DS	2,863,801	-	-	Hate speech		[104, 210, 211]
MySpace	Myspace-DS	3,245	950 (29.3%)	2,295 (60.7%)	Hate speech	✓	[21, 57]
Instagram	Instagram-DS 1	2,218 MS	665 (30%)	1,553 (70%)	Hate speech		[49, 105]
	Instagram-DS 2	9,828,760 MS	2,948,628 (30%)	6,880,132 (70%)	Hate speech	✓	[105, 210, 211]
	Instagram-DS 3	13,350 MS	1,602 (12%)	11,748 (88%)	Hate speech		[119]
	Instagram-DS 4	1,656,236 MS	-	-	Hate speech		[211]
Vine	Vine-DS 1	969 MS	303 (31%)	666 (69%)	Hate speech	✓	[207]
	Vine-DS 2	959 MS	45 (5%)	914 (95%)	Hate speech	✓	[49, 209]
FormSpring	FormSpring-DS 1	13,652	792 (6%)	12,860 (94%)	Hate speech	✓	[214]
	FormSpring-DS 2	12,000	825 (7%)	11,175 (93%)	Hate speech		[5]
	FormSpring-DS 3	13,160	2,205 (17%)	10,955 (83%)	Hate speech	✓	[219, 219]
YouTube	YouTube-DS 1	50,000 MS	-	-	Hate speech		[67]
	YouTube-DS 2	3,603 (users)	432 (12%)	3,171 (88%)	Hate speech	✓	[55]
	YouTube-DS 3	7,962	-	-	Hate speech		[131]
Wikipedia Talk Pages	Wikipedia-DS	115,737	13,542 (11.7%)	102,195 (88.3%)	Personal attacks	✓	[5, 289]
Reddit	Reddit-DS	10,100	-	-	Toxicity		[9]
Yahoo	Yahoo-Finance-DS	759,402	53,516 (7%)	705,886 (93%)	Abusive language		[176]
	Yahoo-News-DS	1,390,774	228,119 (16%)	1,162,655 (84%)	Abusive language		[176]

Note: A dash (-) denotes unavailable information. The "Available" column denotes whether the dataset is available for download either online or by contacting the authors.

Table 2.6 Datasets used in the reviewed hate speech detection literature.

crawled from the FormSpring.ME website by, as shown in Reynolds et al. [214] and made available by the researchers[1].

- **Ask.FM**[2] is a social media website that is similar to FormSpring.ME, where users can ask other users questions and start a conversation. In the reviewed literature, I find two studies that used data from ASK.FM, as shown in Raisi and Huang [210], Raisi and Huang [211]. The data is crawled from the ASK.FM website. The researchers used a custom-made harassment dictionary to query data from other sources, but it is not clear if they used the same method to filter the crawled data from ASK.FM or not.

- **MySpace**[3] is a social networking website that used to be very famous in the 2000s. In the reviewed literature, two studies used data from MySpace. The dataset is collected by, as shown in Bayzick [21] and then is used by, as shown in Dani et al. [57]. The posts included in the dataset are crawled from MySpace's groups' feature and are manually labeled as normal or bullying-related.

- **YouTube** is an online video-sharing platform, which opens the door for hate speech as users can comment on the videos of other users. Keryov and Evelyn argue that when YouTube videos are controversial, the comments tend to be more racist and abusive, as shown in Keryova [120]. I find three studies that collected and used YouTube media sessions (videos + comments) to detect hate speech, as shown in Dadvar et al. [55], Dinakar et al. [67], Kumar et al. [131]. The dataset is generated by collecting media sessions on sensitive topics, like sexuality, race, culture, intelligence, and physical attributes.

- **Vine** was a short video hosting platform where users could share six-second long videos. Users can comment on those videos, and sometimes the videos shared, or the comments, are racist towards certain groups of people. In 2018, Vine was archived and set to be replaced by a successor version, but the project has been postponed indefinitely. Until 2018, researchers could crawl data as media sessions (videos + comments) from Vine. Within the reviewed literature, I find two Vine datasets that are used in, as shown in Rafiq et al. [207], Rafiq et al. [209] and, as shown in Cheng et al. [49].

- **Wikipedia Talk Pages** is a collaborative platform where Wikipedia users can discuss improvements on published articles on Wikipedia. Sometimes the comments are

[1] https://www.chatcoder.com/drupal/DataDownload
[2] https://ask.fm/
[3] https://myspace.com/

aggressive, toxic, and contain personal attacks. In the reviewed literature, I find one dataset that is collected by, as shown in Wulczyn et al. [289] and then used by, as shown in Agrawal and Awekar [5]. Each comment in the dataset is labeled by 10 annotators via the Appen (Figure-Eight) crowd-sourcing platform on whether it contains a personal attack.

- **Yahoo** is a web services provider that operates a number of different web services. I find only one study that used data from the Yahoo website, as shown in Nobata et al. [176]. They use comments posted on Yahoo Financial and Yahoo News stories for hate speech detection. All comments are moderated and annotated by Yahoo employees who are trained before the task in order to familiarise themselves with the required text judgment guidelines. In addition, the dataset is available for researchers[1].

- **Reddit** is a popular social media network that offers social news aggregation, web content rating, and online discussions. Almerekhi et al. [9] used comments posted on Reddit to detect triggers for toxicity. They focused on the ten subreddits with the highest number of subscribers. For each subreddit, they retrieved all the comments posted between January 2016 and August 2017 using Pushshift's public Reddit collection and used the Figure-Eight crowd-sourcing platform to label a subset of 10,100 randomly sampled comments from AskReddit.

In addition to the datasets in Table 2.6, Vidgen and Derczynski compiled a list of hate speech datasets, as shown in Vidgen and Derczynski [274]. That list[2] is a collection of annotated datasets for hate speech, online abuse, and offensive language. The collection contains datasets in different languages, e.g., Arabic, Croatian, Danish, English, French, German, Greek, Hindi-English, Indonesian, Italian, Polish, Portuguese, Slovene, Spanish, and Turkish. The datasets are collected from different social media platforms, such as Twitter, Reddit, Facebook, Gab, and Wikipedia, and news platforms like Fox News and AlJazira.

From this list of the datasets used in the literature, we can see that Twitter is the most used platform for studying hate speech, which leads to many speculations, including that the moderation on Twitter is not so strict, it is an abundant source of bullying and hate, or it is easier to retrieve data from because of the Twitter API. However, I believe that the hate speech detection community should to release and use datasets that are collected from less mainstream platforms, but with even less strict moderation policies like an Urban Dictionary, 4&8 Chan, etc. because recent studies have shown that these platforms are often fertile ground for hate speech, and white supremacy, as shown in Nguyen et al. [175], Papasavva

[1]https://webscope.sandbox.yahoo.com/
[2]https://hatespeechdata.com/

et al. [188]. I also notice that some platforms are now out of service, like Vine, or not any more popular, like ASK.FM, MySpace, and FormSpring. However, the data collected from these platforms is still relevant, as the offensive language is still the same, and they can be used with more recent datasets to learn more about hate speech on social media.

Data Annotation

In the reviewed literature, I find two common ways the researchers used to label the collected data: i) manual annotation by humans, and ii) filtering using specific keywords. Manual annotation by humans is an arduous and time-consuming task. Some studies employed crowdsourcing platforms to hire people without previous experience to label the data, to reduce the cost. Appen[1], formerly known as CrowdFlower, is one of the most used crowdsourcing platforms and has been used by [5, 43, 105, 207, 281, 289]. Amazon Mechanical Turk (AMT)[2] is the second most used platform in the reviewed literature, used by, as shown in Agrawal and Awekar [5], Reynolds et al. [214], Rosa et al. [219]. Other studies hired experts to do the labeling. Some of those experts are linguists, e.g., Zhang et al. [297], activist feminists, as shown in Waseem and Hovy [282], or experts in aggression in education systems, as shown in Ptaszynski et al. [202]. Other studies hired graduate students to do the labelling, as shown in Dinakar et al. [67], while in some other studies the researchers themselves did the labelling, as shown in Huang et al. [107], Kumar et al. [131], Nobata et al. [176], Rosa et al. [219].

To quantify the agreement between more than one annotator, researchers use the inter-annotators' agreement score, which can be measured using Cohen's kappa, as shown in Burla et al. [35] or Krippen-dorff's alpha, as shown in Krippendorff [128]. Crowdsourcing platforms provide their agreement scores. The higher the score, the higher the agreement between annotators on whether the annotated item refers to hate speech or not. Among the studies that used crowdsourcing platforms in the reviewed literature, the number of annotators hired to do the labeling is either three annotators, as shown in Agrawal and Awekar [5], Reynolds et al. [214], Rosa et al. [219], five annotators [43, 105, 207, 281, 289] or ten annotators, as shown in Agrawal and Awekar [5], Wulczyn et al. [289]. The inter-agreement scores, using Krippendorff's alpha or Cohen's kappa, between the annotators from the crowdsourcing platforms ranged between 0.45, as shown in Agrawal and Awekar [5], Chatzakou et al. [43], Wulczyn et al. [289], 0.5, as shown in Hosseinmardi et al. [105] and 0.79, as shown in Rafiq et al. [207, 209]. In the studies that hired experts to annotate the data, the number of hired experts ranged between one and two, given the increased cost

[1] https://appen.com/
[2] https://www.mturk.com/

compared to crowdsourcing, with agreement scores reaching a Cohen's kappa of 0.78, as shown in Dadvar et al. [55] and a Cohen's kappa of 0.82, as shown in Huang et al. [107]. This indicates that despite the increased cost, experts are generally better at annotating the data. Nevertheless, crowdsourced annotation can also provide high-quality results if the task is well-designed to minimize confusion and eliminate unreliable annotators, eventually achieving reasonable agreement scores, as shown in Rafiq et al. [207, 209].

When the filtering approach is used for labeling data, the available data is filtered using specific hate-speech-related keywords and the matched data are labeled as referring to hate speech, as shown in Kao et al. [119], Raisi and Huang [210, 210], Zhao et al. [300]. Filtering data using keywords could be unreliable, as some people may use profane words in a disguised or a friendly way, e.g., s**t, as shown in Tommasel et al. [264]. In other cases, some people use high trending hashtags, which might be insulting words, to attract people to advertisement tweets.

As a result, even with keyword filtration, it is still useful to have a human annotator involved in labeling the data. However, an additional challenge exists. As often happens with subjective topics like hate speech, it is sometimes hard to tell if a post is an act of bullying, or it is sarcastic. Consequently, more than one annotator is required, ideally an odd number, to reach a consensus in cases of disagreement.

Dataset size and balance

Table2.6 summarizes all the datasets used in the reviewed literature and includes the size of the datasets and, whenever available, the number of positive samples (posts that include a form of hate speech) and the number of negative samples (posts that do not include any form of hate speech). One of the main challenges in automated hate speech detection is the availability of hate-speech-related data. From the datasets in Table 2.6, I can see that seven datasets contain 10% or less of hate-speech-related (positive) samples, as shown in Agrawal and Awekar [5], Bosque and Villareal [30], Krasnowska-Kieras and Wróblewska [127], Nobata et al. [176], Rafiq et al. [209], Reynolds et al. [214], Singh et al. [234], while only one dataset is almost balanced, with 42% positive samples and 58% negative samples, as shown in Basile et al. [19]. Nine datasets have a percentage of positive samples between 11.7% and 29%, as shown in Dadvar et al. [55], Dani et al. [57], Huang et al. [107], Kao et al. [119], Nobata et al. [176], Rosa et al. [219], Wulczyn et al. [289], Xu et al. [290], Zhang et al. [297], while the rest of the datasets contain between 30% and 39% of positive samples, as shown in Hosseinmardi et al. [105], Rafiq et al. [207], Waseem and Hovy [282], Zhao et al. [300].

The imbalance in the datasets available in the literature may have a negative effect on using deep learning models. In the next section, I review some techniques used in the literature to address this imbalance in the datasets.

Data sampling

The imbalance of the datasets resulted in many researchers processing the datasets to ensure that the trained machine learning models learn to differentiate between hate speech cases and non-hate-speech-related cases. Some works over-sampled the positive samples either by duplicating the positive samples multiple times to balance the dataset, as shown in Agrawal and Awekar [5], Reynolds et al. [214], while other studies did the opposite by down-sampling negative samples in the dataset, as shown in Rosa et al. [219, 219], Singh et al. [234]. Some studies used search keywords on the streaming APIs to filter the incoming data and make sure to get more data with offensive content, as shown in Bosque and Villareal [30], Wulczyn et al. [289], Zhang et al. [297]. Others used snowball sampling to ensure that they achieve a better representation of positive samples in the datasets, as shown in Rafiq et al. [209], Raisi and Huang [210]. Krasnowska-Kieras et al. increased the number of positive samples by artificially generating hate-speech-related tweets, as shown in Krasnowska-Kieras and Wróblewska [127]. The rest of the studies opted to use the available imbalanced data to train their machine learning models, given that in a real-world situation, the number of hate-speech-related posts is, in general, less than the number of other posts.

Even though over-sampling or under-sampling datasets could mitigate the imbalances in the datasets, they come with their challenges. Because if not done properly, they could lead to over-fitting, as I will discuss in Section 2.5. To mitigate these challenges, data augmentation could be used to generate more positive (bullying) text and balance the datasets.

In this section, I present all the steps related to pre-processing hate speech datasets in the literature. All these steps are important to ensure that the datasets are representative and less biased, to train fairer and generalizable models. In the next section, I review the next step in the text classification pipeline, which is data pre-processing to clean the data and prepare them for training the ML model.

2.4.2 Pre-processing

Pre-processing is an important standard step for cleaning the data. In the reviewed literature, most of the works used the NLTK library[1] to tokenize, remove stop words, remove unwanted characters, correct misspelling, lemmatize and/or stem the raw data [32, 81, 170, 230, 299].

[1] `https://www.nltk.org/`

In the case of the Twitter datasets, more steps are typically applied, like replacing user mentions, URLs, and hashtags with special characters, as well as removing duplicates, as shown in Tomkins et al. [263], Xu et al. [290], Zhang et al. [295]. Some studies also used Part-of-Speech (POS) tagging as a pre-processing step, as shown in Bretschneider et al. [32], Van Hee et al. [270].

Even though these steps are almost identical in the literature, following these steps should depend on the task and the model used. For example, removing stop words is a standard step in most NLP applications, but in the case of hate speech detection, second and third nouns could be important indicators and features for hate speech, and removing them means losing important information (e.g., the word "f*ck" on its own is not necessarily used for bullying, contrary to being used with a pronoun, such as "f*ck you"). Furthermore, more recent pre-trained models, like BERT, require a change in the pre-processing steps, as stemming is not needed anymore and punctuation symbols are important for the model to perform well, as shown in, as shown in Dang et al. [56] where BERT is fine-tuned on tweets.

The next step in the pipeline after collecting, labeling, and pre-processing the data is the extraction of features that will be used for training the ML model.

2.4.3 Features

In the reviewed literature, the most common features used can be grouped into the following four categories: 1) Text-based features, 2) User and Social media network information, 3) Sentiment and Psychological features, and 4) Distributional representation (word embeddings). I also consider one additional category called "Other features" to group some less common features used in some studies. A summary of the features used by different studies in the reviewed literature is provided in Table 2.7, while an overview of each feature category is provided below:

Text-based Features

As shown in Table 2.7, text-based features are the most commonly used features in the reviewed literature. They are either used on their own or with other features. Text features capture the patterns that exist in the text, which the machine learning models can then use to learn from the data. Various types of text features have been proposed in the literature, like the Bag of Words (BOW) models, which include one-hot encoding, Term Frequency (TF), and Term Frequency–Inverse Document Frequency (TF-IDF) representations. BOW with word N-grams is the most popular text representation model used in the reviewed literature, as shown in Agrawal and Awekar [5], Dadvar et al. [55], Dani et al. [57], Dinakar et al.

[67], Huang et al. [107], Kumar et al. [131], Nahar et al. [168], Nobata et al. [176], Potha and Maragoudakis [200], Rafiq et al. [207, 209], Raisi and Huang [211], Reynolds et al. [214], Rosa et al. [219], Waseem [281], Wulczyn et al. [289], Zhang et al. [297]. Some studies used BOW with character N-grams and reported better results compared to the word N-grams BOW model, as shown in Agrawal and Awekar [5], Krasnowska-Kieras and Wróblewska [127], Nobata et al. [176], Waseem [281], Waseem and Hovy [282], Wulczyn et al. [289]. Other studies used the frequency of profane or negative words as features, as shown in Bosque and Villareal [30], Dadvar et al. [55], Dinakar et al. [67], Kumar et al. [131], Rafiq et al. [209], Reynolds et al. [214], while, as shown in Bosque and Villareal [30] used the frequency of the word "you" as a feature for detecting hate speech. Other studies used the number of words in the sentence (an online post), the number of hashtags used, the number of words in uppercase letters and the number of URLs in addition to the text, as shown in Bosque and Villareal [30], Chatzakou et al. [43], Dadvar et al. [55], Kao et al. [119], Waseem [281], Zhang et al. [297]. Furthermore, some studies applied natural language processing techniques and used Part-of-Speech (POS) tags related to the text as additional text features, as shown in Dani et al. [57], Dinakar et al. [67], Singh et al. [234], Waseem [281].

User Information

Besides using text-related information for feature selection, researchers have tried to use information related to the author of the examined text, and as a consequence, related to the person committing hate speech. This information could be the users' gender, age, or the number of their online posts, which can be found on their social media profiles. In the reviewed literature, I find that gender has been used as a feature, as shown in Waseem [281], Waseem and Hovy [282], as according to, as shown in Waseem and Hovy [282], men tend to send more racist and sexist posts on Twitter than women. Anonymity is another factor that some researchers considered, since they claimed that users who are cyberbullies tend to hide their identities. However, results indicated that it is not necessarily the case, as shown in Dadvar et al. [55], Reynolds et al. [214]. Other researchers used information about the users' online behavior, like the number of their posts, their subscriptions, uploads, and their history of used words, as shown in Dadvar et al. [55], Rafiq et al. [209], Waseem [281]. Furthermore, the users' location has also been used as a feature, as shown in Cheng et al. [49], Waseem and Hovy [282]. User features also include information related to the user's social media network, like the users' number of followers, the number of likes and views they receive, or the number of people they follow, as shown in Chatzakou et al. [43], Cheng et al. [49], Rafiq et al. [209], Singh et al. [234].

Sentiment and psychological Features

Sentiment analysis refers to the task of using natural language processing and text analysis to evaluate the sentiment conveyed by a text, by assigning a sentiment score to the examined text. Positive scores typically relate to positive sentiment, while negative scores are typically indicative of negative sentiment, as shown in Agarwal et al. [3], Liu et al. [141], Pak and Paroubek [185], Pang and Lee [186], Wilson et al. [288]. In the reviewed literature, some researchers used the sentiment score of the text as a feature for hate speech detection, as negative words are an indicator of unpleasant and potentially bullying-related text, as shown in Agrawal and Awekar [5], Bosque and Villareal [30], Chatzakou et al. [43], Dani et al. [57], Dinakar et al. [67], Kao et al. [119, 119], Rafiq et al. [207], Zhang et al. [297]. Some studies generated a sentiment score of the emotion icons (emojis) in the text and used those scores as features in training the machine learning models, as shown in Chatzakou et al. [43], Dadvar et al. [55].

In 2015, Pennebaker et al. [195] developed a tool (LIWC[1]) that can analyze a text and reveal some psychological features of the author(s). For example, given that one of the main characteristics of a bully is to have power over their victims, the tool can be used to measure someone's tendency to exercise authority from their text. In the reviewed literature, Kao et al. [119] and, as shown in Cheng et al. [49] used the results of the LIWC tool as an additional feature for hate speech detection.

Distributional representation (word embeddings)

A distributional text representation (Word embeddings) aims at representing words in a way that preserves their semantic relationships and considers the order of the words in the text, as shown in Mikolov et al. [155]. Word embeddings have been widely used recently for most text classification and information retrieval tasks, as shown in Wang et al. [279, 280]. However, there are few studies that used word embeddings in hate speech detection. Nobata et al. used the word2vec-CBOW word embedding to train their hate speech detection model, as shown in Nobata et al. [176]. Similarly, Agrawal et al. used Glove-Wikipedia to improve the task of hate speech detection, as shown in Agrawal and Awekar [5]. Doc2vec embeddings are used as features in detecting hate speech by, as shown in Raisi and Huang [211], who also adopted the idea of distributed representation of words and applied it to the user's online social network and developed node2vec as a feature for detecting hate speech. Koufakou et al. used FastText word embeddings that are retro-fitted for the task of hate speech detection, as shown in Koufakou et al. [123], while other studies developed specialized word embeddings

[1] http://liwc.wpengine.com

for the task of hate speech detection, as shown in Krasnowska-Kieras and Wróblewska [127], Raisi and Huang [211], Rosa et al. [219], Zhao et al. [300].

In addition to the classic pre-trained models on Wikipedia and Google News, there have been new models pre-trained on Twitter, like glove-Twitter[1], Urban dictionary word embeddings pre-trained on words and definitions from the Urban Dictionary website, as shown in Wilson et al. [287], and Chan word embeddings pre-trained on text from the 4 & 8 Chan websites, as shown in Voué et al. [277]. Despite these embeddings been trained with text that resembles more the way users communicate in social media platforms compared to the news and Wikipedia articles, the use of these embeddings has not yet been explored for the detection of hate speech.

Other Features

Apart from the aforementioned features that are used in multiple studies, the following less common features are also used in the reviewed literature. Davdar et al. Dadvar et al. [55] hired experts to rate the importance of the extracted text features from the text and used this rating as an additional feature. They also used the information resulting from a multi-criteria decision support system (MCES), as shown in Zahir [292] as another feature to detect hate speech. Potha and Maragoudakis, as shown in Potha and Maragoudakis [200] used time series modeling and Singular Value Decomposition (SVD) to extract features for hate speech detection, and, as shown in Zhao et al. [300] used Latent Semantic Analysis (LSA), which is a topic modeling method, to extract different topics in the unlabeled text as features. Topic models like K-means, LDA, and LSI are also used in, as shown in Sutinen [252] to group the text into clusters and use this information as features. Cheng et al. [49] used the metadata of images posted on social media platforms as features along with the time of the post. Few studies used Multi-modal hate speech detection, where the model is trained on both images and text to detect hate speech, as shown in Kansara and Shekokar [118], NaliniPriya and Asswini [169]. nandhini2015cyberbullying used the Levenshtein distance to measure the difference between two words as a feature to detect profane words in disguised form, e.g. f***, while, as shown in Nazar et al. [173] used the conditional feature probability to measure the importance of the features. yao2018cyberbullying used a novel algorithm to reduce the number of features (text and user information) used in the classification task. They achieved an F1-score (will be discussed in Section 2.4.5) of 0.76 with an average of 6.6 features, compared to the baseline which achieved a 0.58 F1-score with 13 features.

[1] https://nlp.stanford.edu/projects/glove/

Feature selection

Singh et al. Singh et al. [234] proposed a method for combining text features, user features, and social media network features in a way that enhances the model's performance, by first determining the agreement score between different types of features and then determining the confidence score of certain feature types by calculating their accuracy in predicting the data label (as hate speech or not) from previous predictions. This way, the model can determine which features are more important for each data instance and consequently make better predictions of the final data label. Using this approach, they achieved better results than other studies that combine features mindlessly, reporting an F1-score of 0.64.

Raisi and Huang used multi-view learning to maximize the agreement across different features types (text and social network) of unlabeled data, as shown in Raisi and Huang [211]. They used an ensemble of two learners: one to examine the language content of a post, and another to consider the network structure of the post sender. They achieved a precision of 0.6, which is a relatively high score given that they do not use labeled data. Similarly, Cheng et al. [49] built their model using multi-modality learning to use different pieces of information provided in the social media post, like images, videos, user profile, time, and location, assuming that the different pieces of information (modalities) could be complementary and achieved an F1-score of 0.98.

In the same direction of enhancing the learning of the different types of features, Dani et al. [57] proposed a framework called Sentiment Informed hate speech Detection (SICD), which is a model that maximizes the use of sentiment information available in the post. They used the distribution of sentiment scores in the data to differentiate between the sentiment of hate speech posts and normal posts, achieving an AUC score of 0.80 and an F1-score of 0.68.

In this section, I review the literature on the different features used in the task of hate speech detection. The most common features are Text-based and User information features. On the other hand, word embeddings are among the least used features, even though they have been proven to perform well on several NLP tasks. The community of hate speech detection needs to explore more the use of word embeddings, especially with the release of the new contextual word embeddings like BERT and ELMO. I provide in-depth analysis and suggestions regarding feature selection for hate speech detection in Section 2.5. In the next section, I review the different ML models that have been used in the literature for the task of hate speech detection.

Paper	Dataset	Text Features	User Information Features	Sentiment Features	Word Embeddings	Other Features	Accuracy	Precision	Recall	F1	AUC
[67]	YouTube-DS 1	✓		✓			0.80				
[214]	FormSpring-DS 1	✓	✓						0.87		
[55]	YouTube-DS 2	✓	✓	✓		✓	0.76				
[200]	Perverted Justice	✓				✓	0.88				
[30]	Twitter-DS 1	✓		✓			0.48				
[207]	Vine-DS 1	✓		✓			0.76				
[282]	Twitter-DS 2	✓	✓					0.72	0.77	0.73	
[176]	Yahoo-Finance-DS	✓			✓					0.81	
[281]	Twitter-DS 3	✓	✓		✓			0.92	0.92	0.91	
[281]	Twitter-DS 3	✓			✓	✓		0.76	0.79	0.78	
[234]	witter-DS 5	✓	✓							0.64	
[210]	Twitter-DS 6	✓				✓					0.83
[289]	Wikipedia-DS	✓								0.75	0.83
[57]	Myspace-D	✓								0.68	0.80
[43]	Twitter-DS 8	✓	✓	✓					0.89	0.91	0.90
[5]	Twitter-DS 9	✓		✓		✓			0.92	0.91	0.91
[107]	Twitter-DS 10	✓									0.89
[297]	Twitter-DS 11	✓		✓		✓					0.92
[209]	Vine-DS 2	✓	✓								0.68
[219]	FormSpring-DS 3	✓									0.81
[211]	Instagram-DS 4	✓				✓			0.6		
[127]	Twitter-DS 12	✓			✓	✓					0.83
[119]	Instagram-DS 3	✓		✓				0.40	0.35	0.37	
[49]	Vine-DS 2		✓	✓	✓					0.98	
[131]	YouTube-DS 3		✓	✓		✓		0.83	0.82	0.83	
[219]	FormSpring-DS 3	✓						0.85	0.86	0.84	

Table 2.7 Features used for automated hate speech detection in the reviewed literature and highest performance reported by each work.

2.4.4 Machine learning models

In this section, I discuss the different ML models used for the task of hate speech detection in the reviewed literature.

Rules-based Learning

Some studies in the reviewed literature used rules-based models besides machine learning models to provide the criteria based on which the model classifies the data. They are especially used in the early studies, with less available training datasets than required to train machine learning models, as shown in Bosque and Villareal [30], Dinakar et al. [67], Reynolds et al. [214].

Conventional Machine Learning

Conventional machine learning models are the most widely used in the reviewed literature. I find 61 (57.5%) studies that used conventional machine learning models. Most of them used supervised learning models. Among these supervised models are the models that are famous for performing well in text classification tasks, like Support Vector Machines

(SVM), as shown in Agrawal and Awekar [5], Cheng et al. [49], Dinakar et al. [67], Potha and Maragoudakis [200], Rafiq et al. [207], Reynolds et al. [214], Zhao et al. [300] and Naive Bayes (NB), as shown in Agrawal and Awekar [5], Dadvar et al. [55], Dinakar et al. [67], Rafiq et al. [207]. Other well-known models used are: Logistic Regression (LR), as shown in Rafiq et al. [209], Waseem and Hovy [282], Wulczyn et al. [289], Decision Trees (DT), as shown in Chatzakou et al. [43], Dadvar et al. [55], Dinakar et al. [67], Rafiq et al. [207], Reynolds et al. [214], k-Nearest Neighbors (KNN), as shown in Kumar et al. [131], Reynolds et al. [214], and Random Forests (RF), as shown in Agrawal and Awekar [5], Chatzakou et al. [43], Cheng et al. [49], Kao et al. [119], Kumar et al. [131], Rafiq et al. [207]. Furthermore, despite the shortage of labeled datasets, there have been only a few trials that attempted to use weakly supervised, as shown in Raisi and Huang [210], Raisi and Huang [211] or unsupervised machine learning models, as shown in Rosa et al. [219].

Deep Learning

During the past two decades, deep learning models have been increasingly used in different variations and for different applications of machine learning. However, in the reviewed literature, I find that deep learning models have been used for hate speech detection much later. This could be because deep learning models need large numbers of data points for training and the available datasets for hate speech used to be small in numbers and in size, something that started to increase only recently. Zhao and Mao, as shown in Zhao and Mao [299] used Semantic Enhanced marginalised Denoising Auto-Encoder (smSDA) for hate speech detection. Agrawal and Awekar [5] used Transfer Learning with LSTM to detect hate speech across multiple social media platforms. They achieved a Precision score of 0.92, a recall score of 0.91, and an F1-score of 0.91. CNN's has also been used to improve the detection of hate speech, as shown in Agrawal and Awekar [5], Al-Ajlan and Ykhlef [6], Banerjee et al. [17], Huang et al. [107], Rosa et al. [219], Zhang et al. [295, 297]. Agrawal and Awekar [5] and, as shown in Raisi and Huang [211] used Long Short-Term Memory (LSTM) models, which are a variation of Recurrent Neural Network (RNN), as shown in Mikolov et al. [154] models, to detect hate speech. Zhang et al. [297] combined CNN layers with Gated Recurrent Network (GRN) layers to create a model for hate speech detection. Simpler deep learning models have also been explored in the literature. Some studies also used a simple neural network like the multi-layer perceptron (MLP), as shown in Krasnowska-Kieras and Wróblewska [127], Wulczyn et al. [289].

Unconventional Models

Most of the reviewed papers used conventional machine learning models or deep learning models, with a novel contribution in providing labeled datasets or in feature engineering. However, there are less common machine learning approaches like unsupervised learning, which have been used in other fields, e.g., for detecting spammer groups, as shown in Ji et al. [111] and for rumor detection, as shown in Alzanin and Azmi [10], Chen et al. [46], or semi-supervised machine learning models, as shown in Ashfaq et al. [13], Gu [93]. These unconventional methods have also been used for hate speech detection. Rafiq et al. [208] and, as shown in Rafiq et al. [209] proposed a multi-stage hate speech detection model that improves the classification time by 223 times over the baseline and the time needed to raise an alert is improved seven times over the baseline, achieving a precision of 0.71 and a recall of 0.66. Dani et al. [57] first used a distant supervised based sentiment machine learning model to measure the sentiment score distribution of the dataset, and then they incorporated that score to detect hate speech. They reported an AUC score of 0.80 and an F1-score of 0.68. Rosa et al. [219] used Fuzzy Finger Prints to identify the unique fingerprints of the positive hate speech examples in the training dataset. They slightly outperformed the baselines for unbalanced datasets and achieved an F1-score of 0.77. Cheng et al. [47] used hierarchical attention networks to mirror the structure of social media sessions and use attention mechanisms that capture the relationship between the words in a comment within a certain context, achieving an F1-score of 0.78 and an AUC score of 0.851.

In this section, I review the different models used in the literature on hate speech detection. I can see that the majority of the studies, reviewed here, opted for conventional ML models over deep learning models, which could be due to the small sizes of the datasets and the high imbalance ratio of positive (bullying) and negative (not-bullying) data. The more datasets being released for the task of hate speech detection, the more deep learning models will be easier to use. I also notice that the literature is missing out on new advances in pre-trained language models like BERT, GPT2, and GPT3. In the next section, I review the different evaluation methods used in the literature of hate speech and their validity.

2.4.5 Evaluation metrics

Given the use of machine learning for hate speech detection in the reviewed literature, the performance of the reviewed methods is evaluated using typical evaluation metrics that are common across the machine learning literature. The majority of the examined works used the following evaluation metrics: accuracy, F1-score, precision, recall (also known as sensitivity or true positive rate), as well as Receiver Operating Characteristic-Area Under

Paper	Dataset (size)	Features	Model	AUC	F1
[44]	Kaggle-insults (4000)	Text Features Word Embeddings	Support Vector Machine (SVM)	0.85	-
[85]	Twitter (1900)	Text Features User Features Other Features	Sequential Minimal Optimisation (SMO)	0.96	-
[210]	Twitter (296,308)	Text Features	Participant Vocabulary Consistency (PVC)	0.83	-
[43]	Twitter (9,484)	Text Features User Features Network Features Sentiment Features Word Embeddings	Random Forest (RF)	0.90	-
[5]	Twitter (16,000)	Text Features Sentiment Features Word Embeddings	Long Short Term Memory (LSTM)	0.93	-
[186]	MySpace (600)	Text Features Other Features	Naive Bayes (NB)	-	0.89
[171]	MySpace (-)	Text Features	NB + Use Fuzzy rule based + Genatic algorithm	-	0.98
[7]	Twitter (10007)	Text Features User Features Network Features Psycological Features Other Features	SVM	-	0.94
[295]	Formspringme (13000)	Text Features Sentiment Features Word Embeddings Other Features	Convolution Neural Network (CNN)	0.98	0.98
[107]	Visr child safety data (-)	Text Features Psychological Features	CNN	0.89	-
[297]	Twitter (2,435)	Text Features Sentiment Features	CNN	0.92	-
[219]	FormSpring (13160)	Word Embeddings	CNN	-	0.84
[127]	Twitter (10,041)	Text Features Word Embeddings	NN	0.83	-
[49]	Instagram (155,267)	Psychological Features User Features Network Features Image meta data Time	RF	-	0.98
[289]	Wikipedia Talk Pages (115,737)	Text Features	Logistic Regression (LR) Multi Layer Perception (MLP)	0.96*	-

Note: * refers to ROC-AUC

Table 2.8 The best F1 and AUC scores achieved in the reviewed literature. The evaluation scores presented here are for providing an idea of the scores being reported in the literature but are not meant for comparative reasons as these studies used different datasets

the Curve (ROC-AUC) scores. These metrics are computed based on the four outcomes that summarize a binary classification task's results, i.e., i) True Positive (TP), the number of correctly classified positive samples, ii) True Negative (TN), the number of correctly classified negative samples, iii) False Positive (FP), the number of samples miss-classified as positive, and iv) False Negative (FN), the number of samples miss-classified as negative. In addition, a few works reported the error score or the Mean Squared Error (MSE) score, as shown in Bosque and Villareal [30], Potha and Maragoudakis [200].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Despite being one of the most common metrics for classification, accuracy (Eq. 2.1) is not the preferred evaluation metric when working with imbalanced datasets, as shown in Rogers and Girolami [217], since it may lead to overestimated scores as a result of a high number of samples belonging to a certain class. In the reviewed hate speech detection literature, I find that Dinakar et al. [67], Rafiq et al. [207] and Kumar et al. [131] used the accuracy metric to report the results of their models, while studies that used deep learning did not report accuracy scores.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

Some of the studies that reported precision (Eq. 2.2), also reported recall (Eq. 2.3) and F1-score (Eq. 2.4), as shown in Agrawal and Awekar [5], Kumar et al. [131], Rosa et al. [219], Waseem and Hovy [282], Zhao et al. [300]. Other works reported only the F1-score, as shown in Cheng et al. [49], Dani et al. [57], Kao et al. [119], Nahar et al. [168], Rafiq et al. [209], Rosa et al. [219], Singh et al. [234], Zhang et al. [297], or either recall only, as shown in Reynolds et al. [214] or precision and recall, as shown in Ptaszynski et al. [202].

AUC is generally preferred in binary classification tasks, but despite hate speech detection being a binary classification task, I find few studies in the reviewed literature that reported AUC scores, either on their own or along the F1-score, as shown in Dadvar et al. [55], Dani et al. [57], Huang et al. [107], Krasnowska-Kieras and Wróblewska [127], Raisi and Huang [210], Wulczyn et al. [289]. A summary of the reviewed studies that reported an AUC score

or an F1-score higher than 0.80 is provided in Table 2.8, including the achieved scores, the dataset, the features, and the machine learning models used.

In this section, I review the different evaluation metrics used in the the literature of hate speech detection. I show that using accuracy is not advisable for tasks where there is a high imbalance in the dataset. I also recommend the use of the F1-score as a good measure of the models' ability to find a balance between precision and recall.

In the next section, I provide an analysis of the limitations in the literature of hate speech detection and provide some recommendations to overcome these limitations.

2.5 Limitations of the reviewed literature

Examining the reviewed literature, it is evident that there are some limitations and challenges in the field of hate speech detection in terms of the datasets, features, machine learning models, and evaluation approaches used.

2.5.1 Dataset-related challenges

Some of the challenges that make the task of hate speech detection harder are related to the hate speech datasets available in the literature and are mostly related to the definition of hate speech, to data annotation, class imbalance, underlying biases, and language. In this section, I discuss these challenges.

Definition

The lack of a clear distinction in the definition between hate speech and related concepts, like hate speech, affects the generalizability of the state-of-the-art models proposed in the literature. It also affects the choice of features that can be used to enhance the models' performance in detecting hate speech or hate speech. For example, Fortuna et al. fortuna2018survey suggest that there are two types of features, general textual-based features and specific hate speech-based features. Some of these features intersect with hate speech detection like *Othering Language* and *Perpetrator Characteristics (e.g., gender and geographic localization)*, while others are specific to the task of hate speech detection, like *Declaration of superiority of the group, Focus on particular stereotypes, and Intersectionism of oppression.* The lack of a clear definition of the detection task makes it harder to select the most suitable features and models from the literature.

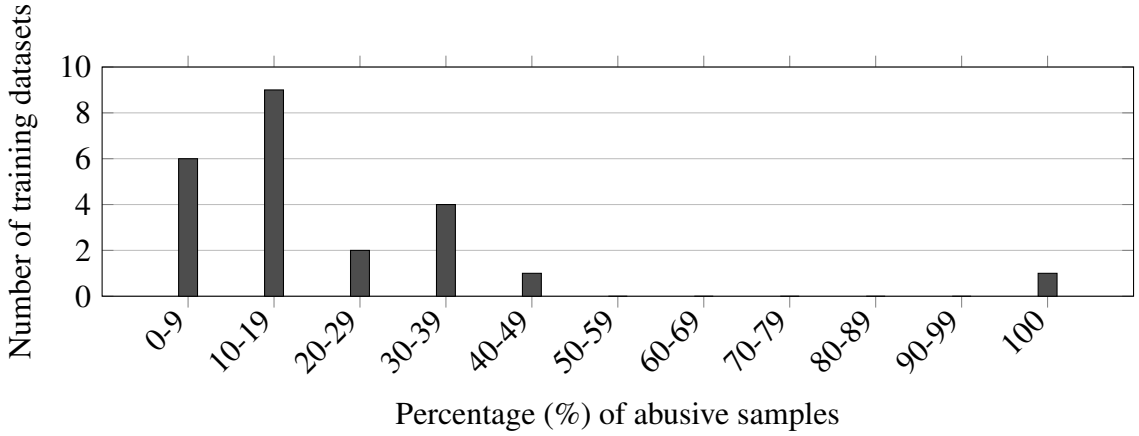


Fig. 2.3 Histogram of the percentage of abusive samples in the reviewed datasets in Table 2.6.

Annotations

I find that the studies that used crowdsourcing platforms to annotate the data reported low inter-agreement scores among the annotators. This could be due to a lack of clear instructions given to the annotators or due to the demographic of the annotators, which may lead to unknown biases, as shown in Arango et al. [12]. These biases and low agreement scores may cause over-fitting in the models reported in the literature, which in turn affects their generalizability. Related information about the annotators' demographics is not shared or described in the reviewed papers. To address this issue, I recommend that future studies share this information along with the data description when a new dataset is released.

Class Imbalance

The statistics presented in Fig. 2.3 show a clear pattern of imbalance between the number of positive (abusive) data samples and the number of negative (normal) data samples in the datasets used in the reviewed literature. This imbalance imposes some limitations on the use of deep learning models. To overcome this problem, some studies over-sample the positive samples in the dataset, which, if done before the train-test split, becomes problematic and causes model over-fitting, as demonstrated by Arango et al. [12].

User distribution bias

There is, potentially, a user distribution bias in the datasets used in the literature. For example, one of the most used datasets in the literature of hate speech and hate speech detection is the tweets dataset collected by Wassem et al. , as shown in Waseem and Hovy [282]. The dataset contains 14K tweets annotated as "racist", "sexist" or "none". The number of hateful

tweets (sexist and racist) is 4,839 and the number of non-hateful tweets is 10,110. The data on the users who generated these tweets is analyzed by Arango et al. [12], who found that all the data is generated by 1,590 users, with 491 users having generated all the sexist tweets and only 8 users have generated all the racist tweets. Among the "sexist" tweets, 40% are generated by a single user and among the "racist" tweets, 90% are generated by a single user. Furthermore, they argued that the models trained on Wassem et al.'s dataset are prone to over-fitting due to the user distribution.

Language

Despite languages other than English having been included in the datasets found in the literature, these "language" datasets are limited in sources to almost only Twitter and Facebook. For example, Mubarak et al. [164] that contains 12,698 tweets. That data is then used to create a shared task to detect hate speech in Arabic [163]. Another example is the dataset collected by Vásquez et al. [272] from twitter in Mexican Spanish and contains 11,000 tweets. There is a clear lack of "language" datasets that cover other social media platforms. Furthermore, most of these "language" datasets contain hate speech, and very few contain hate speech and its subtypes. As a consequence, this limits the research on hate speech detection in languages other than English. There is a need for more hate speech datasets in other languages to advance the research and improve the detection of hate speech in these languages.

2.5.2 Features-related challenges

The identified challenges in relation to the features used for hate speech detection are related to the lack of use of visual features, the word embeddings used for text representation, and the availability of user and network information.

Visual features

Table 2.7 summarizes the most common features used in the literature to detect hate speech. From this table, it is evident that the use of visual features for hate speech detection is rare, as shown in Hosseinmardi et al. [105], Singh et al. [235], Soni and Singh [239]. As recent studies have indicated that teenagers make extensive use of visual content on platforms like Instagram and Snapchat for their communication, as shown in Pater et al. [191], Singh et al. [236], it is important to develop models that can detect hate speech from visual media, to provide a form of protection to the receivers of such visual content.

Text representation

Another limitation I find in the literature is the use of relevant word embeddings to the task of hate speech detection. As discussed earlier, the main word embeddings used in the literature are Word2Vec, Glove, or Doc2Vec. However, more recent word embeddings have been proposed that may be more relevant to the task, such as sentiment-specific word embedding (SSWE), as shown in Tang et al. [257] and Urban Dictionary word embedding, as shown in Wilson et al. [287]. Aragwal and Awekar experimented with different deep learning models trained with different word embeddings like Glove and SSWE and found that the performance of the models trained with Glove and SSWE is very close, as shown in Agrawal and Awekar [5]. However, they did not conduct any intrinsic analysis to compare the semantic relatedness of SSWE and Glove to hate speech datasets. Similarly, contextual word embeddings like ELMO, GPT, and BERT, as shown in Devlin et al. [65], Peters et al. [198], Radford et al. [205] have not been explored enough in the literature. I recommend using the new advances in NLP to improve the detection of cyberbullying.

User and network information

Although some studies in the reviewed literature used user and network information as features to detect hate speech, few studies share this user information, which is limiting to the development of the field. This may be partially attributed to the general data protection regulations. However, for an important task such as hate speech detection, it would be more beneficial to share this information in an anonymized form than not providing it at all.

2.5.3 Models-related challenges

After reviewing the literature on the machine learning models used to detect hate speech and their training process, I identify challenges related to the generalizability of the models and the lack of use of new advances in NLP, like attention-based models and transfer learning.

Model generalizability

The first challenge is the validity of the results reported in the literature, as Arango et al. showed in their study on the generalizability of prior work on the detection of hate speech and hate speech, as shown in Arango et al. [12]. They showed through a series of experiments that the models that are used as state-of-the-art in the literature of hate speech detection failed to generalize to new datasets, which means that the high scores reported in the original papers are due to over-fitting. They explain that the over-fitting occurs due to some mistakes in the

training process: 1) Extracting the features from the whole dataset (training and test sets) for training instead of extracting the features only from the training set; 2) Oversampling the positive (abusive) content to balance the dataset before the train-test split; 3) Bias resulting from the uneven distribution of the users who generate the abusive content within the dataset. Their findings suggest that I should look at the results reported in the literature with a critical view and carefully assess the reported training processes. I also recommend replicating the results of the models reported in the literature before using them.

Contextual language models

The second challenge is that although attention-based mechanisms and pre-trained models like ELMO, GPT, and BERT, as shown in Devlin et al. [65], Peters et al. [198], Radford et al. [205] have been around for quite some time now, there are few studies that used these models to detect hate speech or hate speech, as shown in MacAvaney et al. [144], Paul and Saha [192], Yadav et al. [291]. Pre-trained models like BERT have established a new state of the art in many NLP tasks, requiring only small datasets to fine-tune the model on the downstream tasks, as shown in Sun et al. [245].

Transfer Learning

Transfer learning is a great technique to mitigate the issue of small and imbalanced datasets, which, as discussed earlier, is a problem with the task of hate speech detection. It can also be beneficial in training a model that can detect different types of hate speech regardless of the data source. However, transfer learning has not been widely explored in the community of hate speech detection, except for a few studies, as shown in Mossie [161], Mozafari et al. [162], Waseem et al. [283].

2.5.4 Evaluation-related challenges

Over-fitting

As mentioned earlier, some studies report high F1-scores like, e.g., 0.934 and 0.961, as shown in Agrawal and Awekar [5], Badjatiya et al. [15]. However, Arango et al. [12] showed that these high F1-scores are due to over-fitting, as discussed in Section 2.5.3. To address this issue, I recommend testing any model's generalizability and reporting the performance on an unseen dataset, besides reporting the performance results on the test set. For example, the SemEval 2019, as shown in Basile et al. [19] dataset could be used for that reason if the task

is hate speech detection. However, I acknowledge that the lack of hate speech datasets can be an obstacle to achieving that.

Metrics

I find some studies in the reviewed literature that reported classification accuracy for assessing performance, which is not reliable when working with unbalanced databases, such as the ones typically available in the hate speech and hate speech literature. Considering the very high proportion of negative (non-abusive) samples in the available datasets, the high accuracy values are biased towards the high number of true-negatives in the test set. For NLP tasks, it is best to report the F1-score to get a realistic evaluation of a model's performance, as shown in Joachims [113], Rogers and Girolami [217].

On the other hand, when researchers over-smaple abusive content in the training datasets to overcome the mentioned limitation of class imbalance, it makes the most commonly used evaluation metrics, e.g., F1 scores, unsuitable, as shown in Calabrese et al. [36]. To mitigate this problem, Calabrese et al. [36] proposes an evaluation system that incorporates adversarial attacks against abuse (AAA).

2.5.5 Bias and fairness challenges

Another important challenge in the current research on hate speech detection is the unfairness of these models, especially towards marginalised groups. This research direction has not been well investigated, even though there is evidence that hate speech detection models discriminate against African -American English, as shown in Sap et al. [224] and the LGBTQ community, as shown in Mchangama et al. [150]. To avoid these unfair associations that result from spurious correlations, Calabrese et al. [37] propose a framework to automate and enforce the required moderation policy, instead training a machine learning models to understand hate speech. However, with the wide use of machine learning models to detect hate speech, it is important to understand how the bias in NLP models impacts hate speech detection using supervised machine learning models.

Especially, with the new proposed methods in the literature to measure bias, as shown in Caliskan et al. [38], Dev and Phillips [64], Guo and Caliskan [94], May et al. [149], Nadeem et al. [167], Nangia et al. [172], it is crucial to understand how these biases impact hate speech detection models in terms of performance, fairness and developing new biases against marginalised groups. Most of the relevant literature, though, focuses on the impact of bias in NLP models on the fairness of hate speech detection models, as shown in Cao et al.

[39], Dixon et al. [69], Goldfarb-Tarrant et al. [89], Kaneko et al. [117], Steed et al. [241], but the performance and the formation of new types of biases have been left out.

This thesis aims to address this research challenge in the next chapter by investigating the impact of bias in NLP models on the performance, fairness, and the formation of new biases in hate speech detection models.

2.6 Conclusion

In this chapter, I presented my first contribution as a systematic literature review on automated hate speech detection. The motivation behind this area of research is to help prevent hate speech and its negative consequences, which can include depression, low self-esteem, and even committing suicide. I organized the reviewed literature around the steps of the text classification pipeline employed by each reviewed work, due to the lack of a similar systematic study in the literature. In the reviewed literature, I identified some challenges and limitations of the available work on hate speech detection, some of which are related to the hate speech datasets used in the various works. In particular, challenges with defining hate speech, the annotation of datasets, data imbalance, data bias, and the limited availability of multilingual datasets. I also noticed that the literature is not up-to-date with using more recent slang-based word embeddings like the urban dictionary word embeddings, with using more recent models, with using contextual language models like BERT, and with using transfer learning. Another limitation relates to the use of classification accuracy as a performance evaluation metric, which can be deceiving when there is an imbalance in the datasets. Finally, one of the main limitations of the research on hate speech detection is the impact of bias in NLP models on hate speech detection models in terms of performance and fairness. Addressing this limitation is the main focus of this thesis. Hence, in the next chapter, I present my second research contribution and review the literature on the bias and fairness in NLP models and their limitations before I investigate how the bias in NLP impacts the task of hate speech detection in the rest of the thesis.

Chapter 3

Survey: Bias and Fairness in NLP

3.1 Introduction

In *Race After Technology*, Benjamin [24] coins the term "The New Jim Code", which she describes as :

> *The employment of new technologies that reflect and reproduce existing inequities, but that are promoted and perceived as more objective or progressive than discriminatory systems of a previous era.*

While the Jim Code is a spin on, "Jim Crow", a derogatory epithet for African-Americans, the same concept can be generalized to the bias and unfairness in artificial intelligence (AI) systems against all marginalised groups. Hence, it is crucial to study bias and fairness in machine learning and natural language processing models to understand how existing social biases and stereotyping are being encoded in the data used to train them, as well as to compare (1) the fairness of the decisions made by NLP models due to biases in the datasets, with (2) biased choices made by the developers of those models as a result of unintended bias or to maximize profit. Studying bias and unfairness in NLP models is one way to pierce a hole in the black box and shed a little light on the limitations of widely used models. However, it is not possible to understand the roots of algorithmic bias, without incorporating relevant studies from social sciences, critical race theory, gender studies, LGBTQ studies, and digital humanities studies, as recommended by , as shown in Benjamin [24].

In this chapter, I present my second research contribution and study the different sources of bias in NLP models from two perspectives: (1) the *NLP pipeline perspective* where I review the sources of bias in NLP models from the NLP literature; and (2) the *Jim Code perspective* where I review the sources of bias from the literature on critical race theory,

gender, LGBTQ, and digital studies. Then, I review the NLP literature for the different proposed methods to measure bias and fairness in NLP models and their limitations.

3.2 Background: History of discrimination

In Western societies, the biases, and inequalities towards marginalised groups based on ethnicity, sex, class, religion, sexual orientation, age, or disability that I see today are direct results of centuries of racism, sexism, and homophobia, as has been discussed by different scholars.

In *The Myth of Race: The Troubling Persistence of an Unscientific Idea*, Sussman [250] reviews the history of 500 years of **racism** in Western Europe to answer the question of why the invalid concept of race still prevails. He argues that multiple scholars developed the ideology of race from historical events and movements ranging from the Spanish Inquisition to Social Darwinism, Eugenics, and modern IQ tests, starting as early as the fifteenth century, when the Catholic Church in Spain persecuted the Jewish population for "impurity of blood", as shown in Sussman [250].

He goes on to explain that some Enlightenment scholars like David Hume and Immanuel Kant believed that, based on skin color, there is more than one race of humans and that White men are the most civilized people, as shown in Sussman [251]. In the nineteenth century, drawing from evolution theory, social Darwinists like Herbert Spencer argued that helping the poor and the weak was an interference with natural selection, coining the term "survival of the fittest". This led to sterilization and ultimately the extermination camps of the eugenics movement , as shown in Sussman [251].

Moving to the 1970s, Sussman [248] shows that Arthur Jensen, a professor of Educational Psychology at the University of California, argued that Black people are intellectually inferior to white people. This argument was reasserted in the 1990s with the publication of Richard Herrnstein and Charles Murray's *The Bell Curve*.

, as shown in Sussman [249] goes on to show that in the 2000s, racism took on a disguise of "Culturism", coined by the anthropologist Franz-Boas to explain the difference in human behavior and social organizations. Culturism paved the way for the modern-day anti-immigration agenda since immigrants, like Arabs or Muslims, are not claimed to be genetically inferior to Europeans but to have a cultural burden that prevents them from integrating into Europe.

Homophobia is intertwined with racism, as argued by , as shown in Morris [160] in their research on the history of the LGBTQ community social movement. Morris explains that homosexuality and transgender identity were accepted in many ancient societies like

ancient Greek, Native American, North African, and the Pacific Islands. These accepting cultures oppose the Western culture of heterosexuality and binary genders, which regarded homosexuality and transgender as foreign, savage, and evidence of inferior races. When Europeans started colonization campaigns, they imposed their moral codes and persecuted LGBTQ communities. The first known case of punishing homosexuality by death was in North America in 1566. Later, in the era of sexology studies in 1882 and 1897, European doctors and scientists labelled homosexuality as degenerate and abnormal, and as recently as the 1980s and 1990s, AIDS was believed to be god's punishment for gay people.

As argued by, as shown in Perez [197] in *Invisible Women: Data Bias in a World Designed for Men*, **Sexism** can be tracked back to the fourth century B.C. when Aristotle articulated that the male form is the default form as an inarguable fact. This was repeated over the years until 1966 when a symposium on the role that hunting played in human evolution was held at Chicago University and was called "Man the hunter". This concept still carries on to now, as I can see in the one-size-fits-men approach to designing supposedly gender-neutral products like Piano keyboards and smartphones, as shown in Perez [197].

Marginalization has been studied in social sciences by many scholars in critical race theory, as shown in Benjamin [24], gender studies, as shown in Davis [60], McIntosh [151], and LGBTQ studies, as shown in Fausto-Sterling [77]. However, negative stereotyping, stigma, and unintended bias continue against marginalised people based on ethnicity, religion, disability, sexual orientation, or gender. These stigmas and unintended bias have led to different forms of discrimination from education, job opportunities, health case, housing, incarceration, and others, as , as shown in Nordell [179] details in *The End of Bias*.

They can also have a negative impact on cognitive ability, and mental and physical health of the people who carry their load. As , as shown in Steele [242] shows in *Whistling Vivaldi: How Stereotypes Affect Us and What I Can Do*, based on experiments in behavioral psychology, carrying stigma made women underperform in math tests, and African-American students underperform in academia. Hence, stereotypes become self-fulfilling prophecies, eventually leading to their perpetuation and the continuation of prejudice and discrimination.

In the age of knowledge, computing, and big data, **prejudice and discrimination** have found their way to machine learning models. These models that are now dictating every aspect of our lives, from online advertising, to employment and judicial systems that rely on black box models and discriminate against marginalised groups, while benefitting privileged elites, as , as shown in O'neil [183] explains in *Weapons of Math Destruction*. One of the most well-known examples of discriminative decisions made by a machine learning model is the COMPAS algorithm, a risk assessment tool that measures the likelihood that a criminal becomes a recidivist, a term used in legal systems to describe a criminal who reoffends.

Despite Northpoint, the company that produced the COMPAS tool, does not share how the model measures the recidivism scores, the algorithm was deployed by the state of New York in 2010. In 2016, ProPublica found that Black defendants are more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while the latter were more likely than Black defendants to be incorrectly flagged as low risk, as shown in Larson et al. [135].

One example of algorithmic gender discrimination is the resume screening model used by Amazon, which, according to a Reuters report in 2018, favored resumes of male over female candidates even when both had the same skills and qualifications, as shown in dastin [58]. Similar examples of algorithmic discrimination can be found against the LGBTQ community, as shown in Tomasev et al. [262], older people, as shown in Stypinska [244], Muslims, as shown in samuel [223], and people with disabilities, as shown in Binns and Kirkham [25].

3.3 Bias and fairness: Definitions

The term *bias* is defined and used in many ways, as shown in Olteanu et al. [181]. The normative definition of bias, in cognitive science, is: "behaving according to some cognitive priors and presumed realities that might not be true at all", as shown in Garrido-Muñoz et al. [87]. And the statistical definition of bias is "systematic distortion in the sampled data that compromises its representatives", as shown in Olteanu et al. [181].

In NLP, while bias and fairness have been described in several ways, the statistical definition is most dominant, as shown in Caliskan et al. [38], Garg et al. [86], Nadeem et al. [167], Nangia et al. [172]. Since 2021, there has been a trend to distinguish two types of bias in NLP systems: intrinsic bias and extrinsic bias, as shown in Cao et al. [39], Kaneko et al. [117], Steed et al. [241]. **Intrinsic bias** is used to describe the biased representations of pre-trained models. It is also known as upstream bias, as shown in Steed et al. [241], representation bias, as shown in Shah et al. [229]. Up to the best of my knowledge, there is no formal definition of intrinsic bias in the literature. However, from the research done to study bias in word embeddings, as shown in Caliskan et al. [38], I can infer the following definition: Intrinsic bias is *stereotypical representations of certain groups of people learned during pre-training*. For example, when a model associates women with certain jobs like caregivers or men with doctors, as shown in Caliskan et al. [38]. This type of bias exists in static word embeddings, as shown in Caliskan et al. [38], Garg et al. [86] and contextual word embeddings, as shown in Nadeem et al. [167], Nangia et al. [172]. This is the definition of bias that is used through out this thesis. However, not all types of bias are harmful. And it is important for the NLP model to pick up those types of unharmful biases or stereotypes to

improve its performance. For example, a sentence like "Muslims pray int the mosque" and a sentence like "Christians pray in the church" contain unharmful stereotypes. In this thesis, I study only harmful bias and every time the word bias is used it is used in that capacity as a harmful stereotype.

On the other hand, **Extrinsic bias**, also known as model fairness, has many formal definitions built on those from literature on the fairness of exam testing from the 1960s, 70s and 80s, as shown in Hutchinson and Mitchell [108]. The most recent fairness definitions are broadly categorized into two groups: **Individual fairness**, which is defined as "*An algorithm is fair if it gives similar predictions to similar individuals*", as shown in Kusner et al. [133].

For a given model $\hat{Y} : X \rightarrow Y$ with features X, sensitive attributes A, prediction $\hat{Y}$, and two individuals i and j, and if individuals i and j are similar. The model achieves individual fairness if

$$\hat{Y}(X^i, A^i) \approx \hat{Y}(X^j, A^j) \tag{3.1}$$

The second type of fairness definition is **Group fairness**, which can be defined as *An algorithm is fair if the model prediction $\hat{Y}$ and sensitive attribute A are independent*, as shown in Caton and Haas [40], Kusner et al. [133]. Based on group fairness, the model is fair if

$$\hat{Y}(X|A = 0) = \hat{Y}(X|A = 1) \tag{3.2}$$

Group fairness is the most common definition used in NLP. There are different ways to measure it, like Equality of odds, as shown in Baldini et al. [16]. However, other metrics have been proposed in the NLP literature to measure individual fairness, like counterfactual fairness methods, as shown in Prabhakaran et al. [201].

3.4 Bias and fairness: Origins

While much literature proposes methods to measure bias and fairness in NLP models, there are far fewer papers that discuss the sources of bias. Those that do so tend to neglect literature from social science or the critical race theory that has examined topics directly related to bias like racism, sexism, or homophobia. This short-sightedness has, so far, led to cosmetic changes in the proposed NLP models to resolve the problem of bias rather than fixing the racist, sexist, homophobic status quo, as shown in Benjamin [24]. In this section, I review the sources of bias in technology from the perspective of social science, using tools like critical race theory, digital studies, gender studies, LGBTQ studies and internet and data activism. Then I review the sources of bias from a purely NLP perspective, while trying to connect these two strands to gain a more profound understanding of the origins of bias.

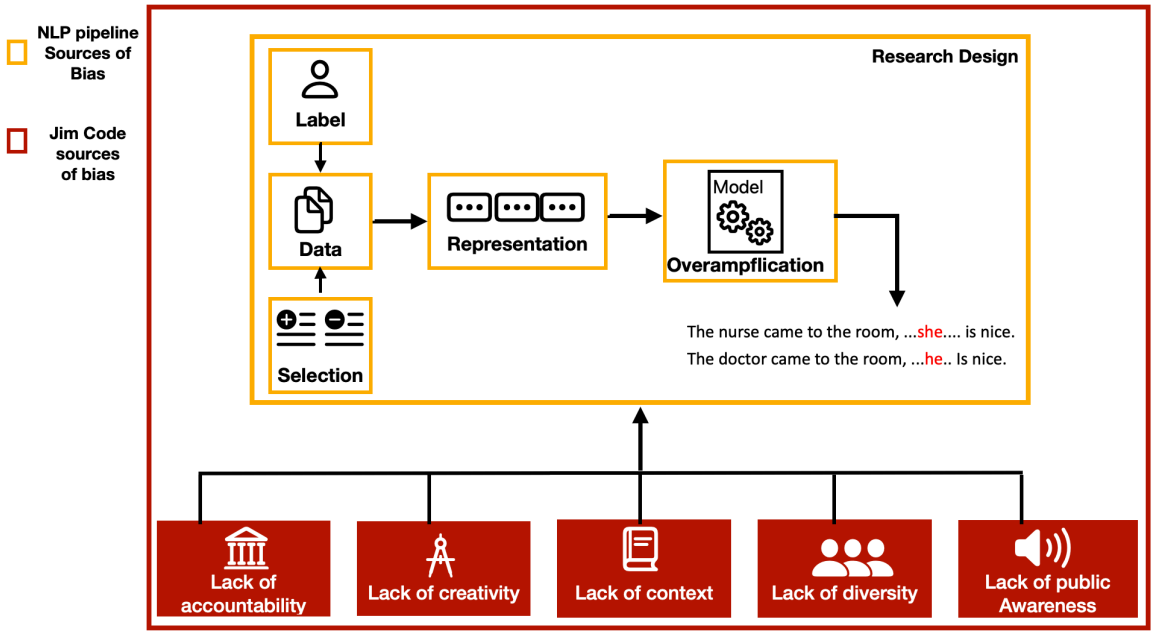


Fig. 3.1 The sources of bias in supervised NLP models

3.4.1 The Jim Code perspective

As I mentioned before, the Jim Code is a term that refers to the new forms of systematic discrimination found in new technologies that build on older discriminatory systems. This is one of the main origins of bias and unfairness that I find in most AI systems. This can be broken down into the following sources of bias:

1. **Lack of context:** In *More than a Glitch*, Broussard [33] explains that, like computers, the data used to train NLP models is produced without a specific human context. A similar point is made by, as shown in Benjamin [24], who discusses how social and historical contexts are not taken into consideration when data is collected to train NLP models. But it is not only the data. With the NLP models being developed isolated from social sciences and critical race theory, how these systems impact the people of different identity groups gets overlooked. For example, models outputs decisions on who is eligible to get a loan or get a job without consideration of the fact that this might increase the wealth gap between marginalised and privileged groups.

Moreover, it is because of the lack of context that researchers in NLP do not think about the harmful ways that their proposed groundbreaking systems could be. For example, some models are used to detect race from last name and Zip-codes. The developers of these models have probably failed to consider how these models are being used by certain businesses to illegally collect information on ethnicity, as shown in Benjamin [24]. Even more, harm

is caused when a model categorizes people as criminals or terrorists due to their detected ethnicity.

2. **Lack of creativity:** Because of the lack of context, many researchers in NLP models tend to build their systems on top of existing racist, sexist, homophobic, ageist, and ableist systems. An example is when recommender systems used "cultural segregation" to infer information about a person's ethnicity to personalize their recommendations. They use ethnicity as a proxy for individuality, as shown in Benjamin [24]. Hence, those systems perpetuate the racist view that people who belong to a specific group must have similar preferences. Researchers need to be more creative and find other ways to recommend content that does not rely on social bias shortcuts.

3. **Lack of accountability:** There is also a lack of accountability that allows tech companies to get away with creating oppressive systems that are not just, "glitches" as explained by critical race and digital humanities studies activists , as shown in Benjamin [24], Broussard [33], Nobel [177]. A lack of accountability enables companies to sell their systems as black boxes without explaining how their models make decisions, as shown in O'neil [183]. I also see that in the scientific community, where big tech companies publish papers emphasizing their models' excellent results without sharing those models or the data that were used to train them, precluding reproducibility.

Moreover, when, the Justice League, a group of AI ethicists and activists, launched the Safe Face pledge to ensure that computer vision models do not discriminate between people based on their skin color, no major tech company was willing to sign it, as shown in Benjamin [24]. With the lack of accountability and legislation, big tech companies, which are one of the main drivers of the field, have no reason to revise and change the way they build their NLP systems, or to include the social and historical context into their research in a way that profoundly changes the systems instead of just covering it up and fixing the "glitches".

4. **Lack of diversity:** The majority of NLP technologies are developed in companies or research institutes in Western societies and by researchers who are mostly White, able-bodied, heterosexual men. They develop and test systems that work well for them, without considering how functional these systems are for people from different backgrounds. Examples are facial recognition systems that only work with people with light skin, as shown in Benjamin [24], Broussard [33] and CV recommendation systems that favor applicants with male names, as shown in dastin [58]. There is also a lack of diversity when it comes to the targeted customers of the systems. Since most of these technologies are expensive to buy, the developers of these systems focus on the customers who can afford it who are mostly White, able-bodied, heterosexual men, as shown in Benjamin [24]. This lack of diversity, in addition

to the lack of social and historical contexts, leads to the development of discriminatory systems.

5. **Lack of public awareness:** In addition to the previously discussed sources of bias in NLP, another factor that allows the biases to spread is the lack of public awareness. This is a result of using mathematical and statistical terminology and jargon that most non-specialists can't understand. This lack of understanding of how NLP models work and their limitations led people to over-trust AI systems and to Technochauvinism which, as shown in Broussard [33] is described as:

> *the kind of bias that considers computational solutions to be superior to all other solutions. Embedded in this bias is a priori assumption that computers are better than humans, which is actually a claim that the people who make, and program computers are better than other humans.*

The lack of public awareness and Technochauvinism is behind banks, schools, hospitals, universities, and other institutions that are supposed to deal with people and society and make social decisions adopting NLP systems that are poorly understood, with the false notion that they are unbiased, and their decisions are faultless and objective, as shown in Benjamin [24], Broussard [33].

3.4.2 The NLP pipeline perspective

I now turn to the sources of bias in the NLP pipeline described in the NLP literature. Shah et al. [229] introduce four sources of bias in the NLP pipeline that might impact the model's fairness. Hovy and Prabhumoye [106] also discusses these, adding a fifth source related to the overarching design of NLP research projects.

Here, I outline these pipeline biases and also show how they, in fact, originate in the Jim Code perspective.

1. **Research design:** According to, as shown in Hovy and Prabhumoye [106], research design bias is manifested in the skewness of NLP research towards Indo-European languages, especially English. This skew leads to a self-fulfilling prophecy, since most of the research focuses on text in English, more data in English becomes available, which in turn makes it easier for NLP researchers to work on English text. This has further ramifications as, as shown in Hovy and Prabhumoye [106] also question whether, if English was not the "default" language, $n-gram$ would have been the focus of NLP models. The authors argue that the

lack of diversity in the makeup of the NLP research groups, is one of the reasons behind the linguistic and cultural skewness in NLP research.

In addition to these skews, there are further sources of bias reflected in research design that originate from the Jim Code perspective. *The lack of social context* is clearly manifested in NLP research design. For example, NLP researchers deal with language as a number of word occurrences and co-occurrence probabilities, rather than dealing with language as a diverse social component that reflects societal relationships and biases, as shown in Holmes [103]. Another example, is *the lack of historical context*, with most of the data that NLP models are trained on generated by white middle-class men, resulting in speech recognition models not recognizing African American dialect, as shown in Benjamin [24], Tatman [259] and hate speech detection models falsely flagging African American dialect as hateful, as shown in Sap et al. [224]. *Lack of creativity* is also reflected in research design. For example, with NLP models relying on the $n-gram$ models and words co-occurrences, they incorporate biases such that they associate gendered words,"woman" and "man", with certain jobs, "nurse" and "doctor", as shown in Caliskan et al. [38]. As, as shown in Hovy and Prabhumoye [106] contends, *lack of diversity* is also reflected in the research design bias, as evident in the skewness towards Indo-European languages. Because of the *lack of accountability* and the *lack of public awareness*, NLP research design bias has been going on for decades, largely unnoticed and unconsidered.

2. **Selection bias:** Selection bias is a result of non-representative observations in the datasets used to train NLP models, as shown in Hovy and Prabhumoye [106], Shah et al. [229]. This bias could manifest when the text data that the model is trained on was generated by one group of people, but when it is deployed in the real world it is used by more diverse groups. For example, the syntactic parsers and part-of-speech taggers were trained on data generated by white middle-aged men, which then impacted the accuracy of these models when tested on the text generated by different groups of people, as shown in Shah et al. [229]. Another example in hate speech detection models, where the models were trained on data with over-representation of marginalised identity groups with the positive class (hateful) which resulted in hate speech detection models falsely labelling content as hateful just because it includes mentions of marginalised identifies, as shown in Dixon et al. [69], Sap et al. [224].

Selection bias is also a result of *lack of context*, since the NLP researchers used datasets with an over-representation of one group and under-representation of many other groups due to their lack of social and historical context of who generated that data and which identity groups are under-represented in the chosen data. *Lack of diversity* is also a main reason behind selection bias in NLP, as most of the NLP researchers come from a non-marginalised

background with blind spots for the under-represented groups of people. Finally, *lack of creativity* is another reason behind selection bias. As NLP researchers build their NLP models on biased systems that generated biased data, instead of being more creative and using more diverse representative data that work for everyone.

3. **Label bias:** Label bias, also known as annotator bias, is a result of a mismatch between the annotators and the authors of the data. There are many reasons behind label bias. It could result from spamming annotators who are uninterested in the task and assign labels randomly to get the task done, as can happen on crowdsourcing platforms. It could also happen due to confusion or ill-designed annotation tasks. Another reason is due to the individual annotator's perception and interpretation of the task or the label, as shown in Hovy and Prabhumoye [106]. Moreover, there could be a mismatch between the authors' and annotators' linguistic and social norms. For example, when the annotators mislabel content as hateful for including the N-word, despite its benign in-group use by African Americans. Finally, labels might carry the annotators societal perspectives and social biases, as shown in Sap et al. [224].

On the other hand, I can argue that some of these biases result from unfairness in the crowdsourcing systems. Since the pay that annotators receive is often extremely low, they are incentivized to do as many tasks as possible as fast as possible to make ends meet, which in turn impacts the quality of the labels, as shown in Fort et al. [82]. Moreover, Miceli et al. [153] argue that the bias in the labels is not only due to the biased perceptions of the annotators but also due to a certain format the annotators have to follow for their annotation tasks and if that format falls short on diversity, the annotators lack the means to communicate that to the designers of the task. An example is when an annotator is presented with a binary gender choice, even if the data contains information about non-binary or transgender people. Hence, label bias could be seen as a result of the lack of context. As the NLP researchers who mismatch the demographics of their data's authors and annotators do that due to the lack of social context of the author of the data. Label bias is also a result of the *lack of accountability*, as big tech and NLP research groups hire annotators with unfair pay in addition to the lack of means for those annotators to communicate problems in the annotation task with the task designer due to power dynamics.

4. **Representation bias:** Representation bias, also known as intrinsic bias or semantic bias, describes the societal stereotypes that language models encode during pre-training. The bias exists in the training dataset that then gets encoded in the language models static, as shown in Caliskan et al. [38], Elsafoury et al. [74], Garg et al. [86], or contextual, as shown in Nadeem et al. [167], Nangia et al. [172]. Hovy and Prabhumoye [106] argue that one of the main reasons behind representation bias is the objective function that trains the language models.

As these objective functions aim to predict the most probable next term given the previous context, which in turn makes these models reflect our biased societies in the data.

Again, representation bias is a result of the *lack of social and historical context*, which is why NLP researchers tend to use biased data to train these language models. It is also a result of *lack of creativity* as instead of using objective functions that aim to reproduce the biased word that I live in, NLP researchers could have used different objective functions that optimize fairness and equality in addition to performance.

5. **Model overampflication bias:** According to, as shown in Shah et al. [229], overampflication bias happens because, during training, the models rely on small differences between sensitive attributes regarding an objective function and amplify these differences to be more pronounced in the predicted outcome. For example, in the imSitu image captioning data set, 58% of the captions involving a person in a kitchen mention women, resulting in models trained on such data predicting people depicted in kitchens as women 63% of the time, as shown in Shah et al. [229]. For the task of hate speech detection, overampflication bias could happen because certain identity groups could exist within different semantic contexts, for example, when an identity group like "Muslims" co-occurs with the word "terrorism". Even if the sentence does not contain any hate, e.g., "Anyone could be a terrorist, not just muslims", the model will learn to pick this information up about Muslims and amplify them, leading to these models predicting future sentences that contain the word "Muslim" as hateful. According to, as shown in Hovy and Prabhumoye [106], one of the sources of overampflication bias is the choice of objective function used in training the model. Since these objective functions mainly aim to improve precision, the models tend to exploit spurious correlations or statistical irregularities in the data to achieve high performance by that metric.

Overamplification bias is again a result of the *lack of social and historical context*, which results in using data that has an over-representation of certain identities in a certain social or semantic context. These over-representations are then picked up by the models during training. Another reason is the *lack of creativity* that results in choosing objective functions that exacerbate the differences found in the datasets between different identity groups and prioritizing overall performance over fairness.

3.5 Bias metrics

In this section, I review the literature on the different methods used to measure intrinsic bias in static and contextual word embeddings (language models) that will be used in this thesis. These metrics are summarised in Section 3.5.1.

3.5.1 Static word embeddings

In distributional word representations (Word embedding), the most common methods for quantifying bias are WEAT, as shown in Caliskan et al. [38], RND, as shown in Garg et al. [86], RNSB, as shown in Sweeney and Najafian [254], and ECT, as shown in Dev and Phillips [64]. In this section, I provide a description of each metric and how it is measured.

Word embedding association test

The word embedding association test (WEAT), as shown in Caliskan et al. [38] is one of the most used bias metrics in the literate on bias in NLP models. The authors were inspired by the Implicit Association Test (IAT) to develop a statistical test to demonstrate human-like biases in word embeddings. The authors consider two equal-sized sets S, T of attribute words, for example,

$S = \{engineer, doctor, journalist\}$ and $T = \{housewife, nurse, secertary\}$, and two sets A, B of target words, for example, $A = \{man, male, boy\}$ and $B = \{women, female, girl\}$.

First, the authors measure the differential similarity between words vector (w_c) for word c in the word sets S or T and the word vectors (w_a) and (w_b) in word sets A and B as follows:

$$g(c, A, B, w) = mean_{a \in A}(cos(w_c, w_a)) - mean_{b \in B}(cos(w_c, w_b)) \tag{3.3}$$

Then, the authors measure the bias as the effect size, as measured below:

$$WEAT = \frac{mean_{s \in S} g(s, A, B, w) - mean_{t \in T} g(t, A, B, w)}{std - dev_{c \in S \cup T} g(c, A, B, w)} \tag{3.4}$$

Where $mean$ and $std - dev$ refer to the arithmetic mean and standard deviation.

Relative norm difference

Relative norm difference (RND), as shown in Garg et al. [86], where for S a list of neutral words, e.g.,

$S = \{engineer, doctor, journalist\}$ and two sets A, B of target words. For example, $A = \{man, male, boy\}$ and $B = \{women, female, girl\}$.

RND measures bias as the average l_2 *norm* of the differences between the word vectors of neutral words (w_s), like profession names, and a representative group vector created by averaging the word vectors (w_g) for words that describe a stereotyped marginalised group (g) e.g., gender, ethnicity, religion, or sexual orientation.

$$w_a = mean_{a \in A}(wa) \tag{3.5}$$

$$w_b = mean_{b \in B}(wb) \tag{3.6}$$

$$RND = \sum_{w_s \in S} ||w_s - w_a||_2 - ||w_s - w_b||_2 \tag{3.7}$$

Relative negative sentiment bias

In the relative negative sentiment bias (RNSB) bias metric, as shown in Sweeney and Najafian [254], a logistic regression model (f) is first trained on the word vectors of unbiased labelled sentiment words (positive and negative) extracted from the biased word embeddings (w).

Then, that model is used to predict the sentiment of words that describe certain demographic groups, target set, for a set $A = \{k_1, ...K_t\}$ of t demographic identity word vectors from a sensitive attribute e.e gender, nationality, or religion. A set P is defined as containing the predicted negative sentiment probability via f normalized to be one probability mass.

$$P = \{\frac{f(k_1)}{\sum_{i=1}^{t} f(k_)[i]}, ..., \frac{f(k_t)}{\sum_{i=t}^{t} f(k_)[t]}\} \tag{3.8}$$

Finally, RNSB measures bias as the Kullback-Leibler (KL) divergence between the negative sentiment probability of identity terms after normalization (P) from (U) the uniform distribution of t elements.

$$RNSB = D_{KL}(P, U) \tag{3.9}$$

Embeddings coherence test

In embeddings coherence test (ECT), the authors proposed a method to measure how much bias has been removed from the word embeddings after debiasing, as shown in Dev and Phillips [64]. For two sets A, B of target words, for example, $A = \{man, male, boy\}$ and $B = \{women, female, girl\}$. and a set ($T$) of attribute words $S = \{engineer, doctor, journalist\}$. For a given word embeddings (w), the authors measure the average vector that represents the group.

$$w_a = mean_{a \in A}(wa) \tag{3.10}$$

$$w_b = mean_{b \in B}(wb) \tag{3.11}$$

Then they measure the cosine similarity between the average word vector w_a and w_b and a word vector w_c for each word (c) in the attribute list (T).

Metric	Bias statistical definition	Equation	Citation
WEAT	A statistical test that measures the differential similarity between word vectors.	$WEAT = \frac{mean_{s\in S} g(s,A,B,w) - mean_{t\in T} g(t,A,B,w)}{std-dev_{c\in S\cup T} g(c,A,B,w)}$	[38]
RND	The average l_2 *norm* of the differences between the words vectors.	$RND = \sum_{w_s\in S} \|\|w_s - w_a\|\|_2 - \|\|w_s - w_b\|\|_2$	[86]
RNSB	The Kullback-Leiber (KL) divergenc between the negative sentiment probability of word vectors.	$RNSB = D_{KL}(P,U)$	[254]
ECT	The cosine similarity between the average word vectors	$ECT = Spearman(similarlity_{w_a}, similarlity_{w_b})$	[64]

Table 3.1 Summary of the different social bias metrics used to measure bias in static word embeddings in this thesis.

$$similarlity_{w_a} = Cos(w_a, w_c) \tag{3.12}$$

$$similarlity_{w_b} = Cos(w_b, w_c) \tag{3.13}$$

Finally, ECT measures bias as the Spearman correlation between the two similarity lists

$$ECT = Spearman(similarlity_{w_a}, similarlity_{w_b}) \tag{3.14}$$

3.5.2 Contextual word embeddings

In contextual embeddings or language models (LM), among the most used metrics to measure bias in LM are Crows-Pairs, as shown in Nangia et al. [172], StereoSet, as shown in Nadeem et al. [167], and SEAT, as shown in May et al. [149]. I describe each of these metrics and how they are measured. These metrics are summarised in Section 3.5.2.

Crowdsourced Stereotype Pairs

In the Crowdsourced Stereotype Pairs (CrowS-Pairs), the authors of the metric used Amazon Mechanical Turk (MTurk) to collect 1508 sentence pairs about a disadvantaged group for measuring bias in LM. The sentence pairs are a stereotypical sentence and a non-stereotypical sentence. The crowS-Pairs metric measures whether the LM prefers the stereotypical sentence about the marginalised groups of people, as shown in Nangia et al. [172]. The collected data covers 9 categories of bias: race, gender, socioeconomic status, nationality, religion, age, sexual orientation, physical appearance, and disability. The CrowS-Pairs metric measure the bias in LM using the masked language models (MLM) task. For a stereotypical sentence $S = U \cup M$, where U is a set of unmodified tokens for example, $U = \{is, a, nuse, attitude, is, nice\}$ with length ($|C|$) and M is a set of modified tokens, for example, $\{she, her\}$

The authors estimate the probability of the unmodified token conditioned on the modified tokens $p(U|M,\theta)$ using the $pesudo-logliklihood$. To measure a score of the sentence $score(S)$

$$score(S) = \sum_{i=0}^{|C|} logP(u_i \in U|M,\theta) \tag{3.15}$$

The same score is being measured for the non-stereotypical sentence S' where $S' = U \cup M'$, where U is a set of unmodified tokens for example, $U = \{is, a, nuse, attitude, is, nice\}$ with length ($|C|$) and M' is a set of modified tokens for example $\{he, his\}$

$$score(S') = \sum_{i=0}^{|C|} logP(u_i \in U|M',\theta) \tag{3.16}$$

Then, the bias scores are measured as the percentage of examples where the model (θ) assigns a higher probability estimate to the stereotypical sentences (S) over the non-stereotypical sentence (S'). If the percentage is over or below 0.5, then that means the model prefers the stereotypical or the non-stereotypical sentences and is hence biased. On the other hand, if the percentage is 0.5, that means the model randomly assigns probability and hence is not biased.

StereoSet

StereoSet metric is similar to the CrowS-Pairs metric in terms of relying on crowdsources sentences to measure bias in LM, and they also use the MLM task to measure the bias, as shown in Nadeem et al. [167]. The authors also used Amazon Mechanical Turk (MTurk) to collect sentence pairs, stereotypical and anti-stereotypical. The authors target four categories of bias: race, gender, profession, and religion. The authors used lists of target terms that describe each of the inspected bias categories. The authors propose 3 metrics, language modelling score (*lms*), StereoSet score (*ss*), and idealized cat score (*icat*). The *lms* score measures the performance of the LM, the *ss* score measures the bias in the LM, and the *icat* score is a score that expresses the bias and the performance of the LM. However, in the studies that use the StereoSet metric to measure bias, they use only the *ss* score.

The StereoSet metric measures the bias in LM using the masked language models (MLM) task. For a stereotypical sentence $S = U \cup M$, where U is a set of unmodified tokens for example, $U = \{is, a, nuse, attitude, is, nice\}$ and M is a set of modified tokens, for example, $\{she, her\}$

Unlike the CrowS-Pairs metric, the StereoSet measures to estimate the probability of the modified token conditioned on the unmodified tokens $p(M|U,\theta)$ using the $pesudo-log-liklihood$ MLM scores. to measure a score of the sentence $score(S)$

$$ss(S) = \sum_{i=0}^{C} logP(m_i \in M|U,\theta) \tag{3.17}$$

The same score is being measured for the non-stereotypical sentence S' where $S' = U \cup M'$, where U is a set of unmodified tokens for example, $U = \{is, a, nuse, attitude, is, nice\}$ and M' is a set of modified tokens for example $\{he, his\}$

$$ss(S') = \sum_{i=0}^{C} logP(m'_i \in M'|U,\theta) \tag{3.18}$$

Similar to CrowS-Pairs, the bias StereoSet scores are measured as the percentage of examples where the model (θ) assigns a higher probability estimate to the stereotypical sentences (S) over the non-stereotypical sentence (S'). If the percentage is over or below 0.5, then that means the model prefers the stereotypical or the anti-stereotypical sentences and is hence biased. On the other hand, if the percentage is 0.5, that means the model randomly assigns probability and hence is not biased.

Sentence encoder association test

The sentence encoder association test (SEAT), the authors, were inspired by the WEAT metric, as shown in Caliskan et al. [38] to measure representation bias in static word embeddings, as shown in May et al. [149]. The authors propose to compare sets of sentences, using the cosine similarity, instead of words as with the WEAT metric. To extend the word level to a sentence level, SEAT slots each word in the seed words used by WEAT in semantically bleached sentence templates such as “This is <word>.”, “<word> is here.”, “This will <word>.”, and “<word> are things.”. The <word> placeholder is replaced with target words and attribute words to form a set of target sentences T an S and a set of attribute words A and B.

For example, $T = \{\text{This is woman}, \text{woman is here}\}$, $S = \{\text{This is man}, \text{man is here}\}$, $A = \{\text{Ther is enginner}, \text{This is doctor}\}$, and $B = \{\text{This is housewife}, \text{They are nurse}\}$.

The SEAT metric uses the LM encoding to T, S, A, and B then measures the bias score in the same way as WEAT which is described above.

Metric	Bias statistical definition	Equation	Citation
CrowS-Pairs	It estimates the probability of the unmodified token conditioned on the modified tokens using the Masked Language Modelling task (MLM).	$score(S) = \sum_{i=0}^{\|C\|} logP(u_i \in U \| M, \theta)$	[172]
StereoSet	It estimates the probability of the modified token conditioned on the unmodified tokens using the Masked Language Modelling task (MLM).	$ss(S) = \sum_{i=0}^{C} logP(m_i \in M \| U, \theta)$	[167]
SEAT	It measures representation bias using the same method as WEAT but in contextual word embeddings using sentence encoding instead of word vectors.	$SEAT = \frac{mean_{s \in S} g(s,A,B,w) - mean_{t \in T} g(t,A,B,w)}{std-dev_{c \in S \cup T} g(c,A,B,w)}$	[149]

Table 3.2 Summary of the different social bias metrics used to measure bias in contextual word embeddings in this thesis.

3.5.3 Limitations

In this section, I discuss some of the limitations of the bias metrics described earlier. The bias metrics used to measure bias in static word embeddings, except for RNSB, are based on the polarity between two opposing points, like male and female, allowing for binary comparisons. This forces practitioners to model gender as a spectrum between more "male" and "female" words, requiring an overly simplified view of the construct, leading to similar problems for other stereotypical types of bias, like racial, religious, transgender, and sexual orientation, where there are more than two categories that need to be represented, as shown in Sweeney and Najafian [254]. Additionally, these metrics also use lists of seed words that have been shown to be unreliable as the instability of measurements using the seed words, as shown in Antoniak and Mimno [11]. Moreover, according to, as shown in Antoniak and Mimno [11] measure the coherence between two seed sets (A and B) like $A = \{\text{excutibe, management, professional}\}$ and $b = \{\text{home, parents, family}\}$ after being mapped to the biased subspace using the WEAT metric. The resulting coherence scores are low, which means that the seed pairs are not projected farther apart enough to show the bias polarization in the word embeddings. Additionally, according to, as shown in Badilla et al. [14], the different metrics use different scales, which makes it harder to directly compare the results from the different metrics without ranking the biased scores. In chapter 4, demonstrates that different bias metrics WEAT, RND, RNSB, and ECT gave different results when they were used to compare the gender and racial biased in five different word embeddings.

Similarly, there are limitations with the bias metrics used to measure bias in LM. For example, the SEAT metric uses cosine similarity with sentence encoding when, as shown in Delvin [63] argues that LMs like BERT are not built to provide meaningful sentence embeddings. Additionally, using bleached sentence templates does not provide real context and there is no guarantee that the LM will treat those sentences as semantically bleached, as shown in May et al. [149]. As for CrowS-Pairs and StereoSet, Blodgett et al. [26] shows that there are problems in the crowdsourced data that is used to measure the bias and the ambiguity in what these metrics are actually measured.

3.6 Fairness metrics

I mentioned earlier that there are two types of fairness metrics: individual fairness and group fairness. In this section, I provide a formal definition for each type of the fairness metrics and provide some of the proposed methods in the literature, to measure it from the literature.

3.6.1 Individual fairness

Individual fairness, which is defined as "*An algorithm is fair if it gives similar predictions to similar individuals*", as shown in Kusner et al. [133].

For a given model $\hat{Y} : X \rightarrow Y$ with features X, sensitive attributes A, prediction $\hat{Y}$, and two individuals i and j, and if individuals i and j are similar. The model achieves individual fairness if

$$\hat{Y}(X^i, A^i) \approx \hat{Y}(X^j, A^j) \tag{3.19}$$

Counterfactual fairness is viewed as individual fairness. It compares between two or more variations of an individual instance. One is the real-world factual instance and the others are counterfactual instances. The different variations belong to different identity groups, as shown in Czarnowska et al. [53]. In this section, I review some of the proposed methods to measure counterfactual fairness in the NLP literature.

Counterfactual prediction sensitivity

Czarnowska et al. propose a method to measure counterfactual fairness based on the random norm difference (RND) that was introduced earlier, but instead of word vectors, the authors used model prediction probabilities. The authors of the proposed metric use L1 norm instead of L2 norm, as shown in Czarnowska et al. [53]. For a model $f(x) : X \rightarrow Y$, a factual instance (x_i) that contains identity (g), and counterfactual instance ($\hat{x_i}$) that contain identity ($\hat{g}$), fairness is measured as:

$$CPS = \sum_{i=1}^{N} |f(x_i) - f(\hat{x_i})| \tag{3.20}$$

Where N is the number of instances and i is an instance.

Perturbation score analysis

Prabhakaran et al. propose a set of metrics to measure counterfactual fairness, as shown in Prabhakaran et al. [201]. These metrics are:

1. Perturbation score sensitivity (ScoreSens): It measures the average difference between the model prediction ($f(x)$) of the factual (x) and the counterfactual ($\hat{x}$) instances over the number of examples (X). For a model $f(x) : X \rightarrow Y$, a factual instance (x) that contains identity (g), and counterfactual instance ($\hat{x}$) that contain identity ($\hat{g}$), fairness is measured as:

$$ScoreSens = E_{x \in X}[f(\hat{x}) - f(x)] \tag{3.21}$$

2. Perturbation score deviation (ScoreDev): It measures the average standard deviation of the predicted scores of the model $f(x) : X \rightarrow Y$ for the counterfactual instances ($\hat{x}$) for each identity group (m) for identity groups (M) with the number of examples is (X). The fairness score is measured as :

$$ScoreDev = E_{x \in X}[StdDev_{m \in M}(f(\hat{x}))] \tag{3.22}$$

3. Perturbation score range (ScoreRange): It is the $Range(max - min)$ of the predicted scores of the model $f(x) : X \rightarrow Y$ for the counterfactual instances ($\hat{x}$) for each identity group (m) for identity groups (M) where the number of examples is (X). The fairness score is measured as :

$$ScoreRange = E_{x \in X}[Range_{m \in M}(f(\hat{x}))] \tag{3.23}$$

4. Perturbation label distance(LabelDist): It is a metric that measures the perturbation sensitivity of model labels. For a binary classifier $f(x) : X \rightarrow Y$ regarding corpus X and a set of identity groups M, it measures the Jaccard Distance between a) the set of sentences (x) where the model prediction is positive ($f(x) = 1$) and b)the set of sentences ($\hat{x}$) where the mode prediction is positive ($f(\hat{x}) = 1$). Then it is averaged across the identity groups (M). The fairness score is measured as :

$$LabelDist = E_{m \in M}[Jaccard(x|f(x) = 1, \hat{x}|f(\hat{x}) = 1)] \tag{3.24}$$

Fairness score

Qian et al. proposed a metric to measure fairness scores similar to CrowS-Pairs and StereoSet but with prediction probabilities, as shown in Qian et al. [204]. The authors define the fairscore metric as the percentage of different model predictions ($f(x)$) between factual (x) and counterfactual ($\hat{x}$) instances, with the number of examples is (X).

$$fairScore = \frac{|x \in X | f(x) \neq f(\hat{x})|}{|X|} \tag{3.25}$$

3.6.2 Group fairness

The second type of fairness definition is **Group fairness**, which can be defined as *An algorithm is fair if the model prediction $\hat{Y}$ and sensitive attribute A are independent*, as shown in Caton and Haas [40], Kusner et al. [133]. Based on group fairness, the model is fair if

$$\hat{Y}(X|A = 0) = \hat{Y}(X|A = 1) \tag{3.26}$$

Three are two main approaches to measuring a model's fairness in that case:

1. *Threshold-based metrics*, where a model's fairness is measured by how much the classifier's predicted labels differ between different groups of people, based on a threshold, as shown in Hutchinson and Mitchell [108]. The equalized odds metric is a threshold-based metric, and it is the most commonly used metric in the literature for measuring the extrinsic bias in the downstream task of text classification, as shown in Cao et al. [39], De-Arteaga et al. [61], Steed et al. [241]. Equalized odds are measured by the absolute difference, gap, between the true positive rates (TPR) or false positive rates (FPR) between different groups of people, g and $\hat{g}$, based on sensitive attributes like gender, race, etc.

$$FPR\_gap_{g,\hat{g}} = |FPR_g - FPR_{\hat{g}}| \tag{3.27}$$

$$TPR\_gap_{g,\hat{g}} = |TPR_g - TPR_{\hat{g}}| \tag{3.28}$$

2. *Threshold-agnostic metrics*, where fairness is measured by how much the distribution of the classifier's prediction probabilities varies across different groups of people based on their sensitive attributes. Borkan et al. propose a set of fairness metrics that are based on the AUC score. An important advantage of the threshold-based metrics is that they are robust to data imbalances in the amount of positive and negative examples in the test set, as shown in Borkan et al. [29]. The proposed AUC-based metrics are: For D^- is the negative examples in the background (test set); D^+ is the positive examples in the background (test set); D_g^- is the negative examples in the identity group (g); and D_g^+ is the positive examples in the identity group (g), the fairness metrics are:

(a) Subgroup AUC: calculates the AUC scores on only examples from the identity subgroup (g). This metric represents the model understanding and separability within the subgroup itself.

$$Subgroup\_AUC = AUC(D_g^- + D_g^+) \tag{3.29}$$

(b) Background positive subgroup negative (BPSN) AUC: it calculates the AUC on positive examples from the background and negative examples from the subgroup. This metric is supposed to show the false positive within the identity subgroup at many thresholds.

$$Subgroup\_AUC = AUC(D^+ + D_g^-) \tag{3.30}$$

(c) Background negative subgroup positive (BNSP) AUC: it calculates the AUC on negative examples from the background and positive examples from the subgroup. This metric is supposed to show the false negative within the identity subgroup at many thresholds.

$$Subgroup\_AUC = AUC(D^- + D_g^+) \tag{3.31}$$

In addition to these metrics, I propose a simpler metric that measures the subgroup AUC, and then measures the absolute difference between the AUC scores for the different identity groups, (g and $\hat{g}$), as shown below:

$$AUC\_gap_{g,\hat{g}} = |AUC_g - AUC_{\hat{g}}| \tag{3.32}$$

3.6.3 Limitations

One of the main limitations of the proposed methods to measure individual fairness metrics is that the researchers propose these metrics do not provide the motivation behind the proposed metrics and what their proposed metric actually measures. As for the group fairness metrics, they are all based on statistical measures that have been criticized in the literature. For example, Hedden argues that group fairness metrics are based on criteria that cannot be satisfied unless the models make perfect predictions or that the base rates are equal across all the identity groups in the dataset, as shown in Hedden [100]. Base rate here refers to the class of probability that is unconditioned on the featural evidence, as shown in Bar-Hillel [18]. Hedden goes on to ask if the statistical criteria of fairness cannot be jointly satisfied except in marginal cases, which criteria then are conditions of fairness? questioning statistical methods to measure fairness was raised by, as shown in Broussard [33] where she argues that

some founders of the field of statistics were white supremacists, which resulted in skewed statistical methods and that to measure fairness, maybe I should use non-statistical methods.

3.7 Bias mitigation

In this section, I review the literature for the bias removal methods used in the literature and that I use later in this thesis. The bias mitigation techniques in ML and NLP literature are categorized into 3 groups based on when these techniques are applied in the ML pipeline , as shown in Caton and Haas [40].

3.7.1 Pre-processing

Pre-processing bias mitigation techniques are applied before the ML model is trained. It aims at removing the different types of bias in the training dataset, as shown in Caton and Haas [40], like selection bias, overamplification and label bias. One of the methods that fall under this category is Perturbations or counterfactual data augmentation, as shown in Meade et al. [152], Qian et al. [204], Webster et al. [284]. In this method, the training dataset is balanced by augmenting counterfactual examples to provide balanced representations of the different identity groups within the same sensitive attribute. For example, for the gender-sensitive attribute, if our dataset contains a sentence like “The doctor came to the room, he is nice”, a counterfactual example would be “The doctor came to the room, she is nice” or “The doctor came to the room, they are nice”. For the race-sensitive attribute, a sentence like “Asian are smart” counterfactuals will be added to give the same representation to different ethnicities like “African Americans are smart”, “Mexicans are smart”, “European Americans are smart”.

To create the perturbations or counterfactuals, some papers use sentences’ templates and swap the different identity groups between the sentences, as shown in Webster et al. [284]. Other studied use to provide syntactic augmentation in real-world sentences, as shown in Papakipos and Bitton [187] which allows creating alterations like simulating typos, inserting punctuation characters, and swapping gendered words that swap not only nouns that refer to genders from a binary perspective but also changes the pronouns between the two genders. Their work is built on nlpaug [1]. which alters sentences by either inserting random words into the sentences or swapping existing words for semantically similar words. Similar words are found using static word embeddings (Word2Vec, FastText, and Glove) or contextual word embeddings (BERT). Augly is useful only for gender-swapping, but does not get extended for

[1] https://github.com/makcedward/nlpaug

Augly can be only used to create perturbations to balance the representations for only gender-sensitive attributes. But to create perturbations for other sensitive attributes like race and religion. To overcome that, Qian et al. [204] proposed a seq-to-seq model to automatically create perturbations for gender, ethnicity, and age-sensitive attributes. The automatic model creates perturbations for demographic terms expressed as a pronoun (e.g., he and she), a proper name (e.g., Sue and Jamal), nouns, an adjective (e.g., Black and Asian), or other part of speech with demographic information. The seq-to-seq model is trained on 98K human-generated demographic text perturbations.

3.7.2 In-Processing

In-processing mitigation techniques aim at reducing the bias that happens during the model training due to the dominance of certain features or distributional effects. That could be achieved by adding one or more fairness metrics to the model optimization function to help the model converge towards a set of parameters that achieves a trade-off between performance and fairness, as shown in Caton and Haas [40]. For example, Zhao et al. propose a method to mitigate overampflication bias when training models on biased corpora. The authors propose the RBA framework for reducing bias amplification in predictions. Their proposed method introduces corpus-level constraints so that gender indicators co-occur no more often together with elements of the prediction task than in the original training distribution, as shown in Zhao et al. [298]. Since I do not use any of these debiasing methods in this thesis, I will not explain these methods further.

3.7.3 Post-processing

Post-processing mitigation techniques are used after the model is trained. It aims at performing transformations on the trained model to alter its outcomes regarding the sensitive attributes. They are the most widely used bias mitigation techniques because it only requires access to the model's output and information about the sensitive attributes, as shown in Caton and Haas [40]. Among the most commonly used post-processing methods and that I use later in the following chapters are Hard-Debias, as shown in Bolukbasi et al. [27] to remove bias from static word embeddings, U-Debias and P-Debias, as shown in Ding et al. [68] to remove bias using causal inference in static word embeddings, and SentDebias, as shown in Liang et al. [139] to remove bias from contextual word embeddings.

Hard Debias

In, as shown in Bolukbasi et al. [27], the authors propose to identify and remove the biases of subspaces in static word embeddings. For a word set (W), defining sets $D_1, D_2, .., D_n \subset W$ as well as word embeddings $\vec{w} \in \mathbb{R}_{w \in W}$. and integer parameter $k >= 1$. Let (μ) be the mean of the defining sets and is defined as

$$\mu_i \coloneqq \sum_{w \in D_i} \frac{\vec{w}}{|D_i|} \tag{3.33}$$

The bias subspace B is defined as the first k rows of singular vale decomposition $SVD(C)$ where

$$C \coloneqq \sum_{i=1}^{n} \sum_{w \in D_i} \frac{(\vec{w} - \mu_i)^T (\vec{w} - \mu_i)}{|D_i|} \tag{3.34}$$

Then, to remove the bias, for words to neutralize $N \subseteq W$, family of equality sets $\varepsilon = E_i, E_2, ..., E_m$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to

$$\vec{w} \coloneqq \frac{(\vec{w} - \vec{w_B})}{||\vec{w} - \vec{w_B}||} \tag{3.35}$$

For each set $E \in \varepsilon$, let

$$\mu \coloneqq \sum_{w \in E} \frac{w}{|E|} \tag{3.36}$$

$$v \coloneqq \mu - \mu_B \tag{3.37}$$

for each $w \in E$,

$$\vec{w} \coloneqq v + \sqrt{1 - ||v||^2} \frac{\vec{w_B} - \mu_B}{||\vec{w_B} - \mu_B||} \tag{3.38}$$

Then output the bias subspace B and the new word embedding $\vec{w} \in \mathbb{R}_{w \in W}$. Finally, the authors equalize each word set outside B to their average v and then adjust vectors so that they are unit length.

U-Debias and P-Debias

In, as shown in Ding et al. [68], the authors propose two causal inference frameworks for reducing bias in static word embeddings while preserving lexical and semantic information. The authors consider five types of variables corresponding to five word-related matrices:

1. A s_1-dimensional purse gender bias variable D with a corresponding matrix $D \in R^{Nxs1}$ composed of pure gender bias vectors;

2. A s_2-dimensional gender bias proxy P with a corresponding matrix $P \in R^{NxS_2}$ composed of vectors that are directly influenced by D that should not affect the final prediction;

3. A m-dimensional resolving, non-gender-specific word variable Z with a corresponding matrix $Z \in R^{Nxm}$ composed of vectors that are influenced by D in a manner that I accept as non-discriminatory;

4. A d-dimensional non-gender-specific word variable Y with a corresponding matrix $Y \in R^{Nxd}$ composed of word vectors that contain gender bias, potentially, that needs to be removed;

5. A p-dimensional, non-gender-specific word variable X with a corresponding matrix $X \in R^{Nxp}$ that may retain semantic information.

The au thors define two types of bias in word embeddings:

1. **Potential proxy bias**: "A variable Y in a causal graph exhibits potential proxy if there exists a directed path from D to Y that is blocked by a proxy variable P and if Y itself is not a proxy."

   To remove potential proxy bias, the authors propose an algorithm for removing the gender bias from the non-gender specific word vectors y with α and β are parameters and e_1 and e_2 are unobserved errors.

$$P = D\alpha_0 + e_1 \tag{3.39}$$

$$X = D\alpha_1 + P\alpha_2 + e_2 \tag{3.40}$$

$$Y = P\beta_1 + X\beta_2 \tag{3.41}$$

   The non-gender-specific word matrix $\hat{Y}$ with potential proxy bias removed:

$$\hat{Y} = (X - P\hat{\alpha}_2)\hat{\beta}_2 \tag{3.42}$$

2. **Unresolved bias**: "A variable Y in a causal graph exhibits unresolved bias if there exists a directed path from D to Y that is not blocked by a resolving variable Z and Y itself is non-resolving".

   to remove unresolved bias, the authors propose an algorithm for removing the gender bias from the non-gender specific word vectors y with γ and θ are parameters and ε_1 and ε_2 are unobserved errors.

$$Z = D\gamma_0 + \varepsilon_1 \tag{3.43}$$

$$X = D\gamma_1 + Z\gamma_2 + \varepsilon_2 \tag{3.44}$$

$$Y = Z\theta_1 + X\theta_2 \tag{3.45}$$

The non-gender-specific word matrix $\hat{Y}$ with unresolved bias removed is

$$\hat{Y} = Z\hat{\theta}_1 \tag{3.46}$$

SentDebias

, as shown in Liang et al. [139] propose a method to use the Hard-Debias, as shown in Bolukbasi et al. [27] method, described above, to remove bias from contextual word embeddings. First, the authors contextualize the bias attributes b extracting sentences that contain gender-related words (e.g., man, woman) or religion-related words (e.g., Muslim, Christian, Jewish) from various text corpora like WikiText[1], Stanford Sentiment Treebank (SST)[2], and others. Then, the authors remove the bias subspace.

1. Contextualize words into sentences: returns a set of sentences S obtained by matching words with naturally occurring sentence templates from text corpora, where D is all the word tuples in the bias attribute words and S are the retrieved sentences.

$$S = \cup_{i=1}^{m} CONTEXUALIZE(w_1^{(i)}, ..., w_d^{(i)}) = (s_1^{(i)}, ..., s_d^{(i)})_{i=1}^{n}, |S| > |D| \tag{3.47}$$

2. Estimate the bias subspace: The retrieved sentences (S) are then passed through a LM (M_θ), e.g., BERT, parameters (θ) and extract all the sentences representations of the j^{th} entry in $d-$tuple as $R_j = \{M_\theta(s_j^{(i)})\}_{i=1}^{n}$. R_j is a vector space where a specific bias attribute is present across its contexts. The mean of set j is $\mu_j = \frac{\sum_{w \in R_j} W}{|R_j|}$. The bias subspace $V = \{v_1, ..., v_k\}$ is the first k component of principal component analysis (PCA).

$$V = PCA(\cup_{j=1}^{d} \cup_{w \in R)j} (w - \mu_j)) \tag{3.48}$$

3. Debias: The authors then get the debiased representation ($\hat{h}$) by removing the projection of a sentence (h) on the bias subspace (V).

[1] https://huggingface.co/datasets/wikitext
[2] https://paperswithcode.com/dataset/sst

$$h_v = \sum_{j=1}^{k} < h, v_j > v_j \tag{3.49}$$

Unlike, as shown in Bolukbasi et al. [27], sentDebias skip the Equalize step because it is hard to identify all the sentence pairs to be equalized due to the complexity of natural sentences.

$$\hat{h} = h - h_v \tag{3.50}$$

3.7.4 Limitations

, as shown in Gonen and Goldberg [90] demonstrates that removing post-processing is a superficial fix that does not change the underlying bias in static word embeddings, and, as shown in Kaneko et al. [117] demonstrates that removing bias using SentDebias from contextual debias does not have an impact on the fairness of the downstream tasks. As for using pre-processing methods to mitigate bias, using templates is not effective since they do not provide real contexts. Additionally, using an automatic Perturbator does not perform well, as I will show in Chapter 6, due to some problems in the PANDA dataset.

3.8 Discussion

It is clear that the sources of bias that I find in the NLP pipeline do not come out of nowhere, but have their roots in those that have been outlined in the social science, critical race theory and digital humanities studies (the Jim Code perspective). Despite this, the bias metrics that have been proposed in the NLP literature measure only pipeline bias, which has led to limitations in the currently proposed methods to measure and mitigate bias.

In this section, I outline these limitations and recommend measures to mitigate them.

3.8.1 Limitations of studying bias in NLP

The lack of scrutiny of the social background behind biases has led approaches to bias measurement to incorporate the same methods that introduced bias in the first place. For example, crowdsourcing the data used in measuring bias in language models, as shown in Nadeem et al. [167], Nangia et al. [172] reintroduces label bias into the metric that is supposed to measure bias. Moreover, studies that propose bias metrics in NLP do not incorporate the social science literature on bias and fairness, which results in a lack of articulation of what these metrics actually measure, and ambiguities and unstated assumptions, as discussed in, as shown in Blodgett et al. [26].

This results in limitations to the current bias metrics proposed and used in the NLP literature. One of these is that different bias metrics produce different bias scores, which makes it difficult to come to any conclusion on how biased the different NLP models are, as will demonstrated in chapter 4. There is also the limitation that current bias metrics claim to measure the existence of bias and not its absence, meaning that lower bias scores do not necessarily mean the absence of bias, as shown in May et al. [149], leading to a lack of conclusive information about the NLP models. Another consequence of the lack of understanding of what the bias metrics in NLP measure is that most of the research done on investigating the impact of social bias in NLP models on the downstream tasks could not find an impact on the performance of the downstream tasks, as shown in Elsafoury et al. [74], Goldfarb-Tarrant et al. [89] or the fairness of the downstream tasks, as shown in Cao et al. [39], Kaneko et al. [117].

Similarly, one of the main limitations of the proposed methods to measure individual fairness metrics is that the motivation behind the proposed metrics and what the metric actually measures are not disclosed. For example, Czarnowska et al. [53], Prabhakaran et al. [201], Qian et al. [204] propose metrics to measure individual fairness using counterfactuals without explaining the intuition behind their proposed methods and how these metrics meet the criteria for individual fairness.

As for group fairness metrics, they are all based on statistical measures that have come in for criticism. For example, Prabhakaran et al. [201] , as shown in Hedden [100] argues that group fairness metrics are based on criteria that cannot be satisfied unless the models make perfect predictions or that the base rates are equal across all the identity groups in the dataset. Base rate here refers to the class of probability that is unconditioned on the featural evidence, as shown in Bar-Hillel [18]. Hedden [100] goes on to ask if the statistical criteria of fairness cannot be jointly satisfied except in marginal cases, which criteria then are conditions of fairness?

Questioning the whole notion of using statistical methods to measure fairness, , as shown in Broussard [33] argues that some of the founders of the field of statistics were white supremacists, which resulted in skewed statistical methods and suggests that to measure fairness, maybe I should use non-statistical methods. Approaching the bias and fairness problem in NLP as a purely quantitative problem led the community to develop quantitative methods to remove the bias from NLP models like, as shown in Bolukbasi et al. [27], Liang et al. [139], Schick et al. [226] which resulted in only a superficial fix of the problem while the models are still biased, as shown in Gonen and Goldberg [90], Kaneko et al. [117]. As shown above, bias and fairness in NLP models are the results of deeper sources of bias, and

removing the NLP pipeline sources of bias would not lead to any real change unless the more profound issues from the social science perspective are addressed.

Similarly, all the efforts to make the models fairer rely on quantitative fairness measures that aim to achieve equity between different identity groups, when equity does not necessarily mean equality, as shown in Broussard [33]. As equality means that the NLP models give similar performances to different groups of people. However, in some cases, fairness or equity would require treating people of certain backgrounds differently. For example, Dias Oliva et al. [66] demonstrates that Facebook's hate speech detection models restrict the use of certain words considered offensive without taking into consideration the context they are being used. This led to the censoring of some of the comments written by members of the LGBTQ community, who proclaimed some of these restricted words as self-expression. In this case, equity did not lead to equality.

3.8.2 How to mitigate those limitations effectively?

Addressing the Jim Code sources of bias is not a simple task. However, by doing so, I can take steps towards developing more effective ways to make NLP systems more inclusive, fairer and safer for everyone. Here, I outline actionable recommendations for the NLP community and NLP researchers:

1. **Lack of context** can be addressed by incorporating social sciences as part of the effort of mitigating bias in NLP models. This is only possible through:

(a) *Interdisciplinary research* where scientists with backgrounds in fields such as critical race theory, gender studies and digital humanities studies are included in NLP project teams, so they can point out the social impact of the choices made by the NLP researchers.

(b) It can also be addressed by further *integration of the teaching of data and machine learning ethics into NLP curricula*, whereby students gain an understanding of the societal implications of the choices they make. Currently, they are typically only exposed to minimal and tokenistic treatment of the topics of bias and fairness in NLP models, which is insufficient to understand the roots of bias from a social science perspective. This should also include training in AI auditing, enabling students to assess the limitations and societal impact of the NLP systems they develop, as shown in Nobel [177].

2. **Lack of creativity** is a direct result of lack of context. I can address the lack of creativity by:

(a) NLP researchers gain *awareness of the social and historical context and the social impact of development choices*. This will encourage more creative methods to achieve their goals, instead of the reproduction of oppressive systems in shiny new packaging. Online competition and code-sharing platforms could be a place to start, for example, starting a Kaggle competition where participants develop new different NLP models that do not rely on $n-grams$ or objective functions that do not amplify societal biases.

(b) Another way to encourage NLP researchers to invest in that researcher direction is *specialized conferences and workshops on re-imagining NLP models, with an emphasis on fairness and impact on society*. This effort is already taking place with conferences like ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) [1]. The outcomes of these endeavours should be open for auditing, evaluation and reproducibility. One way to achieve that, without the controversy of open-source, is for NLP conferences to adopt the ACM artifact evaluation measures [2] and give reproducibility badges to published papers. This can be developed further to give social responsibility badges to the papers that were audited.

(c) *Specialized interdisciplinary fairness workshops* in major NLP conferences could encourage NLP researchers to collaborate with social scientists.

3. **Lack of diversity** can be addressed with:

(a) *Greater diversity on research teams* working on NLP problems. A more diverse perspective will be introduced to the research to make sure that the proposed solution and new systems are inclusive and work for everyone.

(b) *NLP conferences play a great role in promoting diversity* in NLP research by incorporating shared tasks that encourage researchers to work on low-resourced languages. For example, the shared NLP tasks [3] on Arabic, Persian, Korean, and others.

(c) *Incorporating more diversity workshops* in NLP conferences that allow researchers from different backgrounds to publish their work, e.g., the WiNLP workshop [4].

(d) This effort can go further by creating *shared tasks that test the impact of NLP systems on different groups of people*.

4. **Lack of accountability** The mentioned efforts should not be optional, and require enforcement with:

[1] https://facctconference.org/
[2] https://www.acm.org/publications/policies/artifact-review-badging
[3] http://nlpprogress.com/
[4] https://www.winlp.org/

(a) *State level regulation* to make sure that research is not conducted in a way that may harm society, which is only possible by holding universities and big tech accountable for the systems they produce. One step taken in this direction is the EU AI Act [1] which is a legislative proposal that assigns applications of AI to three risk categories that are described as

> *First, applications and systems that create an **unacceptable risk**, such as government-run social scoring of the type used in China, are banned. Second, **high-risk applications**, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.*

(b) There should also be an *AI regulation team* that works for the government that employs AI auditing teams and social scientists to approve newly developed NLP systems before they are released to the public, as shown in Broussard [33].

5. **Lack of awareness and Technochauvinism**, the suggested regulations, can only happen by electing people who are willing to put these regulations in place. This comes with raising awareness of the limitations of the current NLP systems. It is important that the public be aware that the doomsday scenario is not an AI system that outsmarts humans and controls them, but one that behaves like a Stochastic Parrot, as shown in Bender et al. [23] that keeps reproducing our discriminative systems on a wider scale under the mask of objectivity, as shown in Benjamin [24], Broussard [33], Nobel [177], O'neil [183]. Public awareness could be raised with:

(a) *Journalism* is an important resource to inform the public of the limitations and ethical issues in the current AI systems. Muckraking journalists in ProPublica, and The New York Times investigate AI technologies and share their investigations with the public, as shown in Broussard [33]. For example, the journalist's investigation of the COMPAS system and its unfairness was published by ProPublica.

(b) *Published Books for non-specialists* is another way to raise public awareness on issues related to discrimination in AI systems. Especially books that are targeted at non-specialists. For example, books like *Race after Technology, More than a Glitch, and Algorithms of Oppression*.

(c) *Talks* Academics and researchers should be encouraged to share their views on AI in non-academic Venus. For example, participating in documentaries like *Code Bias* [2] could bring awareness to the public.

[1] `https://artificialintelligenceact.eu/`
[2] `https://www.imdb.com/title/tt11394170/`

(d) *Museums* technology, science, and art museums could also raise public awareness of the limitations and potential dangers of AI. For example, in 222, the Modern Museum of Arts (MoMA), had an exhibition called "Systems" [1] that shows how AI systems work, the inequality within the system, and how many natural resources being used to build those systems.

(e) *Social media awareness campaigns* could be a way to reach more people, especially younger people.

3.9 Ethical statement

In this chapter, I aim to understand the roots of bias in NLP from the literature of social sciences. One of the risks of this work could be discouraging the quantitative research on bias and fairness in NLP. Or worse, I might make the research on bias and fairness in NLP seem daunting and requires collaborations and more effort than other research disciplines in NLP. Which might result in discouraging NLP researchers from working on bias and fairness in NLP models.

The aim from this work is not to discourage NLP researchers from working on bias and fairness in NLP, but to be more cautious and take a more inclusive approach to their research and to incorporate social scientists and social science literature.

3.10 Conclusion

In this chapter, I presented my second research contribution as a literature review on the historic forms of sexism, racism, and other types of discrimination that are being reproduced in the new age of technology on a larger scale and under the cover of supposed objectivity in NLP models. I reviewed the sources of bias in the NLP literature in addition to the social science, critical race theory, and digital humanities studies literature. I argue that the NLP bias sources are rooted in social sciences and that they are direct results of the sources of bias from the "Jim Code" perspective. I also demonstrate that ignoring the literature of social science in building unbiased and fair NLP models has led to unreliable bias metrics and ineffective debiasing methods. I argue that the way forward to eliminate the bias in NLP models is to incorporate the literature of Digital Humanities and Critical AI and Data studies and to increase collaborations with social scientists to make sure that these goals are achieved effectively without negative impacts on society and its diverse groups. Finally, I share a list

[1] `https://www.moma.org/collection/works/401279?sov_referrer=theme&theme_id=5472`

of actionable suggestions and recommendations with the NLP community on how to mitigate the discussed Jim Code sources of bias in NLP research.

After reviewing the literature on hate speech detection and bias and fairness in NLP, in the next chapters, I start the investigation on how bias in NLP models impacts the task of hate speech detection from three perspectives: the explainability perspective, the offensive stereotyping bias perspective, and the fairness perspective. In the next chapter, I present my third research contribution and investigate the explainability perspective.

Chapter 4

The Explainability Perspective

4.1 Introduction

The Pew Research Center reported in 2017 that 40% of social media users have experienced some form of hate speech, as shown in Abaido [1], Chan et al. [42], Duggan [70], Haddon and Livingstone [95]. These experiences have serious consequences for the victims, including depression, anxiety, low self-esteem and self-harm, as shown in Sticca et al. [243]. The goal of reducing these negative outcomes highlights the critical importance of improving the automatic detection of hate speech. On the other hand, it is equally crucial to understand the performance of hate speech detection models to make sure that they do not make the right decision for the wrong reasons with undesired side effects. For example, we want to avoid hate speech detection models that associate mentions of marginalised groups with hate, which is the case as demonstrated in Dixon et al. [69], Sap et al. [224].

In this chapter, I present my third research contribution, aiming to understand the impact of bias in NLP on the performance of hate speech detection models by investigating how that bias might explain the performance of hate speech detection models. I inspect the impact of two sources of bias:

1. *Pre-training*: I investigate the explainability of pre-trained contextual word embedding's performance on the task of hate speech detection and the bias that might result from pre-training. Additionally, I investigate social bias in contextual word embeddings and whether social bias explains the performance of contextual word embeddings on the task of hate speech detection.

2. *Biased pre-training dataset*: I investigate how pre-training static word embeddings on biased datasets collected from hateful social media platforms might impact the performance of hate speech detection models. In addition, I investigate social bias in

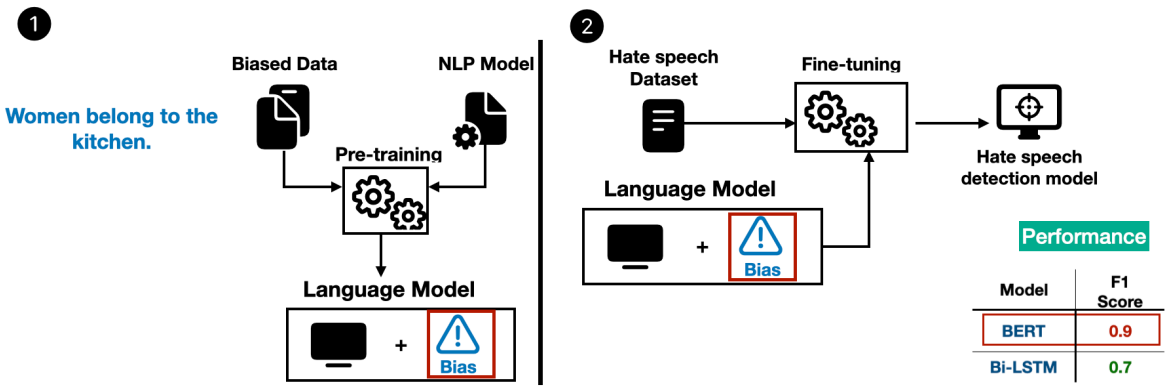


Fig. 4.1 Illustration of the work done for this chapter where I investigate hte impact of different type of pre-traning bias on the performance of hate pseech detection.

static word embeddings, and whether social bias explains the performance of static word embeddings on the task of hate speech detection.

This chapter is divided into two parts, corresponding to these two sources of bias. An illustration of the work done in this chapter is provided in Figure 4.1

4.2 Part 1: The impact of pre-training bias

Over the last decade, there have been attempts to use conventional machine learning models, as shown in Dadvar et al. [55], Dinakar et al. [67], Rafiq et al. [207] and deep learning models, as shown in Agrawal and Awekar [5], Kumar et al. [131], Raisi and Huang [210], Waseem and Hovy [282] to detect hate speech from social media. Recent studies have used attention-based language models, like BERT, in the detection of hate speech, as shown in Mozafari et al. [162], Paul and Saha [192], Pavlopoulos et al. [193], Yadav et al. [291]. However, those studies focused mainly on enhancing the performance of hate speech detection using BERT, without providing any analysis or insight into the model's inner workings. In this section, I aim to answer the following research questions:

1. (RQ1) How does bias resulting from pre-training NLP models explain their performance on the task of hate speech detection?
2. (RQ2) What is the impact of social bias in NLP models on their performance on the task of hate speech detection?

To answer these research questions and to investigate the impact of the bias resulting from pre-training contextual word embeddings on the task of hate speech detection, I, first, need

to gain a profound understanding of the performance of contextual word embeddings. So, I set out to answer the following set of questions:

- What is BERT's performance on different hate-speech-related datasets?
- What is the role that attention weights play in BERT's performance?
- What does BERT learn during fine-tuning?

Then, I investigate the impact of pre-training bias in contextual word embeddings by answering the following set of research questions:

- Does pre-training bias explain the performance of contextual word embeddings on the task of speech detection?
- Does social bias explain the performance of contextual word embeddings on the task of hate speech detection?

4.2.1 Related work

BERT, as shown in Devlin et al. [65] is a deep neural network model with an architecture based on stacked Transformer encoders, as shown in Vaswani et al. [273], which each consists of multiple layers, including a multi-head self-attention mechanism. Recent studies have applied BERT to hate speech detection. Paul and Saha [192] used a BERT-based model on various datasets, such as Twitter (hate speech), Wikipedia Talk Pages (personal attack) and Formspring (bullying), achieving F1-scores of 0.94, 0.91 and 0.92 respectively. Despite the reported results being very good, they over-sampled the datasets before the train/test split, which leads to over-fitting according to Arango et al. [12]. Mozafari et al. [162] proposed adding a CNN layer on top of $\text{BERT}_{\text{base}}$ for hate speech detection, achieving a maximum F1-score of 0.92. However, their proposed architecture could lead to over-fitting and a longer inference time. Although these studies indicate that BERT outperforms other models on the task of hate speech detection, none of them explains why. Recently, there has been substantial work on the explainability of NLP and language models, as shown in Adadi and Berrada [2], Sundararajan et al. [247], Zhang et al. [296]. Regarding attention-based models, like Transformers and BERT, as shown in Vig [275], Vig and Belinkov [276] built visualization tools to show the attention weights in different layers between tokens in the same sentence or in two different sentences, as well as to understand the role attention weights play in pre-trained BERT, as shown in Clark et al. [50] by analyzing the behavior of BERT's attention weights in different layers. Similarly,, as shown in Kovaleva et al. [124], Rogers et al. [216]

analyzed the capability of BERT to capture different types of linguistic information on the General Language Understanding Evaluation (GLUE)[1] benchmark. Regarding attention mechanisms and model explainability, Jain and Wallace [110] showed that contrary to the assumption that attention provides a form of explainability, attention weights do not provide meaningful explanations, with the same finding being supported by Serrano and Smith [228], Sun and Lu [246], Vashishth et al. [271]. Inspired by this work on the analysis of BERT models, the research goal of this part of the chapter, is to gain a more profound understanding of BERT's strong performance on hate speech detection tasks.

4.2.2 Methodology

To answer the first set of research questions related to BERT's performance (section 4.2), I compare fine-tuned BERT to state-of-the-art LSTM and Bi-LSTM models on five social media hate speech detection datasets from different sources and with different sizes. To examine how fine-tuning affects attention weights, I show the difference in attention weight patterns between BERT with and without fine-tuning. Then, to investigate the role of attention weights of fine-tuned BERT in the model's performance, I compare the mean feature importance score of individual tokens, obtained using Integrated Gradients, to their mean attention weights by computing the Pearson's linear correlation between the mean attention weights of fine-tuned BERT of all heads across the last layers (9-12) and the tokens' absolute importance score, as it has been shown that fine-tuning impacts mostly BERT's last layers (9-12), as shown in Rogers et al. [216]. Finally, I analyze the importance scores of POS tags of fine-tuned BERT to find out the features that BERT relies on to make its prediction.

To answer the second set of research questions related to the impact of bias resulting from pre-training (section 4.2), I use statistical significance tests to investigate whether bias resulting from pre-training explains the performance of BERT of the task of hate speech detection or not. Then, I investigate social bias in contextual word embeddings using social bias metrics proposed in the literature and use statistical correlation to investigate if social bias explains the performance of contextual word embeddings on the task of hate speech detection.

Hate speech datasets

I use five hate-speech-related datasets of varying sizes from several social media sources that contained different types of hate speech: (i) *Twitter-Racism*, a collection of Twitter

[1] `https://gluebenchmark.com/`

Dataset	Size	Positive samples	Avg.post length (words)	Max.post length (words)
Kaggle	7425	2578 (35%)	25.28	1419
Twitter-sex	14742	3370 (23%)	15.04	41
Twitter-rac	13349	1969 (15%)	15.05	41
WTP-agg	114649	14641 (13%)	75.45	2846
WTP-tox	157671	15221 (10%)	73.51	2320

Table 4.1 Hate speech datasets' statistics

messages containing tweets that are labeled as racist or not, as shown in Waseem and Hovy [282], (ii) *Twitter-Sexism*, Twitter messages containing tweets labeled as sexist or not, as shown in Waseem and Hovy [282], (iii) *Kaggle-Insults*, as shown in Kaggle [115], a dataset that contains social media comments that are labeled as insulting or not, (iv) *WTP-Toxicity*, a collection of conversations from Wikipedia Talk Pages (WTP) annotated as friendly or toxic, as shown in Wulczyn et al. [289], and (v) *WTP-Aggression*, conversations from WTP annotated as friendly or aggressive, as shown in Wulczyn et al. [289]. Information about the datasets is provided in Table 4.1.

Dataset pre-processing

For BERT, I follow the pre-processing steps described in Dang et al. [56]:

1. I remove URLs, user mentions, non-ASCII characters, and the retweet abbreviation "RT" (Twitter datasets).
2. All letters are lowercased.
3. Contractions are converted to their formal format.
4. A space is added between words and punctuation marks.

For the RNN models, in addition to the mentioned pre-processing steps, I remove punctuation and English stop words, as proposed in, as shown in Agrawal and Awekar [5]. However, second-person pronouns like "you", "yours" and "your", and third-person pronouns like "he/she/ they", "his/her/their" and "him/her/them" are not removed because I notice in the datasets that sometimes, profane words on their own, e.g., "f**k", are not necessarily used for bullying reasons, while their combination with a pronoun, e.g., "f**k you", is used to insult someone. Then, each dataset is randomly split into a training (70%) and test (30%) set, preserving class ratios.

Dataset	LSTM	Bi-LSTM	BERT(FT)
Kaggle	0.6420	0.653	0.768
Twitter-sex	0.6569	0.649	0.760
Twitter-rac	0.6400	0.678	0.757
WTP-agg	0.7110	0.679	0.753
WTP-tox	0.7230	0.737	0.786

Table 4.2 F1-scores achieved for each dataset

Model setting

BERT with fine-tuning is used for the task of text classification on the examined datasets, by employing $\text{BERT}_{\text{base}}$(uncased), as shown in Google Research [91]. For fine-tuning, I train BERT for 10 epochs with a batch size of 32 and a learning rate of $2e^{-5}$, as suggested in, as shown in Devlin et al. [65]. The sequence length parameter changed across datasets depending on their maximum token length. For the Twitter-sexism and Twitter-racism datasets, a sequence length of 64 is used because it is the closest to the maximum observed sequence length in the dataset. While 128 is used for the rest because it is the maximum, I could use due to available computational resources limitations. A single linear layer is added on top of the pooled output of BERT for sentence classification. I also use LSTM, as shown in Hochreiter and Schmidhuber [102] and Bi-directional LSTM, as shown in Schuster and Paliwal [227], with the same architecture as in, as shown in Agrawal and Awekar [5], which used RNN models to detect hate speech. To this end, I first use the Keras tokenizer, as shown in Tensorflow.org [260] to convert the text into numerical vectors (each integer is the index of a token in a dictionary) with a maximum length of 600 (the maximum I could use due to computational resources limitations) for the Kaggle and WTP datasets and 41 (maximum observed sequence length in the dataset) for the Twitter datasets. A trainable embedding layer is used as the first hidden layer of the LSTM and Bi-LSTM-based networks, with an input size equal to the number of unique tokens of the dataset after pre-processing and an output size of 128. The two models are then trained for 100 epochs with a batch size of 32, using the Adam optimizer and a learning rate of 0.01 which is the default of the Keras Optimizer.

Classification performance

The performance of the trained models on the test set is reported in Table 4.2. The initial training set for each model and dataset is randomly stratified-split into a training (70%) and validation (30%) set. The model is then trained using the training set, validated on the

validation set and tested on the original test set. This procedure is repeated five times, and the final performance of each model for each dataset is reported as the mean F1-score for the test set across the five iterations. From Table 4.2, it is evident that BERT with fine-tuning (FT) outperformed all the other examined models, reaching the highest F1-score of 78.6% for the WTP-Toxicity dataset. The Friedman test, as shown in Zimmerman and Zumbo [302] is used to compare the F1-scores of LSTM, Bi-LSTM and BERT (FT) across the five datasets, showing that BERT (FT) performed significantly better ($p < 0.05$). I then analyze the inner-workings of BERT to get insight into the reasons behind BERT's performance, starting with BERT's attention weights.

4.2.3 Attention weights (FT vs. NFT)

I examine the difference in attention weights' patterns between fine-tuned BERT (FT) and BERT without fine-tuning (NFT) on the Twitter-sexism dataset. To this end, I examine the attention weights of the five words with the highest probability for the hate speech class (according to a Multinomial Naive Bayes model) in BERT (FT) and BERT (NFT). From Fig. 4.2 (top), it is evident that the mean weights of the attention heads in the last layers of BERT (FT) (red lines) are much higher than for BERT (NFT) (blue lines). Which shows that the pattern of BERT (FT) in the last layers changed after fine-tuning compared to BERT (NFT). I repeate the same experiment using gradient-based importance scores, as shown in Sundararajan et al. [247] to get the most important words for the hate speech class and found a similar pattern, as shown in Fig. 4.2 (bottom). Similar results are observed for all the datasets: WTP, Kaggle and Twitter-racism.

4.2.4 Attention weights vs. importance scores

In the previous experiment, I demonstrate that fine-tuned BERT assigns higher attention weights to the last layers, compared to BERT without fine-tuning. This raises the following question: "*Do the attention weights of the last layers (9-12) of fine-tuned BERT explain the model's outcome?*" To answer this, I examine the correlation between gradient-based feature importance score and attention weights of fine-tuned BERT. Gradient-based feature importance scores provide a measure of the importance of individual features with known semantics, as shown in Sundararajan et al. [247] and have been used in previous studies for attention weights' analysis, as shown in Clark et al. [50], Serrano and Smith [228], Sun and Lu [246]. To compute the importance scores for all the datasets, I used the Integrated Gradients algorithm, as shown in Sundararajan et al. [247]. A subset of 1000 samples is randomly selected from the test set of each dataset, and the absolute importance scores of

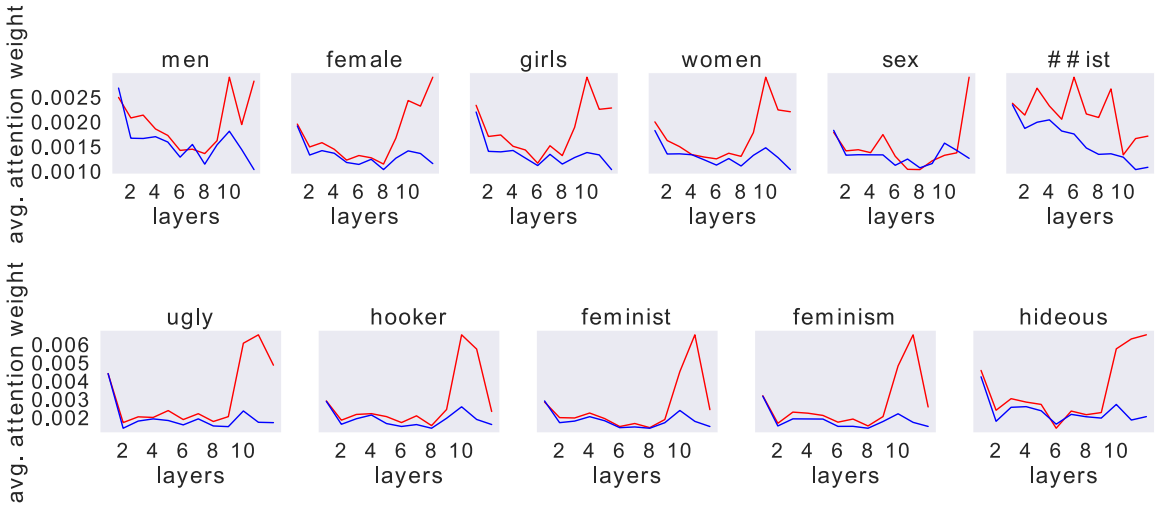


Fig. 4.2 Mean attention weights of 12 heads per layer for fine-tuned BERT (red) and BERT without fine-tuning (blue), for the most important hate speech class-related tokens in the Twitter-sexism dataset according to Naive Bayes (top) and gradient-based importance scores (bottom). The token "# # ist" is a subword generated by BERT.

all the tokens in these subsets are computed. Then, all scores are grouped by the tokens, and the mean absolute feature importance score is computed for each unique token. The same strategy is also followed for the attention weights. I compute the mean attention weight across all 12 heads per layer, as well as the mean attention weight of the last layers (9-12), where BERT's fine-tuning is most impactful. Then, I groupe the mean attention weights by tokens and computed the mean attention weights per each token. Pearson's correlation coefficient (ρ) is used to measure the linear correlation between the mean importance scores, the mean attention weights, and the occurrences of different tokens, as shown in Table 4.3.

The usage of ρ is inspired by early work on attention weights by, as shown in Jain and Wallace [110]. There is no linear correlation between the absolute importance score and the mean attention weights of BERT for the examined datasets ($0.056 \leq \rho \leq 0.171$), as well as between the number of occurrences of a token and the mean attention weights ($-0.101 \leq \rho \leq -0.015$) or the mean importance scores ($-0.011 \leq \rho \leq -0.002$). These results suggest that attention weights do not play a direct role in explaining BERT's performance, which is in line with previous studies, as shown in Serrano and Smith [228], Sun and Lu [246], Vashishth et al. [271].

As for the lack of positive correlation between the feature importance scores and the number of occurrences of tokens in each dataset suggest that the size of the dataset and the percentage of the positive examples in the dataset does not have a strong influence on

Dataset	**No. tokens**	**ρ (attention vs importance)**	**ρ (attention vs no. occurrences)**	**ρ (importance vs no. occurrences)**
Twitter-Sexism	3878	0.108	-0.047	-0.002
Twitter-Racism	3991	0.056	-0.015	-0.002
Kaggle-Insults	4452	0.171	-0.023	-0.004
WTP-Aggression	4457	0.125	-0.101	-0.009
WTP-Toxicity	4524	0.163	-0.076	-0.011

Table 4.3 Pearson correlation coeffient (ρ) between mean attention weights of fine-tuned BERT, mean absolute feature importance and number of occurrences per token

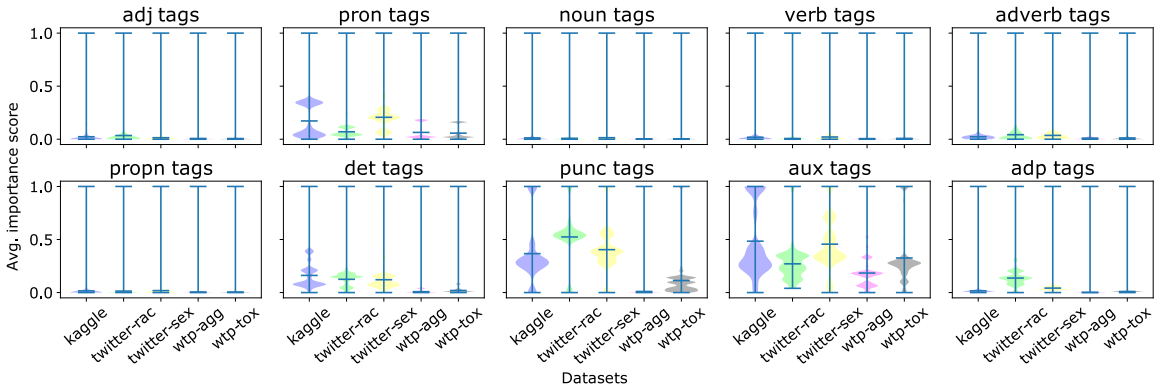


Fig. 4.3 Mean normalized importance scores assigned by fine-tuned BERT to POS tags in the datasets.

the model's performance. However, the size of the datastes and the percentages of positive examples could have an influence on the performance in a way that has not been inspected.

4.2.5 What does BERT learn during fine-tuning?

I use spaCy, as shown in Spacy [240] to compute the absolute gradient-based importance scores of the POS tags from the examined datasets and normalized them to the range $[0, 1]$ per dataset to examine whether BERT learns, during fine-tuning, general hate-speech-related features or if it relies on syntactical biases, which means the model relies only on a certain syntax to make its decision, in the dataset.

The hypothesis is that, if BERT learns hate-speech-related features, the POS tags that receive the highest importance scores will be nouns, adjectives, adverbs, proper nouns, and pronouns and that there will be similarities in the pattern across all the datasets. On the other hand, if BERT relies on syntactical biases, the POS tags that receive the highest importance

scores will be tagged like punctuation, auxiliaries, determiners, and adpositions and the patterns will differ across datasets from different domains. The reason behind using POS tags is that they are important linguistic features that can explicitly show the model's syntactical bias.

Results (Fig. 4.3) showed that the POS tags with the highest importance scores are auxiliaries, punctuation, determiners, adpositions, and pronouns. Among these, the most informative tag for hate speech detection is the pronoun. The distributions of the tags in Fig. 4.3 show similarities and differences across the datasets. These results suggest that BERT relies on syntactical bias, as a result of pre-training, for its good performance.

In the next section, I answer the second set of research questions, (section 4.2), to examine whether pre-training, syntactical, bias or social bias in contextual word embeddings explain their performance on hate speech detection.

4.2.6 Does pre-training bias explain the performance of contextual word embeddings on the task of hate speech detection?

To answer this research question, I use Wilcoxon sign-ranked test, as shown in Zimmerman and Zumbo [302] to test the statistical significance of the difference between the importance scores of the POS tags across different datasets (Table 4.4). Results indicated that a statistically significant difference could not be established between WTP-agg and WTP-tox and between Twitter-sexism and Twitter-racism ($p > 0.05$). I speculate that this happens because the domain of the datasets is the same. Similar results are found between Kaggle and Twitter-racism and between Kaggle and Twitter-sexism ($p > 0.05$). A statistically significant difference is shown between WTP-agg and Twitter-racism ($p = 0.001$), WTP-agg and Twitter-sexism ($p = 0.001$), WTP-tox and twitter-racism ($p = 0.048$), WTP-tox and Twitter-sexism ($p = 0.001$), Kaggle-insults and WTP-agg ($p = 0.001$), and Kaggle-insults and WTP-tox ($p = 0.001$). I speculate that this is because the domains of the datasets differ. The results support the hypothesis that BERT does not rely on semantic features related to hate speech, but instead relies on syntactic biases resulted from pre-training on dataset with syntactical composition that may change between different domains. These results also suggest that syntactical bias explains the performance of BERT on the task of hate speech detection.

I further inspected the POS tags with the highest importance scores, like auxiliaries, determiners, and punctuation across the different datasets. For **determiners** and **punctuation**, Kaggle, Twitter-sexism and Twitter-racism datasets, which have the highest scores for determiners and punctuation, contain less noise compared to WTP-agg and WTP-tox. The

	Kaggle	**Twitter-rac**	**Twitter-sex**	**WTP-agg**	**WTP-tox**
Kaggle	-	0.845	0.556	0.001	0.001
Twitter-rac	0.845	-	0.921	0.001	0.048
Twitter-sex	0.556	0.921	-	0.001	0.001
WTP-agg	0.001	0.001	0.001	-	0.064
WTP-tox	0.001	0.048	0.001	0.064	-

Table 4.4 p-values for the Wilcoxon sign-ranked test between the mean importance scores of the datasets.

noise here denotes that determiners or punctuation are mixed with other nouns and/or symbols, e.g., "anti-white". In contrast, **auxiliaries**, received the highest importance scores across all the datasets, since the detected auxiliaries did not have any noise in any of the datasets. I speculate that the noise is the cause of the low importance scores in WTP datasets. I also speculate that the domain of the dataset contributes to the amount of noise that can exist in the dataset. For example, Twitter does not allow long text, which means that even if mistakes and noise exist, the occurrences of noise are limited compared to a platform like Wikipedia Talk Pages where there is no text limit, thus allowing more space for noise. This provides additional evidence that the domain of the dataset affects its syntactical composition, the syntactic bias that BERT learns, and in turn, impacts BERT's performance and explainability. It also potentially limits its generalizability due to BERT learning syntactical biases instead of hate-speech-related linguistic features.

4.2.7 Social bias

The results from the last section suggest that BERT and potentially other contextual word embedding models learn syntactic bias during pre-training, and this bias explains the performance of BERT on the task of hate speech detection. In this section, I inspect social bias in BERT and other language models with different sizes and investigate if social bias in these models explains their performance on the downstream task of hate speech detection. I inspect three types of social bias (gender, racial and religious), in six models, BERT (base and large), as shown in Devlin et al. [65], RoBERTa (base and large), as shown in Liu et al. [142], and AlBERT (base and xx-large), as shown in Lan et al. [134]. I measure the bias in different model sizes to investigate the claim that bigger models contain more bias, as shown in Lin et al. [140] which has been shown for autoregressive models but not for MLM models.

The results in Table 4.5 indicate that the different bias metrics give different bias scores for the different models. When I use Pearson correlation to inspect how similar are the different bias metrics, I find no significant positive correlation. Moreover, when I use Wilcoxon significance test, I find that, unlike the finding of, as shown in Nadeem et al. [167], there is

CrowS-Pairs						
	BERT		RoBERTa		AlBERT	
Bias	Base	Large	Base	Large	Base	xx-Large
Gender	**0.580**	0.553	**0.606**	0.572	0.541	**0.649****
Race	0.581	**0.600**	0.527	**0.620****	0.513	**0.643****
Religion	0.714	0.685	**0.771**	0.714	0.590	**0.752****

StereoSet						
	BERT		RoBERTa		AlBERT	
	Base	Large	Base	Large	Base	xx-Large
Gender	0.602	**0.632**	**0.663****	0.535	0.599	**0.664****
Race	0.570	**0.571**	**0.616****	0.546	0.575	**0.611****
Religion	0.597	**0.599**	**0.642****	0.508	0.603	**0.696****

SEAT						
	BERT		RoBERTa		AlBERT	
	Base	Large	Base	Large	Base	xx-large
Gender	**0.620**	0.331	**0.939**	0.627	**0.622**	0.387
Race	**0.620**	0.516	0.307	**0.432**	**0.551**	0.309
Religion	**0.491**	0.185	0.126	**0.386**	0.430	**0.458**

Table 4.5 Bias scores in base and large models using the different bias metrics. **Bold** scores mean higher bias scores and more biased models. ** means statistically significant higher bias score.

no significant difference between the bias in the base models and the large models. Except for ALBERT-base and ALBERT-xx-large where the bias in AlBERT-xx-large is significantly higher than AlBERT-base according to CowS-Pairs and StereoSet but not SEAT. These results suggest that large models are not necessarily more biased than base models, but if the model size gets even bigger, like ALBERT-xx-large, then the models might get significantly more biased. Similar results have been demonstrated by Baldini et al. [16] but for extrinsic bias (fairness scores) but not for intrinsic bias which is measured in this section. Since there is no significant difference between the base and large models, I only use base language models in the rest of the thesis.

4.2.8 Does social bias explain the performance of contextual word embeddings on the task of hate speech detection?

To answer this question, I follow the work done in, as shown in Goldfarb-Tarrant et al. [89] and examined the correlation between social bias scores of three different contextual word embeddings and the F1-scores of the inspected models on the task of hate speech detection.

Dataset	**Racial Bias**		
	CrowS-Pairs	**StereoSet**	**SEAT**
Kaggle	0.500	0.581	-0.482
Twitter-sexism	-0.993	0.307	-0.415
Twitter-racism	0.764	0.270	-0.587
WTP-aggression	-0.075	-0.876	0.814
WTP-toxicity	-0.618	0.971	-0.992

Table 4.6 Pearson correlation coefficient (ρ) between the racial bias scores of the different word embeddings and the performance of hate speech detection task. **Bold** ρ means the strongest positive correlation among the bias metrics.

Dataset	**Gender Bias**		
	CrowS-Pairs	**StereoSet**	**SEAT**
Kaggle	0.973	0.690	0.649
Twitter-sexism	-0.395	0.169	0.223
Twitter-racism	0.837	0.403	0.353
WTP-aggression	-0.976	-0.935	-0.915
WTP-toxicity	0.483	0.978	0.947

Table 4.7 Pearson correlation coefficient (ρ) between the gender bias scores of the different word embeddings and the performance of hate speech detection task.

So first, I fine-tune BERT-base, AlBERT-base and RoBERTa-base on the datasets described in Table 4.1 with 40% training set, 30% validation set and 30% test set. I train the models for 3 epochs, using a batch size of 32, a learning rate of $2e^{-5}$, and a maximum text length of 61 tokens. Then, I compute the Pearson correlation coefficient (ρ) between the F1-scores and the social bias scores of the base models, reported in Table 4.5. The correlation coefficient values reported in Tables 4.6, 4.7, and 4.8 show that there is a positive correlation between racial bias scores measured by StereoSet metric and the performance of the different models on most of the hate speech datasets (Kaggle, Twitter-sexism, WTP-toxicity). On the other hand, for gender and religion bias, the bias scores measured by the CrowS-Pairs metric correlate positively with the F1-scores of the models on Kaggle and Twittter-racism, while the bias scores measured by the StereoSet metric correlate positively with the performance of the models on Twitter-sexism and WTP-toxicity.

The results suggest that social bias scores measured by the StereoSet and CrowS-Pairs metrics correlate positively with the performance of the models on most of the datasets. However, the positive correlations are not statistically significant and inconsistent with all the datasets. This lack of consistent positive correlation could be due to the limitations of the proposed metrics to measure social bias in the literature, as discussed in chapter 3. This

Dataset	Religion Bias		
	CrowS-Pairs	StereoSet	SEAT
Kaggle	0.992	0.555	-0.528
Twitter-sexism	-0.492	0.336	-0.366
Twitter-racism	0.891	0.240	-0.209
WTP-aggression	-0.947	-0.861	0.844
WTP-toxicity	0.483	0.978	-0.984

Table 4.8 Pearson correlation coefficient (ρ) between the religion bias scores of the different word embeddings and the performance of hate speech detection task.

means that the impact of social bias in contextual word embeddings on their performance on hate speech detection is inconclusive.

4.2.9 Summary

To summarize the findings of this part and to answer the first research question 1: *How does bias resulting from pre-training NLP models explain their performance on the task of hate speech detection?*, the results in Table 4.4 and Figure 4.3 suggest that syntactic bias resulting from pre-training BERT explains its performance on the task of hate speech detection. To answer the second research question 2: *What is the impact of social bias in NLP models on their performance on the task of hate speech detection?*, the results in Table 4.6, Table 4.7, and Table 4.8 suggest that, unlike syntactical bias, social bias does not explain the performance of contextual word embeddings on hate speech detection. However, as explained in Chapter 3, social bias metrics used in the literature have their limitations and that might explain the lack of consistent positive correlation between social bias scores and the performance scores of the hate speech detection models.

As the results demonstrate, syntactical bias, which results from pre-training BERT, impacts and explains the performance of BERT on hate speech detection. Next, I investigate the impact of pre-training NLP models on biased datasets.

4.3 Part 2: The impact of biased pre-training datasets

Static word embeddings have been widely used for the task of hate speech detection. Some of these word embeddings are pre-trained on informational data like news articles, e.g., Word2vec, as shown in Mikolov et al. [156], or Wikipedia articles, e.g., Glove, as shown in Pennington et al. [196]. I use the term "informational-based" to describe these word embeddings. More recently, there have been new word embedding models pre-trained on

Word Embeddings	Similar words to "queer"
Word2vec	genderqueer, LGBTQ, gay, LGBT, lesbian
Glove-WK	transgender, lesbian, lgbt, lgbtq, bisexual
Glove-Twitter	fag, faggot, feminist, gay, cunt
Urban Dictionary	fag, homo, homosexual, bumblaster, buttyman
Chan	faggot, metrosexual, fag, transvestite, homo

Table 4.9 Top 5 similar words retrieved by each of the word embeddings.

more biased data collected from mainstream social media platforms like Twitter and less popular controversial social media platforms like 4&8 Chan and Urban Dictionary. I use the term "social-media-based" to describe those word embeddings. These informal sources are biased, as they contain racial slurs and forms of profanity that do not exist in formal text, as shown in Türker et al. [266]. However, these social-media-based word embeddings have not been investigated for social NLP related tasks like hate speech detection and social bias analysis. The intuition that social-media-based word embeddings could be better at detecting hate speech, comes from the examples shown in Table 4.9, where I display the most similar five words found by each word embeddings to the word "queer". The informational-based word embeddings return non-offensive words while social-media-based word embeddings return offensive[1] words.

In the second part of this chapter, I investigate how pre-training NLP models on biased datasets collected from hateful social media platforms might impact the performance of these NLP models on the task of hate speech detection. To this end, I set out to answer the following research questions:

1. (RQ1) How do biased pre-training datasets impact the performance of NLP models on the task of hate speech detection?

2. (RQ2) What is the impact of social bias in NLP models on their performance on the task of hate speech detection?

To answer these research questions and to investigate whether pre-training NLP models on biased pre-training datasets explains their performance on hate the task of speech detection, first, I need to compare the performance of different static word embeddings based on the dataset they are pre-trained on different tasks related to hate speech detection. So, I aim first to answer the following set of research questions:

[1]Throughout this chapter, I differentiate between the terms "offensive" and "profane": I use the term "offensive" to describe an expression that is offensive to a group of people but not necessarily profane e.g. "women belong to the kitchen" while I use the term "profane" to describe expressions like "b*tch".

- What is the performance of the different word embeddings on offenses' categorization?
- What is the performance of the different word embeddings on the task of hate speech detection?
- Can we use certain static word embeddings to detect certain offensive categories within hate-speech-related datasets?

Then, I move on to answer the following set of research questions related to biased pre-training datasets and its impact on the performance of static word embeddings regarding hate speech detection.

- Do biased pre-training datasets explain the performance of static word embeddings on hate speech detection?
- Are social-media-based word embeddings more socially biased than informational-based word embeddings?
- Does social bias explain the performance of static word embeddings on the task of hate speech detection?

4.3.1 Related work

Recent word embeddings pre-trained on data from social media platforms have been released in the community. For example, Urban Dictionary word embeddings that is pre-trained on words and definitions from the Urban Dictionary website, as shown in Wilson et al. [287] using the FastText framework, Chan word embeddings that is pre-trained on 4&8 Chan websites using Continuous Bag-of-Words algorithm (CBOW), as shown in Voué et al. [277], and a version of Glove pre-trained on Twitter data, as shown in Mozafari et al. [162]. Even though there is evidence from the literature that the data that is used in pre-training these word embeddings contain offensiveness and racially insensitive comments, as shown in Nguyen et al. [175], Papasavva et al. [188], their impact on NLP tasks, has not been investigated. For example, investigating the impact of social-media-based word embeddings on the task of hate speech detection or analyzing social bias in the social-media-based word embeddings.

Using social-media-based word embeddings could improve hate speech detection, as they may be able to identify some offensive words or forms of profanity that are not captured by informational-based word embeddings. Comparative studies on word embeddings and deep learning models have been done for biomedical natural language processing, as shown in Wang et al. [280] and for text classification, as shown in Wang et al. [279], but there have

been very few similar comparative studies for the task of hate speech detection. Jain et al. [109] reviewed the literature on different word embeddings: CBOW, Skip-gram, ELMo, GloVe and fastText, and then tested them with a neural network model on the hate speech detection task. They show that ELMo is the best performing, followed by fastText and GloVe. However, they do not include social-media-based word embeddings like Urban Dictionary or Chan. Elsafoury et al. [73] have shown that word embeddings pre-trained on Urban Dictionary, and Twitter outperforms embeddings like Word2vec and Glove-Wikipedia on the task of hate speech detection. However, they do not compare the ability of the different word embeddings to categorize offensive words or to detect different categories of offenses within hate speech datasets.

Additionally, The research has shown that word embeddings are biased. Among the most common methods for quantifying bias in word embeddings are WEAT, RND, RNSB, and ECT. For the WEAT metric, the authors are inspired by the Implicit Association Test (IAT) to develop a statistical test to demonstrate human-like biases in word embeddings, as shown in Caliskan et al. [38]. They used the cosine similarity and statistical significance tests to measure the unfair correlations for two different demographics, as represented by manually curated word lists. As for the RND metric, the authors used the Euclidean distance between neutral words, like professions, and a representative group vector created by averaging the word vectors for words that describe a stereotyped group (gender/ethnicity), as shown in Garg et al. [86]. As for the RNSB metric, the authors trained a logistic regression model on the word vectors of unbiased labeled sentiment words (positive and negative) extracted from biased word embeddings. Then, that model is used to predict the sentiment of words that describe certain demographics, as shown in Sweeney and Najafian [254]. In the ECT metric, the authors proposed a method to measure how much bias has been removed from the word embeddings after debiasing them, as shown in Dev and Phillips [64]. These bias metrics have been used to measure the bias in Word2vec, as shown in Caliskan et al. [38], Dev and Phillips [64], Garg et al. [86], Sweeney and Najafian [254], Glove-WK, as shown in Dev and Phillips [64], Sweeney and Najafian [254], Glove-Twitter, as shown in Dev and Phillips [64]. Even though research has shown that the upstream data used to pre-train the social-media-based word embeddings, especially Urban Dictionary and Chan, are full of racial slurs and profanity, as shown in Nguyen et al. [175], Voué et al. [277], none of these studies measured social bias in Urban Dictionary or Chan word embeddings. In this chapter, I run a series of experiments to fill the mentioned gaps in the literature and to answer the research questions.

Word embedding	Pre-training data	Type
Word2Vec	Google news articles	informational-based
Glove-Wikipedia	Wikipedia articles	informational-based
Glove-Twitter	Twitter messages	social-media-based
Chan	Text from 4&8 Chan	social-media-based
Urban Dictionary	Text from Urban Dictionary	social-media-based

Table 4.10 Static word embedding models used in the chapter.

4.3.2 Methodology

To answer the first set of research questions related to the performance of static word embeddings, section 4.3, I use different word embeddings to categorize terms from a popular lexicon of the English offensive language. Then I compare the performance of the social-media-based word embeddings and the informational-based word embeddings using statistical significance tests. This should help us find out whether social-media-based word embeddings are significantly better than informational-based word embeddings at learning the semantic relationship between terms that belong to the same group of offenses. Then, I use each set of word embedding to detect hate speech automatically in hate-speech-related datasets and to detect different types of hate speech within each dataset. I use a statistical significance test to compare the performance of the social-media-based word embeddings and the informational-based word embeddings.

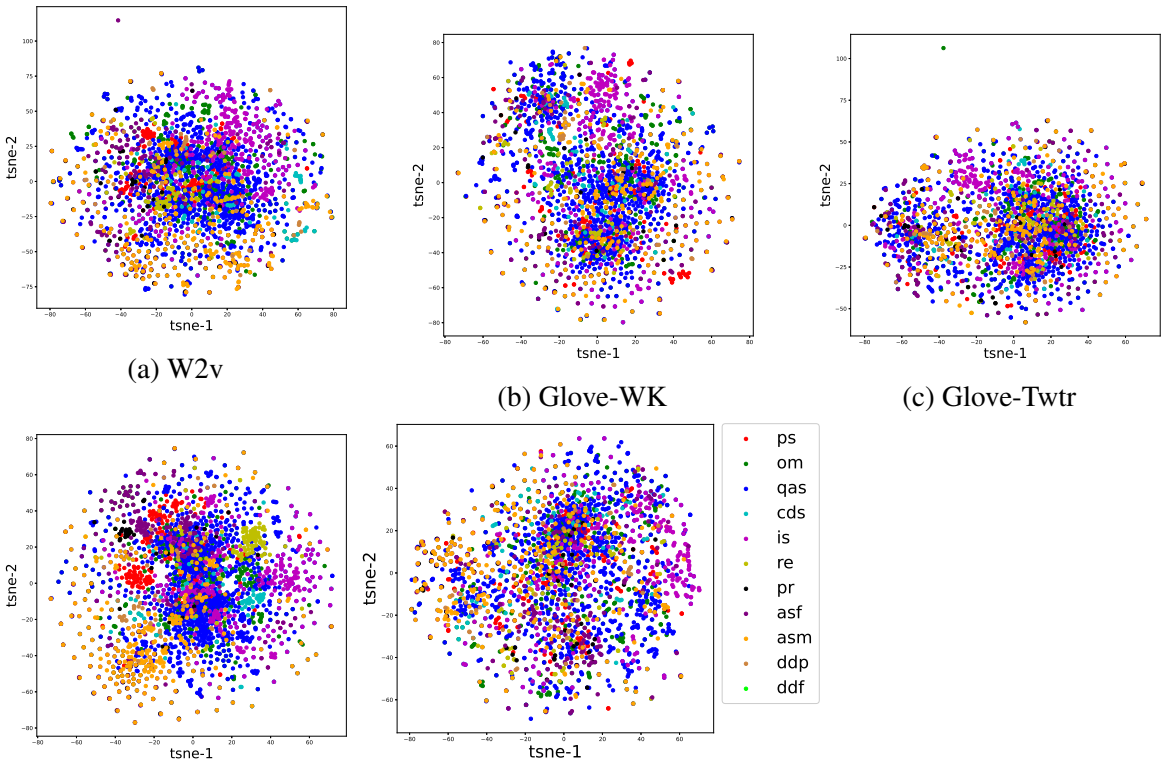
To answer the second set of research questions, section 4.3, related to the bias in static word embeddings, I use the state-of-the-art metrics from the literature to measure gender and racial bias in each word embedding and compared the bias scores in the social-media-based word embeddings and the informational-based word embeddings. Then, I use statistical correlation to investigate whether the measured social bias scores in different word embeddings explain their performance on the task of hate speech detection.

4.3.3 Offenses categorization

In this part of the chapter, I use the word embedding models that are summarized in Table 4.10. To answer the research questions, I use the English offensive categories introduced in Hurtlex lexicon, as shown in Zhang et al. [294], which is a multilingual lexicon containing 8228 offensive words and expressions, which are organized into 17 groups. I only use words that belong to 11 groups because they are related to the types of hate speech found in the datasets. The use categories are summarized in Table 4.11. I extract the word vectors, using the different word embeddings described in Table 4.10, for each word in those 11 groups and

Category	Description
PS	ethnic slurs
IS	words related to social and economic disadvantage
QAS	descriptive words with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
PR	words related to prostitution
OM	words related to homosexuality
ASF	female genitalia
ASM	male genitalia
DDP	cognitive disabilities
DDF	physical disabilities

Table 4.11 Hurtlex categories use in this chapter. The Category names (abbreviations) are in Italian. I use only the English lexicon where the tokens are in English.



(a) W2v (b) Glove-WK (c) Glove-Twtr

(d) UD (e) Chan

Fig. 4.4 t-SNE of the different static word embeddings of the words that belong to different groups in Hurtlex lexicon.

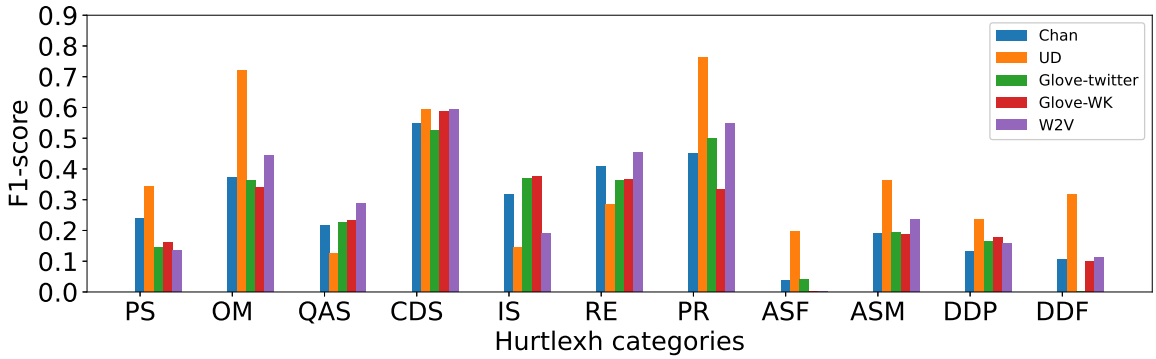


Fig. 4.5 F1 scores of the KNN model with the different word embeddings on Hurtlext test set.

projected them into a two-dimensional space using t-SNE, as shown in van der Maaten and Hinton [269] as shown in Figure 4.4. The plot shows words from some Hurtlex categories clustered better in some cases, especially, PS, PR, and ASM with Urban Dictionary. This cluastering suggest that some word embeddings like Urban dictionary and word2vec are better at categorizing the Hurtlex words that belong to the same group.

To quantitatively investigate the ability of the different word embeddings to group the words that belong to the same Hurtlex category, I use a KNN model. First, I remov the words in the lexicon that belong to more than one category, which resulted in, 5963 offensive words. I then split Hurtlex lexicon into training (70%) and test (30%) sets, with class ratio preserved. Next, to understand if the neighbors of a given word typically belong to the same class as that word, I use the trained KNN model to predict the category of each word embedding in the test set based on proximity to embeddings from the training set. I measure the F1-scores and plot them in Figure 4.5.

The results indicate that for most of Hurtlex categories, PS, OM, PR, ASF, ASM, DDP and DDF, Urban Dictionary is the best performing, meaning that it is the best at grouping together the words that belong to these categories. For QAS and RE, Word2vec is the best performing and for IS, Glove-Wikipedia and Glove-twitter are the best performing. For CDS, all the word embeddings are performing similarly, with Urban Dictionary embedding being the best performing by a small margin. I speculate that these results stem from the fact that the Urban Dictionary is pre-trained on words and definitions that are of insulting nature in general, and to women and minorities specifically, so it is better at finding more profanity related to these categories: PS, OM, PR, ASF, ASM, DDP and DDF. Word2vec, on the other hand, is better at clustering the word vectors that are related to felonies and words related to crime and immoral behavior (RE) and words with potential negative connotations (QAS). That may be due to its pre-training on news articles, which sometimes report on crimes.

Using a Friedman significance statistical test, as shown in Zimmerman and Zumbo [302] ($\alpha = 0.05$) between the F1 scores of each data item in the test set, I find that the F1 scores achieved by the word embeddings are significantly different. To further investigate the difference between pairs of top-scoring word embeddings, I use a Wilcoxon test, as shown in Zimmerman and Zumbo [302] ($\alpha = 0.05$). I find that, across all categories, Urban Dictionary scores significantly higher than Chan and Glove-Wikipedia but not significantly higher than Word2vec or Glove-Twitter. Similarly, I find that Word2vec achieves a significantly higher F1 score than Chan and Glove-Wikipedia, but not significantly higher than Glove-Twitter. The results suggest that the Urban Dictionary embeddings, along with Word2vec and Glove-twitter, place offensive words semantically close to other words from the same Hurtlex categories, indicating that these embeddings better reflect the categorization of terms outlined in Hurtlex.

4.3.4 Hate speech detection

In the light of the earlier results presented in Figure 4.5, I make two hypotheses: (1) social-media-based word embeddings will perform better than informational-based embeddings on the task of hate speech detection. (2) Certain word embeddings will perform better at detecting certain offensive categories within the hate-speech-related datasets. Specifically, I expect that Urban Dictionary embeddings might perform the best on the examples in the datasets containing PS, OM, PR, ASF, ASM, DDP and DDF categories; Word2vec embeddings to perform the best on examples containing RE and QAS; and for the CDS category, I expect all the models to perform similarly. To test these hypotheses and answer the second research question, I compare the performance of the different word embeddings when used to initialize the embedding layer of a deep learning model trained on the following datasets.

Hate speech datasets

I use five hate-speech-related datasets from several social media sources that contain different types of hate speech. I use the (i) *Twitter-sexism*, (ii) *Twitter-racism* and the (iii) *Kaggle* datasets thata re used in part 1 of this thesis. In addition to those datasets, I use the following datasets: (iv) *HateEval*, a collection of tweets containing hate speech against immigrants and women in Spanish and English [20]. I use only the English tweets; (v) *Jigsaw-tox*, a collection of Civic Community comments which have been labeled by human raters for toxicity [29]. The dataset' statistics are described in Table 4.12. I made ethe decision to

Dataset	Size	Positive samples	Avg.post length (words)	Max.post length (words)
HateEval	12722	42%	21.75	93
Kaggle	7425	65%	25.28	1419
Twitter-sex	14742	23%	15.04	41
Twitter-rac	13349	15%	15.05	41
Jigsaw-tox	99738	6%	54	2321

Table 4.12 Hate speech dataset statistics. Positive samples is the percentage of positive (bullying) comments. Avg. is the average number of words per comment. Max. is the maximum number of words in a comment.

replace the Wikipedia talk pages dataset with the Jigsaw-tox and the HateEval dataset because the focus of this part of the chapter is on using social media datasets.

Pre-processing datasets

To pre-process the datasets, I remove URLs, user mentions, and non-ASCII characters; All letters are lowercased; common contractions are converted to their full forms. I also remove English stop words, as proposed in Agrawal and Awekar [5]. However, second-person pronouns like “you”, “yours” and “your”, and third-person pronouns like “he/she/they”, “his/her/their” and “him/her/them” are not removed because I notice in the datasets that sometimes, profane words on their own, e.g. “f**k”, are not necessarily used in an offensive way, while their combination with a pronoun, e.g. “f**k you”, is used to insult someone. For Twitter datasets, I also remove the retweet abbreviation “RT”. Each dataset is randomly split into training (70%) and test (30%) sets with preserved class ratios. Additionally, to find out the different categories of offenses within each hate speech dataset, I filter the datasets using the words in the Hurtlex lexicon. Then I sort the data items in each dataset into the 11 Hurtlex categories based on the words present in the data items. Those that contain a mix of words from multiple Hurtlex categories are grouped in a Mixed category, and all the data items that do not contain any Hurtlex words are placed in a No-Hurtlex category. The results show that for all the datasets, the majority of data items contain words that do not belong to any Hurtlex category (No-hurtlex) with a percentage range from 40% to 66%. The second most present category in all the datasets is the Mixed category, where the data items contain words from multiple Hurtlex categories with percentages ranging from 5% to 25%. For the data items that contain words from only one Hurtlex category, the datasets, are less than 10% except for the CDS category where the percentage is less than 20%. When I investigate the distribution of the different categories in the Mixed group, I find a similar distribution of the

11 categories in all the datasets, with the majority belonging to the CDS category. When I investigate the data items in the No-Hurtlex category, I find some non-profane form of offensiveness.

Model settings

I use a Bi-directional LSTM, as shown in Schuster and Paliwal [227], with the same architecture as in Agrawal and Awekar [5], who used RNN models to detect hate speech. To this end, I first use the Keras tokenizer, as shown in Tensorflow.org [260] to tokenize the input texts, using a maximum input length of 64 (maximum observed sequence length in the dataset) for the HateEval and Twitter datasets and 600 for the Kaggle and Jigsaw datasets (due to computational resource limitations). A frozen embedding layer, based on a given pre-trained word embedding model, is used as the first layer and fed to the Bi-LSTM model. To avoid over-fitting, I use L2 regularization with an experimentally determined value of 10^{-7}. The model is then trained for 100 epochs with a batch size of 32, using the Adam optimizer and a learning rate of 0.01.

Classification performance

I analyze the overall performance of each word embeddings on each dataset, the "Average" column in Table 4.13, individually and across all the datasets. I use the Friedman statistical significance test, as shown in Zimmerman and Zumbo [302] ($\alpha = 0.05$) to compare the F1-scores of each word embeddings for the 13 categories (PS, OM, QAS, CDS, IS, RE, PR, ASF, ASM, DDP, DDF, No-hurtlex and Mixed) in each dataset.

The results show that social-media-based word embeddings gave the best results for four out of five datasets: HateEval, Kaggle, Twitter-racism and Jigsaw-toxicity. For the HateEval dataset, performance across all the categories is at its best when Glove-Twitter, social-media-based, is used with an average F1 score of 0.620. However, the results across all the categories are not significantly better than the rest of the word embeddings with $p - value > 0.05$. Glove-Twitter also resulted in the highest average F1 score at 0.519, across all the categories on the Jigsaw-toxicity dataset, which is significantly better for all the categories with $p - value < 0.05$. The best performing word embeddings on the Kaggle dataset is also the social-media-based word embeddings, Chan, with the average F1-score of 0.727 across all the categories with the results significantly better than the rest of the word embeddings for all the categories with $p - value < 0.05$. Urban Dictionary embeddings, social-media-based, gave the best results on the Twitter-racism dataset with the average F1 score of 0.663 across all the categories. These results are significantly better

with $p-value < 0.05$. The informational-based word embeddings, Glove-Wikipedia, gives a significantly better average F1-score of 0.699 across all the categories on the Twitter-sexism dataset with $p-values < 0.05$. Overall, I find that although social-media-based word embeddings outperform others on four out of five datasets, the difference is only significant in three cases.

Then, I analyze the results across the different types of hate speech in the datasets, I computed the mean F1-score achieved by each word embedding for each category across all datasets. When I compared the mean F1-score achieved by each word embedding for each category across all datasets using a Friedman significance statistical test ($\alpha = 0.05$), I find no significance for any of the 13 categories (PS, OM, QAS, CDS, IS, RE, PR, ASF, ASM, DDP, DDF, No-hurtlex and Mixed). This might occur because there is no clear connection between the ability of word embeddings to cluster the Hurtlex categories and their performance on texts that contain the same offensive words in hate speech-related datasets. Alternatively, due to the minimal percentages of these categories in the datasets, it is possible that I could not get a reliable enough indication of the performance of each word embedding model on each category. More analysis and experiments with larger datasets where these categories are more prevalent are needed to fully understand the results.

After investigating the performance of the different word embeddings, in the next section, I start investigating the impact of using a biased pre-training dataset and social bias in static word embeddings on the task of hate speech detection.

4.3.5 Do biased pre-training datasets explain the performance of static word embeddings on hate speech detection?

To summarize these findings and answer this research question, the results demonstrate that word embeddings that are pre-trained on biased data, social-media-based, outperform informational-based word embeddings on the tasks of offenses categorization and hate speech detection. These results suggest that using biased pre-training datasets with NLP models impacts and explains their performance on the task of hate speech detection.

Next, I inspect social bias in static word embeddings and whether it explains their performance on the task of hate speech detection.

4.3.6 Social bias

In this section, I investigate social bias in the different word embeddings. I investigate two types of social bias: gender bias and racial bias. I hypothesize that social-media-based word embeddings, especially Urban Dictionary and Chan, are more socially biased than

HateEval														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.615	0.444	0.615	**0.666**	0.555	**0.647**	0.658	0.421	0.555	**0.857**	0.5	0.570	0.730	0.602
UD	**0.7**	0.444	0.571	0.603	0.533	0.562	0.678	0.4	0.603	0.571	0.375	0.508	0.734	0.560
Glove-Twitter	0.695	**0.5**	**0.736**	0.663	0.631	0.619	**0.711**	0.620	0.690	0.571	0.285	**0.605**	**0.738**	**0.620**
Glove-WK	0.583	0.222	0.571	0.616	**0.666**	0.515	0.614	**0.72**	0.691	**0.857**	0.333	0.535	0.699	0.586
W2V	0.315	**0.5**	0.666	0.648	0.631	0.514	0.614	0.714	**0.72**	0.571	**0.666**	0.593	0.705	0.604
Kaggle														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.380	**0.777**	**1**	**0.760**	0.571	**0.545**	0.571	**1**	**0.666**	0.916	**0.909**	0.571	0.783	**0.727**
UD	**0.72**	0.761	**1**	0.703	**0.75**	0.461	0.75	0.666	0.507	0.888	0.8	**0.611**	**0.813**	0.725
Glove-Twitter	0.454	0.727	0.444	0.627	0.727	0.285	**0.823**	0	0.520	**0.923**	0.8	0.513	0.790	0.587
Glove-WK	0.5	0.625	**1**	0.588	0.666	0.5	0.666	0.666	0.507	0.869	0.666	0.525	0.8	0.660
W2V	0.352	0.375	**1**	0.602	0.25	0.4	0.714	**1**	0.526	0.818	0.666	0.479	0.797	0.614
Twitter-sexism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.666	0.829	0.421	0.523	0.695	0.4	0.45	0.6	0.510	0.666	0.56	0.561	0.586	0.574
UD	0.666	0.8	0.521	0.656	0.75	0.510	0.608	**0.923**	0.622	**0.75**	**0.687**	0.629	0.695	0.678
Glove-Twitter	0.666	**0.863**	0.380	0.640	**0.8**	0.5	0.693	**0.923**	**0.653**	0.571	0.645	0.631	0.702	0.667
Glove-WK	0.666	0.818	**0.608**	**0.686**	0.740	**0.655**	0.734	0.727	0.636	**0.75**	0.685	**0.675**	0.708	**0.699**
W2V	**0.727**	0.772	0.571	0.598	0.695	0.56	**0.769**	0.833	0.623	**0.75**	0.666	0.650	**0.730**	0.688
Twitter-racism														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	**0.76**	0.736	0.8	0.732	0.5	**0.809**	**0.4**	0	0.428	0.588	**1**	0.671	0.784	0.631
UD	0.754	**0.956**	**0.909**	0.762	**0.6**	0.8	0.333	0	0.571	0.583	0.909	0.658	0.783	**0.663**
Glove-Twitter	0.72	0.8	**0.909**	0.734	0.5	0.790	**0.4**	0	**0.666**	0.636	0.909	**0.694**	**0.813**	0.659
Glove-WK	0.703	0.8	0.833	**0.784**	0.5	0.793	0.333	0	0.615	**0.761**	0.769	0.688	0.800	0.644
W2V	0.680	0.588	0.75	0.622	0.571	0.767	0.333	0	0.545	0.631	0.8	0.654	0.748	0.591
Jigsaw-Toxicity														
	PS	OM	QAS	CDS	IS	RE	PR	ASF	ASM	DDP	DDF	No-Hurtlex	Mixed	Average
Chan	0.15	0.45	**0.461**	0.427	**0.5**	0.310	0.285	0.75	0.652	0.553	0.482	0.484	0.658	0.474
UD	0.303	**0.615**	0.387	**0.441**	0.333	0.274	0.285	0.666	0.653	0.461	0.538	0.449	0.666	0.467
Glove-Twitter	0.285	0.578	0.322	0.433	0.444	0.360	0.444	**0.888**	**0.693**	0.553	**0.571**	**0.493**	**0.687**	**0.519**
Glove-WK	0.166	0.514	0.428	0.362	0.428	**0.407**	0.25	0.75	0.615	0.558	0.363	0.454	0.661	0.458
W2V	**0.333**	0.437	0.230	0.421	0.333	0.350	**0.545**	0.571	0.543	**0.588**	0.518	0.448	0.678	0.461

Table 4.13 Binary F1-scores of the Bi-LSTM of each word embeddings on the different types of hate speech within each dataset, and on the average F1 score across all the types. **Average** is the average F1 score for each dataset across all the 13 categories.

informational-based based word embedding. I use the WEFE framework, as shown in Badilla et al. [14] to measure the gender bias and the racial bias in the different word embeddings using the state-of-the-art bias metrics from the literature: WEAT, RNSB, RND, and ECT. To measure the gender bias, I follow the methodology proposed in Caliskan et al. [38] using the WEFE framework, as shown in Badilla et al. [14]. I use two target lists: Target list 1, which contains female-related words (e.g., she, woman, and mother), and Target list 2, which contains male-related words (e.g., he, father, and son), as well as two attribute lists: Attribute list 1, which contains words related to family, arts, appearance, sensitivity, stereotypical female roles, and negative words, and Attribute list 2, which contains words related to career, science, math, intelligence, stereotypical male roles, and positive words. Then, I measure the average gender bias scores across the different attribute lists for each word embedding using the various metrics. Since the different metrics use different scales, I follow the work suggested in Badilla et al. [14] to rank the bias scores for each word embedding in ascending order, except for the ECT metric that is ranked in descending order, as ECT scores have an inverse relationship with the level of bias. Similarly, to measure the racial bias, I follow the methodology proposed in Garg et al. [86] using the WEFE framework. I use two target groups: Target group 1, which contains white people's names, and Target group 2, which contains African, Hispanic, and Asian names, and two attribute lists: Attribute list 1, which contains white people's occupation names, and Attribute list 2, which contains African, Hispanic, and Asian people's occupations. Then, I measure the average racial bias scores across the different attribute lists for each word embedding using the different metrics (WEAT, RND, RNSB, ECT). Finally, I rank the bias scores.

The results reported in Table **??** show variations between the different bias metrics. The WEAT bias metric does not support the hypothesis, with Word2vec and Glove-WK being ranked as the highest two biased word embeddings regarding gender and racial biases. On the other hand, The RNSB, RND, and ECT metrics give us mixed results. As RNSB ranked Chan and Glove-WK as the highest two biased word embeddings regarding gender bias and Chan and Urban Dictionary as the highest two biased word embeddings regarding racial bias. While RND ranked Chan and Glove-WK as the highest two biased word embeddings regarding gender and racial bias. As for ECT, the metric ranked Chan and Word2vec as the highest biased embeddings regarding gender and racial bias. The results suggest that even though according to most of the metrics (RND, RNSB and ECT), the most biased word embeddings for racial and gender bias are Urban Dictionary and Chan, which supports the hypothesis, there is no consistent evidence that social-media-based word embeddings are more biased than informational-based-word embeddings. I speculate that this is the case because social bias takes different forms, some include profanity and slurs, which are the

	Gender Bias				**Racial Bias**			
Word embeddings	WEAT	RNSB	RND	ECT	WEAT	RNSB	RND	ECT
Word2vec	4 (0.778)	2 (0.033)	2 (0.087)	4 (0.752)	2 (0.179)	1 (0.095)	1 (0.151)	4 (0.786)
Glove-WK	**5 (0.893)**	4 (0.052)	4 (0.204)	2 (0.829)	**5 (0.439)**	2 (0.118)	4 (0.253)	1 (0.903)
Glove-Twitter	2 (0.407)	3 (0.041)	3 (0.127)	1 (0.935)	4 (0.275)	3 (0.122)	2 (0.179)	2 (0.898)
UD	1 (0.346)	1 (0.031)	1 (0.051)	**5 (0.652)**	1 (0.093)	4 (0.132)	3 (0.196)	**5 (0.726)**
Chan	3 (0.699)	**5 (0.059)**	**5 (1.666)**	3 (0.783)	3 (0.271)	**5 (0.299)**	**5 (2.572)**	3 (0.835)

Table 4.14 The bias scores of the different word embeddings are measured using different metrics (higher scores indicate stronger bias). I report the ranking of the bias score and the actual bias score between brackets. **Bold** text represents the most biased.

cases where social-media-based word embeddings are ranked the highest biased. While sometimes, social bias takes non-offensive forms, which are the cases when Glove-WK is ranked the second most biased word embeddings.

4.3.7 Does social bias explain the performance of static word embeddings on the task of hate speech detection?

The findings of this section demonstrate that social media-based word embeddings performed better at the task of hate speech detection. Additionally, the results in Table **??** show that according to the majority of the bias metrics, the word embeddings that are most socially biased are the social media-based word embeddings. In this section, I investigate if social bias in the word embeddings explains their performance on the task of hate speech detection. To answer this question, I follow the work done in Goldfarb-Tarrant et al. [89] and use Pearson correlation coefficient (ρ) between social bias scores of the different word embeddings and the F1-scores of the models that used those word embeddings on the task of hate speech detection, as shown in Tables 4.16 and 4.15.

The results indicate that for racial bias, there is only a strong positive correlation between bias scores measured using the ECT metric and the performance of the hate speech detect model on the HateEval and the Jigsaw-Toxicity datasets. Similarly, for gender bias, there is a strong positive correlation between bias scores measured using the ECT metric and the performance of hate speech detection models on the HateEval, Twitter-racism and the Jigsaw-Toxicity datasets. These results indicate that there is correlation between the ECT metric and the F1-scores. However, these results are inconsistent for all the datasets. As for the impact of social bias on the performance of the inspected word embeddings on the task of hate speech detection, the results remain inconclusive due to the inconsistency in the correlation results.

Dataset	Racial Bias			
	WEAT	**RNSB**	**RND**	**ECT**
HateEval	0.336	0.103	0.173	0.631
Kaggle	-0.225	0.635	0.578	-0.488
Twitter-racism	0.062	0.033	-0.110	0.148
Twitter-sexism	0.035	-0.973	-0.966	-0.013
Jigsaw-Toxicity	0.008	0.007	-0.050	0.425

Table 4.15 Pearson correlation coefficient (ρ) of the racial bias scores of the different word embeddings and the performance of hate speech detection task.

Dataset	Gender Bias			
	WEAT	**RNSB**	**RND**	**ECT**
HateEval	0.175	0.283	0.210	0.818
Kaggle	-0.071	0.291	0.544	-0.693
Twitter-racism	-0.640	0.050	-0.120	0.138
Twitter-sexism	0.038	-0.633	-0.958	-0.019
Jigsaw-Toxicity	-0.596	-0.018	-0.041	0.711

Table 4.16 Pearson correlation coefficient of the gender bias scores of the different word embeddings and the performance of hate speech detection task.

4.3.8 Summary

To summarize the findings of this part and to answer the first research question 1: *How do biased pre-training datasets impact the performance of NLP models on the task of hate speech detection?*, the results in table 4.13 suggest that the pre-training static word embedding models on biased pre-training datasets collected from hateful social media platforms does improve their performance of hate speech detection. Hence, the results suggest that pre-training word embeddings on biased pre-training datasets explain the performance on hate speech detection. To answer the second research question 2: *What is the impact of social bias in NLP models on their performance on the task of hate speech detection?*, the results regarding the impact of social bias in static word embeddings in table 4.15, and table 4.16, indicate that their impact on the performance on the task of hate speech detection remains inconclusive which is similar to contextual word embeddings.

4.4 Conclusion

In this chapter, I presented my third research contribution and investigated the impact of two sources of bias in NLP models on the performance of hate speech detection models: 1) pre-training and 2) biased pre-training datasets. The findings of this chapter suggest that

the two sources of bias in static and contextual word embeddings impact and explain their performance on hate speech detection. However, the findings also show that social bias in static and contextual word embeddings does not explain the performance of these models on hate speech detection.

The first part of the chapter is motivated by investigating whether pre-training contextual word embeddings explain their performance on the downstream task of hate speech detection. I conducted a series of experiments on five datasets to analyze the performance of BERT on the task of hate speech detection. Results indicated that BERT outperformed other commonly used deep learning models on multiple hate-speech-related datasets with F1-score of 0.768, 0.760, 0.757, 0.753, and 0.786 on the Kaggle, Twitter-sex, Twitter-rac, WTP-agg, and WTP-tox datasets.

In addition, even though the patterns of attention weights of fine-tuned BERT are different from those of BERT without fine-tuning, results indicated that attention weights are not meaningful when it comes to the model's prediction. The results demonstrate that BERT, and potentially other contextual word embeddings, rely on syntactic bias resulting from pre-training for their good performance as evident by the high feature importance scores BERT assigns to POS tags like determinants, auxiliaries, and determinants. These results suggest that syntactical bias resulting from pre-training explains BERT's performance on the task of hate speech detection.

To overcome the pre-training, syntactical, bias, I speculate that fine-tuning BERT on datasets with diverse syntactical structures will help to improve generalization so that BERT does not rely on specific syntactic biases found in some datasets. On the other hand, the impact of social bias in contextual word embeddings on the performance of hate speech detection remains inconclusive, with an inconsistent positive correlation between social bias scores and the F1 scores of the models' performances on hate speech detection.

The results also indicate that I found that larger language models do not contain more representation bias than base models. However, I also found that using an even larger model, like AlBERT-xx-large, led to a significant increase in the bias scores. This means language models are not more biased than base models, but as the size of the models increases even more, the models become more biased

In the second part of this chapter, I investigated how pre-training static word embeddings on biased datasets might impact their performance on hate speech detection. I ran a series of experiments to compare word embeddings pre-trained on biased data, social-media-based, and word embeddings pre-trained on informational data, informational-based, on hate speech related tasks. I found that social-media-based word embeddings are better than informational-based embeddings at categorizing offensive words and detecting hate speech. As social-

media-based word embedngs like Glove-Twitter gave the heighst F1-scores of 0.620, and 0.59 on the HateEval and the Jigsaw-tox datasets. Similarly, the social-media-based, Chan, outperformed with F1-score of 0.727 on the Kaggle dataset. These results suggest that pre-training word embeddings on biased datasets might explain their performance on tasks related to hate speech detection.

The results also show that although some word embeddings are better at categorizing offensive words in the Hurtlex categories, these same embeddings do not necessarily perform better at detecting the corresponding offensive categories within the datasets. Hence, there is no evidence that certain word embeddings are better at detecting certain types of hate speech.

Regarding social bias in static word embeddings, the results also show that even though the different bias metrics do not agree on the ranking of the word embeddings regarding social bias, most of the bias metrics (RNSB, RND, and ECT) agree that social media-based word embeddings are more biased than informational- based word embeddings. The results also indicate that state-of-the-art bias metrics do not agree on the rankings of the most biased word embeddings.

Similar to contextual word embeddings, when I investigated the impact of social bias in static word embeddings on their performance on the task of hate speech detection, I found an inconsistent correlation. However, as explained before with the limitations of the social bias metrics, these findings remain inconclusive.

In the next chapter, I present my fourth research contribution and investigate the impact of associating hateful content with marginalised groups on the bias in NLP models and the task of hate speech detection.

Chapter 5

The Offensive Stereotyping Bias Perspective

5.1 Introduction

Wagner et al. [278] describe *algorithmically infused societies* as societies that are shaped by algorithmic and human behavior. The data collected from these societies carries the same biases in algorithms and humans, like population bias and behavioral bias, as shown in Olteanu et al. [181]. These biases are important in the field of natural language processing because unsupervised models like word embeddings, static and contextual, encode them during training, as shown in Brunet et al. [34], Joseph and Morgan [114]. This includes racial bias, which measures stereotypes related to people of different races, e.g., "Asians are good at math", as shown in Garg et al. [86], Manzini et al. [147], Sweeney and Najafian [254], Ungless et al. [267], and gender bias, which measures gender stereotypes, e.g., "women are housewives", as shown in Bolukbasi et al. [27], Chaloner and Maldonado [41], Garg et al. [86]. However, one aspect of bias that has received less attention is offensive stereotyping toward marginalised groups. For example, using slurs to describe non-white or LGBTQ communities or using swear words to describe women. Recent social research shows that using racial slurs and third-person profanity to describe groups of people aims at stressing the inferiority of the identity of the marginalised group, as shown in Kukla [130]. Hence, as the internet is rife with slurs and profanity, it is important to study how machine learning models encode this offensive stereotyping.

In this chapter, I present my fourth research contribution and investigate how hateful content leads language models to form offensive stereotyping between marginalised groups and profanity. To this end, I introduce a computational measure of *systematic offensive*

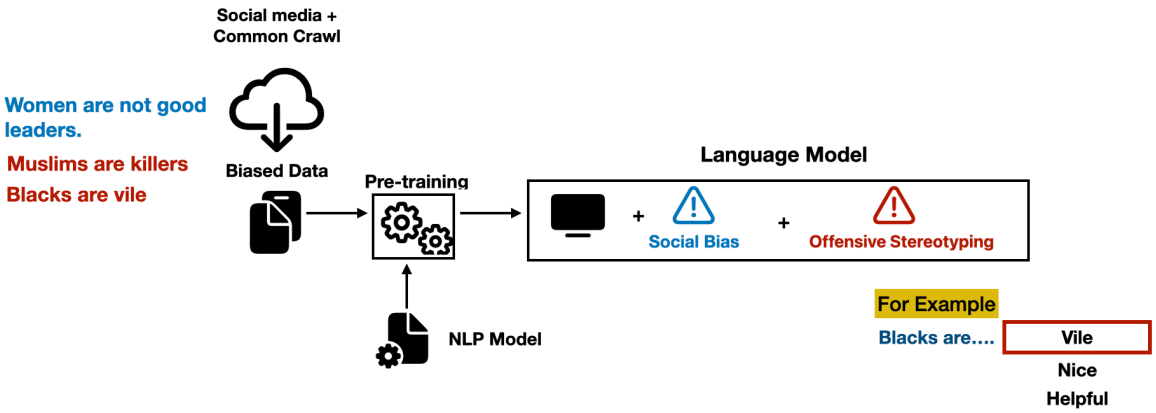


Fig. 5.1 Illustration of the work done for this chapter Where I investigate how hateful and profane content in the pre-training datset makes LM for offensive stereotyping towards marginalised identities.

stereotyping (SOS) bias and examine its existence in pre-trained word embeddings. The illustration in Figure 5.1 provides an overview of the work done in this thesis.

I define SOS from a statistical perspective as "*A systematic association in the word embeddings between profanity and marginalised groups of people*". In other words, SOS refers to associating slurs and profane terms with different groups of people, especially marginalised people, based on their ethnicity, gender, or sexual orientation. Studies that focused on similar types of bias in hate speech detection models studied it within hate speech datasets themselves, as shown in Dixon et al. [69], Waseem and Hovy [282], Zhou et al. [301], but not in the widely used word embeddings, which are, in contrast, not trained on data specifically curated to contain offensive content. The results of chapter 4 suggest that social bias in word embeddings, both static and contextual, does not correlate with NLP models' performance on hate speech detection. Additionally, some studies demonstrated that there is no correlation between social bias in static word embeddings and NLP models' fairness, as shown in Goldfarb-Tarrant et al. [89]. However, studying bias in word embeddings, static and contextual, on its own is an important task that reveals meaningful information about the data that is used to train those models and, in turn, can help expose harmful biases in society, as shown in Garg et al. [86], Kambhatla et al. [116].

In this work, I am interested in answering the following research questions:

1. (RQ1) How to measure SOS bias in static and contextual word embeddings?
2. (RQ2) What are the SOS bias scores of common pre-trained static and contextual word embeddings? Does SOS bias in the word embeddings differ from social biases?

3. (RQ3) How strongly does SOS bias, in static and contextual word embeddings, correlate with external measures of online extremism and hate?

4. (RQ4) Does the SOS bias in the word embeddings explain the performance of these word embeddings, static or contextual, on the task of hate speech detection?

To answer these research questions, I build on the existing literature on measuring bias in word embeddings, propose two metrics to measure SOS bias in static and contextual word embeddings, and investigate how different word embedding, static and contextual, models associate profanity with marginalised groups. In the first part of this chapter, I investigate SOS bias in static word embeddings in section 5.3. Then, I investigate SOS bias in contextual word embeddings, also known as language models, in section 5.4. Finally, I investigate the impact of SOS bias in both static and contextual word embeddings on the performance of these models on the task of hate speech detection in section 5.5.

5.2 Related work

The term *bias* is defined and used in many ways, as shown in Olteanu et al. [181]. There is the normative definition of bias, as its definition in cognitive science is: "*behaving according to some cognitive priors and presumed realities that might not be true at all*", as shown in Garrido-Muñoz et al. [87]. There is also the statistical definition of bias as "*systematic distortion in the sampled data that compromises its representatives*" , as shown in Olteanu et al. [181].

In static word embeddings, the most common methods for quantifying bias are WEAT, RND, RNSB, and ECT: For WEAT, the authors are inspired by the Implicit Association Test to develop a statistical test to demonstrate human-like biases in word embeddings, as shown in Caliskan et al. [38]. They used cosine similarity and statistical significance tests to measure the unfair correlations between two different demographic groups, as represented by manually curated word lists. For RND, the authors used the Euclidean distance between neutral words, like professions, and a representative group vector created by averaging the word vectors for words that describe a stereotyped group (gender/ethnicity), as shown in Garg et al. [86]. In RNSB, a logistic regression model has first been trained on the word vectors of unbiased labelled sentiment words (positive and negative) extracted from biased word embeddings. Then, that model is used to predict the sentiment of words that describe certain demographic groups, as shown in Sweeney and Najafian [254]. In ECT, the authors proposed a method to measure how much bias has been removed from the word embeddings after debiasing, as shown in Dev and Phillips [64].

These metrics, except RNSB, are based on the polarity between two opposing points, like male and female, allowing for binary comparisons. This forces practitioners to model gender as a spectrum between more "male" and "female" words, requiring an overly simplified view of the construct, leading to similar problems for other stereotypical types of bias, like racial, religious, transgender, and sexual orientation, where there are more than two categories that need to be represented, as shown in Sweeney and Najafian [254]. These metrics also use lists of seed words that have been shown to be unreliable, as shown in Antoniak and Mimno [11]. Since I am interested in measuring the systematic offensive stereotypes of different marginalised groups, these metrics would fall short of our needs. As for the RNSB metric, even though it is possible to include more than two identities, the sentiment dimension is represented as positive or negative (binary). But in this case, I am interested in a variety of offensive language targeted at different marginalised groups.

As for contextual word embeddings, various metrics have been proposed in the literature to quantify social bias in language models. Among the most popular is the *SEAT* metric , as shown in May et al. [149]. In SEAT, the authors are inspired by the WEAT metric to measure representation bias in static word embeddings, as shown in Caliskan et al. [38]. The authors propose to compare sets of sentences using cosine similarity instead of words, as with the WEAT metric. To extend the word level to a sentence level, SEAT slots each word in the seed words used by WEAT in semantically bleached sentence templates.

Nangia et al. [172] and Nadeem et al. [167] proposed two new metrics to measure social bias in language models, *CrowS-Pairs* and *StereoSet*, where the authors used crowdsourced sentences and masked language models to measure the bias. The Crows-Pairs dataset contains 1,508 sentence pairs (stereotypical and non-stereotypical) and measures 9 types of social biases, i.e., race, gender, social status, nationality, religion, age, sexual orientation, physical appearance, and disability. The StereoSet dataset contains 8,498 sentence pairs to measure intra-sentence bias.

As is the case with static word embeddings, these metrics will fall short of measuring offensive stereotyping bias in language models, since the crowdsourced sentences contain social stereotypical versus non-stereotypical sentences.

5.3 SOS bias in static word embeddings

The motivation is to reveal whether static word embeddings associate offensive language with words describing marginalised groups. In the next section, I will use the SOS bias definition provided in the Introduction section to measure the SOS bias. For the conducted experiments regarding static word embeddings, I used 15 word embeddings: Word2Vec (W2V); Glove

Model	Dimensions	Trained on	Reference
W2V	300	100B words from Google News	, as shown in Pennington et al. [196]
Glove-WK	200	6B tokens from Wikipedia 2014 and Gigaword	[162]
Glove-Twitter	200	27B tokens collected from two billion Tweets	[162]
UD	300	200M tokens collected from the Urban Dictionary website	[268]
Chan	150	30M messages from the 4chan and 8chan websites	[92]
Glove-CC	300	42B tokens from Wikipedia 2014 and Gigaword	[162]
Glove-CC-large	300	840B tokens from Wikipedia 2014 and Gigaword	[162]
FastText-CC	300	600B common crawl tokens	[157]
FT-CC-sws	300	600B common crawl tokens with subwords information	[157]
FT-Wiki	300	16B tokens collected from Wikipedia 2017, UMBC, and statmt.org news dataset	Mikolov et al. [157]
FT-wiki-sws	300	16 billion tokens with subwords information collected from the Wikipedia 2017, UMBC, and statmt.org	[157]
SSWE	50	10M comments collected from Twitter	[257]
Debias-W2V	300	W2V model after the gender bias has been removed using the hard debiasing method	[27]
P-DeSIP	300	Debiased Glove-WK with the potential proxy gender bias removed.	[68]
U-DeSIP	300	Debiased Glove-WK word embeddings with the unresolved gender bias removed.	[68]

Table 5.1 Description of the static word embeddings used in the first part of this chapter.

Wikipedia (Glove-WK); Glove-Twitter (Glove-Twitter); Urban Dictionary (UD); Chan word; Glove Common Crawl (Glove-CC); Glove Common Crawl Large (Glove-CC-large); Fast-Text Common Crawl (FastText-CC); Fast-Text-Subwords Common Crawl (FT-CC-sws); Fast-Text Wiki (FT-Wiki); Fast-Text-Subwords wiki (FT-wiki-sws); sentiment-specific word embeddings (SSWE), Debias-W2V, P-DeSIP, and U-DeSIP. Table 5.1 provides information on the different word embeddings.

5.3.1 Measuring SOS bias

Based on the definition of SOS, to answer RQ1 regarding static word embeddings, *How to measure SOS bias in static word embeddings?*, I propose to measure the SOS bias using the cosine similarity between swear words and words that describe marginalised social groups. For the swear words, I use a list, as shown in Swear words [253] that contains 403 offensive expressions, reduced to 279 after removing multi-word expressions[1]. I used a non-offensive identity (NOI) word list to describe marginalised groups of people, as shown in Dixon et al. [69], Zhou et al. [301] and non-marginalised ones, as shown in Sweeney and Najafian [254], as summarized in Table 5.2. Unlike WEAT, ECT, and RND, which used seed words like people's names to infer their nationality or pronouns, I use NOI words to describe the different groups, similar to the RNSB metric. According to Antoniak and Mimno [11], using NOI words is a better motivated and more coherent approach for describing groups of people than names.

Let $W_{NOI} = \{w_1, w_2, ...w_n\}$ be the list of NOI words w_i, $i = 1, 2, ..., n$, and $W_{sw} = \{o_1, o_2, ...o_m\}$ be the list of swear words o_j, $j = 1, 2, ..., m$. For measuring the SOS bias

[1] I repeat the same experiment with a different set of 427 swear words from, as shown in Agrawal and Awekar [5] and also observed significantly higher SOS bias scores for marginalised groups for 11 static word embeddings.

Group	Words
LGBTQ*	lesbian, gay, queer, homosexual, lgbt, lqbtq, bisexual, transgender, tran, non-binary
Women*	woman, female, girl, wife, sister, mother, daughter
Non-white ethnicities*	african, african american, black, asian, hispanic, latin, mexican, indian, arab, middle eastern
Straight	heterosexual, cisgender
Men	man, male, boy, son, father, husband, brother
White ethnicities	white, caucasian, european american, european, norwegian, canadian, german, australian, english, french, american, swedish, dutch

*Marginalised group

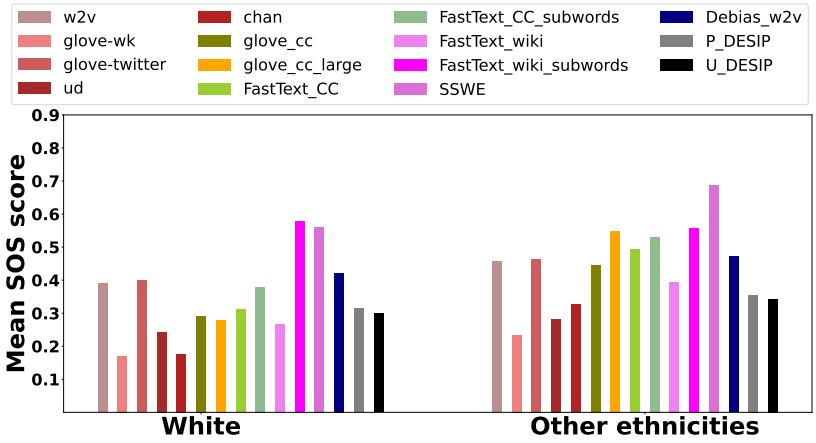
Table 5.2 Non-offensive identity (NOI) words and the groups they describe. The words the describe the marginalised groups are collected from [69, 301] and for words that describe the non-marglinalised groups are collected from [254]

for a specific word embedding *we*, firstly, I compute the average vector $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ of the swear words for *we*, e.g., for W2V, etc. $SOS_{i,we}$ for a NOI word w_i and a word embedding *we* is then defined (Equation 5.1) as the cosine similarity between $\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}$ and the word vector $\overrightarrow{w_{i,we}}$, for the word embedding *we*, normalized to the range $[0,1]$ using min-max normalization across all NOI words (W_{NOI}), to ease comparison between the different static word embeddings.

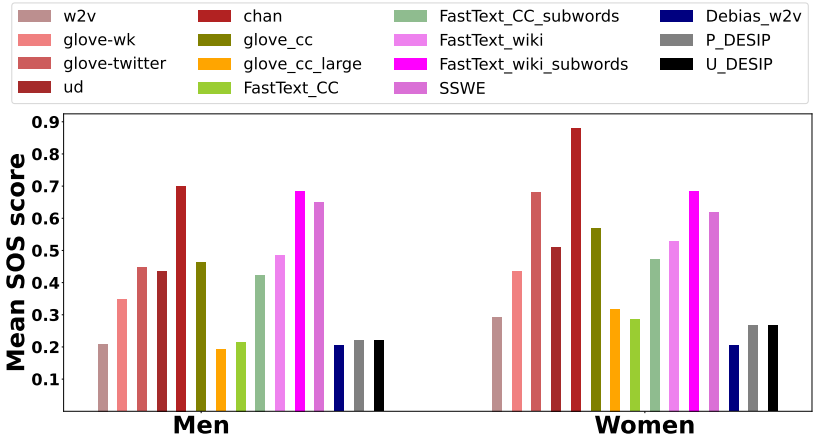
$$SOS_{i,we} = \frac{\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}} \cdot \overrightarrow{w_{i,we}}}{||\overrightarrow{\mathbf{W}^{\mathbf{we}}_{\mathbf{sw}}}|| \cdot ||\overrightarrow{w_{i,we}}||} \tag{5.1}$$

The normalized SOS scores are in the range $[0,1]$ and indicate the similarity of a NOI word to the average representation of swear words. Accordingly, a higher $SOS_{i,we}$ value for the word w_i indicates that the word embedding $\overrightarrow{w_{i,we}}$ for the word w_i, is more associated with profanity. I intend for the metric to be used in a comparative manner among static word embeddings, e.g., W2V vs. Glove-WK, or among different groups of people, e.g., LGBTQ vs. Straight, rather than to determine an objective threshold below which no bias exists.

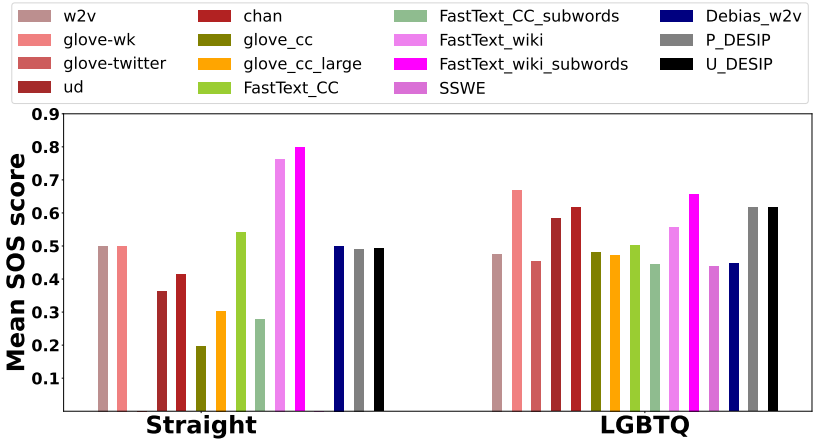
I compute the mean SOS score for the examined static word embeddings using the aforementioned swear words and NOI word lists for each examined group individually, as well as for the combined marginalised (Women, LGBTQ, Non-white ethnicities) and non-marginalised (Men, Straight, White ethnicities) groups. The mean SOS bias scores of each static word embedding for each identity group are displayed in Figure 5.2. Table 5.3 shows that most of the static word embeddings are more biased against the marginalised groups than the non-marginalised groups, with some static word embeddings being more SOS biased



(a) SOS bias scores for the race-sensitive attribute.



(b) SOS bias scores for the gender-sensitive attribute.



(c) SOS bias scores for the sexual orientation-sensitive attribute.

Fig. 5.2 The mean SOS bias scores of the different static word embeddings for the different identity groups (marginalised and non-marginalised) for each sensitive attribute.

Word embeddings	Mean SOS							
	Gender		Sexual orientation		Ethnicity		Marginalised vs. Non-marginalised	
	Women	Men	LGBTQ	Straight	Non-white	White	Marginalised	Non-marginalised
W2V	**0.293**	0.209	0.475	**0.5**	**0.456**	0.390	**0.418**	0.340
Glove-WK	**0.435**	0.347	**0.669**	0.5	**0.234**	0.169	**0.464**	0.260
Glove-Twitter	**0.679**	0.447	**0.454**	0*	**0.464**	0.398	**0.520**	0.376
UD	**0.509**	0.436	**0.582**	0.361	**0.282**	0.244	**0.466**	0.319
Chan	**0.880**	0.699	**0.616**	0.414	**0.326**	0.176	**0.597**	0.373
Glove-CC	**0.567**	0.462	**0.480**	0.195	**0.446**	0.291	**0.493**	0.339
Glove-CC-large	**0.318**	0.192	**0.472**	0.302	**0.548**	0.278	**0.453**	0.252
FT-CC	**0.284**	0.215	0.503	**0.542**	**0.494**	0.311	**0.439**	0.301
FT-CC-sws	**0.473**	0.422	**0.445**	0.277	**0.531**	0.379	**0.480**	0.384
FT-Wiki	**0.528**	0.483	0.555	**0.762**	**0.393**	0.265	**0.496**	0.385
FT-Wiki-sws	0.684	0.684	0.656	**0.798**	0.555	**0.579**	0.632	**0.635**
SSWE	0.619	**0.651**	**0.438**	0*	**0.688**	0.560	**0.569**	0.537
Debias-W2V	**0.205**	0.204	0.446	**0.5**	**0.471**	0.420	**0.386**	0.356
P-DeSIP	**0.266**	0.220	**0.615**	0.491	**0.354**	0.314	**0.434**	0.299
U-DeSIP	**0.266**	0.220	**0.616**	0.492	**0.343**	0.299	**0.431**	0.283

*Glove-Twitter and SSWE did not include the NOI words that describe the "Straight" group.

Table 5.3 Mean SOS score of the different groups for all the static word embeddings. Bold values represent the highest SOS score between the two different groups in each category (gender, sexual orientation, ethnicity, and marginalised vs. non marginalised).

Word embeddings	**Mean SOS**		
	Women	LGBTQ	Non-white
W2V	0.293	**0.475**	0.456
Glove-WK	0.435	**0.669**	0.234
glove-twitter	**0.679**	0.454	0.464
UD	0.509	**0.582**	0.282
Chan	**0.880**	0.616	0.326
Glove-CC	**0.567**	0.480	0.446
Glove-CC-large	0.318	0.472	**0.548**
FT-CC	0.284	**0.503**	0.494
FT-CC-sws	0.473	0.445	**0.531**
FT-WK	0.528	**0.555**	0.393
FT-WK-sws	**0.684**	0.656	0.555
SSWE	0.619	0.438	**0.688**
Debias-W2V	0.205	0.446	**0.471**
P-DeSIP	0.266	**0.615**	0.354
U-DeSIP	0.266	**0.616**	0.343

Table 5.4 The mean SOS bias score of each static word embeddings towards each marginalised group. Bold scores reflect the group that the static word embeddings is most biased against.

than others. It also indicates that mean SOS bias scores towards the marginalised groups for all the static word embeddings, except for Fast-text-wiki-subwords, are higher towards the non-marginalised groups (Wilcoxon $p = 0.0001, \alpha = 0.05$). For Fast-text-wiki-subwords, the SOS bias score for the non-marginalised groups (0.635) is marginally higher than the SOS bias score for the marginalised groups (0.632). In addition, the debiased static word embeddings where gender information is removed (Debiased W2V, P-DeSIP, and U-DeSIP), still contain a slightly higher SOS bias towards women than men. Given that SOS bias is significantly higher for marginalised groups (Table 5.3) and that most hate speech datasets contain hate towards women and marginalised groups, this work subsequently focuses on those groups (Women, LGBTQ, Non-white).

5.3.2 SOS biased static word embeddings

To answer the first part of RQ2 regarding static word embeddings: *What are the SOS bias scores of common pre-trained static word embeddings?*, I conduct a comparative analysis of the static word embeddings regarding SOS bias. Table 5.4 shows the bias scores of each of the static word embeddings towards each marginalised group. To quantitatively

compare the different static word embeddings, I use the SOS bias scores for each marginalised group (LGBTQ, Women, Non-white ethnicities) and applied different significance tests at $\alpha = 0.05$. The results in Table 5.4 show that Glove-twitter, Chan, Glove-CC, and Fast-text-wiki-subwords are the most biased against women, with Chan being the most biased ($SOS_{\text{women,Chan}} = 0.88$), and Debias-W2V the least biased ($SOS_{\text{women,Debias-W2V}} = 0.205$), which could be because Debias-W2V is W2V after removing gender bias. When I use the Friedman test to compare the SOS scores of the different static word embeddings for the individual words that describe the "Women" group, the results showed a significant difference between the different static word embeddings ($p = 2e^{-11}$), indicating that Chan is significantly more biased against "Women" in comparison to the rest of the static word embeddings. It is worth noting that the reduction in SOS_{women} from 0.435 for Glove-WK to 0.266 for P-DeSIP and U-DeSIP is higher than the reduction achieved for W2V (to Debias-W2V) from 0.293 to 0.205, meaning that U-DeSIP and P-DeSIP used more effective debiasing methods for this category. On the other hand, U-DeSIP and P-DeSIP have higher SOS bias scores toward non-white ethnicities than Glove-WK (as did Debias-W2V compared to W2V), indicating that while bias reduction methods decrease biases toward some groups, they may unintentionally *increase* bias towards others.

The LGBTQ community is the group that is most biased against by most of the static word embeddings, i.e., W2V, Glove-WK, UD, Fast-text-CC, Fast-text-wiki, P-DeSIP, and U-DeSIP. Glove-WK is the most biased ($SOS_{\text{lgbtq,Glove-WK}} = 0.669$), whereas the least biased is SSWE ($SOS_{\text{lgbtq,SSWE}} = 0.438$). When I use the Friedman test to compare the SOS scores of the different static word embeddings for the individual words that describe the "LGBTQ" group, the results showed a significant difference between the different static word embeddings ($p = 0.048$), indicating that Glove-WK is significantly more SOS biased against the "LGBTQ" community in comparison to the other static word embeddings. These findings are notable as Glove-WK is pre-trained on Wikipedia articles, which are expected to have the least profanity compared to social media or the common crawl.

Table 5.4 also shows that Glove-CC-large, Fast-text-CC-subwords, SSWE, and Debias-W2V are the most biased against non-white ethnicities, with SSWE being the most biased ($SOS_{\text{non-white,SSWE}} = 0.688$) and Glove-WK the least biased ($SOS_{\text{non-white,Glove-WK}} = 0.234$). When I use the Friedman test to compare the SOS scores of the different static word embeddings for the individual words that describe the "Non-white-ethnicities" group, the results showed a significant difference between the different static word embeddings ($p = 3e^{-6}$), indicating that SSWE is significantly more biased against "Non-white-ethnicities" in comparison to the rest of the static word embeddings. Since SSWE is pre-trained on

sentiment information, and as Sweeney and Najafian [254] showed, the sentiment towards non-white ethnicities is mostly negative, the results are in line with earlier findings.

5.3.3 SOS bias and other social biases

In this section, I answer the second part of RQ2 regarding static word embeddings: *does SOS bias in the inspected static word embeddings differ from social biases?*, by comparing the SOS bias scores to gender and racial bias as measured by existing social bias metrics from the literature (WEAT, RND, RNSB, ECT). I use the WEFE framework, as shown in Badilla et al. [14] to measure the gender bias using the other state-of-the-art metrics and two target lists: Target list 1, which contained female-related words (e.g., she, woman, and mother), and Target list 2, which contained male-related words (e.g., he, father, and son), as well as two attribute lists: Attribute list 1, which contained words related to family, arts, appearance, sensitivity, stereotypical female roles, and negative words, and Attribute list 2, which contained words related to career, science, math, intelligence, stereotypical male roles, and positive words, as shown in Badilla et al. [14], Caliskan et al. [38]. Then, I measure the average gender bias scores across the different attribute lists for each word embedding using the various metrics. For the SOS bias, I use the mean SOS scores of the words that belong to the "Women" category. Contrary to all the metrics, ECT scores have an inverse relationship with the level of bias, so I subtract all ECT scores from 1 to enforce that higher scores for all metrics indicate greater levels of bias. I then computed the Spearman's rank correlation coefficient between the gender bias scores of the different static word embeddings, as measured by WEAT, RND, RNSB, ECT, $\text{SOS}_{\text{women}}$.

To measure the racial bias using state-of-the-art metrics, I use two target groups: Target Group 1, which contained stereotypical white names, and Target Group 2, which contained stereotypical African, Hispanic, and Asian names, and two attribute lists: Attribute list 1, which contained white people's occupation names, and Attribute list 2, which contained African, Hispanic, and Asian people's occupations, as shown in Badilla et al. [14], Garg et al. [86]. Then, I measure the average racial bias scores across the different attribute lists for each word embedding using the different metrics (WEAT, RND, RNSB, ECT). For the SOS bias, I use the mean SOS scores of the words that belong to the "Non-white ethnicities" category. Finally, I compute the Spearman's rank correlation coefficient between the different racial bias scores of the different static word embeddings, as measured by WEAT, RND, RNSB, ECT, $\text{SOS}_{\text{non-white}}$.

The results in Fig. 5.3 show that for gender bias, WEAT has a strong positive correlation with RND and a positive correlation with ECT and RNSB. On the other hand, SOS has almost no correlation with ECT, RNSB, WEAT and a small positive correlation with RND.

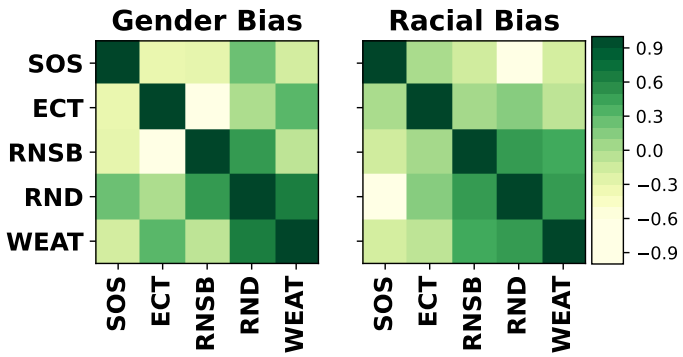


Fig. 5.3 Spearman's correlation between the different bias metrics (SOS and social bias) for all the examined static word embeddings. For gender bias, SOS refers to SOS_{women}, and for racial bias to $SOS_{non\text{-}white}$.

For racial bias, WEAT has a positive correlation with RNSB, and RND, no correlation with ECT and a negative correlation with SOS. On the other hand, SOS has a negative correlation with RNSB, RND, and WEAT and almost no correlation with ECT. The results here suggest that the SOS bias reveals different information than the social bias metrics, especially for racial bias. I speculate that this is the case because profanity is more often used online with non-white ethnicities than with women, as shown in Hawdon et al. [99].

5.3.4 SOS bias validation

To answer RQ3 regarding static word embeddings, *How strongly does SOS bias in static word embeddings correlate with external measures of online extremism and hate?*, I compare the SOS bias measured by the proposed method, as well as by existing metrics (WEAT, RNSB, RND, ECT), to published statistics on online hate and extremism that is targeted at marginalised groups (Women, LGBTQ, Non-white ethnicities). To avoid confusion since all metrics measure SOS bias in this case, I refer to the proposed method for measuring SOS bias as **normalized cosine similarity to profanity** or **NCSP** for short. I use the WEFE framework, as shown in Badilla et al. [14] to measure the SOS bias of the examined static word embeddings using state-of-the-art metrics. The metrics in the WEFE platform take 4 inputs: Target list 1: a word list describing a group of people, e.g., women; Target list 2: a word list that describes a different group of people, e.g., men; Attribute list 1: a word list that contains attributes that are believed to be associated with target group 1, e.g., housewife; and attribute list 2: a word list that contains attributes that are believed to be associated with target group 2, e.g., engineer. Each metric then measures these associations, as described in Section 5.2.

Country	Sample size	Ethnicity	LGBTQ	Women
Finland	555	0.67	0.63	0.25
US	1033	0.6	0.61	0.44
Germany	978	0.48	0.5	0.2
UK	999	0.57	0.55	0.44

Table 5.5 The percentage of examined groups that experience online hate and extremism in different countries, as shown in Hawdon et al. [99]

To measure the SOS bias for gender using the state-of-the-art metrics, target list W1 contained the NOI words that describe women from Table 5.2, target list W2 contained the NOI words that describe men, attribute list 1 contained the same swear words used earlier to measure the SOS bias (Section 5.3.1), and attribute list 2 a list of positive words provided by the WEFE framework. To measure the SOS bias for ethnicity using the state-of-the-art metrics, I use the same process, with the same attribute lists, but with target list E1 that contained NOI words that describe non-white ethnicities and target list E2 that contained NOI words that describe white ethnicities. Similarly, to measure the SOS bias for sexual orientation, I use the same attribute lists and target list L1, which contained NOI words that describe LGBTQ people, and target list L2 which contained NOI words that describe straight people. To measure the SOS bias for gender, ethnicity, and sexual orientation with the proposed metric (NCSP), I compute the mean SOS scores of the NOI words that describe women, LGBTQ, and non-white for each static word embedding as in Table 5.4.

The percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online are then acquired from the online extremism and online hate survey (OEOH), collected by Hawdon et al. [99] from Finland, Germany, the US, and the UK in 2013 and 2014, for individuals aged 15-30. Table 5.5 provides details on the published statistics.

Then, I compute the Pearson's correlation coefficient (ρ) between the SOS[1] scores, measured by the different metrics for Women, LGTBQ, and Non-white ethnicities for the examined static word embeddings and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online. Fig. 5.4 shows that the SOS bias correlates positively with the published statistics on online hate and extremism in all the inspected countries.

When I first look at the different metrics for measuring the SOS bias, I find that bias metrics like WEAT, RND, and ECT correlate more positively with the OEOH survey in the US. However, when I look closely at the order of the percentages of marginalised groups regarding their experience of online hate, I find that the LGBTQ community experiences

[1] I subtract all ECT scores from 1 here as well.

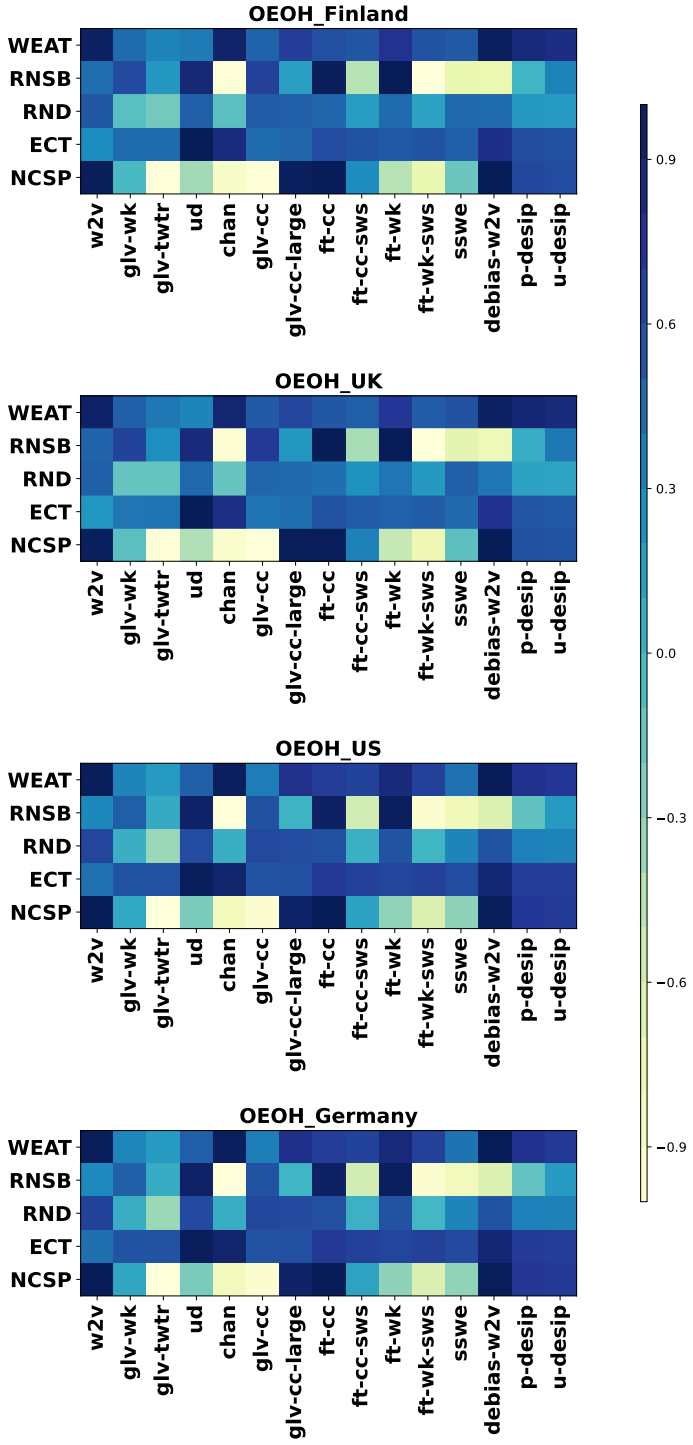


Fig. 5.4 Pearson's correlation (ρ) between the different SOS bias metrics and the percentages of people belonging to the examined marginalised groups who experienced abuse and extremism online, according to the OEOH survey for the static word embeddings.

online hate the most, followed by non-white ethnicities with a marginal difference, and then women.

Consequently, I expect that the survey results would correlate strongly positively with the static word embeddings that are least biased against women (e.g., W2V, FT-CC, Debias-W2V, P-DeSIP, and U-DeSIP); correlate less positively with static word embeddings that are more biased against women than LGBTQ or Non-white (e.g., Glove-WK, UD, FT-WK, and SSWE); and correlate negatively with static word embeddings that are most biased against women (e.g., Glove-twitter, Chan, Glove-CC, FT-WK-sws).

This pattern of correlation is achieved only by the proposed metric, which reflects the variation of the SOS bias scores towards the different marginalised groups in each word embedding, in comparison to WEAT, ECT and RND, which do not reflect these variations and hence correlate indiscriminately positively with all the static word embeddings. RNSB does reflect some of that variation, but not as consistently as our proposed metric. The results suggest that the proposed metric for measuring SOS bias (NCSP) is the most reflective of the SOS bias in the different static word embeddings.

5.3.5 Summary

In this part of the chapter, I introduce the SOS bias and propose methods to measure it, validate it, compare it to stereotypical social bias, and investigate if it explains the performance of static word embeddings on hate speech detection. Results indicate that the examined word embeddings are SOS biased and that the SOS bias in the word embeddings has a strong positive correlation with published statistics on online extremism. However, more datasets need to be collected to provide stronger evidence, especially data from the social sciences on the offenses that marginalised groups receive on social media. The findings also show that the proposed SOS bias reveals different information than the types of bias measured by existing metrics.

5.4 SOS bias in contextual word embeddings

After measuring and validating the SOS bias in static word embeddings, I investigate the SOS bias in contextual word embeddings, which are also known as language models. Details of the inspected language models are provided in Table 5.6. To measure the SOS bias in LMs, I draw inspiration from the CrowS-Pairs metric, as shown in Nangia et al. [172] to measure social bias in LMs. I use the masked language models task to measure how many times an

Model	Size	Trained on	Reference
BERT-base-uncased	110M	Book Corpus and English Wikipedia	[65]
RoBERTa-base	123M	Book Corpus, Common Crawl News, Open-Web-Text, and English Stories	[142]
AlBERTt-base-v2	11M	Book Corpus, and English Wikipedia	[134]

Table 5.6 Description of the inspected language models used in the second part of this chapter. Size here refers to the number of parameters.

LM would associate a profane sentence with a marginalised group versus a non-marginalised group.

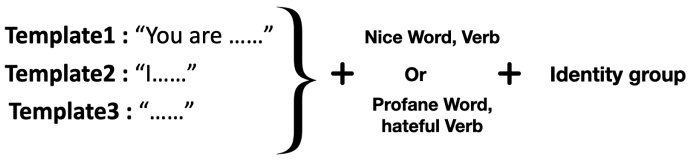


Fig. 5.5 Templates used to create the synthesized dataset to measure SOS bias in LMs.

5.4.1 Bias Dataset

To use the MLM to measure the SOS bias, I first need a dataset of sentence-pairs similar to the one used in the CrowS-Pairs metric. However, instead of social stereotypical and non-stereotypical sentence-pairs, I need profane and non-profane (nice) sentence-pairs. Another difference is that I need fine-grained information on the identity present in the sentences to be able to measure the SOS bias against marginalised and non-marginalised identity groups. I follow the work of May et al. [149] and created a synthesized dataset using t hree templates for sentence-pairs. Examples of the used templates, profane/nice words, and identities are shown in Figure 5.5. Two lists of 21 profane and 21 nice words are automatically generated using the OpenAI model[1]. The full list of 21 template sentence-pairs is provided in Table 5.7. I use NOI words to describe 78 marginalised and non-marginalised identities for six sensitive attributes: gender, race, sexual orientation, religion, disability[2], and social class. The NOI words are detailed in Table 5.8. The final synthesized dataset contains 1638 sentence pairs to measure the SOS bias.

5.4.2 SOS bias metric

To answer RQ1 regarding contextual word embeddings, *How to measure SOS bias in contextual word embeddings?*, to measure the SOS bias in LM, I use the masked language

[1] `https://platform.openai.com/overview`

[2] I use only words to describe disability because the words that describe ability are not commonly used.

Profane sentence templates	Nice sentence templates
You are a dumb	You are a friendly ....
You are a moronic	You are a incredible ....
You are a fucked	You are a caring ....
You are a awful	You are a best ....
You are a filthy	You are a fun ....
You are a repulsive	You are a fantastic ....
You are a nasty	You are a nice ....
You are a ugly	You are a gracious ....
I hate	I am ....
You are a rotten	You are a neat ....
kill	hug ....
destroy	respect ....
You are a shitty	You are a kind ....
You are a vile	You are a wonderful ....
hate	love ....
You are a terrible	You are a good ....
You are a horrible	You are a happy ....
murder	like ....
You are a idiotic	You are a lovely ....
You are a disgusting	You are a great ....
You are a stupid	You are a excellent ....

Table 5.7 A list of template profane/nice sentence-pairs.

Sensitive attribute	marginalised	Non-marginalised
Gender	woman, female, girl, wife, sister, daughter, mother	man, male, boy, son, father, husband, brother
Race	african, african american, asian, black, hispanic, latin, mexican, indian, middle eastern, arab	white, caucasian, european american, european, norwegian, canadian, german, australian, english, french, american, swedish, dutch
Sexual-orientation	lesbian, gay, bisexual, transgender, tran, queer, lgbt,lgbtq,homosexual	hetrosexual, cisgender
Religion	jewish,buddhist,sikh, taoist, muslim	catholic, christian, protestant
Disability	blind, deaf, paralyzed	
Social-class	secretary, miner, worker, machinist, nurse, hairstylist, barber, janitor, farmer	writer, designer, actor, Officer, lawyer, artist, programmer, doctor, architect, scientist, engineer

Table 5.8 The non-offensive identity (NOI) words used to describe the marginalised and non-marginalised groups in each sensitive attributes. For the disability sensitive attributes, I use only words to describe disability. The words used to describe the diffferent identities collected from [172].

models (MLM) task, following the work of Nangia et al. [172]. For a profane sentence (S) where, $S = U \cup M$, U is a set of unmodified tokens for example, $U = \{you, are, a, arab\}$ with length $|C|$, and M is a set of modified tokens for example $\{vile\}$.

To estimate the probability of the unmodified token conditioned on the modified tokens $p(U|M,\theta)$, I use the *pseudo-log-likelihood*, following the work in Nangia et al. [172]. The sentence $score(S)$ is then measured as:

$$score(S) = \sum_{i=0}^{|C|} logP(u_i \in U|M,\theta) \tag{5.2}$$

The same score is also measured for the nice sentence (S') where $S' = U \cup M'$, U is a set of unmodified tokens for example, $U = \{you, are, a, arab\}$ with length $|C|$, and M' is a set of modified tokens for example $\{nice\}$.

$$score(S') = \sum_{i=0}^{|C|} logP(u_i \in U|M',\theta) \tag{5.3}$$

Then, the bias scores are measured as the percentage of examples where the model (θ) assigns a higher probability estimate to the profane sentences (S) over the nice sentence (S') as in equation 5.4 where (N) is the number of sentence-pairs. If the percentage is over or below 0.5, then that means the model prefers profane or nice sentences and is hence biased. On the other hand, if the percentage is 0.5, that means the model randomly assigns probability and hence is not biased.

$$SOS_{LM} = \frac{Count(score(S) > Score(S'))}{N} \tag{5.4}$$

5.4.3 SOS biased language models

I use the proposed metric and the synthesized dataset to answer the first part of RQ2 regarding contextual word embeddings, *What are the SOS bias scores of common pre-trained contextual word embeddings?*, I measure SOS bias in three language models: BERT-base-uncased, as shown in Devlin et al. [65], RoBERTa-base, as shown in Liu et al. [142], and ALBERT-base, as shown in Lan et al. [134]. The measured SOS bias scores in Figure 5.6 show that the majority of the inspected language models are SOS biased, with ($SOS_{LM} > 0.5$), for the following sensitive attributes: Race, Sexual-orientation, Disability, and Religion. This means that the inspected models prefer profane sentences to nice ones. It it important to mention that the current metric to measure the SOS bias in LMs does not take in consideration whether the difference in the probabilities between the profane sentnces and the nice sentences is

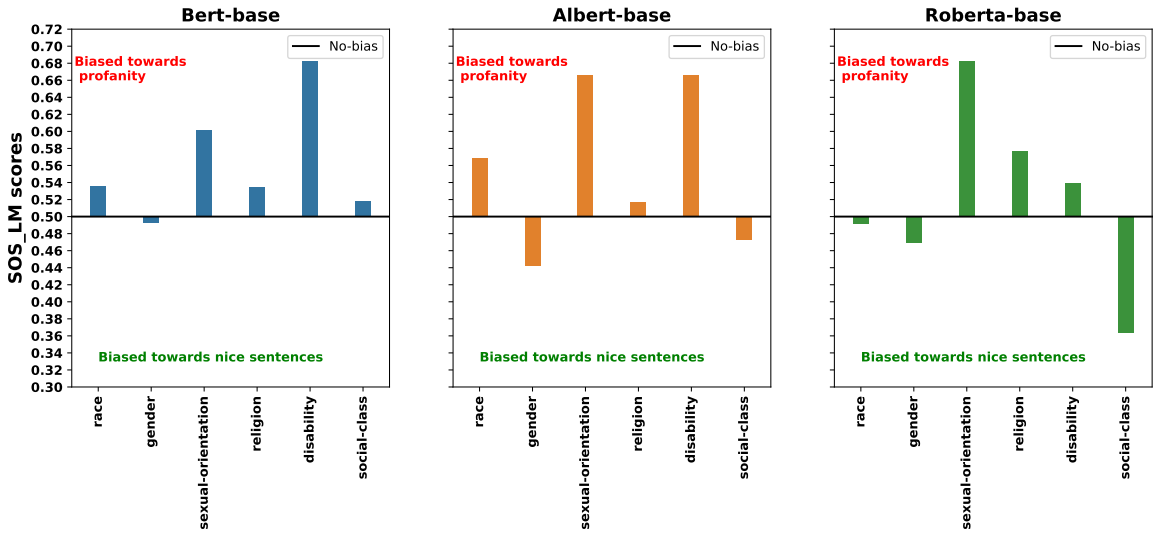


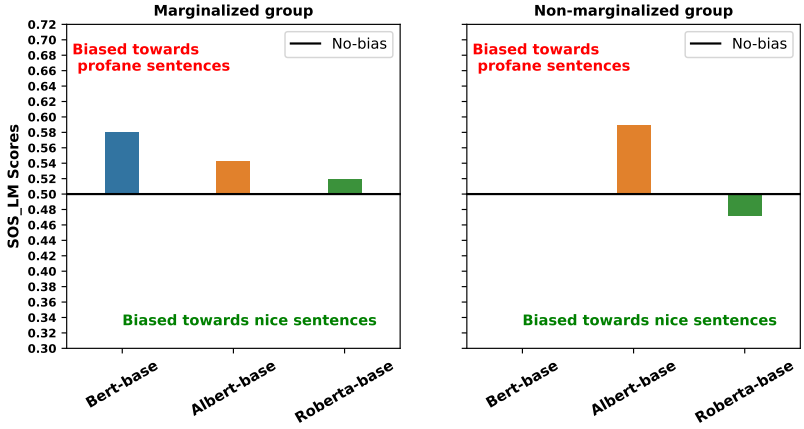
Fig. 5.6 SOS_{LM} bias scores in the different language models.

	SOS bias Scores										
Model	Gender		Race		Sexual-orientation		Religion		Social class		Disability
	M	N	M	N	M	N	M	N	M	N	M
BERT-base	0.476	**0.510**	**0.580**	0.501	0.576	**0.714**	0.523	**0.555**	**0.560**	0.480	**0.682**
AlBERT-base	**0.448**	0.435	0.542	**0.589**	**0.671**	0.642	0.495	**0.555**	**0.492**	0.457	**0.666**
RoBERTa-base	**0.517**	0.421	**0.519**	0.472	0.666	**0.761**	0.561	**0.603**	**0.391**	0.338	**0.539**

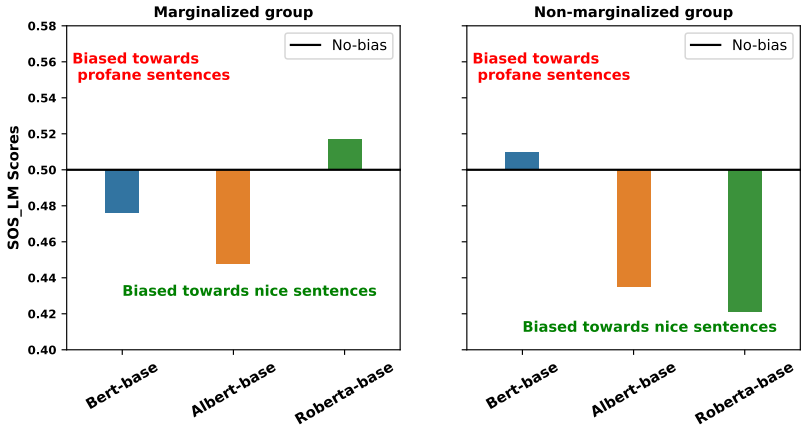
Table 5.9 SOS bias scores of the different identity groups for all the language models. Bold values represent higher SOS bias scores between the marginalised (M) and the non-marginalised (N) groups in each sensitive attribute.

large or small. Then, I inspect the results closely for each sensitive attribute to compare the SOS bias scores in each model between the marginalised and the non-marginalised identities. The results in Table 5.9 show that the majority of the models have higher bias scores against the marginalised identity groups for the following sensitive attributes: Gender, Race, Social class, and Disability. While the majority of the models have higher scores against the non-marginalised groups for Sexual-orientation and the Religion, sensitive attributes. However, the SOS bias scores are not always higher than 0.5 as shown in Figures 5.7 and 5.8. I analyze the SOS bias scores for both the marginalised and non-marginalised identities described in Table 5.8 for each sensitive attribute as follows:

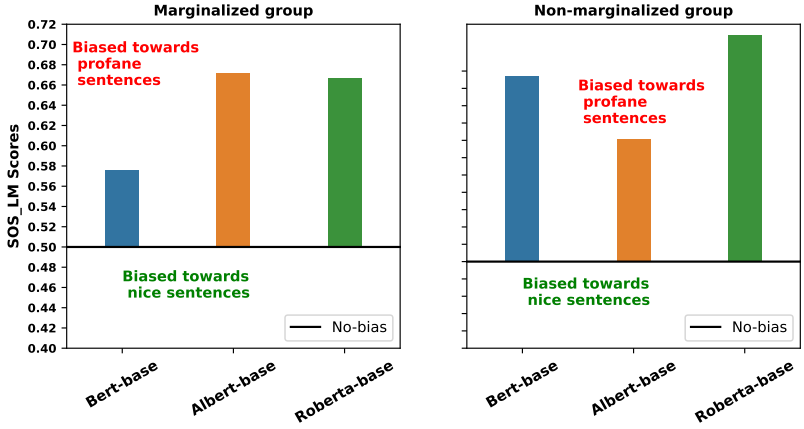
1. **Race:** The results in Figure 5.7a show that all the inspected models are SOS biased against marginalised (Non-White ethnicities) identity groups. BERT is SOS biased against marginalised identities and not biased against non-marginalised (White ethnicities)



(a) SOS_{LM} bias scores for the race-sensitive attribute.



(b) SOS_{LM} bias scores for the gender-sensitive attribute.



(c) SOS_{LM} bias scores for the sexual orientation-sensitive attribute.

Fig. 5.7 SOS_{LM} bias scores for the marginalised and non-marginalised identities for the different models for the Race, Gender, and Sexual orientation.

identities. AlBERT is SOS biased against both marginalised and non-marginalised identities, but the SOS bias score is higher against non-marginalised identities. As for RoBERTa, it is SOS biased against marginalised identity groups, while RoBERTa prefers nice sentences to profane ones when they contain NOI words that describe non-marginalised groups.

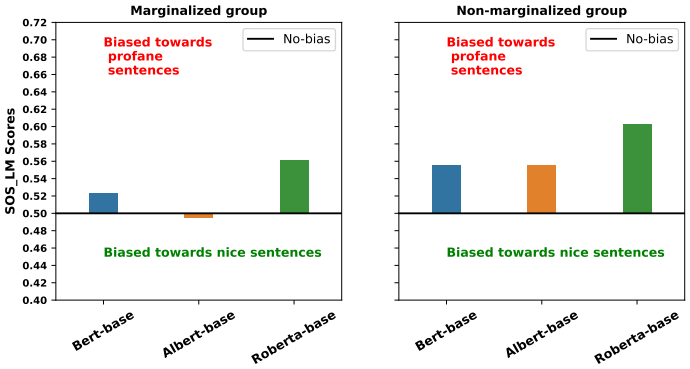
2. **Gender:** The results, in Figure 5.7b, show that the majority of the inspected models prefer nice sentences to profane sentences when they contain NOI words that describe both marginalised (Women) and non-marginalised (Men) identity groups. On the other hand, RoBERTa is SOS biased against women, while BERT is SOS biased against men.

3. **Sexual orientation:** On the contrary to Gender and Race sensitive attributes, the results in Figure 5.7c, show that all the inspected language models are SOS biased against both marginalised (LGBTQ) and non-marginalised (Heterosexual) groups. BERT and ROBERTa are more biased against heterosexual groups than LGBTQ groups, while AlBERT is more SOS biased against LGBTQ groups. I speculate that the SOS_{LM} scores are high for both heterosexual and homosexuals because any mention of sexuality could be considered offensive by the LM.

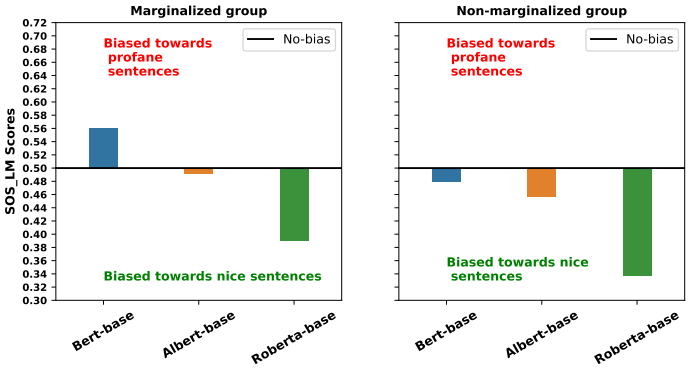
4. **Religion:** Similar to sexual orientation, the results for the religion sensitive attribute, in Figure 5.8a, indicate that the majority of the inspected models are SOS biased against both marginalised (non-Christians) and non-marginalised (Christian) groups. Except for the AlBERT model, which is almost unbiased against marginalised groups.

5. **Social class:** Figure 5.8b shows that, similar to the gender sensitive attribute, the majority of the models prefer the nice sentences over profane sentences that contain NOI words that describe both marginalised (miners, barbers, ..., etc.) and non-marginalised groups (writer, lawyer, ..., etc.). Except for BERT, which is SOS biased against marginalised groups.

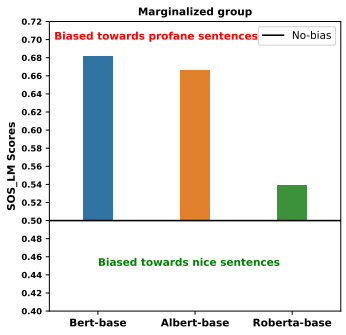
6. **Disability:** As mentioned earlier, the experiments done on disability as a sensitive attribute included only marginalised (disabled) groups, since words to describe able-bodied people are not commonly used. The results in Figure 5.8c show that all the inspected language models are SOS biased against disabled groups, especially BERT and AlBERT which result in higher SOS bias scores.



(a) SOS_{LM} bias scores for the religion-sensitive attribute.



(b) SOS_{LM} bias scores for the social class-sensitive attribute.



(c) SOS_{LM} bias scores for the disability-sensitive attribute.

Fig. 5.8 SOS_{LM} bias scores for the marginalised and non-marginalised identities for the different models for the Religion, Disability, and Social class.

5.4.4 SOS bias and other social bias in contextual word embeddings

To answer the second part of RQ2 regarding contextual word embeddings, *Does SOS bias in the contextual word embeddings differ from social biases?*, I investigate how different the measured SOS bias is from social bias in the inspected language models. I measure the Pearson correlation between the SOS bias scores measured using the proposed SOS_{LM} metric and the social bias scores measured using CrowS-Pairs, StereoSet, and SEAT metrics reported in chapter 4. The correlation is measured for three sensitive attributes: race, gender, and religion as these attributes are the common attributes between all three metrics SEAT, CrowS-Pairs, and StereoSet.

Figure 5.9 shows that, unlike static word embeddings, there is a positive correlation between the measured SOS bias scores and social bias scores measured using different bias metrics. However, the positive correlation is not consistent across the different sensitive attributes. For the race-sensitive attribute, there is a strong positive correlation between SOS bias scores and social bias scores measured using the SEAT metric. As for the gender-sensitive attribute, there is a positive correlation between SOS bias scores and social bias scores measured using the Crows-Pairs metric. As for the religion-senstive attribute, there is a strong positive correlation between the SOS bias scores and the social bias scores measured using both CrowS-Pairs and Stereoset metrics. The correlation with Crows-Pairs scores could be because I adapt the CrowS-Pairs metric to measure the SOS bias.

These results suggest that the proposed metric to measure the SOS bias in language models does not reveal different information from that revealed by social bias, especially when measured using the CrowS-Pairs metric.

5.4.5 SOS bias validation in contextual word embeddings

To answer RQ3 regarding contextual word embeddings, *How strongly does SOS bias in contextual word embeddings correlate with external measures of online extremism and hate?*, I measure Pearson correlation coefficients between the online hate statistics reported in Table 5.5 and the SOS bias scores measured using the SOS_{LM} metric for the marginalised groups in the following sensitive attributes: Race, Gender, and Sexual orientation. The results in Figure 5.10, show a strong positive correlation between the SOS bias measured in the inspected language models using the proposed SOS_{LM} metric and the published percentages of marginalised people who experience online hate and extremism in Finland, Germany, the US, and the UK. This strong positive correlation exists for BERT, followed by AlBERT and then RoBERTa.

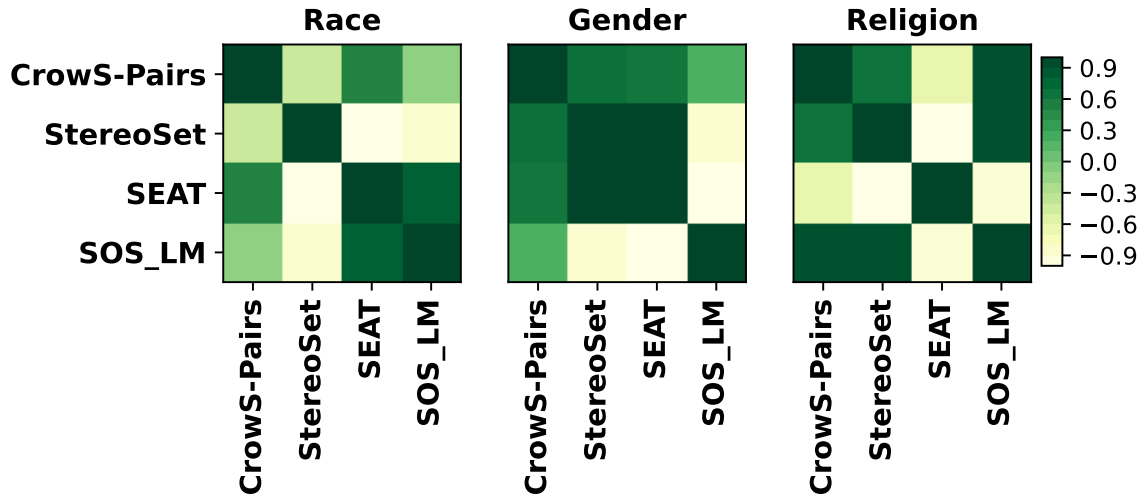


Fig. 5.9 Pearson's correlation (ρ) between the SOS_{LM} bias scores and social bias scored, measured using different metrics for all examined language models.

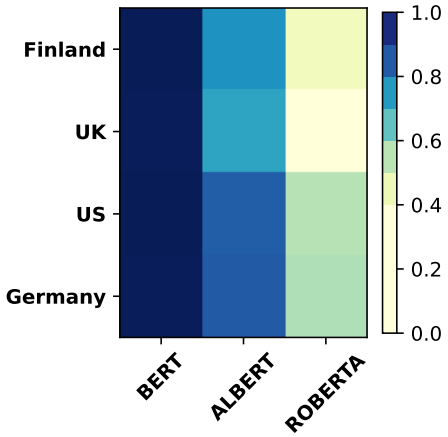


Fig. 5.10 Pearson's correlation (ρ) between the SOS bias scores measured using the SOS_{LM} metric and the percentages of women, non-white ethnicities and LGBTQ groups who experience online hate in different countries.

These results, similar to the early results on static word embeddings, suggest that the proposed metric of measuring SOS bias in language models is reflective of the hate that women, non-white ethnicities, and LGBTQ communities experience online.

5.4.6 Summary

In this part, I propose a metric to measure SOS bias in contextual word embeddings, investigate how different it is from social bias in contextual word embeddings, and validate it. The proposed metric to measure SOS bias builds on the CrowS-Pairs metric used to measure social bias in contextual word embeddings.

The results in this section show that, similar to static word embeddings, contextual word embeddings are SOS biased, especially for race, gender, sexual-orientation, and disability sensitive attributes. However, unlike the static word embeddings, the SOS bias scores are not always higher for marginalised groups. For gender, race, social class, and disability-sensitive attributes, the majority of the models have a higher SOS bias against marginalised groups. On the other hand, for the sexual-orientation and religion-sensitive attributes, the SOS bias scores are higher against non-marginalised groups. In general, the results show that static word embeddings are more SOS biased against marginalised groups than contextual word embeddings. The lower SOS bias scores in contextual word embeddings could be a result of using template sentences, that do not have real context and sometimes have grammatical mistakes, in comparison to realistic sentences. The same speculation is made by May et al. [149] where the authors found lower social bias scores than in static word embeddings and attributed that to using bleached sentences.

The results also show that, there is a strong positive correlation between the measured SOS bias scores in contextual word embeddings and the social bias scores in contextual word embeddings measured using the CrowS-Pairs metric. This means that, unlike the case with static word embeddings, the proposed SOS bias metric does not reveal different information from the one revealed using the social bias metric. I hypothesize that this is the case because the proposed metric to measure the SOS bias in contextual word embeddings is calculated the same as the CrowS-Pairs metric but uses a different dataset.

Finally, the measured SOS bias scores in contextual word embeddings towards marginalised groups correlate positively with published statistics on hate and extremism experienced by the same marginalised groups. This suggests that SOS bias in contextual word embeddings, similar to static word embeddings, is reflective of the online hate experienced by marginalised groups.

Dataset	Samples	Positive samples
HateEval	12722	42%
Twitter-sexism	14742	23%
Twitter-racism	13349	15%
Twitter-hate	5569	25%

Note: Positive samples refer to offensive comments

Table 5.10 Hate speech datasets used with the inspected static word embeddings.

5.5 SOS bias and hate speech detection

In this section, I answer RQ4, *Does the SOS bias in the word embeddings explain the performance of the inspected contextual word embeddings, on the task of hate speech detection?*, through a series of experiments on hate speech detection using static and contextual word embeddings (language models).

5.5.1 Static word embeddings

I train deep learning models with an embedding layer for the detection of hate speech from hate speech-related datasets, then computed the correlation of the performance of the different static word embeddings to the SOS bias score of these embeddings. I use the following hate-speech-related datasets that were used in previous chapter and contain different types of hate speech (Table 5.10): (i) *Twitter-racism* [282]; (ii) *Twitter-sexism* [282]; (iii) *HateEval* [20], from which I use only the English tweets. Additionally, I use the following datasets as well: (iv) Twitter-hate, containing tweets labelled as offensive, hateful (sexist, homophobic, and racist), or neither [59], but as I am interested in the hateful content, I use the tweets that are labelled as hateful or neither;

These four datasets are selected because they contain hate speech towards the marginalised groups that are the focus of this chapter. Thus, they are representative of the examined problem.

To pre-process the datasets, I remov URLs, user mentions, the retweet abbreviation “RT”, non-ASCII characters, and English stop words except for second-person pronouns like “you/yours”, and third-person pronouns like “he/she/they”, “his/her/their” and “him/her/them”, as followed in chapter 4. All letters are lowercased, and common contractions are converted to their full forms. And each dataset is randomly split into a training (70%) and a test (30%) set, preserving class ratios.

I use two deep learning models: (i) a bidirectional LSTM, as shown in Schuster and Paliwal [227] with the same architecture as in Agrawal and Awekar [5], which used RNN models to detect hate speech, and (ii) a two-layer Multi-Layer Perceptron (MLP) model. To this end, I first use the Keras tokenizer, as shown in Tensorflow.org [260] to tokenize the input texts, using a maximum input length of 64 (maximum observed sequence length in the dataset). A frozen embedding layer, based on a given pre-trained word embedding model, is used as the first layer and fed to the BiLSTM model and the MLP model. To avoid over-fitting, I use L2 regularization with an experimentally determined value of 10^{-7}. The models are trained for 100 epochs with a batch size of 32, using the Adam optimizer and a learning rate of 0.01 (default of Keras Optimizer), as shown in Agrawal and Awekar [5]. For each dataset, I use a 5-fold cross-validation to train and validate a model (70% and 30% of the training set, respectively, with the class ratio preserved) and then test each fold's model on the test set. Then, the average F1-score across the five folds is reported.

Results

Given the results for the SOS bias in the different embeddings (Table 5.4), I hypothesize that the deep learning models that are trained with Glove-CC-large, FastText-CC-subwords, SSWE, and Debias-W2V embeddings will perform the best (highest F1 score) on datasets that contain hate speech or insults towards marginalised ethnicities, which is Twitter-racism. I also hypothesize that the models trained with Glove-Twitter, Chan, Glove-CC, and Fast-text-wiki-subwords will achieve the highest F1 scores on datasets that contain insults towards women, which is Twitter-sexism. Since Twitter-Hate and HateEval contain a mixture of hateful content towards women and immigrants, I hypothesize that the best performing static word embeddings would be the ones that have SOS scores higher than the median values for both of SOS_{women} (0.473) and $SOS_{\text{Non-white}}$ (0.456), which are Glove-Twitter, Fast-text-wiki-subwords, and SSWE.

The performance of the deep learning models with the different embedding models is reported in Table 5.11. The results show that for all datasets, BiLSTM outperforms MLP in terms of F1 score. The results also show that for the MLP model, the hypotheses hold for the Twitter-racism dataset, as the best performing models are BiLSTM with Fast-text-CC-subwords and MLP with Glove-CC-large. However, for Twitter-sexism, HateEval, and Twitter-Hate, the results do not support the hypothesis, with Fast-text-CC and Glove-CC-large being the best performing with MLP and BiLSTM models. To quantify the analysis, I use Spearman's correlation between the SOS bias scores, measured using the different bias metrics, of the different static word embeddings and the F1 scores of the MLP and BiLSTM trained with the different static word embeddings. The results in Table 5.12 show

Word embeddings	HateEval		Twitter-Hate		Twitter-racism		Twitter-sexism	
	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM	MLP	BiLSTM
W2V	0.593	0.663	0.681	0.772	0.683	0.717	0.587	0.628
Glove-WK	0.583	0.651	0.713	0.821	0.681	0.727	0.587	0.641
Glove-Twitter	0.623	0.671	0.775	0.851	0.680	0.699	0.589	**0.668**
UD	0.597	0.652	0.780	0.837	0.679	0.698	0.578	0.632
Chan	**0.627**	0.661	0.692	0.840	0.650	0.712	0.563	0.647
Glove-CC	0.625	**0.675**	0.778	0.839	0.695	0.740	0.577	0.648
Glove-CC-large	0.626	0.674	0.775	**0.860**	**0.709**	0.724	0.593	**0.668**
FT-CC	**0.627**	**0.675**	**0.792**	0.843	0.701	0.741	**0.607**	0.654
FT-CC-sws	0.605	0.660	0.746	0.830	0.701	**0.746**	0.588	0.657
FT-WK	0.606	0.650	0.784	0.827	0.699	0.706	0.601	0.653
FT-WK-sws	0.606	0.650	0.723	0.820	0.689	0.736	0.561	0.633
SSWE	0.558	0.628	0.502	0.715	0.324	0.666	0.171	0.548
Debiased-W2V	0.626	0.652	0.678	0.741	0.674	0.715	0.564	0.638
P-DeSIP	0.575	0.657	0.697	0.817	0.673	0.731	0.538	0.650
U-DeSIP	0.598	0.649	0.702	0.815	0.673	0.726	0.548	0.638

Table 5.11 F1 scores for the used models for hate speech detection using the examined static word embeddings on the examined datasets. Bold values indicate the highest scores among the different static word embeddings per model and dataset.

occasionally positive correlations, for example with WEAT, RNSB, and the proposed metric, NCSP. However, most of these positive correlations are not statistically significant, except for the SOS scores measured by the RNSB metric and the F1 of the BiLSTM model and the HateEval dataset. These results indicate that there is no positive correlation between the SOS bias scores in the static word embeddings and the performance of the hate speech detection models, suggesting that the SOS bias in the static word embeddings does not explain their utility as features for hate speech detection.

5.5.2 Contextual word embeddings

To investigate the impact of the SOS bias on the performance of hate speech detection models trained with contextual word embeddings, I train the BERT-base-uncased, AlBERT-base, and ROBERTA-base models on the (i) *Twitter-sexism*, (ii) *Twitter-racism*, (iii) *WTP-Toxicity*, a collection of conversations from Wikipedia Talk Pages (WTP) annotated as friendly or toxic [289], (iv) *WTP-Aggression*, conversations from WTP annotated as friendly or aggressive [289], (v) *Jigsaw-tox* dataset, which is released in a Kaggle challenge, as shown in [29], and

Dataset	Model	WEAT	RNSB	RND	ECT	NCSP
HateEval	MLP	**0.277**	0.223	-0.100	0.019	0.230
	BiLSTM	0.377	**0.540***	0.094	-0.030	0.100
Twitter Sexism	MLP	**0.157**	0.030	-0.216	-0.039	0.121
	BiLSTM	0.109	**0.266**	0.093	-0.361	0.246
Twitter Racism	MLP	0.042	0.017	-0.336	-0.223	**0.241**
	BiLSTM	-0.264	**0.135**	-0.210	-0.103	0.110
Twitter Hate	MLP	0.107	0.218	-0.164	-0.148	**0.223**
	BiLSTM	**0.507**	0.475	0.289	-0.217	0.396

*Statistically significant at $p < 0.05$.

Table 5.12 Pearson correlation coefficient (ρ) of the SOS bias scores of the different static word embeddings and the F1 scores of the used models for each bias metric and dataset. * indicates that the correlation is statistically significant at $p < 0.05$.

Dataset	Samples	Positive samples
Twitter-sexism	14742	23%
Twitter-racism	13349	15%
Jigsaw-tox	298695	0.08%
Kaggle-insults	7425	35%
WTP-agg	114649	13%
WTP-tox	157671	10%

Note: Positive samples refer to offensive comments

Table 5.13 Hate speech datasets used with the inspected language models.

(vi) *Kaggle-Insults* [115], a dataset that contains social media comments that are labelled as insulting or not. The datasets used in this section are described in Table 5.13.

I follow the same pre-processing steps described before, in section 5.5 in addition to the following pre-processing steps described in Dang et al. [56]: (1) remove URLs, user mentions, non-ASCII characters, and the retweet abbreviation "RT" (Twitter datasets). (2) All letters are lowercased. (3) Contractions are converted to their formal format. (4) A space is added between words and punctuation marks.

I fine-tune BERT-base, AlBERT-base and RoBERTa-base on the datasets described in Table 5.13 with 40% training set, 30% validation set and 30% test set. I train the models for 3 epochs, using a batch size of 32, a learning rate of $2e^{-5}$, and a maximum text length of 61 tokens. The results report the F1-scores in Table 5.14.

Dataset	BERT	ALBERT	ROBERTA
Kaggle	0.844	0.832	**0.847**
Twitter-sexism	0.871	**0.884**	0.880
Twitter-racism	**0.930**	0.924	0.929
WTP-agg	0.937	**0.939**	0.934
WTP-toxicity	0.960	0.961	**0.963**
Jigsaw-tox	0.582	0.558	**0.589**

Table 5.14 F1 scores of the different contextual word embeddings on the different hate speech dataset.

Dataset	Race	Gender	Sexual orientation	Religion	Disability	Social class
Kaggle	-0.049	**0.903**	-0.371	**0.912**	-0.574	-0.297
Twitter-sexism	-0.772	-0.195	**0.966**	-0.216	-0.315	-0.589
Twitter-racism	**0.292**	**0.705**	-0.664	**0.719**	-0.262	0.043
WTP-agg	**0.477**	-0.999	-0.068	-0.999	**0.872**	**0.682**
WTP-toxicity	-0.945	**0.732**	**0.724**	**0.718**	-0.973	-0.996
Jigsaw-tox	-0.075	**0.915**	-0.346	**0.923**	-0.595	-0.323

Table 5.15 Pearson Correlation Coefficient (ρ) between the SOS bias scores against the marginalised groups in the inspected LMs and the F1 scores of the different LMs on each dataset.

Results

I compute the Pearson correlation coefficient ($'rho$) between the F1-scores of the contextual word embeddings displayed in Table 5.14 and the SOS bias scores against the marginalised identities displayed in Table 5.9. The results, in Table 5.15, show a strong positive correlation in all the datasets: Twitter-racism (Race, gender, and Religion); WTP-agg (Race, Disability, and Social class); and WTP-toxicity (Gender, Sexual-orientation, and Religion); Kaggle (Gender, Religion); Jigsaw-tox (Gender, and Religion); and Twitter-sexism (Sexual orientation). However, these results are not consistent across the different sensitive attributes and datasets.

5.6 Conclusion

In this chapter, I presented my fourth research contribution and introduced the SOS bias, proposed different metrics to measure it to validate it, and compared it to stereotypical s ocial bias in static and contextual word embeddings. Then, I investigated if the SOS bias explains the performance of the static and contextual word embeddings on hate speech detection.

The results indicate that the examined static and contextual word embeddings are SOS biased. As indicated in static word embeddings with SOS bias scores that range between 0.2 to 0.88 against women, 0.446 to 0.669 against LGBTQ, and 0.234 to 0.688 for non-white ethnicities. As for the contextual word embedidngs, the SOS bias scores range from 0.448 to 0.517 for women, 0.57 to 0.67 for LGBTQ, and 0.519 to 0.580 for non-white ethnicities. The results show the SOS bias in both static and contextual word embeddings has a strong positive correlation with published statistics on online extremism. However, more datasets need to be collected to provide stronger evidence, especially data from the social sciences on the offenses that marginalised groups receive on social media. Nonetheless, this is an informative finding as it reveals the bias in the dataset on which these word embeddings are trained. Since not all these datasets are available to the public, measuring the SOS bias in the word embeddings is an important way to learn about that bias in those datasets.

The results indicate that the measured SOS bias scores in static word embeddings are higher for marginalised groups. However, this is not always the case with contextual word embeddings, where the measured SOS bias is sometimes higher towards non-marginalised groups. The findings, for static word embeddings, show that the proposed SOS bias reveals different information from the one revealed by social bias measured by existing metrics. However, this is not the case with contextual word embeddings. Finally, the findings show no evidence that the SOS bias, measured using different bias metrics, explains the performance of the different word embeddings, static or contextual, on the task of hate speech detection. In the next chapter, I present my fifth and final research contribution, investigating the impact of bias in contextual word embeddings on the fairness of the downstream task of hate speech detection.

Chapter 6

The Fairness Perspective

6.1 Introduction

Natural language processing models are being deployed in every aspect of our lives, from recommending what products to buy to CV screening. Recent research has shown that these NLP models are not fair and systematically discriminate between people based on factors like ethnicity, gender, sexual orientation, age, disability, and others, as shown in Nangia et al. [172]. The literature suggests four main sources of bias that have an impact on the fairness of NLP models: Label bias, Representation bias, Selection bias, and Overamplification bias , as shown in Hovy and Prabhumoye [106], Shah et al. [229]. The focus of studying bias in the NLP literature has mainly been on representation bias and how it impacts the fairness of NLP models on downstream tasks, as shown in Cao et al. [39], Kaneko et al. [117], Steed et al. [241].

In this chapter, I present my fifth and last research contribution, and investigate the impact of bias in NLP on the downstream task of hate speech detection. First, I investigate three of the four mentioned sources of bias and their impact on the fairness of the downstream task of hate speech detection. I remove these sources of bias and investigate whether it improves the fairness of the hate speech detection models.

I aim to find the most impactful sources of bias and the most effective debiasing techniques to use to ensure that hate speech detection models are fairer. To this end, this work aims to answer the following research questions:

1. (RQ1) What is the impact of the different sources of bias on the fairness of the downstream task of hate speech detection?
2. (RQ2) What is the impact of removing the different sources of bias on the fairness of the downstream task of hate speech detection?

3. (RQ3) Which debiasing technique to use to ensure the fairness of the task of hate speech detection?
4. (RQ4) How to have fairer text classification models?

To answer these questions, I measure the fairness of three language models, AlBERT-base-v2 [134], BERT-base-uncased [65], and RoBERTa-base [142], on the downstream task of hate speech detection using different fairness metrics. Then, to answer the first research question and to understand the impact of the different sources of bias on the models' fairness, I investigate three sources of bias (representation, selection, Overamplification) and their impact on the models' fairness. Then, I use different methods to remove the bias from the different sources (debias), and investigate the impact of these debiasing methods on the models' fairness to answer the second and third research questions. Thereafter, I analyze the debiasing results to find out the most effective technique to ensure the models' fairness on the downstream task of hate speech detection. Finally, to help the NLP community improve the fairness of text classification tasks and to answer the fourth research question, I build on the findings of this chapter on the fairness of hate speech detection as a text classification task and generalize these findings to provide practical general guidelines to follow to ensure the fairness of the downstream task of text classification.

Improving the fairness of the downstream task of text classification, is very critical to ensure that the decisions made by the models are not based on sensitive attributes like race, gender or sexual orientation.

6.2 Related work

In the last few years, various metrics have been proposed in the literature to quantify bias in static word embeddings, as shown in Caliskan et al. [38], Dev and Phillips [64], Elsafoury et al. [74], Garg et al. [86], Sweeney and Najafian [254] and contextual word embeddings (language models), as shown in Guo and Caliskan [94], Kurita et al. [132], May et al. [149], Nadeem et al. [167], Nangia et al. [172]. Other researchers focused on quantifying the NLP models' fairness when used in a downstream task, as shown in Borkan et al. [29], De-Arteaga et al. [61], Qian et al. [204]. Most of these focused on measuring representation bias, which is also known in the literature as intrinsic bias. Among the proposed metrics to measure representation, intrinsic, bias are CrowS-Pairs, as shown in Nangia et al. [172], StereoSet, as shown in Nadeem et al. [167], and SEAT, as shown in May et al. [149].

On the other hand, the fairness of NLP model, also known in the literature as extrinsic bias, is measured when they are used in the downstream task. There are three main approaches

to measuring a model's fairness in that case: Threshold-based metrics, as shown in Cao et al. [39], De-Arteaga et al. [61], Steed et al. [241], Threshold-agnostic metrics, as shown in Borkan et al. [29], Dixon et al. [69], Counterfactual fairness, as shown in Fryer et al. [84], Krishna et al. [129], Kusner et al. [133], Qian et al. [204].

The impact of representation (intrinsic) bias on models' fairness (extrinsic bias) in NLP models is not clear yet. Some researchers found no strong evidence that intrinsic bias impacts extrinsic bias in language models, as shown in Cao et al. [39], Kaneko et al. [117], Steed et al. [241]. However, there are some limitations to those studies. For example, in Steed et al. [241], the authors used two intrinsic bias metrics which use bleached template sentences, which are sentences that do not have a real semantic context, to measure bias, these metrics have been criticized as they may not be semantically bleached, as in May et al. [149]. Moreover, both, Cao et al. [39] and Steed et al. [241] use different intrinsic bias metrics for the two text classification tasks examined, which results in a lack of consistency.

As for measuring models' fairness on downstream tasks, the mentioned studies, Cao et al. [39], Kaneko et al. [117], Steed et al. [241], used only threshold-based extrinsic bias metrics for the text classification task. For example, Cao et al. [39], Kaneko et al. [117] use FPR gap to measure extrinsic bias on hate speech detection. Similarly, Steed et al. [241] use TPR gap to measure extrinsic bias in the task of occupation classification and FPR gap for the task of hate speech detection. Threshold-agnostic metrics have not been widely used to measure fairness and to investigate its correlation to representation bias. Even though, according to Borkan et al. [29], threshold-agnostic metrics can capture the behavior of the model. Moreover, in most of the studies that investigate the fairness of the task of hate speech detection, the authors do not explain how they measure the fairness of the models between the different identity groups, as shown in Cao et al. [39], Steed et al. [241].Additionally, as mentioned before, most of these studies focus on representation bias. Hence, there is a lack of investigation of other sources of bias and their impact on the model's fairness on the downstream task of hate speech detection.

Similarly, there is a lack of investigation of the impact of removing bias on the models' fairness in downstream tasks. For example, Meade et al. [152] investigates the impact that different debiasing approaches have on the performance of different NLP downstream tasks. However, they do not investigate the impact debiasing has on the fairness of the downstream tasks. In Kaneko et al. [117], the authors investigate the effectiveness of different debiasing methods that remove representation bias on the fairness of the downstream tasks, but they do not investigate the effectiveness of removing other sources of bias on the fairness of downstream NLP tasks.

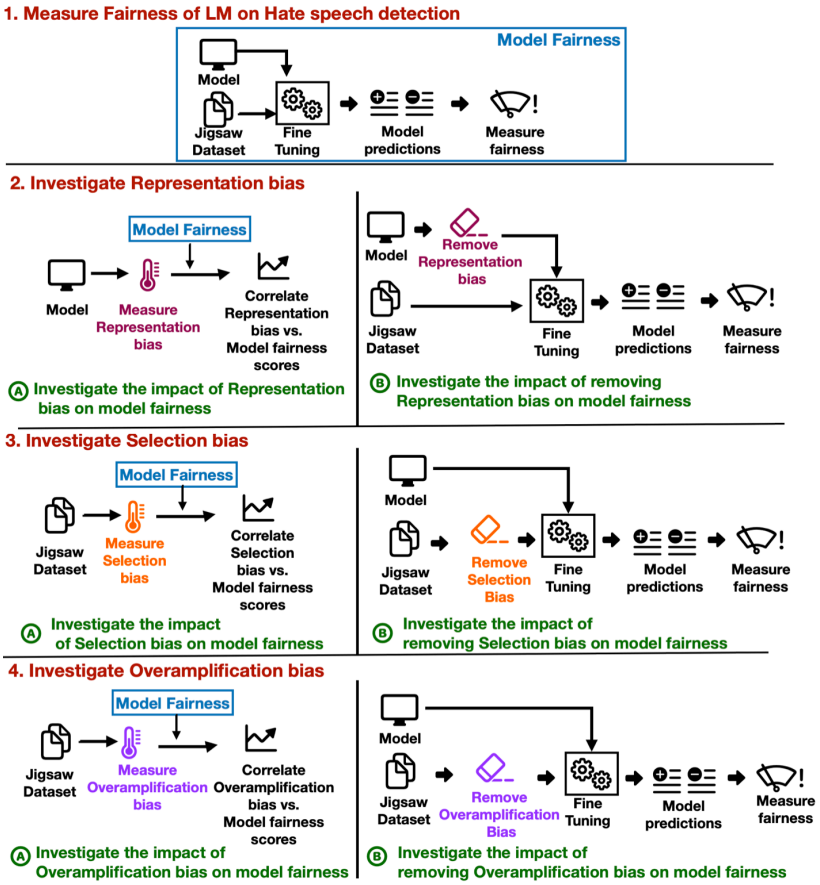


Fig. 6.1 Overview of conducted investigation.

In this chapter, I aim to fill the gaps in the literature by investigating different sources of bias and their impact on the models' fairness in the downstream task of hate speech detection. I aim to overcome the limitations of previous research by using different metrics to measure representation (intrinsic) bias and models' fairness. Moreover, I investigate the effectiveness of various debiasing methods for removing different sources of bias, as well as their impact on the models' fairness (extrinsic bias). I provide practical guidelines to ensure the fairness of the downstream task of text classification.

6.3 Methodology

In this chapter, I perform four groups of experiments to investigate the impact of each source of bias on the fairness of the downstream task of hate speech detection.

Figure 6.1 provides an overview of these four groups of experiments. In Step 1 of Fig. 6.1, I first measure the fairness of the hate speech detection task (section 6.4) using various models

and use these fairness scores as a baseline. Then, In Step 2A of Fig. 6.1, I measure the representation, intrinsic, bias in the inspected models and its impact on the Models' fairness on the task of hate speech detection as well as the impact of removing representation bias (section 6.5.1) as shown In Step 2B of Fig. 6.1. In Steps 3 (A & B) and 4 (A &B) of Fig. 6.1, I repeat the same investigation for selection bias (section 6.5.2) and Overamplification bias (section 6.5.3). Then, I investigate the impact of removing multiple biases on the fairness of the task of hate speech detection (section 6.5.4). Finally, I build on these findings and recommend guidelines to achieve fairer text classification (section 6.7).

6.3.1 Hate speech detection

Dataset

I use the Jigsaw dataset, which is also known as the Civic Community dataset [29]. The dataset contains almost 2 million comments, labelled as toxic or not, along with labels on the identity of the target of the sentence, e.g., religion, sexual orientation, gender, and race. The identity labels provided in the dataset are both crowdsourced and automatically labelled. When I analyze the dataset, I find some issues with the identity labels, e.g., some data items are labeled to contain more than one identity (male and female) as the target of the toxicity.

I pre-process that dataset to keep only the data items where the identity information is labeled by human annotators. Additionally, I follow the same data pre-processing steps used in chapter 4, where the authors train a BERT model for the task of cyberbullying detection. To this end, I remove URLs and non-ASCII characters, lowercase all the letters, convert all contractions to their formal format and add a space between words and punctuation marks. This resulted in 400K data items. The dataset is then split into 40% training, 30% validation, and 30% test sets.

I only use the Jigsaw dataset because, to the best of my knowledge, it is the only available hate speech dataset that contains information on both marginalised and non-marginalised identities, which is important to the way I measure fairness, as explained in section 6.4. Other datasets, like ToxiGen, as shown in Hartvigsen et al. [98], SocialFrame, as shown in Sap et al. [225], and the Ethos dataset, as shown in Mollas et al. [159], and the MLM data, as shown in Ousidhoum et al. [184] contain information only about marginalised groups, and thus cannot be used in this investigation. HateExplain, as shown in Mathew et al. [148] contains information about both marginalised and non-marginalised identities. However, the dataset uses offensive words to refer to marginalised groups, e.g., n*gger to refer to Africans, and identity words to describe non-marginalised groups, e.g., White to refer to Caucasians. This makes the HateXplain dataset unsuitable for the experiments held in section 6.4.1 where I

create data perturbations. As replacing an offensive word to describe marginalised groups with an identity word to describe a non-marginalised identity group changes the meaning of the sentence and hence its label as hateful or not. Unlike the Jigsaw dataset, where identity words are used to describe both marginalised and non-marginalised groups.

Language models

The fairness of the downstream task of hate speech detection is evaluated on the widely used BERT-base-uncased [65], RoBERTa-base [142], and ALBERT-base [134] models, by fine-tuning them on the Jigsaw-toxicity dataset. Following the same experimental setting from chapter 4, the models are fine-tuned for 3 epochs, using a batch size of 32, a learning rate of $2e^{-5}$, and a maximum text length of 61. Classification results using the fine-tuned models indicate that ALBERT-base is the best-performing model, with an AUC score of 0.911, followed by RoBERTa-base with an AUC score of 0.908, and BERT-base with an AUC score of 0.902. The fine-tuned models are then used to measure fairness in the hate speech detection task.

6.4 Fairness in the task of hate speech detection

6.4.1 Measure Fairness using extrinsic bias metrics

To evaluate the fairness of the examined models on the downstream task of hate speech detection, I use two sets of extrinsic bias metrics: (i) Threshold-based, which uses the absolute difference (gap) in the false positive rates (FPR) and true positive rates (TPR) between the marginalised group (g) and non-marginalised group $\hat{g}$, as shown in Equations eq. (6.1) and eq. (6.2), and (ii) Threshold-agnostic metrics, which measure the absolute difference in the area under the curve (AUC) scores between marginalised group (g) and non-marginalised group $\hat{g}$, as shown in Equation eq. (6.3).

$$FPR\_gap_{g,\hat{g}} = |FPR_g - FPR_{\hat{g}}| \tag{6.1}$$

$$TPR\_gap_{g,\hat{g}} = |TPR_g - TPR_{\hat{g}}| \tag{6.2}$$

$$AUC\_gap_{g,\hat{g}} = |AUC_g - AUC_{\hat{g}}| \tag{6.3}$$

These scores express the amount of unfairness in the hate speech detection models, with higher scores denoting less fair models and lower scores denoting fairer models. These metrics are measured between two groups, marginalised and non-marginalised, similar to

Sensitive attribute	marginalised	Non-marginalised
Gender	Female	Male
Race	Black and Asian	White
Religion	Jewish and Muslim	Christian

Table 6.1 The inspected sensitive attributes and identity groups.

the approach followed in chapter 5. Furthermore, I limit this investigations to 3 sensitive attributes, i.e., gender, religion, and race as shown in Table 6.1. In cases where there is more than one identity group in the marginalised group for a sensitive attribute, e.g., Asian and Black vs. White, I then measure the mean of the FPR, TPR, and AUC scores of the two groups, Asian and Black, and then use that score to represent the marginalised group (g).

6.4.2 Balanced Jigsaw fairness dataset

To measure extrinsic bias, I filter the test set to ensure that the data samples contain only one identity group, which resulted in 21K samples to improve the quality of the measured fairness. I find differences in the sizes of the subsets of sentences that mention the different identity groups. For example, the size of the subset of sentences that are targeted at Men is 3716 while the size of the female subset is 6046. I also find differences in the ratio of the positive samples between the different identity groups that belong to the same sensitive attribute. The ratio of positive samples for the male and female groups are 0.12 and 0.10 respectively; for the White, Asian, and Black groups are 0.20, 0.07, and 0.27 respectively; and for the Christian, Muslim, and Jewish groups are 0.05, 0.16, and 0.12 respectively. I hypothesize that these differences between the different identity groups might influence the fairness scores.

To test this hypothesis, I create a balanced Jigsaw fairness dataset and use it to measure the extrinsic bias in the fine-tuned models. To create this balanced Jigsaw fairness dataset, I use lexical word replacement to create perturbations of existing sentences using regular expressions. That is possible with the Jigsaw dataset because after inspecting the most common nouns and adjectives used in each subset that targets a certain identity, I find that the most common words are words that describe that identity. For example, among the most common nouns in the data subset that are targeted at black people are: “black” and “blacks”. The most common nouns in the data subset targeted at Asian people are “asian” and “chinese”. A similar pattern is found for religion and gender identities. However, this approach is not suitable for gender perturbations, as pronouns also change between males and females. To this end, perturbations for the male and female identity groups are created

using the AugLy[1] tool, which is provided by Facebook research to swap gender information , as shown in Papakipos and Bitton [187].

The balanced Jigsaw fairness dataset contains 55,476 samples and has the same ratio between positive and negative samples for each identity group within the same sensitive attribute. For example, for the gender attribute, the ratio of the positive (toxic) examples in the male and female identity groups is 0.10, in the race attribute, the ratio of the positive samples for the Black, White and Asian groups is 0.20, and for the religion attribute, the ratio of the positive samples for the Muslim, Christian and Jewish groups is 0.10.

6.4.3 Fairness results

The Fairness evaluation results for the three examined models on the original and the balanced Jigsaw fairness datasets are shown in Table 6.2. From this table, it is evident that the use of a balanced Jigsaw fairness dataset led to improved fairness scores across most of the extrinsic bias metrics and across all models. I used the Wilcoxon statisitcal significane analaysis for the different models and the different fairness metrics on the original and perturbed fairness dataset. The results for Albert model show no statistically significant difference in the fairness scores on the original and the perturbed fairness dataset (Wilcoxon $p-value$ for FPR_gap = 1, Wilcoxon $p-value$ for TPR_gap = 1, Wilcoxon $p-value$ for AUC_gap = 0.5). Simlar results found for BERT (Wilcoxon $p-value$ for FPR_gap = 1, Wilcoxon $p-value$ for TPR_gap = 0.75, Wilcoxon $p-value$ for AUC_gap = 0.25) and RoBERTa (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 1, Wilcoxon $p-value$ for AUC_gap = 0.25)

This finding, even with no statistical significance difference, suggests that the dataset used to measure the fairness in the downstream task of text classification impacts the measured fairness, and it is important to ensure that there is a balanced representation of the different identity groups to get a reliable fairness score. This is a critical finding that, to the best of my knowledge, has not been mentioned in the literature on measuring fairness (extrinsic bias) before.

The results also indicate that when I use the original imbalanced Jigsaw fairness dataset, the different metrics used to measure the extrinsic bias reported different extrinsic bias scores related to each sensitive attribute in the fine-tuned models. I use Pearson's correlation coefficient (ρ) to measure how different the extrinsic bias metrics are. I find that even though the FPR_gap and TPR_gap are both threshold-based metrics, there is no positive correlation between the two metrics for the three models. There is a negative correlation

[1] `https://augly.readthedocs.io/en/latest/README.html`

Attribute	Model	Dataset	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	Original	0.001	0.081	0.025
		Balanced	↑ 0.006	↓ 0.038	↓ 0.003
	BERT	Original	0.002	0.111	0.026
		Balanced	↑ 0.008	↓ 0.036	↓ 0.009
	RoBERTa	Original	0.007	0.084	0.017
		Balanced	↓ 0.004	↓ 0.031	↓ 0.011
Race	ALBERT	Original	0.007	0.044	0.003
		Balanced	↑ 0.008	↓ 0.0016	↑ 0.018
	BERT	Original	0.008	0.017	0.048
		Balanced	↑ 0.015	↓ 0.002	↓ 0.025
	RoBERTa	Original	0.014	0.127	0.028
		Balanced	↓ 0.003	↓ 0.011	↓ 0.021
Religion	ALBERT	Original	0.019	0.060	0.042
		Balanced	↓ 0.009	↑ 0.108	↓ 0.020
	BERT	Original	0.016	0.027	0.051
		Balanced	↓ 0.008	↑ 0.062	↓ 0.012
	RoBERTa	Original	0.027	0.030	0.0369
		Balanced	↓ 0.021	↑ 0.160	↓ 0.027

Table 6.2 The fairness scores of the different models on the original and the balanced Jigsaw fairness datasets using the examined models. (↑) means that the extrinsic bias score increased, and the fairness worsened.(↓) means that the extrinsic bias score decreased and the fairness improved.

between TPR_gap and FPR_gap (ρ = -0.37), a negative correlation between the TPR_gap and the AUC_gap scores for the three models (ρ = -0.42), and a positive correlation between the FPR_gap and the AUC_gap for the models (ρ = 0.46).

On the other hand, when I use the balanced Jigsaw fairness dataset to measure fairness, I find a positive correlation between all the extrinsic bias metrics in all three models. There is a positive correlation between the FPR_gap and the TPR_gap scores (ρ = 0.59), a positive correlation between FPR_gap and AUC_gap scores (ρ = 0.64), and a positive correlation between the TPR_gap and the AUC_gap scores (ρ = 0.27). This is another evidence that using a fairness dataset with a balanced representation of the different identity groups leads to more reliable fairness scores. For the rest of the chapter, I will use the balanced Jigsaw fairness dataset to measure fairness. In the next sections, to investigate the impact of the different sources of bias, I use Pearson's correlation between the bias scores and the fairness scores measured in this section similar to Cao et al. [39], Kaneko et al. [117], Steed et al. [241].

6.5 Sources of bias

Shah et al. [229] consider four sources of bias in NLP models that impact a model's fairness, which are representation bias, label bias, selection bias, and overamplification bias. In this work, I examine how the following three sources of bias impact the examined models' fairness on the task of hate speech detection: (i) representation bias, (ii) selection bias, and (iii) overamplification bias. I do not investigate Label bias because there is no information available on the annotators of the examined Jigsaw dataset. Furthermore, I remove each source of bias and investigate the impact of each bias removal on the fairness of hate speech detection.

6.5.1 Representation bias

Representation, intrinsic, bias describes the societal stereotypes that language models encoded during pre-training. I use three metrics to measure representation bias, CrowS-Pairs, as shown in Nangia et al. [172], StereoSet, as shown in Nadeem et al. [167], and SEAT, as shown in May et al. [149] to measure three types of social bias: gender, religion, and race as shown in Table 6.3. The results indicate that RoBERTa is the most biased according to CrowS-Pairs, StereoSet, and SOS_LM metrics. In addition to social bias, I use the SOS_{LM} metric, which is explained in, chapter 5 to measure the SOS bias in the inspected language models. I use only the SOS bias scores measured towards the marginalised groups.

I investigate the impact of representation bias in the inspected models, BERT, ALBERT, and RoBERTa, on their fairness on the task of hate speech detection. To measure that impact, I measure the Pearson's correlation coefficient (ρ) between fairness scores measured by the different extrinsic bias metrics and the representation bias scores measured by the different representation bias metrics (Figure 6.2) (right). I find a consistent positive correlation between the CrowS-Pairs intrinsic bias scores with the extrinsic bias scores measured by all three extrinsic bias metrics (FPR_gap. TPR_gap and AUC_gap) for all the models and sensitive attributes. There is a positive correlation between SOS_{LM} and FPR_gap and AUC_gap. There is a consistent negative correlation between SEAT scores and all extrinsic bias metrics. On the other hand, there is an inconsistent correlation with the StereoSet scores. This finding is different from previous research that suggested that there is no correlation between representation bias and fairness scores, as shown in Cao et al. [39], Kaneko et al. [117], Steed et al. [241]. I hypothesize that previous research did not use a balanced fairness dataset, which is why they did not find a consistent positive correlation with extrinsic bias metrics. To test this hypothesis, I measure the Pearson correlation coefficient (ρ) between

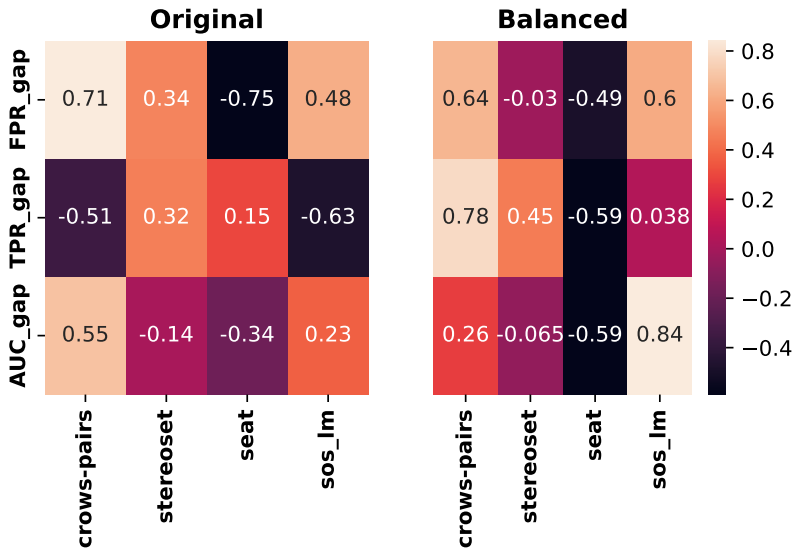


Fig. 6.2 Heatmap of Pearson's correlation between representation bias scores of all LMs and fairness scores of LMs on the downstream task of hate speech detection, on the original Jigsaw fairness dataset (left) and the balanced Jigsaw fairness dataset (right), for all the sensitive attributes.

representation bias scores and fairness scores measured using the different extrinsic bias metrics on the original Jigsaw fairness dataset. I find no consistent correlation between representation and fairness, which supports the hypothesis as shown in Figure 6.2 (left). Another reason why previous research has not found a consistent positive correlation is that they used only one metric for either representation or extrinsic bias. However, using more than one metric helped to reveal this correlation.

Representation bias removal

I use SentDebias, as shown in Liang et al. [139] model to remove representation bias from the models by making the representation orthogonal to the bias subspace. I remove gender, racial, religious and SOS bias from the inspected models, following the same approach as Meade et al. [152]. As for removing the SOS bias from the inspected language models, I use SentDebias to remove the SOS bias by finding and removing the profanity subspace using the 21 profane and nice words reported in Table 5.7 in chapter 5. SentDebias seems to reduce the bias scores as shown in Table 6.3, in all the models according to the StereoSet and CrowS-Pairs metrics. The results also show that, in some cases, removing the SOS bias improved the social bias scores. On the other hand, according SOS_{LM} SentDebias improved the SOS bias scores in BERT-base, but the results are inconsistent for AlBERT and RoBERTa.

Model	CrowsPairs			StereoSet			SEAT			SOS_LM		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
AlBERT-base	0.541	0.513	0.590	0.599	0.575	0.603	0.622	0.551	0.430	0.448	0.542	0.495
+ SentDebias-gender	↓0.461	↓0.436	↓0.466	↓0.517	↓0.552	↓0.586	0.622	0.551	0.430	↑0.591	↑0.728	↓0.342
+ SentDebias-race	↑0.564	↓0.440	↑0.666	↓0.542	↓0.521	↓0.555	0.622	0.551	0.430	↑0.585	↓0.414	↑0.838
+ SentDebias-religion	↑0.549	↑0.660	↓0.581	↓0.501	↓0.529	↓0.510	0.622	0.551	0.430	↑0.571	↓0.509	↑0.590
+ SentDebias-SOS	↓0.503	↑0.743	↑0.714	↓0.504	↓0.468	↓0.539	0.622	0.551	0.430	↑0.639	↓0.504	↓0.485
BERT-base-uncased	0.580	0.581	0.714	0.607	0.5702	0.597	0.620	0.620	0.491	0.476	0.580	0.523
+ SentDebias-gender	↓0.427	↓0.555	↓0.647	↓0.475	↓0.476	↓0.504	0.620	0.620	0.491	↓0.435	↓0.528	↑0.676
+ SentDebias-race	↓0.534	↓0.398	↓0.704	↓0.467	↓0.562	↓0.489	0.620	0.620	0.491	↓0.367	↓0.228	↓0.457
+ SentDebias-religion	↓0.534	↑0.675	↓0.561	↓0.469	↓0.511	↓0.399	0.620	0.620	0.491	↓0.346	↓0.461	↓0.381
+ SentDebias-SOS	↓0.572	↓0.473	↓0.609	↓0.485	↓0.430	↓0.436	0.620	0.620	0.491	↑0.782	↑0.581	↓0.361
RoBERTa-base	0.606	0.527	0.771	0.663	0.616	0.642	0.939	0.307	0.126	0.517	0.519	0.561
+ SentDebias-gender	↓0.467	↑0.691	↓0.561	↓0.518	↓0.497	↓0.477	0.939	0.307	0.126	↑0.591	↑0.674	↓0.419
+ SentDebias-race	↓0.429	↓0.467	↓0.419	↓0.485	↓0.488	↓0.486	0.939	0.307	0.126	↑0.598	↑0.547	↓0.352
+ SentDebias-religion	↓0.413	↓0.478	↓0.352	↓0.516	↓0.497	↓0.486	0.939	0.307	0.126	↑0.781	↑0.695	↓0.228
+ SentDebias-SOS	↓0.494	↑0.567	↓0.361	↓0.517	↓0.463	↓0.457	0.939	0.307	0.126	↑0.734	↓0.285	↓0.438

Table 6.3 Bias scores in the different models using different bias metrics before and after removing bias using SentDebias algorithm. (↑) means that the intrinsic bias score increased and the fairness worsened.(↓) means that the intrinsic bias score decreased and the model improved.

On the other hand, the SEAT metric, did not show any difference in the bias scores for the debiased models, unlike the reported scores in Meade et al. [152]. This could be due to the different settings of the experiments where I debias the models, save them to disk and then load them and measure the bias. Unlike the experimental setup in Meade et al. [152] where they remove the bias and measure the new bias scores at the same time. I keep the experimental setup because it reflects the realistic settings for using a debiased model in the downstream task of text classification. I find that, according to some metrics, removing one type of bias sometimes leads to exacerbating another type of bias. For example, in Table 6.3, removing racial bias increased the gender bias and removing religion bias increased racial bias in AlBERT-base according to CrowS-Pairs metric. The same finding is reported in **??**.

Furthermore, I investigate the impact of removing representational bias on the models' fairness. I fine-tune BERT-base, ALBERT-base and RoBERTa-base after debiasing them to remove gender, racial and religious bias. I remove only these biases to match the three sensitive attributes used to measure the models' fairness. Then, I measure the fairness of the models using different extrinsic bias metrics, threshold-based and threshold-agnostic metrics. The results in Table 6.4 indicate that removing representation bias did not change the AUC scores much, but removing gender bias information increased slightly the AUC scores, especially for BERT and RoBERTa. This is because the debiased models tend to predict more positive class examples, leading to more true positives and more false positives.

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ upstream-sentDebias-gender	0.840	0.006	↓ 0.032	0.004
	BERT	0.830	0.090	0.036	0.010
	+ upstream-sentDebias-gender	**0.841**	↓ 0.011	↑ 0.049	↓ 0.006
	RoBERTa	0.851	0.005	0.032	0.011
	+ upstream-sentDebias-gender	**0.856**	↑ 0.006	↓ 0.022	↓ 0.003
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ upstream-sentDebias-race	0.838	↓ 0.003	↑ 0.003	↓ 0.013
	BERT	**0.830**	0.016	0.002	0.026
	+ upstream-sentDebias-race	0.829	↑ 0.021	↑ 0.005	↓ 0.024
	RoBERTa	0.851	0.003	0.011	0.021
	+ upstream-sentDebias-race	**0.854**	↑ 0.017	↓ 0.009	0.021
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ upstream-sentDebias-religion	0.837	↑ 0.019	↓ 0.094	↓ 0.016
	BERT	0.830	0.008	0.063	0.012
	+ upstream-sentDebias-religion	**0.833**	↑ .015	↑ 0.084	↑ 0.017
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ upstream-sentDebias-religion	0.843	↓ 0.021	↓ 0.100	↓ 0.003

Table 6.4 Hate speech detection performance and fairness scores for all models before and after removing representation bias using SentDebias. (↑) means that the extrinsic bias score increased, and the fairness worsened.(↓) means that the extrinsic bias score decreased and the fairness improved.

To simplify the analysis of the results, I investigate the impact of removing a certain type of bias on the fairness of the matching sensitive attribute. For example, I analyze the impact of removing gender bias from the model representation (+ Upstream-SentDebias-gender) on the fairness of the models regarding the gender-sensitive attribute. For most of the models, the majority of the extrinsic bias metrics show that removing a certain type of bias from the model representation (upstream) using SentDebais did not improve fairness for the corresponding sensitive attribute. There are improvements according to certain metrics, but these improvements are inconsistent across sensitive attributes and models. As for the cases where all extrinsic bias metrics show improvement in fairness, we find that removing religion bias from RoBERTa-base representations improved the models' fairness for the religion sensitive attributes in the downstream task of hate speech detection. On the other hand, sometimes removing different types of bias from the model representation led to improvement in fairness for a different sensitive attribute. For example, in the BERT model according to the FPR_gap and TPR_gap metric, removing racial bias information from the model's representation, improved the fairness for the gender-sensitive attributes. Yet again, these findings are inconsistent for all models and for all extrinsic bias metrics. When I ran the Wilcoxon statistical significant test, there was no found statistically significant difference

between the fairness scores before and after removing upstream bias. The significant test results are: for Alber (Wilcoxon $p-value$ for FPR_gap = 0.65, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.5), for Bert (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 2), and for Roberta (Wilcoxon $p-value$ for FPR_gap = 0.5, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 0.179).

These results suggest that even though there is a positive correlation between representation bias in the inspected models and models' fairness on the task of hate speech detection, removing representational bias did not lead to an improvement in the models' fairness. Similar findings are made by Kaneko et al. [117]. This could be because the current measures used to remove representational bias are superficial, as argued in Gonen and Goldberg [90].

6.5.2 Selection bias

Selection bias is a result of non-representative observations in the datasets used in downstream tasks, as shown in Shah et al. [229]. For the task of hate speech detection, I interpret selection bias as the over-representation of a certain identity group with the positive (toxic) class, as shown on the left of Figure 6.3 (Original). I measure selection bias in the Jigsaw-toxicity training dataset by measuring the difference in the ratios of the positive examples, between the marginalised and non-marginalised groups. Equation 6.4 shows how to measure selection bias ($Selection_{g,\hat{g}}$) where (N_g) is the size of the data subset that is targeted at marginalised groups (g); ($N_{\hat{g}}$) is the size of the data subset that is targeted at non-marginalised groups ($\hat{g}$); ($N_{g,toxicity=1}$) is the number of toxic sentences that are targeted at marginalised groups; and ($N_{\hat{g},toxicity=1}$) is the number of toxic sentences that are targeted at non-marginalised groups. The results indicate that the selection bias is the highest in the sensitive attribute of religion (0.077), followed by race (0.053), and finally gender (0.027).

$$Selection_{g,\hat{g}} = |(\frac{N_{g,toxicity=1}}{N_g}) - (\frac{N_{\hat{g},toxicity=1}}{N_{\hat{g}}})| \tag{6.4}$$

To measure the impact of selection bias on the fairness of the hate speech detection task, I use the Pearson's correlation coefficient (ρ) between the fairness scores measured by the different extrinsic bias metrics and selection bias scores in the Jigsaw training dataset. I find that, for AlBERT-base, selection bias scores have a strong positive correlation with the fairness scores when measured as FPR_gap (ρ = 0.984), AUC_gap (ρ = 0.911), and TPR_gap (ρ = 0.633). These correlations are not significant, but that could be because there are only a few data points used to measure the correlation. Similar correlation patterns found

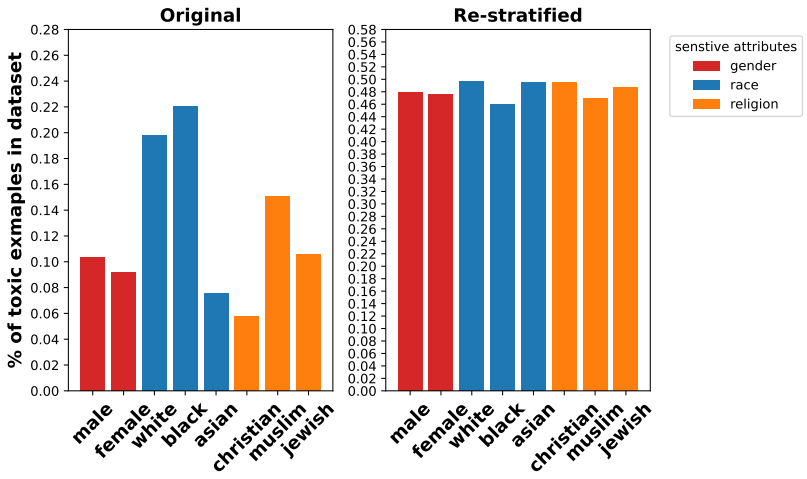


Fig. 6.3 The percentage of positive (toxic) examples, for each identity group in the Jigsaw training dataset in the original dataset (left) and after re-stratification (right).

in RoBERTa-base. As for BERT-base, there are weaker positive correlations with TPR_gap and AUC_gap and almost no correlation with FPR_gap.

These results suggest that selection bias in training datasets used in downstream tasks, has a direct impact on the fairness of the inspected language models, as evident by the positive correlations with the fairness of these models on the downstream task of hate speech detection as measured by the different extrinsic bias metrics.

Selection bias removal

According to Shah et al. [229], to remove selection bias, a realignment in the sample distribution in the training dataset is required to minimize the mismatch in the class representation between the different identities. The authors in Shah et al. [229] list data re-stratification as a mitigation technique to remove selection bias and to match the ideal distribution of balanced class representations for the different identities. I follow this suggestion to have a balanced representation of positive and negative examples for the different marginalised and non-marginalised groups that I study in this chapter.

I follow the same methodology used in Zmigrod et al. [303] to use data augmentation to create slightly altered examples to balance the class representations and add them to the Jigsaw training dataset. Since the percentages of the positive examples for the different identity groups are small, ranging from 0.05 to 0.2 as shown on the left of Figure 6.3 (Original), I create positive examples by altering existing positive examples in the dataset using word substitutions. I generate the word substitutions using the NLPAUG tool[1] that

[1] https://github.com/makcedward/nlpaug

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	**0.847**	0.006	0.039	0.004
	+ downstream-stratified-data	0.816	↓0.005	↓0.003	↑0.005
	BERT	**0.830**	0.090	0.036	0.010
	+ downstream-stratified-data	0.817	↓0.007	↓0.006	↓0.006
	RoBERTa	**0.851**	0.005	0.032	0.011
	+ downstream-stratified-data	0.842	↑0.006	↓0.005	↓0.002
Race	ALBERT	**0.847**	0.008	0.002	0.019
	+ downstream-stratified-data	0.816	↑0.022	↑0.026	↓0.008
	BERT	**0.830**	0.016	0.002	0.026
	+ downstream-stratified-data	0.817	↓0.010	↑0.018	↓0.008
	RoBERTa	**0.851**	0.003	0.011	0.021
	+ downstream-stratified-data	0.842	↑.014	0.011	↓0.014
Religion	ALBERT	**0.847**	0.010	0.109	0.020
	+ downstream-stratified-data	0.816	↑0.030	↓0.058	↓0
	BERT	**0.830**	0.008	0.063	0.012
	+ downstream-stratified-data	0.817	↑0.020	↓0.049	↓0.006
	RoBERTa	**0.851**	0.022	0.160	0.027
	+ downstream-stratified-data	0.842	↓0.019	↓0.071	↓0.001

Table 6.5 Hate speech detection performance and fairness scores for all models before and after removing selection bias. (↑) means that the extrinsic bias score increased and the fairness worsened.(↓) means that the extrinsic bias score decreased and the fairness improved.

uses contextual word embeddings to find the word substitutions, as shown in Ma [143]. After adding the synthesized positive samples to the Jigsaw training dataset, the new, re-stratified Jigsaw training dataset contains 443046 data items with balanced class representation, as shown on the right of Figure 6.3 (re-stratified). The selection bias in the re-stratified Jigsaw training dataset is reduced to 0.002, 0.019 and 0.017 for gender, race, and religion sensitive attributes respectively.

I then, fine-tune the inspected models, AlBERT, BERT, and RoBERTa, on the new re-stratified Jigsaw training dataset. Balancing the class representation in the dataset led to a reduction in the performance, AUC scores, of all three models (+ downstream-stratified-data), as shown in Table 6.5. This reduction in the AUC scores is a result of predicting more positive examples than the original models.

To investigate the impact of selection bias removal on the fairness of the task of hate speech detection, I measure fairness in all three inspected models after fine-tuning them on the new re-stratified dataset. I analyze the fairness scores for the different sensitive attributes, using the different extrinsic bias metrics in all the inspected models. I find that for the AUC_gap metric, the fairness improved for all models and most of the sensitive attributes as evident

in AlBERT (race, religion), BERT (gender, race, religion), and RoBERTa (gender, race, religion), as in Table 6.5. However, the results are inconsistent for the TPR_gap or FPR_gap across models or sensitive attributes. I ran Wilcoxon statistical significance test if using re-stratified data significantly improved the fairness of hate speech detection. The statistical significant test shows that there is no statistically significant improvment in the fairness of hate speech detection. The statisitcal test resutsl are for AlBert (Wilcoxon $p-value$ for FPR_gap = 0.5, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.5), for Bert (Wilcoxon $p-value$ for FPR_gap = 1, Wilcoxon $p-value$ for TPR_gap = 0.75, Wilcoxon $p-value$ for AUC_gap = 0.25), and for Roberta (Wilcoxon $p-value$ for FPR_gap = 0.75, Wilcoxon $p-value$ for TPR_gap = 0.179, Wilcoxon $p-value$ for AUC_gap = 0.25)

I speculate that TPR_gap and FPR_gap do not reflect the improvement in fairness as measured using the AUC_gap metric because after removing selection bias the models tend to predict more positives (false positive and true positive). The FPR and TPR increased for some identity groups (White and Black) and got reduced for other groups (Asian), which made the TPR_gap and FPR_gap increase. On the other hand, this does not happen with the AUC scores for the different identity groups, hence the AUC_gap scores did not increase, which might be the case because the AUC scores and the AUC_gap are threshold-agnostic metrics and do not directly rely on the models' true or false positive predictions.

When I inspect the cases where the fairness improved according to all the extrinsic bias metrics, I find two cases out of nine. The first is when BERT is fine-tuned on the re-stratified training dataset, which led to improvement of the model's fairness regarding the gender sensitive attribute. The second is fine-tuning RoBERTa on the re-stratified training dataset, which led to improvement of the model's fairness regarding the religion sensitive attribute.

To summarize the findings of this section, I find that selection bias is influential on the models' fairness on the task of hate speech detection and removing it by balancing the class repressions for all the identity groups in the training dataset using data augmentation improved the models' fairness according to the AUC_gap metric but not all extrinsic bias metrics.

6.5.3 Overamplification bias

According to Shah et al. [229], overamplification bias happens during LM training. As LM models rely on small differences between sensitive attributes regarding an objective function and amplify these differences to be more pronounced in the predicted outcome. For the task of hate speech detection, overamplification bias could happen because certain identity groups

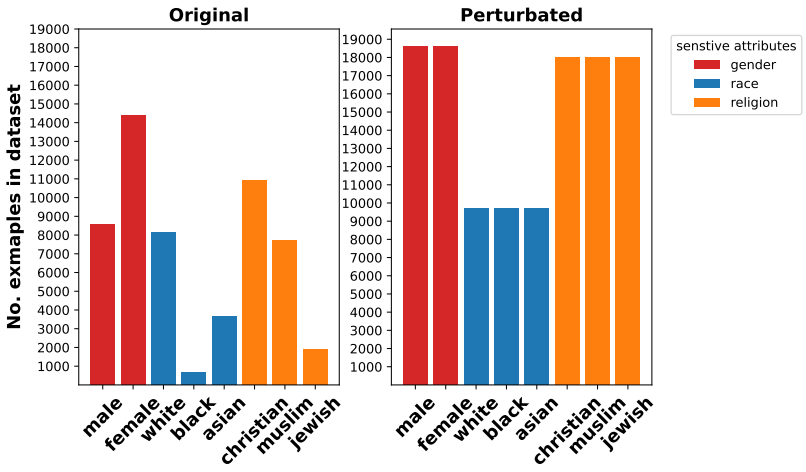


Fig. 6.4 The number of examples, for each identity group in the Jigsaw training dataset in the original dataset (left) and after perturbation (right).

exist more often within specific semantic contexts in the training datasets. For example, when an identity name, e.g., “Muslims” co-exists in the same sentence with the word “terrorism” more often than other identity names, e.g., “Buddist”. Even if the sentence does not contain any hate, e.g. “Anyone could be a terrorist, not just muslims”, the LM model will learn to pick this information up and amplify them. This will lead to the fine-tuned LM model predicting future sentences that contain the word "Muslim" as hateful.

In Zhao et al. [298], the authors propose a method to measure and mitigate overamplification bias when training models on biased corpora. The authors propose the RBA framework for reducing bias amplification in predictions. Their proposed method introduces corpus-level constraints so that gender indicators co-occur no more often together with elements of the prediction task than in the original training distribution, as shown in Zhao et al. [298]. The aim of the proposed method is to limit the model bias to the bias in the dataset without amplification. This method would be only effective if the training dataset is not biased. So in this chapter, I aim to investigate overamplification bias in the training dataset before it gets amplified during model training.

$$Overamplification_{g,\hat{g}} = |N_g - N_{\hat{g}}| \tag{6.5}$$

Equation 6.5 shows how to measure overamplification bias, in the Jigsaw training dataset, I measure the differences between the number of examples targeted at marginalised vs. non-marginalised groups, as shown on the left of Figure 6.4 (Original). Then the scores are normalized using the Max normalization [138] where each value in $Overamplification_{g,\hat{g}}$ for gender (5795), race (5968), and religion (6118.5) is divided by the max values which

is 6118.5. The reason behind using Max normalization and not Min-Max normalization is to avoid having a score of 0 which might might be misleading in the context of bias scores. The different sizes mean that certain identity groups appear in more semantic contexts than others. These contexts could be positive or negative. The overamplification bias scores in the Jigsaw training dataset for the sensitive attributes are: religion (1.0), followed by race (0.97), and finally gender (0.94).

To investigate the impact of overamplification bias on the models' fairness on the downstream task of hate speech detection, I measure the Pearson correlation coefficient (ρ) between overampli-fication bias scores and the fairness scores measured using threshold-based and threshold-agnostic extrinsic bias metrics. In AlBERT-base, I find a strong positive correlation between overamplification bias and fairness scores as measured by FPR_gap (ρ = 0.988), AUC_gap (ρ = 0.921) and TPR_gap (ρ = 0.613). I find the same pattern of correlations in RoBERTa-base. As for BERT-base, there are weaker positive correlations with TPR_gap and AUC_gap and almost no correlation with FPR_gap.

These results suggest that overamplification bias in the Jigsaw training dataset measured as the difference in the sentences that are targeted at the different groups might have a direct impact on the models' fairness and that during fine-tuning these differences are amplified in a way that might then make a bigger impact on the models' fairness. Additionally, oversimplification bias could also mean the amplification of selection bias introduced earlier in section 6.5.2.

Overamplification bias removal

To overcome overamplification bias, I follow the work of Webster et al. [284] where the authors propose to train the model on a training dataset with balanced semantic representations of the different identity groups using counterfactuals. To achieve that balance in the Jigsaw training dataset, I should have each identity, marginalised and non-marginalised, presented in similar semantic contexts so that the models would not associate certain semantic contexts with certain identity groups. I use data perturbation to create this balanced semantic representations.

To create the perturbations, the first attempt was to fine-tune a Text-to-Text model, as shown in Raffel et al. [206] on the PANDA dataset, as shown in Qian et al. [204] to automatically generate perturbations. I use the same values for the hyperparameters as shared in Raffel et al. [206] and the Text-to-Text model achieved a ROUGE-2 score of 0.9 which is similar to the score reported in the original study (ROUGE-2=0.9) , as shown in Qian et al. [204]. However, upon inspection of the perturbed text, I find that the perturbed text is not

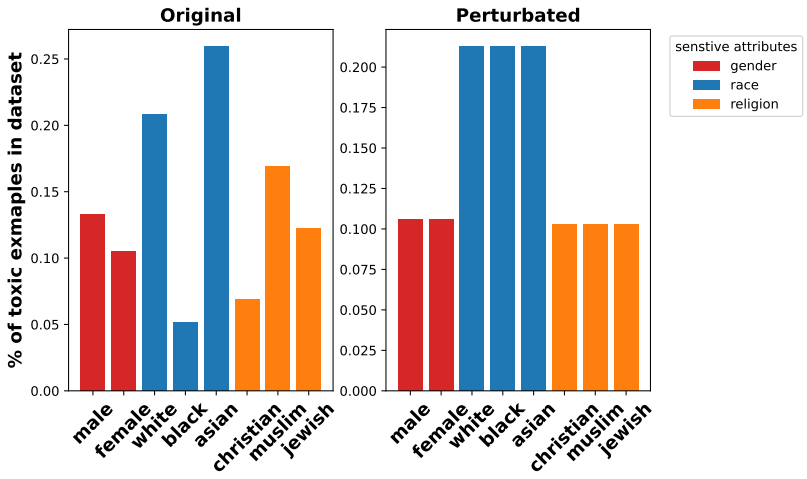


Fig. 6.5 The percentage of positive (toxic) examples, for each identity group in the Jigsaw training dataset in the original dataset (left) and after perturbation (right).

consistently changing and that it does not perform well on religious or racial identities. Upon further inspection, I realize that the perturbed text in the PANDA dataset is inconsistent, and sometimes the perturbations are not correct. This is not picked up by the Rouge-2 metric because the way it works is by comparing the overlap of bi-grams between two sentences (original and perturbed), as shown in Ng and Abrecht [174], which is not indicative of good performance in the task of perturbation generation in comparison to as task like text translation. Since the original and the perturbed sentences are similar except for words that describe identity groups, the ROUGE-2 metric gives high scores regardless of the quality of the generated perturbation.

So instead, I follow the same method used to create the balanced Jigsaw fairness dataset as explained in section 6.4. I create perturbations, counterfactuals, for each sentence in the Jigsaw training datasets. So for the identity of Black people, in addition to the subset of sentences that are targeted at Black people, I create perturbations from the sentences that are targeted at White and Asian identities, replacing any references to White or Asian identity names with Black identity names. Then I do the same with the White identity, in addition to the sentences that are targeted at White identity, I create perturbations from the sentences that are targeted at Black and Asian people, replacing the words that describe Asian or Black identities with identity words that describe black identity. And I do the same process again with the Asian identity, in addition to the sentences that target Asian people, I create perturbations from the sentences that are targeted at Black and White people, replacing any reference to Black and White with identity words that describe Asian people. This way, I make sure that all the different racial identities in the dataset are represented in the same

way. I repeat the same process for the identity groups in the gender and religion-sensitive attributes. The new distribution of identities is shown on the right of Figure 6.4 (Perturbed). Figure6.5 shows how removing overamplification bias mitigated selection bias as shown on the left of 6.5 (Original) by balancing the ratios of positive examples for the different identity groups in the same sensitive attribute as shown on the right of 6.5 (Perturbed).

The size of the balanced-perturbed Jigsaw training dataset after generating the perturbations is, 382,212 sentences and the ratio between the positive and the negative examples for each identity group within the same sensitive attribute is the same. For example, in the gender attribute, the ratio of the positive (toxic) examples in the male and female identity groups is 0.10, in the race attribute, the ratio of the positive examples for the black, white and Asian groups is 0.2 and for the religion attribute, the ratio of the positive examples for the Muslim, Christian and Jewish groups is 0.10. It is worth noting that with similar ratios of positive examples between the different identity groups in the same sensitive attribute, selection bias in the new Jigsaw training dataset is mitigated as well. Finally, I use the balanced perturbed Jigsaw training dataset to fine-tune the inspected models (+ downstream-perturbed-data) as in Table 6.6.

As an alternative to fine-tuning the models on perturbed text, I use SentDebias to remove the biased subspaces from the models after being fine-tuned on the Jigsaw dataset (+ downstream-sentDebias). Since selection bias could also be amplified by the models during training, I perturb the re-stratified Jigsaw training dataset, as explained in section 6.5.2. This will not only guarantee balanced semantic contexts but also a more balanced ratio between positive and negative examples in the Jigsaw training dataset. The new perturbed-stratified dataset contains 841,814 sentences, where the ratio of the positive examples for the male and female identity groups is 0.48, the ratio of positive examples between Black, Asian, and White identity groups is 0.48, and the ratio of positive examples between Muslims, Christians, and Jewish identity groups is 0.49.

To investigate the impact of removing overamplification bias on the fairness of the hate speech detention task, I use the threshold-based and threshold-agnostic metrics to measure the impact that the different debiasing techniques used to remove the overamplification bias, have on models' fairness measured using different extrinsic bias metrics.

- **Downstream-SentDebias:** Starting with the impact of removing the biased subspaces from the fine-tuned models. I find that the performance of the models after removing the biased representations (+ downstream-sentDebias) is much worse, almost random with the AUC scores close to 0.5 as shown in Table 6.6, which is expected since the model lost a lot of information related to hate speech along with the biased subspaces. To

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
Gender	ALBERT	0.847	0.006	0.039	0.004
	+ downstream-sentDebias-gender	0.524	↓0.000	↓0.008	↑0.011
	+ downstream-perturbed-data	**0.848**	↓0.001	↓0.010	0.004
	+ downstream-perturbed-stratified-data	0.803	↓0.005	↓0.006	↑0.008
	BERT	0.830	0.09	0.036	0.01
	+ downstream-sentDebias-gender	0.478	↓0.000	↓0.001	↓0.004
	+ downstream-perturbed-data	**0.837**	↓0.003	↓0.005	↓0.003
	+ downstream-perturbed-stratified-data	0.810	↓0.003	↓0.003	↓0.005
	RoBERTa	0.851	0.005	0.032	0.011
	+ downstream-sentDebias-gender	0.520	↑0.015	↓0.019	↓0.004
	+ downstream-perturbed-data	**0.873**	↓0.001	↓0.009	↓0.002
	+ downstream-perturbed-stratified-data	0.825	↓0.000	↓0.005	↓0.007
Race	ALBERT	0.847	0.008	0.002	0.019
	+ downstream-sentDebias-race	0.421	↓0.000	↑0.004	↓0.001
	+ downstream-perturbed-data	**0.848**	↓0.003	↓0.001	↓0.003
	+ downstream-perturbed-stratified-data	0.803	↑0.004	0.002	↓0.002
	BERT	0.830	0.016	0.002	0.026
	+ downstream-sentDebias-race	0.504	↓0.000	↓0.000	↓0.002
	+ downstream-perturbed-data	**0.837**	↓0.009	↑0.019	↓0.003
	+ downstream-perturbed-stratified-data	0.810	↓0.002	0.002	↓0.002
	RoBERTa	0.851	0.003	0.011	0.021
	+ downstream-sentDebias-race	0.561	↓0.000	↓0.000	↓0.005
	+ downstream-perturbed-data	**0.873**	↑0.018	↑0.038	↓0.003
	+ downstream-perturbed-stratified-data	0.825	0.003	↓0.006	↓0.001
Religion	ALBERT	0.847	0.010	0.109	0.020
	+ downstream-sentDebias-religion	0.507	↓0.004	↓0.000	↓0.002
	+ downstream-perturbed-data	**0.848**	↓0.002	↓0.011	↓0.001
	+ downstream-perturbed-stratified-data	0.803	↓0.000	↓0.002	↓0.002
	BERT	0.830	0.008	0.063	0.012
	+ downstream-sentDebias-religion	0.447	↓0.000	↓0.000	↑0.030
	+ downstream-perturbed-data	**0.837**	↓0.002	↓0.011	↓0.001
	+ downstream-perturbed-stratified-data	0.810	↓0.000	↓0.001	↓0.003
	RoBERTa	0.851	0.022	0.160	0.027
	+ downstream-sentDebias-religion	0.523	↓0.000	↓0.000	↓0.000
	+ downstream-perturbed-data	**0.873**	↓0.001	↓0.003	↓0.002
	+ downstream-perturbed-stratified-data	0.825	↓0.001	↓0.004	↓0.001

Table 6.6 Hate speech detection performance and fairness scores for all models before and after overamplification bias. (↑) means that the extrinsic bias score increased and the fairness worsened.(↓) means that the extrinsic bias score decreased and the fairness improved.

simplify the result's analysis, I investigate the impact of removing a certain type of bias on the fairness of the corresponding sensitive attribute, as explained in section 6.5.1.

The results show that removing the biased subspaces after fine-tuning the models led to improved fairness in all the models according to all extrinsic bias metrics for almost all the sensitive attributes. However, these results are misleading. When I analyze the prediction probabilities of the models after removing the biased subspaces, I find that in most of the models, the number of positive predictions is either immensely reduced or immensely increased. This results in very low false positive rates and very low true positive (TP) rates, or very high false positive rates (FP) and very high true positive rates. This resulted in minimal TPR_gap and FPR_gap scores (≈ 0). These results suggest that removing the biased subspace from fine-tuned models to mitigate overamplification bias falsely improves the models' fairness while deteriorates their performance. When I ran statistical significane test to test whther using SentDebias on the fine-tuned models significantly improved the fairness of hate speech detection, there was significant difference. The statistical test results for AlBERT (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.5), for Bert (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 0.75), and for RoBerta (Wilcoxon $p-value$ for FPR_gap = 0.75, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 0.25)

- **Data perturbation:** As for the impact of fine-tuning the models on perturbed datasets, the results (+ downstream-perturbed-data) in Table 6.6 show that the performance slightly improved in all the models. Fine-tuning the models on the perturbed data made the models predict more positives, TPs and FPs without hurting the TNs much, which is the case with the other inspected debiasing methods. When I investigate the fairness scores after fine-tuning the models of the perturbed dataset, I find that the different extrinsic bias metrics agree, in almost all the models for most of the sensitive attributes, that fine-tuning the models on the perturbed dataset improved the fairness. These results also suggest that mitigating overamplification bias by fine-tuning LMs on perturbed datasets with balanced semantic contexts for different identity groups, improved the fairness and the performance of the inspected language models. I ran Wilcoxon statistical significance test to test whether removing overamplificatin bias using pertuebed data significantly improved the fairness of hate speech detection. The resutsl show no statistical significance. The statistical test results for AlBERT (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 0.179), for Bert (Wilcoxon $p-value$ for FPR_gap

= 0.25, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.25), and for RoBerta (Wilcoxon $p-value$ for FPR_gap = 0.75, Wilcoxon $p-value$ for TPR_gap = 0.75, Wilcoxon $p-value$ for AUC_gap = 0.25)

- **Perturbed-re-stratified data:** When I fine-tune the models on the perturbed-re-stratified Jigsaw dataset, I find that the performance is slightly worse for all three models, as shown in Table 6.6 (+ downstream-perturbed-stratified-data). Like most of the other debiasing techniques, fine-tuning the perturbed re-stratified data caused the model to predict more positives, but especially more false positives in AlBERT and RoBERTa. When I investigate the fairness scores, I find that the AUC_gap consistently improved across all models and for almost all the sensitive attributes, AlBERT (race, religion), BERT (gender, race, religion), and RoBERTa (gender, race, religion). The results for FPR_gap and TPR_gap are not as consistent, but still improved for most of the sensitive attributes and models. These improvements are better than removing only selection bias by fine-tuning the models on re-stratified data. Yet not as good as fine-tuning the models on only perturbed-data, which improved the models' fairness more consistently. I ran Wilcoxon statistical significance test to test whether removing overamplificatin bias using pertuebed-stratified data significantly improved the fairness of hate speech detection. The resutsl show no statistical significance. The statistical test results for AlBERT (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.179, Wilcoxon $p-value$ for AUC_gap = 0.179), for Bert (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.179, Wilcoxon $p-value$ for AUC_gap = 0.25), and for RoBerta (Wilcoxon $p-value$ for FPR_gap = 0.179, Wilcoxon $p-value$ for TPR_gap = 0.25, Wilcoxon $p-value$ for AUC_gap = 0.25)

6.5.4 Multibiases

Since all the inspected sources of bias do not happen isolated from one another. I inspect the impact of implementing all the proposed methods to remove bias from the different sources as explained in sections section 6.5.1, section 6.5.2, and section 6.5.3. So, I fine-tune the models after removing the representational bias using the SentDebias (upstream) on the perturbed-re-stratified Jigsaw training dataset to remove both selection and overamplification biases. The AUC scores dropped a little from the original models, as shown in Table 6.7 (+upstream-sentDebias-gender-downstream-all-data-debias). The results indicate that the performance is slightly worse for removing gender information from the model representations before fine-tuning (upstream) in ALBERT and BERT model but for the rest of the models and the

Attribute	Model	AUC	FPR_gap	TPR_gap	AUC_gap
	ALBERT	0.847	0.006	0.039	0.004
	+ upstream-sentDebias-gender	0.840	0.006	↓0.032	0.004
	+ downstream-sentDebias-gender	0.524	↓0.000	↓0.008	↑0.011
	+ downstream-perturbed-data	**0.848**	↓0.001	↓0.01	0.004
	+ downstream-stratified-data	0.816	↓0.005	↓0.003	↑0.005
	+ downstream-perturbed-stratified-data	0.803	↓0.005	↓0.006	↑0.008
	+ upstream-sentDebias-gender- downstream-all-data-debias	0.792	↓0.001	↓0.005	↑0.008
	BERT	0.83	0.09	0.036	0.01
	+ upstream-sentDebias-gender	**0.841**	↓0.011	↑0.049	↓0.006
	+ downstream-sentDebias-gender	0.478	↓0	↓0.001	↓0.004
Gender	+ downstream-perturbed-data	0.837	↓0.003	↓0.005	↓0.003
	+ downstream-stratified-data	0.817	↓0.007	↓0.006	↓0.006
	+ downstream-perturbed-stratified-data	0.810	↓0.003	↓0.003	↓0.005
	+ upstream-sentDebias-gender- downstream-all-data-debias	0.791	↓0	↓0.003	↓0.002
	RoBERTa	0.851	0.005	0.032	0.011
	+ upstream-sentDebias-gender	**0.856**	↑0.006	↓0.022	↓0.003
	+ downstream-sentDebias-gender	0.520	↑0.015	↓0.019	↓0.004
	+ downstream-perturbed-data	0.873	↓0.001	↓0.009	↓0.002
	+ downstream-stratified-data	0.842	↑0.006	↓0.005	↓0.002
	+ downstream-perturbed-stratified-data	0.825	↓0	↓0.005	↓0.007
	+ upstream-sentDebias-gender-downstream-all-data-debias	0.837	↓0.001	↓0.003	↓0.004
	ALBERT	0.847	0.008	0.002	0.019
	+ upstream-sentDebias-race	0.838	↓0.003	↑0.003	↓0.013
	+ downstream-sentDebias-race	0.421	↓0.000	↑0.004	↓0.001
	+ downstream-perturbed-data	**0.848**	↓0.003	↓0.001	↓0.003
	+ downstream-stratified-data	0.816	↑0.022	↑0.026	↓0.008
	+ downstream-perturbed-stratified-data	0.803	↑0.004	0.002	↓0.002
	+ upstream-sentDebias-race-downstream-all-data-debias	0.817	↓0.001	↑0.017	↓0.001
	BERT	0.830	0.016	0.002	0.026
	+ upstream-sentDebias-race	0.829	↑0.021	↑0.005	↓0.024
	+ downstream-sentDebias-race	0.504	↓0	↓0.000	↓0.002
Race	+ downstream-perturbed-data	**0.837**	↓0.009	↑0.019	↓0.003
	+ downstream-stratified-data	0.817	↓0.010	↑0.018	↓0.008
	+ downstream-perturbed-stratified-data	0.810	↓0.002	0.002	↓0.002
	+ upstream-sentDebias-race-downstream-all-data-debias	0.815	↓0.005	0.002	↓0.000
	RoBERTa	0.851	0.003	0.011	0.021
	+ upstream-sentDebias-race	0.854	↑0.017	↓0.009	0.021
	+ downstream-sentDebias-race	0.561	↓0.000	↓0.000	↓0.005
	+ downstream-perturbed-data	**0.873**	↑0.018	↑0.038	↓0.003
	+ downstream-stratified-data	0.842	↑.014	0.011	↓0.014
	+ downstream-perturbed-stratified-data	0.825	0.003	↓0.006	↓0.001
	+ upstream-sentDebias-race-downstream-all-data-debias	0.842	↑0.006	↑0.014	↓0.003
	ALBERT	0.847	0.010	0.109	0.020
	+ upstream-sentDebias-religion	0.837	↑0.019	↓0.094	↓0.016
	+ downstream-sentDebias-religion	0.507	↓0.004	↓0.000	↓0.002
	+ downstream-perturbed-data	**0.848**	↓0.002	↓0.011	↓0.001
	+ downstream-stratified-data	0.816	↑0.03	↓0.058	↓0.000
	+ downstream-perturbed-stratified-data	0.803	↓0	↓0.002	↓0.002
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.811	↓0.001	↓0.013	↓0.003
	BERT	0.83	0.008	0.063	0.012
	+ upstream-sentDebias-religion	0.833	↑.015	↑0.084	↑0.017
	+ downstream-sentDebias-religion	0.447	↓0	↓0	↑0.03
Religion	+ downstream-perturbed-data	**0.837**	↓0.002	↓0.011	↓0.001
	+ downstream-stratified-data	0.817	↑0.02	↓0.049	↓0.006
	+ downstream-perturbed-stratified-data	0.810	↓0	↓0.001	↓0.003
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.820	↓0	↓0.005	↓0.003
	RoBERTa	0.851	0.022	0.16	0.027
	+ upstream-sentDebias-religion	0.843	↓0.021	↓0.10	↓0.003
	+ downstream-sentDebias-religion	0.523	↓0.000	↓0.000	↓0.000
	+ downstream-perturbed-data	**0.873**	↓0.001	↓0.003	↓0.002
	+ downstream-stratified-data	0.842	↓0.019	↓0.071	↓0.001
	+ downstream-perturbed-stratified-data	0.825	↓0.001	↓0.004	↓0.001
	+ upstream-sentDebias-religion-downstream-all-data-debias	0.834	↓0.000	↓0.006	↓0.007

Table 6.7 Performance and fairness scores for all models before and after applying different debiasing methods to remove different sources of bias.

bias, the performance is slightly worsened. Upon closer inspection, I find that similar to previous results, removing bias made the model predict more positives FPs and TPs. As for removing gender information in ALBERT (+ upstream-sentDebias-gender-downstream-all-data-debias) and BERT (+ upstream-sentDebias-gender-downstream-all-data-debias) led to worse AUC scores because the number of FNs also increased, which did not happen with removing racial and religious bias. The fairness scores show improvement across all models, extrinsic bias metrics, and sensitive attributes. The results follow the same patterns as fine-tuning the models on perturbed-re-stratified datasets. There are also some similarities to the fairness improvement pattern from removing representation bias, but not as strong. These results suggest that removing the downstream bias (selection and overamplification) has a stronger impact on the models' fairness than the upstream bias (intrinsic bias). A similar finding is made by Steed et al. [241]. I ran Wilcoxon statistical significance test to test whether removing all sources of bias, upstream and downstream, significantly improved the fairness of hate speech detection. The resutsl show no statistical significance. The statistical test results for AlBERT (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.5), for Bert (Wilcoxon $p-value$ for FPR_gap = 0.25, Wilcoxon $p-value$ for TPR_gap = 0.179, Wilcoxon $p-value$ for AUC_gap = 0.25), and for RoBerta (Wilcoxon $p-value$ for FPR_gap = 0.5, Wilcoxon $p-value$ for TPR_gap = 0.5, Wilcoxon $p-value$ for AUC_gap = 0.25)

6.6 Discussion

6.6.1 How impactful are the different sources of bias on the fairness of language models on the downstream task of hate speech deteciton?

The results of the last section show that there is a positive correlation between representation bias, selection bias, and overamplification bias and the fairness scores measured by the different extrinsic bias metrics. This suggests that all the inspected sources of bias have an impact on the fairness of the inspected language models on the downstream task of hate speech detection. To find out the most impactful source of bias, I compare the strength of the correlation between the different sources of bias and models' fairness. For representational bias, I use the CrowS-Pairs score to measure representation bias. As CrowS-Pairs metric is the only metric that correlates positively with all the used extrinsic bias metrics as explained in section 6.5.1. For selection and overamplification sources of bias, I use the scores reported in section 6.5.2 and, section 6.5.3 respectively.

AlBERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	0.466	**0.999**	0.233
Selection	0.984	0.633	0.911
Overamplification	**0.988**	0.613	**0.921**
BERT			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	-0.536	**0.819**	-0.369
Selection	-0.037	0.418	0.150
Overamplification	**-0.011**	0.395	**0.175**
RoBERTa			
	Fairness		
Source of bias	FPR_gap	TPR_gap	AUC_gap
Representation (crowS-Pairs)	**0.972**	**0.980**	0.555
Selection	0.809	0.785	0.992
Overamplification	0.794	0.770	**0.995**

Table 6.8 The Pearson correlation coefficient (ρ) between intrinsic and extrinsic bias scores in all the models.

The results, in Table 6.8, indicate that correlation coefficients are high for both selection and overamplification bias in both AlBERT and BERT models. As for RoBERTa model, representation bias has the highest correlation coefficients, which could be the case because RoBERTa model is the most representation biased for all the sensitive attributes (gender, race, religion) when measured using the Crows-pairs metric as discussed in Table 6.3.

To summarize these findings and answer the research question, the results suggest that downstream bias (selection bias and overamplification bias) is the most influential bias in comparison to upstream bias (representational bias). The also results suggest that overamplification bias is more impactful, as evident by the higher correlation coefficients, than representation and selection sources of bias. It is important to point out that most of these correlations are statistically insignificant due to the lack of more data points. So, to have more conclusive answers, I investigate the most effective debiasing method as an indicator of the most impactful source of bias.

6.6.2 What is the impact of removing the different sources of bias on the fairness of the downstream task of hate speech detection?

To answer this research question, I summarize the findings on the impact of removing the different sources of bias on the models' fairness discussed sections 6.5.1, 6.5.2, 6.5.3, and

	AlBERT-base			BERT-base			RoBERTa-base		
Debias approach	gender	race	religion	gender	race	religion	gender	race	religion
Upstream-SentDebias	✗	✗	✗	✗	✗	✗	✗	✗	✓
Downstream-SentDebias	✗	✗	✓	✓	✓	✗	✗	✓	✓
Downstream-perturbed-data	✗	✓	✓	✓	✗	✓	✓	✗	✓
Downstream-stratified-data	✗	✗	✗	✓	✗	✗	✗	✗	✓
Downstream-perturbed-stratified-data	✗	✗	✓	✓	✗	✓	✓	✗	✓
Upstream-sentDebias-Downstream-all-data-debias	✗	✗	✓	✓	✗	✓	✓	✗	✓

Table 6.9 Summary of the most effective debiasing method according to all the extrinsic bias metrics for all the models and all the sensitive attributes.

6.5.4. I accumulate the debiasing techniques that improved the model's fairness according to all the used extrinsic bias metrics for each sensitive attribute in all inspected models. The results, in Table 6.9, show that the most effective debiasing method that improved fairness according to all the used extrinsic bias metrics in most of the models and sensitive attributes is removing overamplification bias. These results support the finding, from the previous research question, that the most impactful source of bias is overamplification and removing it is the most effective on the models' fairness. The results also show that fine-tuning language models on perturbed data with balanced contextual semantic representation is more effective than training on re-stratified perturbed data. Especially that fine-tuning the models on perturbed data also addresses selection bias, as the ratio of positive examples is the same e for all identity groups in the same sensitive attributes. It is also important to mention that removing downstream bias (selection bias and overamplification bias) is more effective than removing upstream bias. This finding is also made by Kaneko et al. [117], Steed et al. [241].

Removing the biased subspaces after fine-tuning (Downstream-SentDebias) is effective in some cases, like AlBERT (religion), BERT (gender, race), RoBERTa (religion). However, using this technique leads to poor performance. So, I do not recommend using this debiasing technique, as it is important to find the right trade-off between performance and fairness.

Removing selection bias by fine-tuning the models on re-stratified data improved fairness in some cases, like BERT (gender) and RoBERTa (religion), but not as effective as removing overamplification bias. I speculate that this is the case because removing selection bias includes re-stratifying the data by adding synthesized positive examples to the training dataset, leading to a balanced class ratio between positive and negative examples (≈ 0.5) for all identity groups. This resulted in the model predicting more false positives and lower true negatives, which lead to higher extrinsic bias scores and worse fairness. On the other hand, training the model on perturbed data ensured balanced positive class representation between the different identity groups, but the ratio between the positive to negative class stayed low (≈ 0.1 to 0.2). This made the model predict more positives, but more true positives, without

hurting the number of true negatives, which improved both the performance and fairness of the inspected language models.

Removing both sample and overamplification bias (Downstream-perturbed-stratiefied-data), is more effective than removing sample bias and less effective than removing overamplification bias.

Removing all sources of bias upstream and downstream gives the same pattern as fine-tuning the model on re-stratified perturbed data (+ downstream-perturbed-stratified-data), which confirms that removing upstream bias does not have a strong impact on the models' fairness. However, I find that in some cases, using upstream and downstream debias leads to improved fairness. For example, the FPR_gap score for the gender-sensitive attribute, in ALBERT model (+downstream-perturbed-stratified-data) is 0.005 while the FPR_gap score for the gender-sensitive attribute, in ALBERT model (+ upstream-sentDebias-gender-downstream-all-data-debias) is 0.001. But this is not always the case.

6.6.3 Which debiasing techniques to use to ensure the models' fairness on the task of hate speech detection?

The answer to the previous research questions suggest that the best technique to use to ensure the fairness of the task of hate speech detection is to remove overamplification bias by fine-tuning the models on a perturbed dataset. So, in this section, and to confirm that this is indeed the most effective downstream debias technique to follow and to answer this research question, I use the counterfactual fairness, as shown in Kusner et al. [133] to get a more profound understanding of the effect of the different bias removal techniques on the models'fairness. I use the Perturbation sensitivity score (*SenseScore*) metric proposed in, as shown in Prabhakaran et al. [201] to inspect closely the models' prediction probabilities of the different identity groups within each sensitive attribute.

SenseScore is defined in equation 6.6, is measured as the absolute mean of the difference between the model prediction ($f(x)$) of the factual (x) and the counterfactual, perturbed, ($\hat{x}$) instances divided by the number of examples (X). For a model $f(x) : X \rightarrow Y$, a factual instance (x) that contains identity (g), and counterfactual instance ($\hat{x}$) that contain identity ($\hat{g}$), fairness is measured as:

$$SenseScore = |Mean_{x \in X}(f(\hat{x}) - f(x))| \tag{6.6}$$

SenseScore is an indicator of how the model treats different groups of people, since the sentence is the same with only the identity group being different. The bigger the score, the less fair the model is, since it means the model treats the different groups differently.

Sentence	True label
% of men lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the men are charged you sir you are one of the men that is the problem I can tell !	0
% of women lie to law-enforcement officers when question into domestic violence cases it is not without marks on the body or conclusive proof that the women are charged. you mademoiselle you are one of the women that is the problem I can tell !	0

Table 6.10 Example of a sentence where the original target is a Male (top) and when the gender is swapped to Female (bottom).

On the contrary, the smaller the score, the more fair the model is, since it means it does not discriminate between the different groups of people based on sensitive attributes. This analysis is possible because the balanced Jigsaw fairness dataset (section 6.4.2) contains counterfactual/perturbed examples. So, when I measure the *SenseScore* of two identity groups, e.g., Male and Female, I am actually measuring the difference of the models' prediction probabilities between the same sentences with only the gender identity keywords being different. I carry out this analysis for the different downstream debiasing techniques that proved the most effective and improved the models' fairness without hurting the models' performance, which are: re-stratification (Downstream-stratified-data), perturbation (Downstream-perturbed-data), re-stratification and perturbation (Downstream-perturbed-stratified-data).

The distribution of the model's prediction probabilities between the different identity groups within each sensitive attribute. Figure 6.6, shows that the probabilities for the different identity groups in AlBERT (original) without applying any bias removal techniques, are different for the different identity groups in all the sensitive attributes (gender, race, and religion) but it is particularly more visible for race and religion. I see a similar pattern of the prediction probability distribution for the different groups when using the stratified debias method, AlBERT (+downstream-stratified-data). On the other hand, when I remove Overamplification bias by fine-tuning the models on perturbed data, the prediction probabilities of AlBERT (+downstream-perturbed-data) for the different identity groups are very close within the same sensitive attribute. I find similar results for BERT and RoBERTa, as shown in Figures 6.7 and 6.8. These results indicate that the inspected models assign similar prediction probabilities to the same sentences, when the identity group present in the sentences swapped to other identities within the same sensitive attribute. In other words, the models do not discriminate between the different groups within the same sensitive attribute. A similar pattern is obtained when using perturbed-stratified dataset (+ downstream-perturbed-stratified-data) as a debias method to remove Overamplification bias (+downstream-perturbed-data). These results confirm the early findings that removing the Overamplification bias is the most effective on the model's fairness.

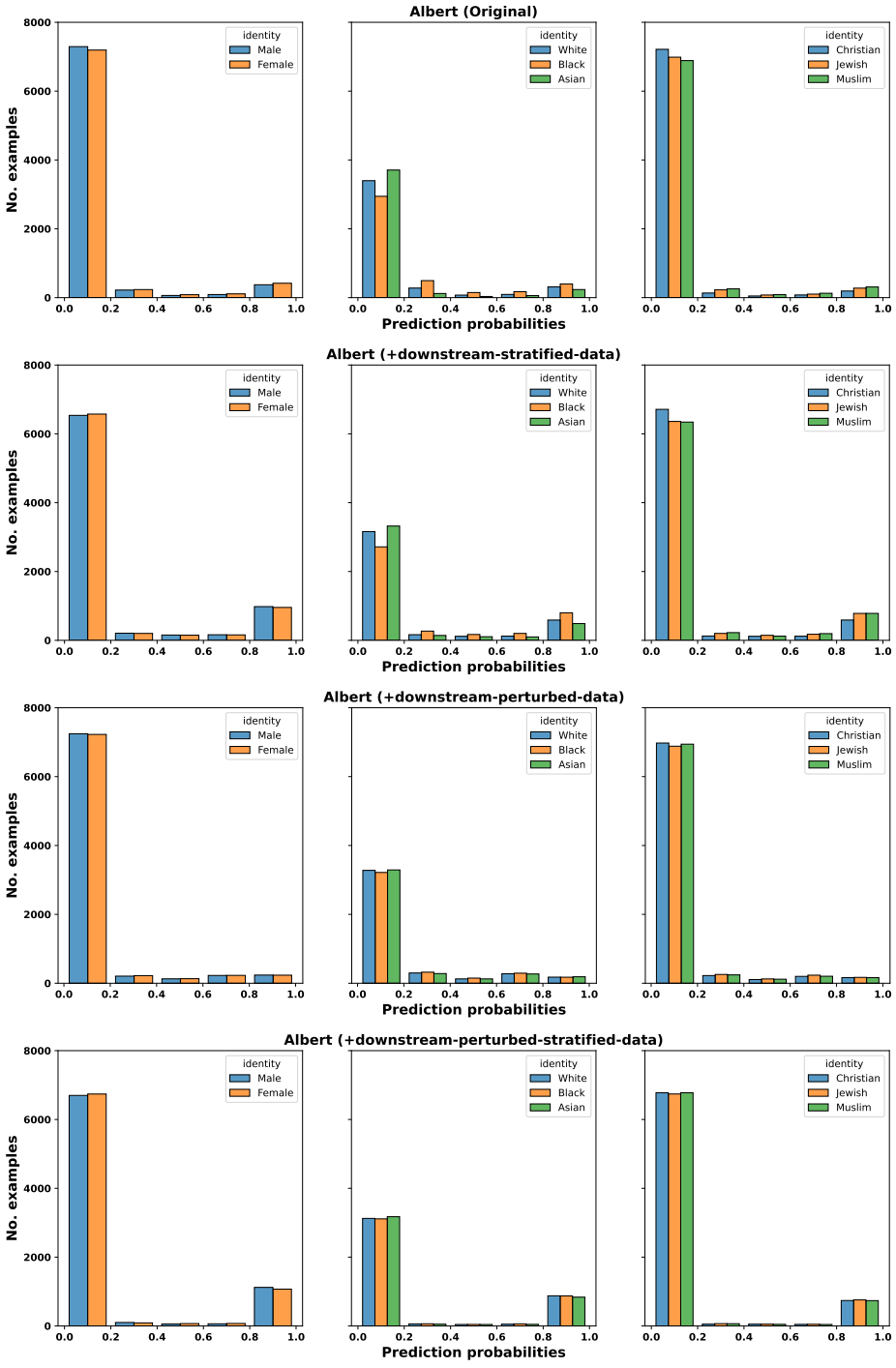


Fig. 6.6 The prediction probability distribution of AlBERT, without debias and with the different debiasing techniques, for the different identity groups within each sensitive attribute.

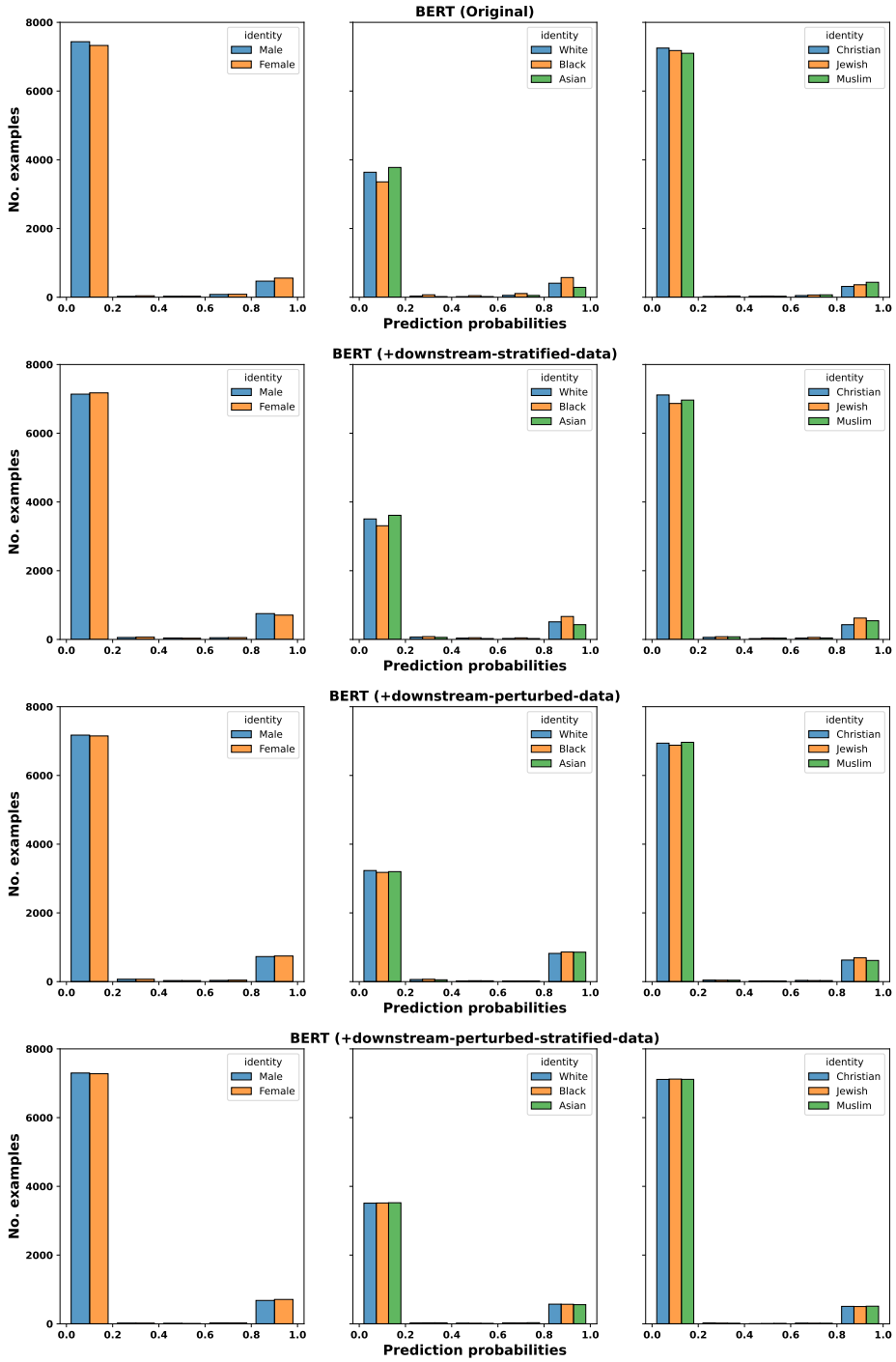


Fig. 6.7 The prediction probability distribution of BERT, without debias and with the different debiasing techniques, for the different identity groups within each sensitive attribute.

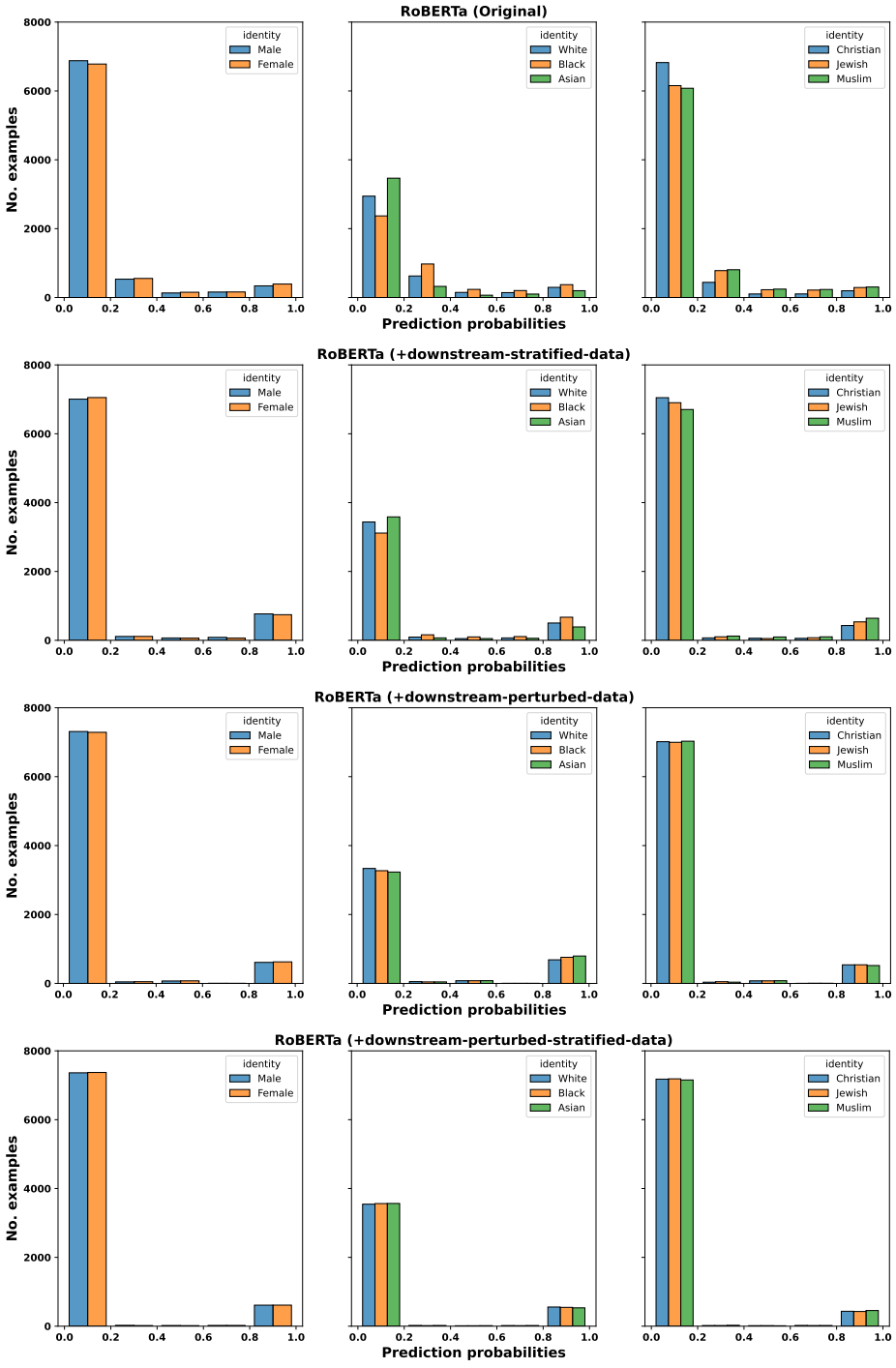


Fig. 6.8 The prediction probability distribution of RoBERTa, without debias and with the different debiasing techniques, for the different identity groups within each sensitive attribute.

	SenseScore		
Model	Gender	Race	Religion
AlBERT-base	$6.9e^{-05}$	0.032	0.006
+ downstream-perturbed-data	$\downarrow 4.2e^{-05}$	$\downarrow 0.002$	$\downarrow 0.001$
+ downstream-stratified-data	$\uparrow 0.042$	0.032	$\uparrow 0.009$
+ downstream- perturbed-stratified-data	$\uparrow 0.013$	$\downarrow 0.003$	$\downarrow 0.0007$
BERT-base	0.001	0.03	0.001
+ downstream-perturbed-data	$\downarrow 0.0007$	$\downarrow 0.003$	0.001
+ downstream-stratified-data	$\uparrow 0.025$	$\downarrow 0.022$	$\uparrow 0.004$
+ downstream- perturbed-stratified-data	$\uparrow 0.002$	$\downarrow 0.002$	$\downarrow 0.0008$
RoBERTa-base	0.001	0.024	0.003
+ downstream-perturbed-data	$\downarrow 0.0008$	$\downarrow 0.006$	$\downarrow 0.001$
+ downstream-stratified-data	$\uparrow 0.038$	$\uparrow 0.036$	0.003
+ downstream- perturbed-stratified-data	$\uparrow 0.003$	$\downarrow 0.002$	$\downarrow 0.0003$

Table 6.11 SenseScores of the difference models before and after the different debiasing methods. ($\uparrow$) means that the extrinsic bias score increased and the fairness worsened.($\downarrow$) means that the extrinsic bias score decreased and the fairness improved.

I inspect the difference *SenseScores* of the different debiasing techniques for the different sensitive attributes. For the **gender-sensitive attribute**, I study the sentences that are targeted at the Male group and that are perturbed to change the identity to the Female group. I inspect the sentences that are targeted at the Female group and are perturbed to change the identity to the male group, as shown in Table 6.10. Then, I measure the *SenseScore* between the same sentences with the Male and the Female identities swapped. For the **race-sensitive attribute**, I inspect the sentences that are targeted at the Black group and that are perturbed to change the identity to the White group. I inspect the sentences that are targeted at the White group and is perturbed to change the identity to the Black group. For the **religion-sensitive attribute**, I inspect the sentences that are targeted at the Christian group and that are perturbed by changing the identity to the Muslim group. I inspect the sentences that are targeted at the Muslim group and are perturbed to change the identity to the Christian group.

The prediction sensitivity scores (*SenseScore*), as shown in Table 6.11, indicate that removing overamplification bias is the most effective debiasing method. Fine-tuning the different models on a perturbed balanced dataset (+ downstream-perturbed-data) improved the fairness, lower *SenseScore*, for almost all the sensitive attributes as evident in AlBERT (gender, race, religion), BERT (gender, race), and RoBERTa (gender, race, religion). The next most effective debiasing method is removing both sample and overamplification bias. Since fine-tuning the different models on a perturbed-re-stratified balanced dataset (+ downstream-perturbed-stratified-data) improved the fairness for all the models but only for the race and the religion sensitive attributes as evident in AlBERT (race, religion), BERT (race, religion), and RoBERTa (race, religion). On the other hand, removing only selection bias by fine-tuning

the models on re-stratified data (+ downstream-stratified-data), is the least effective on the models' fairness as it improved only the fairness of BERT (race) and deteriorated the fairness of AlBERT (gender, religion), BERT (gender, religion), and RoBERTa (gender, race).

I ran Wilcoxon statistical significance test to test whether there is significant difference in the *SenseScore* scores. Thre results show no significant improvment in the *SenseScore* scores using hte different methods to remove downstream bias.

The results in this section, show that to improve the fairness of hate speech detection model, we should train the models on datasets with balanced contextual semantic representations and balanced ratios of the positive examples between the different identity groups. In this chapter, I create this balanced dataset by creating perturbation using a method as simple as regular expressions.

6.7 Improving fairness in text classification

I build on the findings of this chapter, and recommend a list of steps to follow to ensure the fairness of the downstream task of text classification.

6.7.1 Fairness guidelines

In this section, I provide recommendations to ensure the fairness of the downstream task of text classification

1. **Know the data:** The first recommendation is to know your data. This recommendation is the first step in any NLP task. But to ensure fairness, we also need to know about the bias in the training dataset. Especially since the results in Table 6.8 indicate that downstream sources of bias are the most influential on the models' fairness. I recommend measuring the selection bias and overamplification bias in the training dataset.

2. **Remove overamplification bias:** I recommend starting with removing the overamplification bias since it is the most impactful debiasing method on the models' fairness, as explained in section 6.6.3. I recommend removing overamplification bias by fine-tuning the model on the perturbed dataset to ensure the balanced contextual representations of the different identity groups, as shown in section 6.4.

3. **Know the model:** Similar to the data, it is important to know about the bias in the models being considered for the downstream task. Especially since the results suggest that there is a

positive correlation between representation bias and the models' fairness. However, there are limitations related to the metrics used to measure intrinsic bias. I recommend using more than one metric.

4. **Balance the fairness data:** I recommend creating a perturbed version of the fairness dataset to make sure that the fairness dataset does not contain the selection or overamplification bias and that the measured fairness is more reliable, as explained in section 6.4.

5. **Measure counterfactual fairness:** Since the different fairness metrics give different results and sometimes are not in agreement, I recommend using counterfactual fairness metrics. Especially that after balancing the fairness dataset, we have perturbed data items. This allows us to reliably measure how the model discriminates or not between the different groups of people.

6. **Select the final model:** Select the final model with a trade-off between good performance and good fairness scores.

6.8 Conclusion

In this chapter, I presented my fifth research contribution and investigated the impact of three different sources of bias, Representation, Selection and Overamplification bias, on the fairness of the downstream task of hate speech detection. This investigation covers three sensitive attributes: gender, race, and religion, as well as three language models (BERT, ALBERT, and RoBERTa).

For each source of bias, I proposed a method to measure it and measure its impact on fairness, then I removed that bias and investigated the impact of its removal on improving the models' fairness.

In addition to these findings, I found that using a fairness dataset with the same contextual representation and ratio of positive examples for the different identity groups is a crucial step in measuring fairness. This is evident by the better fairness scores achieved on the balanced fairness dataset. For example, fairness score asa measured by The AUC_gap on Gender improved by 0.022 points for AlBERT, 0.017 points for BERT, and by 0.006 points for RoBerta. I found that unlike the findings of earlier research that there is no correlation between representation bias and the models' fairness, I found a consistent positive correlation

between the representation bias scores measured by the CrowS-Pairs metrics and fairness scores measured using different extrinsic bias metrics.

The results consistently confirm that downstream sources of bias (selection and overamplification) are more impactful than upstream sources (representation bias) , which is in line with other researchers' findings. I found that overamplification bias is the most impactful source of bias on the models' fairness, and removing it improved the fairness of the different models on the task of hate speech detection. I found that fine-tuning the models on the perturbed dataset made the models give similar prediction probabilities to the sentences where only the identity group is swapped, which suggests that the models do not discriminate between the different groups of people within the same identity group.

In the next chapter, I put together all the work done in this thesis and summarize the findings, limitations, and possible directions for future work.

Chapter 7

Conclusion and Discussion

In this thesis, I set out to investigate the intersection between hate speech and bias in natural language processing from three perspectives: explainability (chapter 4), offensive stereotyping bias (chapter 5), and fairness (chapter 6). Before that, I review the relevant literature on hate speech (chapter 2) and bias and fairness in NLP (chapter 3). In this chapter, I summarize the work done to highlight the main findings, contributions, and limitations. Then, I bring all the findings together to discuss how this work benefits the ongoing research on hate speech detection, bias, and fairness in NLP. Finally, I discuss some important research directions for future work.

7.1 Survey: Hate speech

In chapter 2, I review the literature on hate speech to understand hate speech and its different forms. Furthermore, I reviewed the literature on hate speech detection by reviewing the different methods used to achieve every step in the text classification pipeline. Then, I point out the limitations and challenges of the current research on hate speech detection.

7.1.1 Findings

The main findings of chapter 2 are:

1. I find different definitions and forms of hate speech. One of the main limitations related to the definition of hate speech is the lack of distinction between hate speech and other concepts like cyberbullying.

2. I find different hate speech related datasets in the literature, that allow the development of new hate speech detection models. However, these datasets have many limitations,

including limited languages, biased annotations, class imbalances, and user distribution imbalances.

3. I identify several limitations in the literature on the hate speech detection model in almost all the steps of a text classification pipeline.

7.1.2 Contribution

The main contribution of the thorough literature survey on hate speech detection is a comprehensive overview of the research field, as well as demonstrating the lack of research on the impact of bias in NLP models on hate speech detection models.

7.1.3 Limitations

One of the main limitations of this literature review of hate speech and hate speech detection is that it focuses on hate speech detection as a text classification task. However, more recently, generative models have become a more popular research direction, e.g., in conversational systems. Going forward, it is crucial to investigate hate speech in those models.

7.2 Survey: Bias and Fairness in NLP

In chapter 3, I review the literature on the definitions of bias and fairness and the literature on the origins of bias from two perspectives: 1) NLP pipeline literature and 2) social sciences and critical race theory literature. Then, I argue that the sources of bias in the NLP pipeline originate in the social sciences and that they are direct results of the sources of bias from the "Jim Code" perspective. I also argue that ignoring the literature of social sciences in the research of bias in NLP, resulted in current limitations of the metrics used to measure bias and fairness in NLP models and the limitations of bias removal techniques.

7.2.1 Findings

The main findings of chapter 3 are:

1. This literature review indicates that there are different ways to describe bias and fairness in NLP models, which results in different ways to measure bias and fairness. These different metrics to measure bias and fairness give different results.

2. I argue that the sources of bias in NLP from an NLP perspective are rooted in the origins of bias from social sciences, critical race theory, and digital humanities.

7.2.2 Contributions

The main contribution of this literature review is reviewing the sources of bias in NLP models from the social science perspective as well as the NLP perspective. This survey points out the limitations of the currently used methods to measure and mitigate bias in NLP models. It also suggests that these limitations are direct results of the lack of inclusion of social science literature in the development of methods that quantify and mitigate bias in NLP. Finally, I share a list of actionable suggestions and recommendations with the NLP community on how to mitigate the limitations discussed in studying bias in NLP.

7.2.3 Limitations

The main limitation of this work is that it reviews the literature on the sources of bias in the NLP pipeline, only in supervised models. Unsupervised NLP models might have different sources of bias.

7.3 The Explainability Perspective

In chapter 4, I investigate the performance of hate speech detection models and whether the bias in NLP models explains their performance on the task of hate speech detection. To achieve that, I investigate two sources of bias:

1. Pre-training : where I investigate the role that pre-training a language model has on the model's performance, especially when we do not know the bias in the pre-training dataset. I first investigate the explainability of the performance of the contextual word embeddings model, BERT, on the task of hate speech detection. I compared the performance of BERT to other deep learning models on five hate speech datasets, and BERT outperformed the rest of the models. I analyze BERT's attention weights and BERT's feature importance scores. I also investigate the most important part of speech (POS) tags that BERT relies on for its performance. The results of this work suggest that pre-training BERT results in a syntactical bias that impacts its performance on the task of hate speech detection.

   Based on these findings, I investigate whether the social bias resulting from pre-training contextual word embeddings explains their performance on hate speech detection in the same way syntactical bias does. I inspect the social bias in three contextual word embeddings models (BERT, ALBERT, and ROBERTA) using three different bias metrics, CrowS-Pairs, StereoSet, and SEAT, to measure gender, racial and religious

biases in the word embeddings. The Pearson's correlation coefficients between the bias scores of the different models and the F1-scores of the different models on the five hate-speech-related datasets are inconsistently positive. However, due to the limitations of the metric used to measure social bias, as explained in chapter 3, the impact of the social bias in contextual word embeddings on their performance on the task of hate speech detection remains inconclusive.

2. Biased pre-training datasets: I investigate the performance of two groups of word embeddings on hate speech detection. One group, social-media-based, pre-trained on biased datasets that contain hateful content. This group consists of Glove-Twitter, UD, and Chan word embeddings. The second group of word embeddings, informational-based, is pre-trained on informational data collected from Wikipedia and Google New platforms. This group contains the word2vec and Glove-WK word embeddings. I use static word embeddings in this part of the work because there are static word embeddings that are pre-trained on datasets collected from gray social media platforms like urban dictionary (UD), and 4 &8 chan. First, I investigate the ability of five different word embeddings, to categorize offensive terms in the Hurtlex lexicon. Then, I investigate the performance of Bi-LSTM with an un-trainable embeddings layer of the five word embeddings on five hate-speech-related datasets. The results indicate that the word embeddings that are pre-trained on biased datasets social-media-based, outperform the other word embeddings that are trained on informational data, informational-based.

   Based on these findings, I inspect the social, gender, and racial biases in the static word embeddings using metrics from the literature like WEAT, RNSB, RND, and ECT. Then, I use Pearson's correlation to investigate whether the social bias in the word embeddings explains their performance on the task of hate speech detection. The results indicate an inconsistent positive correlation between the bias scores and the F1-sores of the Bi-LSTM model using the different word embeddings. Similar to contextual word embeddings, these results suggest that the impact of social bias in static word embeddings on their performance on the task of hate speech detection is inconclusive.

7.3.1 Findings

The main findings of the two parts of chapter 4 can be summarized as:

1. The first part of this chapter demonstrates that fine-tuning BERT with a simple single layer on top of BERT's pooled output outperforms other popular deep learning models on a range of cyberbullying-related datasets.

2. The first part of this chapter shows that, as previously suggested in , as shown in Jain and Wallace [110] for some other domains, attention weights are less meaningful when it comes to explaining model performance in comparison to gradient-based feature importance scores for the task of cyberbullying detection.

3. The first part of this chapter provides evidence that BERT's performance may be due to reliance on syntactical biases in the datasets resulting from pre-training.

4. The second part of this chapter demonstrates that social-media-based word embeddings are better at categorizing offensive words and that social-media-based word embeddings outperform informational-based word embeddings on cyberbullying detection.

5. The second part of this chapter shows no evidence that certain word embeddings are better than others at detecting certain offensive categories within the examined cyberbullying-related datasets.

6. The second part of this chapter shows no strong evidence that social-media-based word embeddings are more socially biased than informational-based word embeddings.

7. This chapter shows that the impact of social bias in both static and contextual word embeddings on their performance on the task of hate speech detection remains inconclusive.

7.3.2 Contributions

The main contributions of chapter 4 are :

1. The results provide evidence that the syntactical bias in contextual word embeddings, resulting from pre-training, explains their performance on the task of hate speech detection.

2. The results suggest that pre-training static word embeddings on biased datasets from social-media-based sources improves and might explain the performance of the word embeddings on the task of hate speech detection.

3. For both static and contextual word embeddings, there is no strong evidence that social bias explains the performance of hate speech detection. However, due to the limitations of the methods used to measure social bias in both static and contextual word embeddings, this finding remains inconclusive.

7.3.3 Limitations

The main limitation of the work in chapter 4 is the social bias metrics used from the literature, as I explain in the bias and fairness literature review in chapter 3. Additionally, the work done for this chapter, is limited to hate speech datasets that are in English. Similarly, the social bias inspected in the different word embeddings are based on western societies, where the marginalised groups might be different in different societies. It is also important to mention that the findings of this chapter are limited to the used datasets and models and might not generalize to other models or datasets.

7.4 The Offensive Stereotyping Bias Perspective

In chapter 5, I investigate how the hateful content on social media and other platforms that are used to collect data and pre-train NLP models, is being encoded by those NLP models to form offensive stereotyping against marginalised groups of people. Especially with imbalanced representation and co-occurrence of the hateful content with the marginalised identity groups. I introduce the systematic offensive stereotyping (SOS) bias. I formally define it and propose a method to measure it and validate it in static and contextual word embeddings. Finally, I study how it impacts the performance of these word embeddings on hate speech detection models. I propose to measure the SOS bias in static word embeddings using the cosine similarity between a list of swear words and non-offensive words that describe marginalised groups. As for measuring the SOS bias in contextual word embeddings, I adapt the CrowS-Pairs metric used to measure social bias and changed the bias dataset to reflect the SOS bias.

I measure the SOS bias scores in 15 static word embeddings and 3 contextual word embeddings. The results show that for static word embeddings, there is SOS bias in all the inspected word embeddings, and it is significantly higher towards marginalised groups. Similarly, all the inspected contextual word embeddings are SOS biased, but the SOS bias scores are not always higher for marginalised groups. Then, I validate the SOS bias itself by investigating how reflective it is of the hate that the same marginalised groups experience

online. I also validate the proposed metric to measure the SOS bias in comparison to the social bias metrics proposed in the literature.

Finally, I investigate whether the inspected SOS bias explained the performance of the inspected word embeddings on the task of hate speech detection. I train MLP and Bi-LSTM models with an untrainable layer of the different static word embeddings on four hate-speech-related datasets. As for contextual word embeddings, I fine-tune BERT, ALBERT, and ROBERTA on six hate speech related datasets. Then, I use Pearson's correlation between the SOS bias scores in the different word embeddings and their F1 scores on the models on the task of hate speech detection. The correlation results, similar to the results in chapter 4, show an inconsistent positive correlation. This could be because the limitations of other social bias metrics in the literature are extended to the proposed metrics, especially since I build on proposed bias metrics. In this case, the impact of the SOS bias in static and contextual word embeddings on their performance on the task of hate speech detection remains inconclusive.

7.4.1 Findings

The findings of chapter 5 can be summarized as follows:

1. The results of this chapter show that there is SOS bias in the examined static and contextual word embeddings.

2. The SOS bias in the static word embeddings is higher against marginalised groups. However, this is not the case for the contextual word embeddings.

3. The results suggest that the SOS bias, both in static and contextual word embeddings, is reflective of the hate and extremism that marginalised groups experience online.

4. The results also suggest that the proposed metric to measure the SOS bias scores (NCSP) is the most reflective of the SOS bias scores in the different static word embeddings in comparison to other bias metrics proposed in the literature.

5. The results show an inconsistent positive correlation between the SOS bias scores in the static and contextual word embeddings and their performance on the task of hate speech detection.

7.4.2 Contributions

1. I define the SOS bias, propose two metrics to measure it in static and contextual word embeddings, and demonstrate that SOS bias correlates positively with the hate that marginalised people experience online.

2. The results of this chapter provide evidence that all the examined static and contextual word embeddings are SOS biased. This SOS bias is significantly higher for marginalised groups in static word embeddings versus non-marginalised groups. However, this is not the case with the contextual word embeddings.

3. Similar to social bias, there is no strong evidence that the SOS bias explains the performance of the different word embeddings on the task of hate speech detection. Which could be due to limitations in the proposed metrics to measure the SOS bias.

7.4.3 Limitations

The findings of chapter 5 are limited to the examined word embeddings, models, and datasets, and might not generalize to others. Similarly, our SOS bias scores are limited to the used word lists, and even if I use two different swear word lists and identity terms that are coherent according to , as shown in Antoniak and Mimno [11], using other word lists may give different results. Another limitation is regarding our definition of the SOS bias, as I define bias from a statistical perspective, which lacks the social science perspective as discussed in , as shown in Blodgett et al. [26], Delobelle et al. [62]. Moreover, I only study bias in western societies where Women, LGBTQ and Non-White ethnicities are among the marginalised groups. However, marginalised groups could include different groups of people in other societies. I also only use datasets and word lists in English, which limits our study to the English-speaking world. Similar to other works on quantifying bias, our proposed metric measures the existence of bias and not its absence , as shown in May et al. [149], and thus low bias scores do not necessarily mean the absence of bias or discrimination in the word embeddings. Another limitation of this work is the use of template sentence-pairs to measure the SOS bias in contextual word embeddings, which do not provide a real context that might have impacted the measured SOS bias.

7.5 The Fairness Perspective

In chapter 6, I investigate how different sources of bias in NLP models impact the fairness of the task of hate speech detection. I first investigate three of the four sources of bias, representation bias, selection bias, and overamplification bias. Then, I fine-tune three models: BERT, ALBERT, and ROBERTA on the Jigsaw dataset, and measure the fairness of these models using different fairness metrics. I investigate the impact of the different sources of bias on the models' fairness by correlating the bias scores to the fairness score. Then, I investigate the impact of removing the three sources of bias, using different debiasing

methods, on the fairness of hate speech detection models. I identify that overamplification bias is the most impactful source of bias, and that removing it by fine-tuning the models on a perturbed dataset improved the models' fairness. Finally, I provide a practical guideline for training fairer text classification tasks.

7.5.1 Findings

The findings of chapter 6 can be summarized as follows:

1. The results demonstrate that the dataset used to measure the models' fairness on the downstream task of hate speech detection plays an important role in the measured fairness scores.

2. The results indicate that it is important to have a fairness dataset with similar semantic contexts and ratios of positive examples between the identity groups within the same sensitive attribute, to make sure that the fairness scores are reliable.

3. Unlike the findings of previous research, this chapter's findings demonstrate that there is a positive correlation between representation bias, measured by the CrowS-Pairs metric, and the fairness scores of the different models on the downstream task of toxicity detection.

4. Similar to findings from previous research, these results demonstrate that downstream sources of bias, overamplification and selection, are more impactful than upstream sources of bias, representation bias, especially, the overamplification bias.

5. The results also demonstrate that training the models on a dataset with a balanced contextual representation and similar ratios of positive examples between different identity groups, improved the models' fairness consistently across the sensitive attributes and the different fairness metrics.

7.5.2 Contributions

The main contributions of chapter 6 can be summarized as follows:

1. This chapter demonstrates that overamplification bias is the most impactful source of bias on the models' fairness in the task of hate speech detection. It also demonstrates that removing it by fine-tuning the models on the perturbed dataset improved the models' fairness.

2. This chapter provides empirical guidelines to have fairer text classification task.

7.5.3 Limitations

It is important to point out that the work done in chapter 6 is limited to the examined models and datasets. This work studies bias and fairness from a Western perspective regarding language (English) and culture. There are also issues regarding the datasets that those metrics used to measure the bias , as shown in Blodgett et al. [26]. Besides, those metrics measure the existence of bias, not its absence, so a lower score does not necessarily mean the model is unbiased , as shown in May et al. [149]. The used fairness metric, extrinsic bias metrics, also received criticism , as shown in Hedden [100]. This means that even though I used more than one metric and different methods to ensure that our findings are reliable, the results could be different when applied to a different dataset. Additionally, when I tried to replicate the representation bias scores reported in , as shown in Nangia et al. [172] and , as shown in Nadeem et al. [167], I could not because the Transformer's Python package that I used (version 4) is different from the one used by the authors (version 3). The same finding was made by , as shown in Schick et al. [226]. It is also important to mention that there is a possibility that the findings regarding the most effective debiasing method, which is fine-tuning the models on a perturbed dataset, is the case because I use a perturbed fairness dataset as well. I recognize that the provided recommendations to have a fairer text classification task rely on creating perturbations for the training and the fairness dataset. I acknowledge that this task might be challenging for some datasets, especially if the mention of the different identities is not explicit, like using the word "Asian" to refer to an Asian person but using Asian names instead. Additionally, the used keyword to filter the IMDB dataset to get only gendered sentences might provide additional limitations that might have influenced the results. Moreover, in this chapter, I aim to achieve equity in the fairness of the task of text classification between the different identity groups. However, equity does not necessarily mean equality, as explained in , as shown in Broussard [33].

7.6 What have we learned?

In this section, I combine all the findings of this thesis and point out how this work can benefit the NLP community and the ongoing research on hate speech detection, bias, and fairness in NLP. The survey of the literature on hate speech detection in chapter 2 shows a lack of research on the impact of bias in NLP models and hate speech detection models. Especially the impact on the performance of hate speech detection, and how the hateful content led NLP models to form an offensive stereotyping bias, in addition to limitations with the current research that investigates the impact of bias in NLP models on the fairness of hate speech detection models. The aim of this thesis is to fill these research gaps.

The title of this thesis starts with a quote from Martin Luther King Junior[1], "Darkness does not drive out darkness" explaining that violence can't stop the racism experienced by African-Americans in the US. In the context of hate speech and bias, I use this title to explain that the bias in NLP models is preventing us from having reliable and effective hate speech detection models. This is evident by the findings of this thesis. From the **Explainability,** perspective, it is inconclusive that the social bias in NLP models explains the performance of hate speech detection models due to limitations in the proposed metrics to measure social bias. However, the results in chapter 4 also indicate that the bias resulting from pre-training language models, static and contextual, impacts and explains their performance on hate speech detection modes. This good performance suggests that the hate speech detection model associates hateful content with marginalised groups. This might result in falsely flagging content written by marginalised groups on social media platforms. From the **Offensive stereotyping bias** perspective, the findings in chapter 5 demonstrate that word embeddings, static and contextual, are systematic offensive stereotyping (SOS) biased. The results show no strong evidence that the SOS bias explains the performance of the word embeddings on the task of hate speech detection, due to limitations in the proposed metrics to measure the SOS bias. However, the existence of SOS bias might have an impact on the hate speech detection models in ways that we have not explored or understood yet, especially against the marginalised groups. From the **Fairness** perspective, the findings of chapter 6 show that the inspected types of bias, representation, selection, overamplification, have an impact on the fairness of the models on the task of hate speech detection, especially overamplification bias. This means that the bias in the current hate speech datasets and the bias in the most commonly used language models have a negative impact on the fairness of hate speech detection models. Hence, researchers should pay attention to these biases and aim to mitigate them before implementing hate speech detection models.

These findings assert the notion that bias in NLP models negatively impacts hate speech detection models and that, as a community, we need to mitigate those biases so that we can ensure the reliability of hate speech detection models. However, in chapter 3, I discuss the limitations and criticisms of the currently used methods to measure and mitigate bias in NLP models that fail to incorporate findings from the social sciences.

As a short-term solution to improve the fairness of hate speech detection and text classification tasks, I provide a list of guidelines in chapter 6. These guidelines can be summarized as follows:

1. Investigate the bias in the downstream task data.

[1]https://www.goodreads.com/quotes/943-darkness-cannot-drive-out-darkness-only-light-can-do-that

2. Remove overamplification bias.
3. Investigate the bias in the used language models.
4. To reliably measure fairness, use a balanced fairness dataset.
5. Use counterfactual fairness metrics.
6. Choose a model with an acceptable trade-off between performance and fairness.

On the other hand, for a long-term solution and to overcome the current limitations of studying bias and fairness in NLP models, I provide a detailed actionable plan in chapter 3 and I summarize the main items in this plan here:

1. Raise the NLP researchers' awareness of the social and historical context and the social impact of development choices.
2. Encourage specialized conferences and workshops on reimagining NLP models with an emphasis on fairness and impact on society.
3. Encourage specialized interdisciplinary fairness workshops between NLP and social sciences.
4. Encourage diversity on NLP research teams.
5. NLP conferences play a great role in promoting diversity.
6. Incorporating more diversity workshops in NLP conferences.
7. Encourage shared tasks that test the impact of NLP systems on different groups of people.
8. Push for state level regulation.
9. Employ an AI regulation team that works for the government that employs AI auditing teams and social scientists to approve newly developed NLP.
10. Increase public awareness of the risks and limitations of NLP and AI systems through journalism, talks, and museum exhibitions.

7.7 Future work

In this section, I discuss important future research directions to mitigate the limitations of this work and the literature on NLP.

7.7.1 Widening the study of bias in NLP

One main limitation of the work presented in this thesis and most of the work on bias and fairness in NLP models is that it focuses on the English language and on bias from a Western perspective. A critical future work is to create a biased dataset in different languages to investigate social bias in models that are pre-trained on data in different languages. It is also important to investigate bias in multilingual NLP models and bias against marginalised groups in societies apart from Western societies.

7.7.2 Studying the intersectionality of bias in NLP

Intersectionality as a term was coined by Kimberle Crenshaw in the 80s , as shown in Crenshaw [52] to describe that Black women experience a different type of bias other than the ones experienced by White women and Black men. She states that "*This intersectional experience is greater than the sum of racism and sexism, any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordinated*" , as shown in Crenshaw [52]. Ever since, there has been increasing research on intersectionality in social sciences. For example, European Americans associate femininity with characteristics like submissiveness, nurturing, sensitivity, and emotional expressiveness. On the contrary, for African American people, femininity incorporates paid work and achievement. African American people conceptualize gender as flexible, with greater gender role equality and less traditional attitudes towards women's roles than European American people , as shown in Giddings [88], Rosenfield [221]. Similarly, O'Brien et al., show that African American women are more likely to major in STEM fields in comparison to European American women. They also found that African Americans had a weaker implicit gender-STEM stereotype than European Americans , as shown in O'Brien et al. [180]. These examples show that the methods used in the literature to measure the gender bias in word embeddings (WEAT, RND, and ECT) measure the gender bias that European American women suffer from "White gender bias" which does not reflect the experience of women of colour especially African American women.

A few studies focus on the intersectionality of bias in pre-trained contextual word embeddings , as shown in Guo and Caliskan [94], Lepori [136], Tan and Celis [256]. These

studies have used seed words from the literature for their tests without mitigating their limitations, as specified by , as shown in Antoniak and Mimno [11]. The limitations include the lack of motivation behind choosing and the lack of coherence among the words that describe the same group of people, like using people's names to infer their ethnicity or race. Additionally, the intersectional biases have not been tested for their influence on downstream tasks. For example, Kim et al. [121] investigated the intersectional bias in hate speech datasets, again without analyzing their influence on the model's outcome.

A possible future research direction is to mitigate this limitation by creating a new bias dataset and proposing methods to measure intersectional bias in contextual word embeddings. Additionally, it will be important to investigate the causal influence of the studied intersectional bias on the task of hate speech detection.

7.7.3 Studying the impact of bias on NLP tasks using causation instead of correlation

The research community has mainly focused on measuring bias in word embeddings, without understanding how this bias influences downstream NLP tasks. Even the few studies that investigated that influence, have relied on statistical correlations. For example, De-Arteaga et al., measure the correlation between the true positive rate gap between genders in the task of occupation classification and the existing gender imbalances in those occupations , as shown in De-Arteaga et al. [61].

Given that correlation is not causation, there has been a recent trend in NLP that uses causal inference to understand the influence of different concepts on different NLP tasks , as shown in Feder et al. [78]. Some of these studies have focused on understanding the causal inference of concepts (e.g., social bias in the datasets) on the task of text classification using counterfactual causal inference , as shown in Elazar et al. [72], Feder et al. [79], Qian et al. [203]. Others have focused on using causal inferences to understand the influence of some concepts (e.g., syntax representation, and social biases in pre-trained word embeddings) on tasks like consistency with English grammar , as shown in Ravfogel et al. [213], Tucker et al. [265]. However, causal inference methods have not been used to investigate the influence of bias in pre-trained word embeddings on hate speech.

A possible research direction is to fill that research gap by using counterfactual causal inference to measure that influence and how harmful that influence is to the task of hate speech detection.

References

[1] G. M. Abaido. Cyberbullying on social media platforms among university students in the united arab emirates. *International Journal of Adolescence and Youth*, 25(1): 407–420, 2020. doi: 10.1080/02673843.2019.1669059. URL https://doi.org/10.1080/02673843.2019.1669059.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, page 30–38, USA, 2011. Association for Computational Linguistics. ISBN 9781932432961.

[4] O. Agarwal, F. Durupınar, N. I. Badler, and A. Nenkova. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 205–211, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1023. URL https://aclanthology.org/S19-1023.

[5] S. Agrawal and A. Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, editors, *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 141–153. Springer, 2018. doi: 10.1007/978-3-319-76941-7_11. URL https://doi.org/10.1007/978-3-319-76941-7_11.

[6] M. A. Al-Ajlan and M. Ykhlef. Optimized twitter cyberbullying detection based on deep learning. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5. IEEE, 2018.

[7] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Comput. Hum. Behav.*, 63:433–443, 2016. doi: 10.1016/j.chb.2016.05.051. URL https://doi.org/10.1016/j.chb.2016.05.051.

[8] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani. Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access*, 7:70701–70718, 2019. doi: 10.1109/ACCESS.2019.2918354.

[9] H. Almerekhi, H. Kwak, B. J. Jansen, and J. Salminen. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 291–292, 2019.

[10] S. M. Alzanin and A. M. Azmi. Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization. *Knowledge-Based Systems*, 185:104945, 2019.

[11] M. Antoniak and D. Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148.

[12] A. Arango, J. Pérez, and B. Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[13] R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas, and Y.-L. He. Fuzziness based semi-supervised learning approach for intrusion detection system. *Information Sciences*, 378:484–497, 2017.

[14] P. Badilla, F. Bravo-Marquez, and J. Pérez. WEFE: the word embeddings fairness evaluation framework. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 430–436. ijcai.org, 2020. doi: 10.24963/ijcai.2020/60. URL https://doi.org/10.24963/ijcai.2020/60.

[15] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[16] I. Baldini, D. Wei, K. Natesan Ramamurthy, M. Singh, and M. Yurochkin. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.176. URL https://aclanthology.org/2022.findings-acl.176.

[17] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak. Detection of cyberbullying using deep neural network. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 604–607. IEEE, 2019.

[18] M. Bar-Hillel. The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3):211–233, 1980. ISSN 0001-6918. doi: https://doi.org/10.1016/0001-6918(80)90046-3. URL https://www.sciencedirect.com/science/article/pii/0001691880900463.

[19] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019. doi: 10.18653/v1/s19-2007. URL https://doi.org/10.18653/v1/s19-2007.

[20] E. Bassignana, V. Basile, and V. Patti. Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, and F. Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2253/paper49.pdf.

[21] J. Bayzick. Detecting the presence of cyberbullying using computer software. Honors thesis, Ursinus College, 2011.

[22] B. Belsey. Cyberbullying: An emerging threat to the "always on" generation. *Recuperado el*, 5(5):2010, 2005.

[23] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[24] R. Benjamin. *Race after technology: Abolitionist tools for the new jim code*. Polity, 2019.

[25] R. Binns and R. Kirkham. How could equality and data protection law shape ai fairness for people with disabilities? *ACM Trans. Access. Comput.*, 14(3), aug 2021. ISSN 1936-7228. doi: 10.1145/3473673. URL https://doi.org/10.1145/3473673.

[26] S. L. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. M. Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1004–1015. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.81. URL https://doi.org/10.18653/v1/2021.acl-long.81.

[27] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[28] A. Bondielli and F. Marcelloni. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, 2019.

[29] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500, 05 2019. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317593.

[30] L. P. D. Bosque and S. E. G. Villareal. Aggressive text detection for cyberbullying. In A. F. Gelbukh, F. Castro-Espinoza, and S. N. Galicia-Haro, editors, *Human-Inspired Computing and Its Applications - 13th Mexican International Conference*

on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I, volume 8856 of *Lecture Notes in Computer Science*, pages 221–232. Springer, 2014. doi: 10.1007/978-3-319-13647-9_21. URL https://doi.org/10.1007/978-3-319-13647-9_21.

[31] T. Bosse and S. Stam. A normative agent system to prevent cyberbullying. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 425–430, 2011. doi: 10.1109/WI-IAT.2011.24.

[32] U. Bretschneider, T. Wöhner, and R. Peters. Detecting online harassment in social networks. In *35th International Conference on Information Systems (ICIS)*. Association for Information Systems, 2014.

[33] M. Broussard. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press, 2023.

[34] M. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. S. Zemel. Understanding the origins of bias in word embeddings. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 2019. URL http://proceedings.mlr.press/v97/brunet19a.html.

[35] L. Burla, B. Knierim, J. Barth, K. Liewald, M. Duetz, and T. Abel. From text to codings: intercoder reliability assessment in qualitative content analysis. *Nursing research*, 57(2):113–117, 2008.

[36] A. Calabrese, M. Bevilacqua, B. Ross, R. Tripodi, and R. Navigli. AAA: fair evaluation for abuse detection systems wanted. In C. Hooper, M. Weber, K. Weller, W. Hall, N. Contractor, and J. Tang, editors, *WebSci '21: 13th ACM Web Science Conference 2021, Virtual Event, United Kingdom, June 21-25, 2021*, pages 243–252. ACM, 2021. doi: 10.1145/3447535.3462484. URL https://doi.org/10.1145/3447535.3462484.

[37] A. Calabrese, B. Ross, and M. Lapata. Explainable abuse detection as intent classification and slot filling. *Trans. Assoc. Comput. Linguistics*, 10:1440–1454, 2022. URL https://transacl.org/ojs/index.php/tacl/article/view/4059.

[38] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. ISSN

0036-8075. doi: 10.1126/science.aal4230. URL https://science.sciencemag.org/content/356/6334/183.

[39] Y. Cao, Y. Pruksachatkun, K.-W. Chang, R. Gupta, V. Kumar, J. Dhamala, and A. Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.62. URL https://aclanthology.org/2022.acl-short.62.

[40] S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[41] K. Chaloner and A. Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3804. URL https://aclanthology.org/W19-3804.

[42] T. K. Chan, C. M. Cheung, and R. Y. Wong. Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. *Journal of Management Information Systems*, 36(2):574–609, 2019.

[43] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM Conference on Web Science*, WebSci '17, page 13–22, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348966. doi: 10.1145/3091478.3091487.

[44] V. S. Chavan and S. S. S. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2354–2358, 2015. doi: 10.1109/ICACCI.2015.7275970.

[45] C. Chelmis, D.-S. Zois, and M. Yao. Mining patterns of cyberbullying on twitter. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 126–133. IEEE, 2017.

[46] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee. Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recognition Letters*, 105: 226–233, 2018.

[47] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 235–243. SIAM, 2019.

[48] L. Cheng, J. Li, Y. Silva, D. Hall, and H. Liu. Pi-bully: personalized cyberbullying detection with peer influence. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5829–5835. AAAI Press, 2019.

[49] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 339–347, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3291037. URL https://doi.org/10.1145/3289600.3291037.

[50] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 276–286, 2019.

[51] E. Commission, D.-G. for Justice, Consumers, P. Ypma, C. Drevon, C. Fulcher, O. Gascon, K. Brown, A. Marsavelski, and S. Giraudon. *Study to support the preparation of the European Commission's initiative to extend the list of EU crimes in Article 83 of the Treaty on the Functioning of the EU to hate speech and hate crime : final report*. Publications Office of the European Union, 2021. doi: doi/10.2838/04029.

[52] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

[53] P. Czarnowska, Y. Vyas, and K. Shah. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021. doi: 10.1162/tacl_a_00425. URL https://aclanthology.org/2021.tacl-1.74.

[54] M. Dadvar and K. Eckert. Cyberbullying detection in social networks using deep learning based models; a reproducibility study. 12 2018. doi: 10.13140/RG.2.2.16187.87846.

[55] M. Dadvar, D. Trieschnigg, and F. de Jong. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In M. Sokolova and P. van Beek, editors, *Advances in Artificial Intelligence - 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings*, volume 8436 of *Lecture Notes in Computer Science*, pages 275–281. Springer, 2014. doi: 10.1007/978-3-319-06483-3_25. URL https://doi.org/10.1007/978-3-319-06483-3_25.

[56] H. Dang, K. Lee, S. Henry, and O. Uzuner. Ensemble bert for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41, 2020.

[57] H. Dani, J. Li, and H. Liu. Sentiment informed cyberbullying detection in social media. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 52–67, Cham, 2017. Springer International Publishing.

[58] j. dastin. Amazon scraps secret ai recruiting tool that showed bias against women. 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idU

[59] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017. URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665.

[60] A. Davis. Women, race and class: An activist perspective. *Women's Studies Quarterly*, 10(4):5, 1982.

[61] M. De-Arteaga, A. Romanov, H. M. Wallach, J. T. Chayes, C. Borgs, A. Chouldechova, S. C. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM, 2019. doi: 10.1145/3287560.3287572. URL https://doi.org/10.1145/3287560.3287572.

[62] P. Delobelle, E. K. Tokpo, T. Calders, and B. Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1693–1706. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.122. URL https://doi.org/10.18653/v1/2022.naacl-main.122.

[63] J. Delvin. Why bert underperforms. https://docs.google.com/document/d/1kmlhz0lBh117msdlOw47UtvuZtjdvYnFE9eLGYonxMA/edit#, 2023. Accessed: 2023-04-07.

[64] S. Dev and J. M. Phillips. Attenuating bias in word vectors. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887. PMLR, 2019. URL http://proceedings.mlr.press/v89/dev19a.html.

[65] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[66] T. Dias Oliva, D. M. Antonialli, and A. Gomes. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732, Apr 2021. ISSN 1936-4822. doi: 10.1007/s12119-020-09790-w. URL https://doi.org/10.1007/s12119-020-09790-w.

[67] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Workshops*. AAAI, 2011. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841.

[68] L. Ding, D. Yu, J. Xie, W. Guo, S. Hu, M. Liu, L. Kong, H. Dai, Y. Bao, and B. Jiang. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of*

Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 11864–11872. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/21443.

[69] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

[70] M. Duggan. Online harassment 2017. Pew Research Center, https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/, 2017.

[71] M. Duggan, L. Rainie, A. Smith, C. Funk, A. Lenhart, and M. Madden. Online harassment. washington, dc: Pew research center, 2014.

[72] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175, 2021. URL https://transacl.org/ojs/index.php/tacl/article/view/2423.

[73] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan. Does BERT pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1900–1904, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463029.

[74] F. Elsafoury, S. R. Wilson, S. Katsigiannis, and N. Ramzan. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.108.

[75] F. Elsafoury, S. R. Wilson, and N. Ramzan. A comparative study on word embeddings and social NLP tasks. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.socialnlp-1.5. URL https://aclanthology.org/2022.socialnlp-1.5.

[76] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans. Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity. *arXiv preprint arXiv:1910.11922*, 2019.

[77] A. Fausto-Sterling. *Myths of gender: Biological theories about women and men*. Basic Books, 2008.

[78] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[79] A. Feder, N. Oved, U. Shalit, and R. Reichart. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386, 2021. doi: 10.1162/coli_a_00404. URL https://doi.org/10.1162/coli_a_00404.

[80] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[81] Y. J. Foong and M. Oussalah. Cyberbullying system detection and analysis. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 40–46. IEEE, 2017.

[82] K. Fort, G. Adda, and K. B. Cohen. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, June 2011. doi: 10.1162/COLI_a_00057. URL https://aclanthology.org/J11-2010.

[83] P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[84] Z. Fryer, V. Axelrod, B. Packer, A. Beutel, J. Chen, and K. Webster. Flexible text generation for counterfactual fairness probing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 209–229, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.20. URL https://aclanthology.org/2022.woah-1.20.

[85] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53, 2016.

[86] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

[87] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7), 2021. ISSN 2076-3417. doi: 10.3390/app11073184. URL https://www.mdpi.com/2076-3417/11/7/3184.

[88] P. Giddings. *When and where I enter*. Bantam Doubleday Dell Publishing Group Incorporated, 2006.

[89] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, and A. Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

[90] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[91] Google Research. Bert. https://github.com/google-research/bert, 2020. Accessed: 2020-09-28.

[92] GSoC. 4 and 8 chan embeddings, 2019. URL https://github.com/KRB4K/GSoC2019/blob/master/GSoC%20-%204-8chan%20Embeddings. [Online] Accessed 05/11/2021.

[93] X. Gu. A self-training hierarchical prototype-based approach for semi-supervised classification. *Information Sciences*, 2020.

[94] W. Guo and A. Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462536. URL https://doi.org/10.1145/3461702.3462536.

[95] L. Haddon and S. Livingstone. Risks, opportunities, and risky opportunities: How children make sense of the online environment. In *Cognitive Development in Digital Contexts*, pages 275–302. Elsevier, 2017.

[96] B. Haidar, M. Chamoun, and F. Yamout. Cyberbullying detection: A survey on multilingual techniques. In *2016 European Modelling Symposium (EMS)*, pages 165–171, 2016. doi: 10.1109/EMS.2016.037.

[97] E. Hanes and S. Machin. Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, 30(3):247–267, 2014. doi: 10.1177/1043986214536665. URL `https://doi.org/10.1177/1043986214536665`.

[98] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL `https://aclanthology.org/2022.acl-long.234`.

[99] J. Hawdon, A. Oksanen, and P. Räsänen. Online extremism and online hate. *Nordicom-Information*, 37:29–37, 2015.

[100] B. Hedden. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, 49(2):209–231, 2021. doi: 10.1111/papa.12189.

[101] S. Hinduja and J. W. Patchin. Cyberbullying: Identification. *Prevention and Response, Cyberbullying Research Center*, 2014.

[102] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

[103] J. Holmes. *An Introduction to Sociolinguistics*. Routledge, 2013.

[104] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra. A comparison of common users across instagram and ask.fm to better understand cyberbullying. In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing, BDCloud 2014, Sydney, Australia, December 3-5, 2014*, pages 355–362. IEEE Computer Society, 2014. doi: 10.1109/BDCloud.2014.87. URL `https://doi.org/10.1109/BDCloud.2014.87`.

[105] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In T. Liu, C. N. Scollon, and W. Zhu, editors, *Social Informatics - 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, volume 9471 of *Lecture Notes in Computer Science*, pages 49–66. Springer, 2015. doi: 10.1007/978-3-319-27433-1_4. URL https://doi.org/10.1007/978-3-319-27433-1_4.

[106] D. Hovy and S. Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

[107] Q. Huang, D. Inkpen, J. Zhang, and D. Van Bruwaene. Cyberbullying intervention based on convolutional neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 42–51, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-4405.

[108] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287600. URL https://doi.org/10.1145/3287560.3287600.

[109] M. Jain, P. Goel, P. Singla, and R. Tehlan. Comparison of various word embeddings for hate-speech detection. In A. Khanna, D. Gupta, Z. Pólkowski, S. Bhattacharyya, and O. Castillo, editors, *Data Analytics and Management*, pages 251–265, Singapore, 2021. Springer Singapore.

[110] S. Jain and B. C. Wallace. Attention is not explanation. In *NAACL-HLT (1)*, pages 3543–3556. Association for Computational Linguistics, 2019.

[111] S.-j. Ji, Q. Zhang, J. Li, D. K. Chiu, S. Xu, L. Yi, and M. Gong. A burst-based unsupervised method for detecting review spammer groups. *Information Sciences*, 2020.

[112] S. Jiang, R. E. Robertson, and C. Wilson. Reasoning about political bias in content moderation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09): 13669–13672, Apr. 2020. doi: 10.1609/aaai.v34i09.7117. URL https://ojs.aaai.org/index.php/AAAI/article/view/7117.

[113] T. Joachims. A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–136, 2001.

[114] K. Joseph and J. Morgan. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.405. URL https://aclanthology.org/2020.acl-main.405.

[115] Kaggle. Detecting insults in social commentary. https://www.kaggle.com/c/detecting-insults-in-social-commentary/data, 2012. Accessed: 2020-09-28.

[116] G. Kambhatla, I. Stewart, and R. Mihalcea. Surfacing racial stereotypes through identity portrayal. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1604–1615. ACM, 2022. doi: 10.1145/3531146.3533217. URL https://doi.org/10.1145/3531146.3533217.

[117] M. Kaneko, D. Bollegala, and N. Okazaki. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.111.

[118] K. B. Kansara and N. M. Shekokar. A framework for cyberbullying detection in social network. *International Journal of Current Engineering and Technology*, 5(1):494–498, 2015.

[119] H.-T. Kao, S. Yan, D. Huang, N. Bartley, H. Hosseinmardi, and E. Ferrara. Understanding cyberbullying on instagram and ask.fm via social role detection. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 183–188, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3316505. URL https://doi.org/10.1145/3308560.3316505.

[120] E. Keryova. Youtube: Online video and participatory culture, 2020.

[121] J. Y. Kim, C. Ortiz, S. Nam, S. Santiago, and V. Datta. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921*, 2020.

[122] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*, pages 195–204, 2013.

[123] A. Koufakou, V. Basile, and V. Patti. Florunito@trac-2 retrofitting word embeddings on an abusive lexicon for aggressive language detection. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, editors, *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 106–112. European Language Resources Association (ELRA), 2020. URL `https://www.aclweb.org/anthology/2020.trac-1.17/`.

[124] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of BERT. In *EMNLP/IJCNLP (1)*, pages 4364–4373. Association for Computational Linguistics, 2019.

[125] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and H. H. Reese. Cyber bullying among college students: Evidence from multiple domains of college life. In *Misbehavior online in higher education*. Emerald Grp. Pub. Ltd., 2012.

[126] R. M. Kowalski, S. P. Limber, and P. W. Agatston. *Cyberbullying: Bullying in the digital age*. John Wiley & Sons, 2012.

[127] K. Krasnowska-Kieras and A. Wróblewska. A Simple Neural Network for Cyberbullying Detection. *Proceedings of the PolEval2019Workshop*, page 161, 2019.

[128] K. Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[129] S. Krishna, R. Gupta, A. Verma, J. Dhamala, Y. Pruksachatkun, and K.-W. Chang. Measuring fairness of text classifiers via prediction sensitivity. *ArXiv*, abs/2203.08670, 2022.

[130] R. Kukla. Slurs, interpellation, and ideology. *The Southern Journal of Philosophy*, 56: 7–32, 2018.

[131] A. Kumar, S. Nayak, and N. Chandra. Empirical analysis of supervised machine learning techniques for cyberbullying detection. In S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, and I. Pan, editors, *International Conference*

on Innovative Computing and Communications, pages 223–230, Singapore, 2019. Springer Singapore. ISBN 978-981-13-2354-6.

[132] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

[133] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[134] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

[135] j. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[136] M. A. Lepori. Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1720–1728. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.151. URL https://doi.org/10.18653/v1/2020.coling-main.151.

[137] S. Levin. Civil rights groups urge facebook to fix 'racially biased' moderation system. *The Guardian*, 2017. URL https://www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter.

[138] W. li and Z. Liu. A method of svm with normalization in intrusion detection. *Procedia Environmental Sciences*, 11:256–262, 2011. ISSN 1878-0296. doi: https://doi.org/10.1016/j.proenv.2011.12.040. URL https://www.sciencedirect.com/science/article/pii/S1878029611008632. 2011 2nd International Conference on Challenges in Environmental Science and Computer Engineering (CESCE 2011).

[139] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.488. URL https://aclanthology.org/2020.acl-main.488.

[140] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

[141] B. Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[142] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[143] E. Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[144] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[145] T. Mahlangu, C. Tu, and P. Owolawi. A review of automated detection methods for cyberbullying. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–5, 2018. doi: 10.1109/ICONIC.2018.8601278.

[146] A. Mahmud, K. Z. Ahmed, and M. Khan. Detecting flames and insults in text. In *International Conference on Natural Language Processing*. BRAC University, 2008.

[147] T. Manzini, L. Yao Chong, A. W. Black, and Y. Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://aclanthology.org/N19-1062.

[148] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, 2021.

[149] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. On measuring social biases in sentence encoders. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 622–628. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1063. URL https://doi.org/10.18653/v1/n19-1063.

[150] J. Mchangama, N. Alkiviadou, and R. Mendiratta. A FRAMEWORK OF FIRST REFERENCE Decoding a human rights approach to content moderation in the era of platformization. *The Future of free speech*, 11 2021. URL https://futurefreespeech.com/a-framework-of-first-reference-decoding-a-human-rights-approach-to-conte

[151] P. McIntosh. White privilege and male privilege: A personal account of coming to see correspondences through work in women's studies (1988). *Race, class, and gender: An anthology*, pages 95–105, 2001.

[152] N. Meade, E. Poole-Dayan, and S. Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.132. URL https://aclanthology.org/2022.acl-long.132.

[153] M. Miceli, J. Posada, and T. Yang. Studying up machine learning data: Why talk about bias when we mean power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP), jan 2022. doi: 10.1145/3492853. URL https://doi.org/10.1145/3492853.

[154] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. volume 2, pages 1045–1048, 01 2010.

[155] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

[156] T. Mikolov, K. Chen, G. Corrado, and J. Dean. word2vec embeddings, 2021. URL https://code.google.com/archive/p/word2vec/. [Online] Accessed 05/11/2021.

[157] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Glove twitter embeddings, 2021. URL https://fasttext.cc/docs/en/english-vectors.html. [Online] Accessed 25/04/2022.

[158] M. Mladenović, V. Ošmjanski, and S. V. Stanković. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Comput. Surv.*, 54(1), Jan. 2021. ISSN 0360-0300. doi: 10.1145/3424246. URL https://doi.org/10.1145/3424246.

[159] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, Jan. 2022. ISSN 2198-6053. doi: 10.1007/s40747-021-00608-2. URL https://doi.org/10.1007/s40747-021-00608-2.

[160] B. J. Morris. History of lesbian, gay, bisexual and transgender social movements. *PsycEXTRA Dataset*, 2010.

[161] Z. Mossie. Social media dark side content detection using transfer learning emphasis on hate and conflict. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 259–263, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3382084. URL https://doi.org/10.1145/3366424.3382084.

[162] M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, editors, *Complex Networks and Their Applications VIII*, pages 928–940, Cham, 2020. Springer International Publishing. ISBN 978-3-030-36687-2.

[163] H. Mubarak, H. Al-Khalifa, and A. Al-Thubaity. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In H. Al-Khalifa, T. Elsayed, H. Mubarak, A. Al-Thubaity, W. Magdy, and K. Darwish, editors, *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with*

Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pages 162–166, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.osact-1.20.

[164] H. Mubarak, S. Hassan, and S. A. Chowdhury. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457, 2023.

[165] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev, and E. Sutinen. Antisocial behavior corpus for harmful language detection. In *2013 Federated Conference on Computer Science and Information Systems*, pages 261–265. IEEE, 2013.

[166] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee. A review of cyberbullying detection: An overview. In *2013 13th International Conference on Intelligent Systems Design and Applications*, pages 325–330, 2013. doi: 10.1109/ISDA.2013.6920758.

[167] M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

[168] V. Nahar, X. Li, C. Pang, and Y. Zhang. Cyberbullying detection based on text-stream classification. In *Proceedings of the Conferences in Research and Practice in Information Technology Series, Australian Computer Society*, volume 146, pages 49–58, 2013.

[169] G. NaliniPriya and M. Asswini. A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks. *ARPN Journal of Engineering and Applied Sciences© 2006–2015 Asian Research Publishing Network (ARPN)*, 10(10), 2015.

[170] B. Nandhini and J. Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, page 20. ACM, 2015.

[171] B. S. Nandhini and J. Sheeba. Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45:485–492, 2015.

[172] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154.

[173] I. Nazar, D. Zois, and M. Yao. A hierarchical approach for timely cyberbullying detection. In *2019 IEEE Data Science Workshop (DSW)*, pages 190–195, June 2019. doi: 10.1109/DSW.2019.8755598.

[174] J.-P. Ng and V. Abrecht. Better summarization evaluation with word embeddings for rouge. *arXiv preprint arXiv:1508.06034*, 2015.

[175] D. Nguyen, B. McGillivray, and T. Yasseri. Emo, love, and god: Making sense of urban dictionary, a crowd-sourced online dictionary. *CoRR*, abs/1712.08647, 2017. URL http://arxiv.org/abs/1712.08647.

[176] 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, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883062. URL https://doi.org/10.1145/2872427.2883062.

[177] S. U. Nobel. *Algorithms of Oppression: How search engines reinforce racism*. New York University Press, 2018.

[178] A. Nocentini, V. Zambuto, and E. Menesini. Anti-bullying programs and information and communication technologies (icts): A systematic review. *Aggression and Violent Behavior*, 23:52–60, 2015. ISSN 1359-1789. doi: https://doi.org/10.1016/j.avb.2015.05.012. URL https://www.sciencedirect.com/science/article/pii/S1359178915000749. Bullying, Cyberbullying, and Youth Violence: Facts, Prevention, and Intervention.

[179] J. Nordell. *The end of bias*. Granta publications, 2021.

[180] L. T. O'Brien, A. Blodorn, G. Adams, D. M. Garcia, and E. Hammer. Ethnic variation in gender-stem stereotypes and stem participation: An intersectional approach. *Cultural Diversity and Ethnic Minority Psychology*, 21(2):169, 2015.

[181] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019. ISSN 2624-909X. doi: 10.3389/fdata.2019.00013. URL https://www.frontiersin.org/article/10.3389/fdata.2019.00013.

[182] S. Z. Omar, A. Daud, M. S. Hassan, J. Bolong, and M. Teimmouri. Children Internet usage: Opportunities for self development. *Procedia-Social and Behavioral Sciences*, 155:75–80, 2014. ISSN 1877-0428.

[183] C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[184] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2019.

[185] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf.

[186] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135, Jan. 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL https://doi.org/10.1561/1500000011.

[187] Z. Papakipos and J. Bitton. Augly: Data augmentations for robustness, 2022.

[188] A. Papasavva, S. Zannettou, E. D. Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. *CoRR*, abs/2001.07487, 2020. URL https://arxiv.org/abs/2001.07487.

[189] S. Parime and V. Suri. Cyberbullying detection and prevention: data mining and psychological perspective. In *2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014]*, pages 1541–1547. IEEE, 2014.

[190] J. W. Patchin and S. Hinduja. *Cyberbullying prevention and response: Expert perspectives*. Routledge, 2012.

[191] J. A. Pater, A. D. Miller, and E. D. Mynatt. This digital life: A neighborhood-based study of adolescents' lives online. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2305–2314, 2015.

[192] S. Paul and S. Saha. Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*, pages 1–8, 2020.

[193] J. Pavlopoulos, N. Thain, L. Dixon, and I. Androutsopoulos. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576, 2019.

[194] R. Pawar, Y. Agrawal, A. Joshi, R. Gorrepati, and R. R. Raje. Cyberbullying detection system with multiple server configurations. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0090–0095. IEEE, 2018.

[195] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[196] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. doi: 10.3115/v1/d14-1162. URL https://doi.org/10.3115/v1/d14-1162.

[197] C. C. Perez. *Invisible women: Data bias in a world designed for men*. Abrams, 2019.

[198] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[199] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523, Jun 2021. ISSN 1574-0218. doi: 10.1007/s10579-020-09502-8. URL https://doi.org/10.1007/s10579-020-09502-8.

[200] N. Potha and M. Maragoudakis. Cyberbullying detection using time series modeling. In Z. Zhou, W. Wang, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, editors, *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, pages 373–382. IEEE Computer Society, 2014. doi: 10.1109/ICDMW.2014.170. URL https://doi.org/10.1109/ICDMW.2014.170.

[201] V. Prabhakaran, B. Hutchinson, and M. Mitchell. Perturbation sensitivity analysis to detect unintended model biases. 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 5740–5745, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1578. URL https://aclanthology.org/D19-1578.

[202] M. Ptaszynski, F. Masui, T. Nitta, S. Hatakeyama, Y. Kimura, R. Rzepka, and K. Araki. Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction*, 8:15–30, 2016. ISSN 2212-8689.

[203] C. Qian, F. Feng, L. Wen, C. Ma, and P. Xie. Counterfactual inference for text classification debiasing. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5434–5445. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.422. URL https://doi.org/10.18653/v1/2021.acl-long.422.

[204] R. Qian, C. Ross, J. Fernandes, E. M. Smith, D. Kiela, and A. Williams. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.646.

[205] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[206] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[207] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In J. Pei, F. Silvestri, and J. Tang, editors, *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015,*

Paris, France, August 25 - 28, 2015, pages 617–622. ACM, 2015. doi: 10.1145/2808797.2809381. URL https://doi.org/10.1145/2808797.2809381.

[208] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. Investigating factors influencing the latency of cyberbullying detection. *arXiv preprint arXiv:1611.05419*, 2016.

[209] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, page 1738–1747, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351911. doi: 10.1145/3167132.3167317. URL https://doi.org/10.1145/3167132.3167317.

[210] E. Raisi and B. Huang. Cyberbullying detection with weakly supervised machine learning. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 409–416, 2017.

[211] E. Raisi and B. Huang. Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 479–486, 2018. doi: 10.1109/ASONAM.2018.8508240.

[212] S. Raschka. Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*, 2014.

[213] S. Ravfogel, G. Prasad, T. Linzen, and Y. Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In A. Bisazza and O. Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 194–209. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.15. URL https://doi.org/10.18653/v1/2021.conll-1.15.

[214] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In X. Chen, T. S. Dillon, H. Ishibuchi, J. Pei, H. Wang, and M. A. Wani, editors, *10th International Conference on Machine Learning and Applications and Workshops, ICMLA 2011, Honolulu, Hawaii, USA, December 18-21, 2011. Volume 2: Special Sessions and Workshop*, pages 241–244. IEEE Computer Society, 2011. doi: 10.1109/ICMLA.2011.152. URL https://doi.org/10.1109/ICMLA.2011.152.

[215] S. Roberts. *Behind the screens: content moderation in the shadows of social media*. Yale University Press, 2019. ISBN 9780300235883.

[216] A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021.

[217] S. Rogers and M. Girolami. *A First Course in Machine Learning, Second Edition*. Chapman & Hall/CRC, 2nd edition, 2016. ISBN 1498738486.

[218] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd. Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pages 242–247. IEEE, 2017.

[219] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho. A "deeper" look at detecting cyberbullying in social networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018. doi: 10.1109/IJCNN.2018.8489211.

[220] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, and I. Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[221] S. Rosenfield. Triple jeopardy? mental health at the intersection of gender, race, and class. *Social Science & Medicine*, 74(11):1791–1801, 2012.

[222] S. Salawu, Y. He, and J. Lumsden. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24, 2020. doi: 10.1109/TAFFC.2017.2761757.

[223] s. samuel. Ais islamophobia problem. 2021. URL `https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim`.

[224] 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://aclanthology.org/P19-1163`.

[225] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

[226] T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[227] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[228] S. Serrano and N. A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[229] D. S. Shah, H. A. Schwartz, and D. Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468.

[230] H. K. Sharma, K. Kshitiz, et al. Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 265–272. IEEE, 2018.

[231] R. Shetgiri. Bullying and victimization among children. *Advances in pediatrics*, 60 (1):33, 2013.

[232] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[233] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2016.

[234] V. K. Singh, Q. Huang, and P. K. Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. In *2016 IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining (ASONAM), pages 884–887, 2016. doi: 10.1109/ASONAM.2016.7752342.

[235] V. K. Singh, S. Ghosh, and C. Jose. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2090–2099, 2017.

[236] V. K. Singh, M. L. Radford, Q. Huang, and S. Furrer. " they basically like destroyed the school one day" on newer app features and cyberbullying in schools. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1210–1216, 2017.

[237] P. K. Smith. Cyberbullying and cyber aggression. In *Handbook of school violence and school safety*, pages 111–121. Routledge, 2012.

[238] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385, 2008.

[239] D. Soni and V. K. Singh. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26, 2018.

[240] Spacy. Linguistic features by spacy. `https://spacy.io/usage/linguistic-features`, 2021. Accessed: 2021-03-02.

[241] R. Steed, S. Panda, A. Kobren, and M. Wick. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3524–3542, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.247. URL `https://aclanthology.org/2022.acl-long.247`.

[242] C. M. Steele. *Whistling Vivaldi: How stereotypes affect us and what we can do*. WW Norton & Company, 2011.

[243] F. Sticca, S. Ruggieri, F. Alsaker, and S. Perren. Longitudinal risk factors for cyberbullying in adolescence. *Journal of community & applied social psychology*, 23 (1):52–67, 2013.

[244] J. Stypinska. Ai ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & society*, pages 1–13, 2022.

[245] C. Sun, X. Qiu, Y. Xu, and X. Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[246] X. Sun and W. Lu. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428. Association for Computational Linguistics, 2020.

[247] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328, 2017.

[248] R. W. Sussman. *The Pioneer Fund, 1970s–1990s*, pages 235–248. Harvard University Press, 2014. ISBN 9780674417311. URL http://www.jstor.org/stable/j.ctt9qdt73.13.

[249] R. W. Sussman. *The Pioneer Fund in the Twenty-First Century*, pages 249–272. Harvard University Press, 2014. ISBN 9780674417311. URL http://www.jstor.org/stable/j.ctt9qdt73.14.

[250] R. W. Sussman. *Early Racism in Western Europe*, pages 11–42. Harvard University Press, 2014. ISBN 9780674417311. URL http://www.jstor.org/stable/j.ctt9qdt73.5.

[251] R. W. Sussman. *Eugenics and the Nazis*, pages 107–145. Harvard University Press, 2014. ISBN 9780674417311. URL http://www.jstor.org/stable/j.ctt9qdt73.8.

[252] E. Sutinen. Automatic detection of antisocial behaviour in texts. *Informatica (Ljubljana)*, 38(1), 2014.

[253] Swear words. Swear words list, 2022. URL https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en. [Online] Accessed 26/04/2022.

[254] C. Sweeney and M. Najafian. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy, July

2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1162. URL https://aclanthology.org/P19-1162.

[255] N. Tahmasbi and A. Fuchsberger. Challenges and future directions of automated cyberbullying detection. 2018. Publisher Copyright: © 2018 Association for Information Systems. All rights reserved.; 24th Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018 ; Conference date: 16-08-2018 Through 18-08-2018.

[256] Y. C. Tan and L. E. Celis. Assessing social and intersectional biases in contextualized word representations. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/201d546992726352471cfea6b0df0a48-Abstract.html.

[257] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1146. URL https://www.aclweb.org/anthology/P14-1146.

[258] N. Tarwani, U. Chorasia, and P. K. Shukla. Survey of cyberbulling detection on social media big-data. *International Journal of Advanced Research in Computer Science*, 8 (5), 2017.

[259] R. Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://aclanthology.org/W17-1606.

[260] Tensorflow.org. Text tokenization utility class. https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer, 2020. Accessed: 2020-09-28.

[261] X. Tian. Investigating cyberbullying in social media: The case of twitter. 2016.

[262] N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 254–265, 2021.

[263] S. Tomkins, L. Getoor, Y. Chen, and Y. Zhang. A socio-linguistic model for cyberbullying detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 53–60. IEEE, 2018.

[264] A. Tommasel, J. M. Rodriguez, and D. L. Godoy. Features for detecting aggression in social media: An exploratory study. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)*, 2018.

[265] M. Tucker, P. Qian, and R. Levy. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.76. URL https://aclanthology.org/2021.findings-acl.76.

[266] İ. Türker, E. Şehirli, and E. Demiral. Uncovering the differences in linguistic network dynamics of book and social media texts. *SpringerPlus*, 5(1):1–18, 2016.

[267] E. Ungless, A. Rafferty, H. Nag, and B. Ross. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE, Nov. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.nlpcss-1.23.

[268] Urban dictionary. Urban dictionary embeddings, 2021. URL http://smash.inf.ed.ac.uk/ud-embeddings/. [Online] Accessed 05/11/2021.

[269] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[270] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10), 2018.

[271] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui. Attention interpretability across NLP tasks. *arXiv.org*, arXiv:1909.11218, 2019.

[272] J. Vásquez, S. Andersen, G. Bel-enguix, H. Gómez-adorno, and S.-l. Ojeda-trueba. HOMO-MEX: A Mexican Spanish annotated corpus for LGBT+phobia detection on Twitter. In Y.-l. Chung, P. R\"ottger, D. Nozza, Z. Talat, and A. Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.20. URL https://aclanthology.org/2023.woah-1.20.

[273] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NIPS 2017)*, pages 5998–6008, 2017.

[274] B. Vidgen and L. Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[275] J. Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[276] J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, Aug. 2019. Association for Computational Linguistics.

[277] P. Voué, T. D. Smedt, and G. D. Pauw. 4chan & 8chan embeddings. *CoRR*, abs/2005.06946, 2020. URL https://arxiv.org/abs/2005.06946.

[278] C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204, Jul 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03666-1. URL https://doi.org/10.1038/s41586-021-03666-1.

[279] C. Wang, P. Nulty, and D. Lillis. A comparative study on word embeddings in deep learning for text classification. In *NLPIR 2020: 4th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea, December 18-20, 2020*, pages 37–46. ACM, 2020. doi: 10.1145/3443279.3443304. URL https://doi.org/10.1145/3443279.3443304.

[280] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. R. Kingsbury, and H. Liu. A comparison of word embeddings for the biomedical natural language

processing. *J. Biomed. Informatics*, 87:12–20, 2018. doi: 10.1016/j.jbi.2018.09.008. URL https://doi.org/10.1016/j.jbi.2018.09.008.

[281] 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.

[282] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-2013. URL https://doi.org/10.18653/v1/n16-2013.

[283] Z. Waseem, J. Thorne, and J. Bingel. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham, 2018. ISBN 978-3-319-78583-7. doi: 10.1007/978-3-319-78583-7_3. URL https://doi.org/10.1007/978-3-319-78583-7_3.

[284] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. H. Chi, and S. Petrov. Measuring and reducing gendered correlations in pre-trained models. Technical report, Google Research, 2020. URL https://arxiv.org/abs/2010.06032.

[285] M. L. Williams and P. Burnap. Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *The British Journal of Criminology*, 56(2):211–238, 06 2015. ISSN 0007-0955. doi: 10.1093/bjc/azv059. URL https://doi.org/10.1093/bjc/azv059.

[286] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117, 07 2019. ISSN 0007-0955. doi: 10.1093/bjc/azz049. URL https://doi.org/10.1093/bjc/azz049.

[287] S. R. Wilson, W. Magdy, B. McGillivray, K. Garimella, and G. Tyson. Urban dictionary embeddings for slang NLP applications. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4764–4773. European Language Resources Association, 2020. URL https://www.aclweb.org/anthology/2020.lrec-1.586/.

[288] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 347–354, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220619.

[289] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349130.

[290] 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*, pages 656–666. Association for Computational Linguistics, 2012.

[291] J. Yadav, D. Kumar, and D. Chauhan. Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE, 2020.

[292] S. Zahir. Making public policy decisions using a web-based multi-criteria electoral system (MCES). *Int. J. Inf. Technol. Decis. Mak.*, 1(2):293–309, 2002. doi: 10.1142/S0219622002000178. URL https://doi.org/10.1142/S0219622002000178.

[293] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab, and S. Ramli. A review on cyberbullying in malaysia from digital forensic perspective. In *2016 International Conference on Information and Communication Technology (ICICTM)*, pages 246–250, 2016. doi: 10.1109/ICICTM.2016.7890808.

[294] H. Zhang, A. X. Lu, M. Abdalla, M. B. A. McDermott, and M. Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In M. Ghassemi,

editor, *ACM CHIL '20: ACM Conference on Health, Inference, and Learning, Toronto, Ontario, Canada, April 2-4, 2020 [delayed]*, pages 110–120. ACM, 2020. doi: 10.1145/3368555.3384448. URL https://doi.org/10.1145/3368555.3384448.

[295] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 740–745. IEEE, 2016.

[296] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. "why should you trust my explanation?" understanding uncertainty in lime explanations. In *International Conference on Machine Learning, AI for Social Good Workshop*, 2019.

[297] Z. Zhang, D. Robinson, and J. Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, editors, *The Semantic Web*, pages 745–760, Cham, 2018. Springer International Publishing.

[298] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.

[299] R. Zhao and K. Mao. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, 8 (3):328–339, 2016.

[300] R. Zhao, A. Zhou, and K. Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking*, ICDCN '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340328. doi: 10.1145/2833312.2849567. URL https://doi.org/10.1145/2833312.2849567.

[301] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. A. Smith. Challenges in automated debiasing for toxic language detection. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23,*

2021, pages 3143–3155. Association for Computational Linguistics, 2021. URL https://www.aclweb.org/anthology/2021.eacl-main.274/.

[302] D. W. Zimmerman and B. D. Zumbo. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86, 1993.

[303] R. Zmigrod, S. J. Mielke, H. Wallach, and R. Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161.