Pós-Graduação em Ciência da Computação

FRANCIMARIA RAYANNE DOS SANTOS NASCIMENTO

Hate Speech Detection and Gender Bias Mitigation on Online Social Media

Recife

2024

FRANCIMARIA RAYANNE DOS SANTOS NASCIMENTO

Hate Speech Detection and Gender Bias Mitigation on Online Social Media

Doctoral thesis presented to the Postgraduate Program in Computer Science of the Centre of Informatics of Federal University of Pernambuco as a partial requirement for obtaining the degree of Doctor in Computer Science.

Concentration area: Computational Intelligence

Supervisor: Prof. PhD George DC Cavalcanti

Co-supervisor: Prof. PhD Márjory Cristiany Da Costa Abreu

Recife

2024

Catalogação na fonte
Bibliotecária: Luiza de Oliveira/CRB1316

N244h Nascimento, Francimaria Rayanne dos Santos.
Hate speech detection and gender bias mitigation on online social media / Francimaria Rayanne dos Santos Nascimento.– 2024.
136 f.: il.

Orientador: George D C Cavalcanti.
Coorientadora: Márjory Cristiany da Costa Abreu
Tese (Doutorado) – Universidade Federal de Pernambuco. Centro de Informática. Programa de Pós-graduação em Ciência da Computação, Recife, 2024.
Inclui referências e apêndice.

1. Discurso de ódio. 2. Ensemble learning. 3. Viés de gênero. 4. Multi-view. 5. Redes sociais. I.Cavalcanti, George D C. II. Abreu, Márjory Cristiany da Costa. III. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2024 – 58

FRANCIMARIA RAYANNE DOS SANTOS NASCIMENTO

HATE SPEECH DETECTION AND GENDER BIAS MITIGATION ON ONLINE SOCIAL MEDIAL

Tese de doutorado apresentada ao Programa de Pós Graduação em Ciência da Computação da Universidade Federal de Pernambuco, Centro de Informática, como requisito para a obtenção do título de Doutor em Ciência da Computação. Área de concentração: Inteligência Computacional.

.

Aprovado em: _12_/_03_/_2024_.

BANCA EXAMINADORA

_____
Prof. Dr. Tsang Ing Ren (Examinador Interno)
Centro de Informática / UFPE

_____
Prof. Dr. Rafael Menelau Oliveira e Cruz (Examinador Externo)
Département de génie logiciel et des TI / ÉTC

_____
Profa. Dra. Mirella Moura Moro (Examinador Externo)
Departamento de Ciência de Ciência da Computação / UFMG

_____
Profa. Dra. Lilian Berton (Examinador Externo)
Departamento de Ciência e Tecnologia / UNIFESP

_____
Profa. Dra. Carolina Scarton (Examinador Externo)
Department of Computer Science / University of Sheffield

ACKNOWLEDGEMENTS

Primeiramente agradeço a Deus por ter me guiado nesta caminhada, me dando força e coragem para enfrentar todos os obstáculos que ocorreram durante esta trajetória.

A minha família e em especial a minha amada avó (**Josefa Izabel dos Santos**, *in memoriam*), meu pai (**Josenildo Antônio Fernandes**) e a minha mãe (**Francinete Izabel dos Santos**), pelos valiosos conselhos e por sempre acreditar em mim. A meu amado **João Batista de Oliveira Neto**, pela dedicação e companheirismo.

Agradeço a meu orientador, professor **George DC Cavalcanti**, e minha co-orientadora, professora **Márjory Cristiany da Costa Abreu**, pelo empenho dedicado e orientação. Obrigada pela presteza, pelo incentivo, e principalmente pelo papel de educador.

Ao apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (**CAPES**), que foi de suma importância para conclusão desta tese.

RESUMO

A popularização das redes sociais online permitiu a rápida proliferação de conteúdos gerados pelos usuários. A grande quantidade de conteúdo gerado a cada segundo nas plataformas de redes sociais torna a moderação adequada do seu conteúdo árdua e demorada, resultando em uma fácil disseminação do discurso de ódio. Embora tenham sido feitos avanços significativos na detecção automática de discurso de ódio, preocupações têm surgido a cerca de robustez do modelo de aprendizagem e do seu impacto devido aos seus comportamentos potencialmente tendenciosos, levando a tendências questionáveis baseadas em termos de identidade (por exemplo, mulheres, negros ou gay). Nesta tese, abordamos o preconceito não intencional, especificamente o preconceito de gênero não intencional, na tarefa de detecção de discurso de ódio. Em primeiro lugar, propusemos um estudo abrangente do discurso de ódio, incluindo uma análise crítica das definições de discurso de ódio propostas em múltiplas plataformas e na comunidade científica. Também apresenta uma visão geral das principais abordagens normalmente utilizadas na detecção automática de discurso de ódio. Os resultados apresentaram uma análise crítica dos recursos teóricos e práticos, discutindo oportunidades nesta área e diversos desafios, incluindo questões como o preconceito. Considerar o preconceito não intencional no modelo de detecção automática de discurso de ódio é essencial para prevenir uma potencial discriminação não intencional. Portanto, propusemos uma nova metodologia usando um conjunto com multi-visões (multi-view) para detecção automática de discurso de ódio e mitigação não intencional de preconceitos de gênero. A metodologia proposta consiste em dois módulos: (1) um módulo de mitigação de preconceito de gênero baseado na detecção e substituição de palavras sensíveis ao preconceito e (2) um módulo de detecção de discurso de ódio usando um classificador empilhado de múltiplas visualizações (multi-view stacked). O classificador empilhado multivisualizações combina classificadores básicos treinados com representações de recursos distintas. Resultados experimentais em quatro conjuntos de dados de benchmark demonstram a eficácia da abordagem proposta em comparação com soluções de última geração, reduzindo o viés não intencional sem comprometer o desempenho do modelo. Além disso, existem preocupações se o viés não intencional pode apresentar comportamentos diferentes dependendo da técnica de extração de características utilizada. Portanto, também propusemos uma estrutura para ajudar a analisar comportamentos tendenciosos das técnicas de extração de características. Além disso, foi concebido um novo conjunto de dados abrangente para ajudar na avaliação de preconceitos de gênero não intencionais, denominado

conjunto de dados imparcial. Conduzimos um estudo experimental sobre vários métodos de extração de recursos de última geração, com foco em seu potencial viés em relação aos termos de identidade. Nossas descobertas indicam que a técnica de extração de características pode influenciar o viés encontrado no modelo final, e sua eficácia pode depender do conjunto de dados analisado.

Palavras-chaves: Discurso de ódio; Ensemble learning; Viés de gênero; Multi-view; Redes sociais.

ABSTRACT

The popularisation of online social media has allowed the quick proliferation of user-generated content. The large amount of content generated every second on social media platforms makes the proper moderation of its content arduous and time-consumed, resulting in an easy dissemination of hate speech. Even though significant advances have been made for automatic hate speech detection, concerns have been raised about the robustness of the learning model and its impact due to its potentially biased behaviours, leading to questionable trends based on identity terms (e.g., women, black, or gay). In this thesis, we address unintended bias, specifically unintended gender bias, in the hate speech detection task. Firstly, we proposed a comprehensive study of hate speech, including a critical analysis of definitions of hate speech proposed across multiple platforms and in the scientific community. It also overviews the main approaches typically used in automatic hate speech detection. The results presented a critical analysis of theoretical and practical resources, discussing opportunities in this area and several challenges, including bias issues. Considering the unintended bias in the model for automatically detecting hate speech is essential to prevent potential unintended discrimination. Therefore, we proposed a new methodology using a multi-view ensemble for automatic hate speech detection and unintended gender bias mitigation. The proposed methodology consists of two modules: (1) a gender bias mitigation module based on the detection and replacement of bias-sensitive words and (2) a hate speech detection module using a multi-view stacked classifier. The multi-view stacked classifier combines base classifiers trained with distinct feature representations. Experimental results over four benchmark datasets demonstrate the proposed approach's effectiveness compared to state-of-the-art solutions, reducing the unintended bias without compromising the model performance. Furthermore, there are concerns whether unintended bias may presents different behaviours depend on the feature extraction technique used. Therefore, we also proposed a framework to help analyse the biased behaviour of feature extraction techniques. In addition, a new comprehensive dataset to help the unintended gender bias evaluation is designed, called the Unbiased dataset. We have conducted an experimental study on various state-of-the-art feature extraction methods, focusing on their potential bias towards identity terms. Our findings indicate that the feature extraction technique can influence the bias found in the final model, and its effectiveness can rely on the dataset analysed.

Keywords: Hate speech detection; Ensemble learning; Gender Bias; Multi-view; Social media.

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Δ	Original training set
Δ'	Training set after the gender bias mitigation module
Γ	Validation set
τ	Test set
F	Feature extractor
f_Δ	Feature vector of training set
f_τ	Feature vector of test set
f_Γ	Feature vector of unbiased test set
C	Trained classifier

CONTENTS

1 INTRODUCTION

This chapter introduces hate speech detection and gender bias context, the main motivations for this work, the problem statement being addressed, and objectives. Finally, it describes the contributions of this thesis and the organisation of the rest of it.

1.1 CONTEXT AND MOTIVATION

The increasing popularity of online social media, such as X (new Twitter brand), Instagram, and Facebook, has driven exponential growth in the number of content published (or shared) online, making manual moderation of this content expensive or unsustainable. In this way, it promotes an environment conducive to disseminating abusive content, such as hate speech. Hate speech is a severe problem demonstrating a clear intent to incite hate or promote hostility. This issue requires urgent solutions to prevent this harmful content from being disseminated.

Hate speech detection has been defined in different studies. A precise definition of hate speech is crucial in order to automatically distinguish it from other content (ROSS et al., 2016). In (DAVIDSON et al., 2017), the researchers defined hate speech as dehumanising language or hostile expression against a target group or group members based on specific characteristics through direct attacks or incitements to violence.

Different approaches have been proposed to automatically detect hate speech, including Machine Learning (ML) algorithms combined with Natural Language Processing (NLP) techniques (CRUZ; SOUSA; CAVALCANTI, 2022; KAPIL; EKBAL, 2020; MAZARI; BOUDOUKHANI; DJEFFAL, 2023; RISCH; KRESTEL, 2020; SALMINEN et al., 2020). Even though these tools have presented significant performance, concerns have arisen about intrinsic biases incorporated in the models, resulting in discrimination against certain groups (GARG et al., 2023).

In the learning process of the machine learning model, including hate speech detection, some bias in the data is assumed to perform prediction. This bias can help the model improve performance. However, it is not appropriate for a hate speech detection model to rely on characteristics such as the speaker's gender. If a model demonstrates such bias, it is referred to as unintended bias. The unintended bias can be learned by the over-generalisation of the association of specific terms, such as woman and gay, and the hateful class in different models, increasing false-positive instances (DIXON et al., 2018; MOZAFARI; FARAHBAKHSH; CRESPI,

2020).

Numerous studies have demonstrated the perilous consequences of bias in Artificial Intelligence (AI) systems across various domains over the past few years. In the context of automated hiring, recruiting tools based on AI algorithms have presented sexist behaviours penalising women in hiring processes based only on their gender (DASTIN, 2018). Similarly, in healthcare, an AI system developed to manage the patient's healthcare needs exhibited racial bias and sub-served black patients in the process (OBERMEYER et al., 2019). Clearly, bias must be mitigated in AI systems through design tools.

Different strategies have attempted to mitigate the unintended bias by directly employing statistic correction of the distribution of sensitive data (DIXON et al., 2018; NOZZA; VOLPETTI; FERSINI, 2019) or data correction by changing its representation to a more generalist format (BADJATIYA; GUPTA; VARMA, 2019). However, these strategies can provide unrealistic assumptions about the training data distribution or lose much information about the original data. Another approach proposed in the literature is to mitigate the bias in the model training by optimising the algorithm (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; ZHAO; ZHANG; HOPFGARTNER, 2022). Nevertheless, this strategy depends on the algorithm and requires manipulating the model parameters to reduce correlations based on biased attributes.

The unintended bias in learning models is a substantial problem that can affect the performance of ML models and perpetuate discrimination, reinforcing or amplifying social stereotypes and leading to social injustice. Therefore, dealing with unintended bias during the development process of a AI solution for hate speech detection is crucial.

1.2 PROBLEM STATEMENT

This thesis is focused on hate speech detection and unintended bias mitigation, more specifically on gender bias. Gender bias is a serious concern that can lead to preference or prejudice of gender over the other in the model, reinforcing or amplifying social gender stereotypes in the systems (SUN et al., 2019). The gender bias in hate speech detection models is a serious concern and can increase sexist behaviours, leading the model to perform better for a determinate gender than for others (PARK; SHIN; FUNG, 2018; NOZZA; VOLPETTI; FERSINI, 2019).

The machine learning algorithms are data-driven, specifically, the model learning from the main patterns in the training data. As a result, these algorithms can incorporate intrinsic biases

from the data. The bias can be incorporated through the significant association between a term and a specific class, leading to discrimination (BADJATIYA; GUPTA; VARMA, 2019). For instance, the strong association between specific terms such as "feminism" and "women" and sexist comments in benchmark datasets (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022; PARK; SHIN; FUNG, 2018). This association can contribute to over-fitting the original hate speech detection model, leading to generalisations such as labelling the instances with the word "women" as hateful.

Despite the proposed studies dedicated to unintended bias mitigation in the hate speech detection task, different issues need further exploration. A key issue is to mitigate the unintended bias without compromising the classification performance. For instance, in (PARK; SHIN; FUNG, 2018), although the approaches investigated, including Debiased Word Embeddings (BOLUKBASI et al., 2016), Gender Swap (ZHAO et al., 2018), and Bias fine-tuning, had reduced the bias, these approaches negatively affected the classification performance. Similarly, in (NOZZA; VOLPETTI; FERSINI, 2019), the proposed bias mitigation strategy presented an AUC drop in the misogyny detection context.

Furthermore, proper evaluation of the unintended bias is essential. Traditional metrics, such as accuracy, F1-score, and so on, usually are employed to assess the model performance from the original test set predictions. However, in the context of unintended bias evaluation, the metrics computed from the original test set can present unreliable results due to the possibility of the test set sharing the same biased distribution of identity terms as the training set (DIXON et al., 2018). Moreover, some metrics required an equal distribution of identity terms. Specifically, all terms need to appear in the same context (BORKAN et al., 2019; GARG et al., 2023). Even though different synthetic test set has been proposed in the context of toxicity comments classification (DIXON et al., 2018) and misogyny detection (NOZZA; VOLPETTI; FERSINI, 2019), designing a comprehensive dataset that addresses several test cases is still a significant challenge in the context of hate speech detection.

1.3 OBJECTIVES

The objective of this thesis is to develop a robust framework for hate speech detection that mitigates unintended gender bias without compromising the classification model performance. In addition, since the unintended bias can be introduced in machine learning models in different stages of the development process (LEE; SINGH, 2021), this thesis also proposes a framework

to help analyse the unintended bias at the feature level to evaluate the bias and its effects on the classification performance of machine learning algorithms. Therefore, the following specific objectives are stipulated:

- **Select and analyse different feature extraction approaches for textual data**: Each feature extraction method captures a different abstraction about the data and can present a different classification performance for each dataset;
- **Explore how to deal with gender biases**: We focus on the unintended gender bias problem for the hate speech detection task. Specifically, we address gender bias mitigation in the data;
- **Design a unbiased dataset and evaluate the gender bias in the hate speech context**: Reliable labels across a range of terms are needed to assess bias effectively. Considering that the original test set can follow the same biased distribution of the training set, evaluating the unintended bias with the original training set can compromise the evaluation of the bias metrics. We also assess performance metrics to evaluate the bias mitigation effects on the classification performance using multiple machine learning classifiers on different datasets.

Therefore, this thesis aims to answer the following research questions:

1. *Does the proposed multi-view stacked classifier combined with template-based mitigation outperform current techniques for hate speech detection in the context of unintended gender bias?*

   In Chapter 3, we proposed a multi-view stacked classifier using nine feature extraction methods combined with a template-based strategy for hate speech detection and gender bias mitigation. We performed our experiments in four real-world datasets using different classifiers to analyse whether the proposed ensemble learning model outperforms current techniques for hate speech detection.

2. *Can the bias mitigation method deal with gender biases in datasets without compromising the performance of the ensemble learning model?*

   In Chapter 3, we analysed the proposed ensemble learning model with and without the template-based strategy and in contrast with the best monolithic classifier evaluated. Moreover, we performed case studies to evaluate the effectiveness of the proposed

methodology using different pairs of examples. The examples presented are non-hateful tweets to evaluate the hateful score obtained by each model.

3. *Does the choice of the feature extraction technique impact the presence of unintended gender bias on the model's prediction?*

   In Chapter 4, we investigated the performance of five strategies for feature extraction, including two methods based on the Bag-of-Words strategy, Term Frequency (TF) and Term frequency-inverse document frequency (TF-IDF), and three embedding methods, Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), FastText, and Global Vectors for Word Representation (GloVe). We included the methods based on the Bag-of-Words strategy to evaluate the effectiveness and bias impact of methods less computationally complex. In addition, we also include RoBERTa, a language model developed based on BERT architecture, to contrast bias impact in an optimised version of the BERT model that has been extensively used. We analyse the results obtained with unintended bias metrics using an unbiased test dataset for each feature extraction method.

4. *Do feature extraction techniques tend to present bias when dealing with different datasets?*

   In Chapter 4, we evaluated using three real-world datasets and contrasted the results obtained for each one.

5. *Can the bias affect the performance of the models?*

   In Chapter 4, we investigated the classification performance of several classifiers using the metrics AUC and F1-score.

1.4 CONTRIBUTIONS

The main contribution of this thesis is in hate speech detection and unintended gender bias mitigation, leading to the proposal of a novel multi-view stacked framework and the discovery of several insights to overcome weaknesses related to the specific nuances of this subject and the complexity of this classification task.

As this thesis is manuscript-based, each chapter presents a distinct contribution that aims to address the problem by exploring novel approaches to training machine learning models that are both accurate and unbiased. The contributions are listed below:

- In Chapter 2, a comprehensive study is performed of the main approaches currently explored for hate speech detection, including popular features, datasets, and algorithms. The findings have strong potential to help researchers overcome weaknesses related to the intricacies of hate speech detection on online social media and direct different research opportunities.

- In Chapter 3, a new framework using multi-view data and ensemble learning for hate speech detection while dealing with gender bias is proposed. Experimental results demonstrate the proposal's effectiveness compared to the state-of-the-art solutions, reducing the bias without compromising the classifier performance.

- In Chapter 4, a new framework for evaluating the potential biases in feature extraction methods is proposed. In addition, a new dataset to help in the unintended gender bias evaluation is designed. The proposed framework and unbiased data can enhance our understanding of how these techniques function and assist in developing more fair models.

1.5 ORGANISATION OF THE THESIS

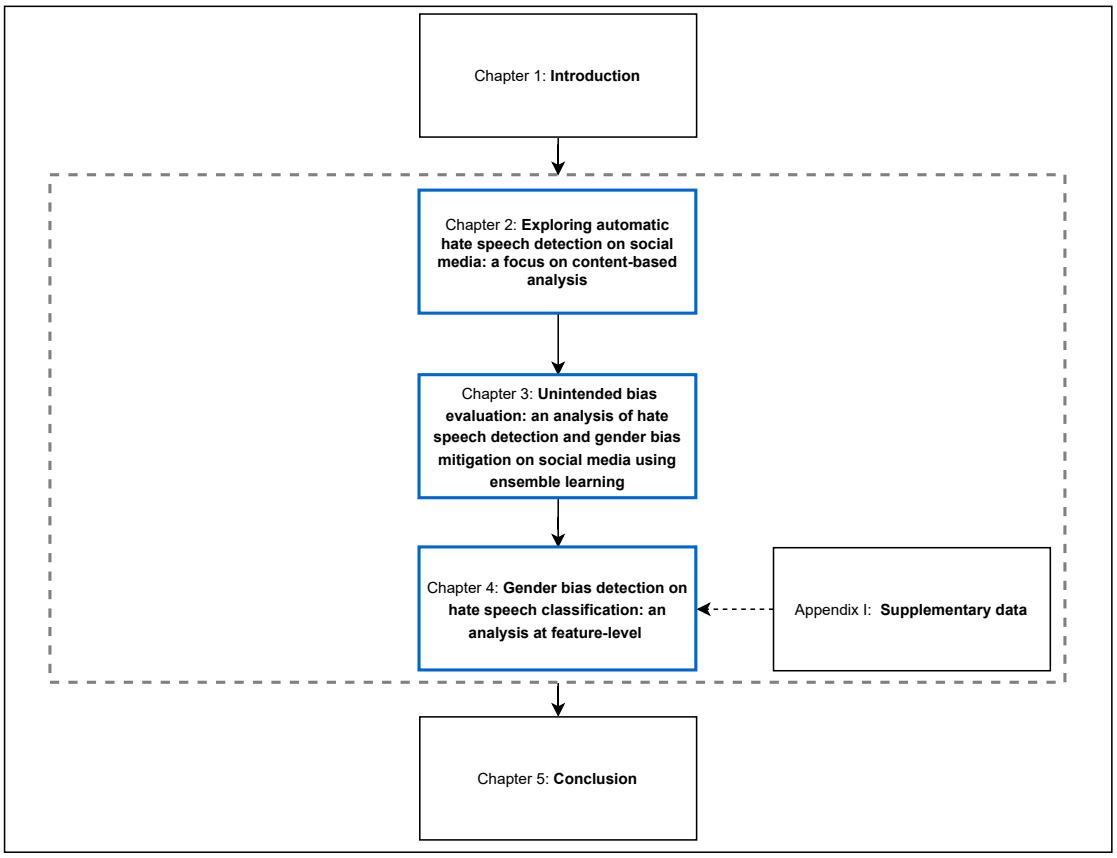
The structure of this manuscript-based thesis is composed of five chapters. Figure 1 provides an overview of this thesis organisation. The following chapters are:

- **Chapter 2** presents a survey on automatic hate detection on online social media. The main concepts related to this subject are discussed, such as benchmark datasets, feature extraction techniques, and classification models. In addition, several definitions of hate speech are analysed to help better understand this problem. The contents of this chapter have been published in the SAGE Open journal (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2023a);

- **Chapter 3** proposed a new framework for hate speech detection and gender bias mitigation. The contents of this chapter have been published in the Expert System with Applications journal (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022);

- **Chapter 4** proposed a new framework to analyse gender bias at the feature level. In addition, a new dataset is designed. The contents of this chapter have been published as

Figure 1 – Overview of the thesis organisation.



Source: Prepared by the author.

pre-print in the SSRN (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2023b) and are under review in the Neural Computing and Applications journal (NCAA);

- **Chapter 5** presents the final remarks about the main content discussed in this thesis.

2 EXPLORING AUTOMATIC HATE SPEECH DETECTION ON SOCIAL MEDIA: A FOCUS ON CONTENT-BASED ANALYSIS

Francimaria RS Nascimento[1], George DC Cavalcanti[1], and Márjory Da Costa-Abreu[2]

[1]Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes s/n, Recife, Brazil

[2] Department of Computing, Sheffield Hallam University, Sheffield, UK

Article Published in ≪ SAGE Open ≫ submitted 2021, published 2023.

Abstract

Hate speech is a challenging problem, and its dissemination can cause potential harm to individuals and society by creating a sense of general unwelcoming to marginalised groups, which are usually targeted. Therefore, it is essential to understand this issue and which techniques are useful for automatic detection. This paper presents a survey on automatic hate speech detection on social media, providing a structured overview of theoretical aspects and practical resources. Thus, we review different definitions of the term 'hate speech' from social network platforms and the scientific community. We also present an overview of the methodologies used for hate speech detection, and we describe the main approaches currently explored in this context, including popular features, datasets, and algorithms. Furthermore, we discuss some challenges and opportunities for better solving this issue.

2.1 INTRODUCTION

Social media platforms allow users to publish content about different subjects quickly and easily. Easy content dissemination and anonymity on social media platforms can increase the published harmful content. Different information types can intentionally or unintentionally harm (GIACHANOU; ROSSO, 2020), including misinformation, disinformation, and mal-information. *Misinformation* (ASWANI; KAR; ILAVARASAN, 2019; KAR; ASWANI, 2021), often defined as satirical, is incorrect or fictional information created and spread, disregarding the proper intention. *Disinformation* (NASIR; KHAN; VARLAMIS, 2021), e.g. fake news is deliberately created to mislead the target users. *Mal-information* (DAVIDSON et al., 2017; GIACHANOU; ROSSO, 2020), e.g., hate speech is created to incite or cause harm. In this survey, we particularly investigate the hate speech detection task.

Hate speech is a challenging problem that demonstrates a clear intention to incite harm or promote hatred against others. This issue is considered a worldwide problem faced by many countries and organisations. With the growth of online social media, millions of users can spread much information every second, and the problem has become quite significant. There is a general understanding that when a person feels physically safe, the person's speech tends to be more aggressive (WATANABE; BOUAZIZI; OHTSUKI, 2018). Moreover, there is a real movement from hate groups to recruit people to create and diffuse hate speech messages (VIGNA et al., 2017).

The easy spread of hate speech on online platforms is a serious concern for our society, considering that the dissemination of hate speech can cause potential harm to individual victims and society, e.g., raising hostility between groups (MIŠKOLCI; KOVÁČOVÁ; RIGOVÁ, 2020; TEH; CHENG; CHEE, 2018). Particularly, repetitive exposure to hate speech can lead to desensitisation to this form of violence, thus lowering the victims' evaluations and increasing the bias against the target groups (MATHEW et al., 2019).

Social media platforms, such as Facebook, Twitter, and YouTube, have claimed they intend to solve this problem, which they present in policies on hate behaviour and attempts to combat hate speech (FACEBOOK, 2020; YOUTUBE, 2020; TWITTER, 2020). Much of this content moderation currently requires manual review of questionable documents (WASEEM; HOVY, 2016). However, the speed with which such messages are transmitted (shared) makes manual control over message content labour-intensive, time-consuming, expensive, and not scalable (ZHANG; ROBINSON; TEPPER, 2018; CAO; LEE; HOANG, 2020).

Furthermore, the hate speech detection task suffers from several weaknesses related to specific nuances of this subject and the complexity of this classification task (POLETTO et al., 2020). A relevant issue consists of clearly defining hate speech to understand the problem better and avoid strong subjective interpretations. As we will present in this survey, several disciplines have different definitions for the term "hate speech", which are complementary.

All the listed issues and limitations of the manual approaches have motivated considerable research. This survey also aims to provide an overview of better aspects of the problem, such as its definition, different features used in this problem, datasets, and methods. Furthermore, we highlight challenges and draw future work directions, obtaining a theoretical starting ground for new scientists on the topic.

Understanding the better aspects of hate speech detection is relevant to dealing with this issue. As a general basis for this area, we found some surveys proposed in this field

exploring different questions. In (SCHMIDT; WIEGAND, 2017) and (FORTUNA; NUNES, 2018), the researchers also survey critical tasks employed for hate speech detection. Nevertheless, it is relevant to note that this field has received increasing attention from the scientific community, and different resources included in the present survey had not been released when these surveys were published or at least when the researchers performed the search. Other works have focused on survey-specific characteristics of hate speech detection, such as multilingual corpus (AL-HASSAN; AL-DOSSARI, 2019), annotated corpora (POLETTO et al., 2020), and hate speech on the social media platform Twitter (AYO et al., 2020).

This contribution aims to complement these works and present a critical analysis of theoretical aspects and practical resources since this field has constantly grown. (i) We overview a general methodology for hate speech detection on social media, focusing on textual data. (ii) Besides, we present a comprehensive overview of recent resources from different social media and languages, such as the datasets, features used, and algorithms. (iii) We describe the advantages and limitations of several feature extraction techniques currently used in the literature. (iv) We point out different open challenges and opportunities in this field.

This paper is organised as follows: We first present an analysis of different definitions for the term “hate speech" based on several sources; Then, we explain the methodology used to select the works for this review; Next, we discuss a general methodology for hate speech detection; Then, an overview of the related datasets; After, we summarise several feature extraction approaches and present the advantages and limitations of the features explored; Then, we discuss several classification methods used in the literature; Furthermore, we present different challenges highlighted in the literature and opportunities in this field; finally, we conclude this survey with the final remarks.

2.2 WHAT IS HATE SPEECH?

Hate speech is a complex phenomenon, and detecting whether a text contains hate speech is not a trivial task, even for humans. Therefore, a precise definition of hate speech is crucial to automatically distinguish hate speech from other content (ROSS et al., 2016). We have seen an increasing number of studies that have addressed hate speech detection with different definitions of the term. This is probably because of the fog limits between hate speech and appropriate freedom of expression (MACAVANEY et al., 2019).

Thus, we have decided to analyse different sources' definitions, considering the wide range

of origins. We have analysed the description of hate speech presented by social media in their 'terms and conditions' contracts (Twitter, Facebook, YouTube) because hate speech often occurs on those platforms and some related studies, including the perspective of the scientific community. Since (COHEN-ALMAGOR, 2013) proposed one popular definition in the communication literature, (FORTUNA; NUNES, 2018) analysed several sources and considered distinct aspects, and (DAVIDSON et al., 2017) annotated a dataset used in several works. Thus, we will be considering those three aspects in our work.

1. Facebook: "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We define attack as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation." (FACEBOOK, 2020)

2. Twitter: "Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories." (TWITTER, 2020)

3. YouTube: "Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status." (YOUTUBE, 2020)

4. Cohen-Almagor: "Hate speech is defined as a bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics." (COHEN-ALMAGOR, 2013)

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

6. Davidson et al.: “Language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." (DAVIDSON et al., 2017)

In some aspects, these definitions can be considered similar. A common theme is that hate speech is used against a specific targeted group or group members. Besides, it has been seen by different sources as an attack or incitement to violence. (DAVIDSON et al., 2017) defined it as a language that is intended to be abusive, derogatory, humiliating, or insulting. While (COHEN-ALMAGOR, 2013) considers hostile and malicious speech based on innate characteristics. In general, these definitions have complementary nuances to each other. In particular, (FORTUNA; NUNES, 2018) specifically considers that hate speech can occur even in subtle forms. The authors argue that subtle forms of discrimination can use humour to reinforce stereotypes and racial discrimination, causing adverse effects for some people.

Considering these definitions, we can point out four main characteristics of hate speech described: (1) promotes attack or incites violence; (2) used against a specific target group or members of the group based on any characteristics such as gender, race, sexual orientation, religion, ethnicity or other aspects; (3) may or may not use 'abusive language' and derogatory terms; (4) can occur in subtle forms, for example, subtle metaphors '*expecting gender equality is the same as genocide*', this example of hate tweet does not contain explicit hateful lexical (ZHANG; LUO, 2019).

2.3 RESEARCH METHODOLOGY

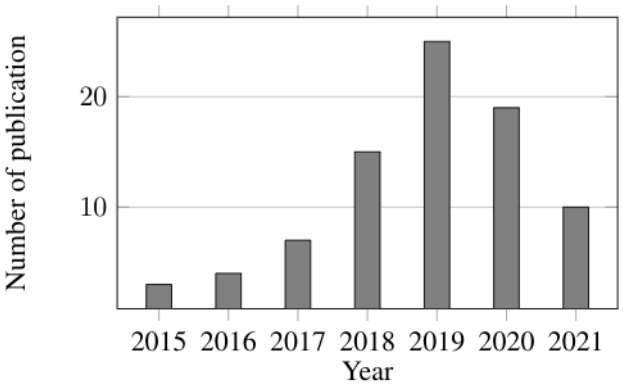
We have surveyed to understand hate speech detection on social media better, focusing on textual data. Our goal is to investigate the most recent studies developed in this field. To limit this research's scope, we have decided to restrict our search to documents published starting in 2015. The reason for this decision is the fact that in (FORTUNA; NUNES, 2018), it was shown that before 2014, this theme received little attention in computer science and engineering research, which is highlighted by the fact that many resources had not been released when previous surveys were published (POLETTO et al., 2020).

We searched the documents in different sources, such as ACM digital library, IEEE, Elsevier, and Springer. The keywords selected were “hate speech detection", “hate speech classification", and also considered the search for “Abusive language", considering that abusive language is

a sub-category of hate speech. The keywords selected were searched in the publication title, abstract and keywords. We also used Google Scholar to search for references that cited the original work. We check on these sets and search for the keyword “hate speech detection" on the titles of the documents. Several entries appeared as results of more than one search string.

We have focused on the field of computer science and engineering research. Also, we only included papers with at least four pages and peer-reviewed scientific resources. Furthermore, we restricted the works as automatic hate speech detection to the only ones performed on social media platforms, particularly from textual data. The text published on these platforms has specific characteristics (e.g., a limited number of characters, URLs, emojis, mentions, and so on). Thus, we have selected a total of 83 papers in the search period. Figure 2 presents the distribution of papers over the selected time interval.

Figure 2 – Number of publications towards the years for hate speech detection from January, 1st 2015 to July, 31st 2021.



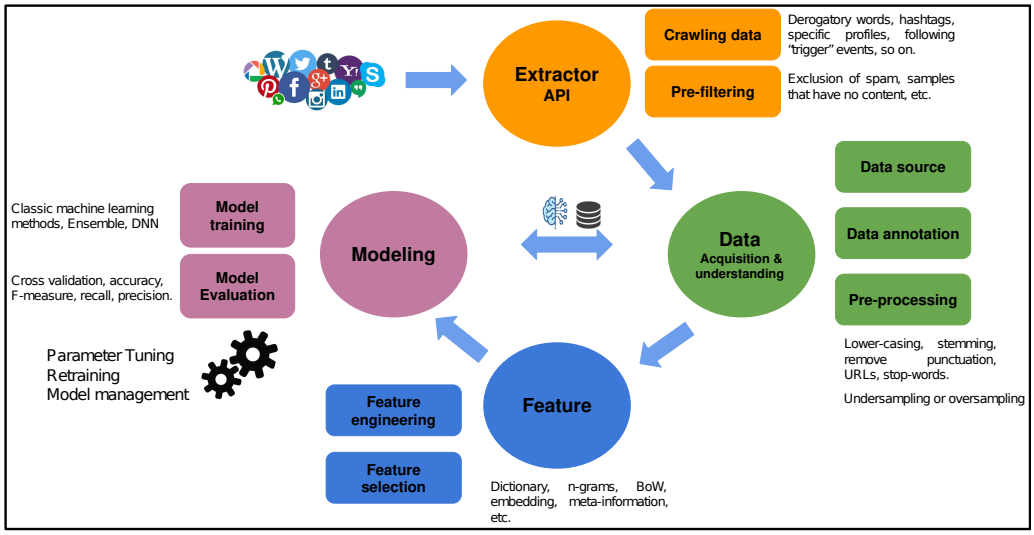
Source: Prepared by the author.

It is quite clear the scientific community's recent efforts towards dealing with automatic hate speech detection relate to the processing and analysis of textual data. The following sections present several automatic hate speech detection techniques that explore this aspect.

2.4 AUTOMATIC HATE SPEECH DETECTION

The automatic hate speech detection process includes tasks such as data collection and processing, feature extraction, detection, and classification. We analyse and summarise the main tasks typically employed in automatic hate speech detection on social media platforms. Figure 3 presents an overview of the architecture for hate speech detection.

Figure 3 – Overview of architecture for hate speech detection on social media platforms.



Source: Prepared by the author.

Social media platforms provide a wide variety of information that can be collected using the programming libraries known as Application Programming Interface (API). The researchers have adopted different strategies to crawl data related to hate speech, such as derogatory words, common slurs, hashtags, specific profiles, following "trigger" events, and so on (DAVIDSON et al., 2017; WASEEM; HOVY, 2016; FOUNTA et al., 2018; FORTUNA et al., 2019; BURNAP; WILLIAMS, 2015). Moreover, several works have used pre-filtering to exclude spam, samples with no content, and samples not in English (FOUNTA et al., 2018; PRATIWI; BUDI; ALFINA, 2018). According to (FOUNTA et al., 2018), abusive tweets are relatively rare, and the percentage can range between 0.1% and 3% of the samples collected.

The methodology employed to collect and annotate the dataset should be carefully chosen to avoid bias in the dataset (WIEGAND; RUPPENHOFER; KLEINBAUER, 2019). The annotation task in different studies used CrowdFlower (CF) workers (CHATZAKOU et al., 2017; DAVIDSON et al., 2017; WASEEM, 2016; FOUNTA et al., 2018; KUMAR et al., 2019), but this approach can be expensive. The authors in (CHATZAKOU et al., 2017; FOUNTA et al., 2018) used a default payment scheme for batch (each with 10 tweets) to minimise costs without compromising the annotation quality. Moreover, the authors also performed the annotation task (WASEEM; HOVY, 2016) or used non-experts and experts annotated (BASILE et al., 2019; FORTUNA et al., 2019; WASEEM, 2016). Another approach employed is active learning annotation (CHARITIDIS et al., 2020) for further annotation and dataset expansion. Several authors (ALSAFARI; SADAOUI; MOUHOUB, 2020; GOLBECK et al., 2017; MOSSIE; WANG, 2020; WASEEM; HOVY, 2016) developed

a coding guideline to help human annotators classify the content due to the subjectivity of the human interpretation of hate speech. The following section presents a further overview of hate speech detection datasets.

In the context of social media platforms, the text used frequently has specific characteristics, such as abbreviations, incorrect spelling, slang, acronyms, URLs, hashtags, emojis, mentions, and so on. The unstructured text and, at times, the informal language can introduce noise in the classification task (NASEEM; RAZZAK; EKLUND, 2020). Several pre-processing methods are explored before the feature extraction task in order to reduce noise in the dataset, such as lower-casing of words, stemming, removing punctuation, URLs, stop-words, replacing emoticons and emojis, elongated characters (DORRIS et al., 2020; PRATIWI; BUDI; ALFINA, 2018; ZHANG; ROBINSON; TEPPER, 2018; SOHN; LEE, 2019; WATANABE; BOUAZIZI; OHTSUKI, 2018; NUGROHO et al., 2019). (NASEEM; RAZZAK; EKLUND, 2020) evaluated twelve different pre-processing techniques and the combination of them in three datasets of hate speech (proposed in (DAVIDSON et al., 2017; GOLBECK et al., 2017; WASEEM; HOVY, 2016)). The authors concluded that the lemmatisation and lower casing of words presented a high performance in most cases. On the other hand, removing punctuation and URLs, user mentions, and Hashtags symbols presented a low performance in most cases. Moreover, some studies focused on techniques to deal with the class imbalance problem, such as oversampling and undersampling. The oversampling technique is applied in the training data to increase the minority class (CHATZAKOU et al., 2017; ELISABETH; BUDI; IBROHIM, 2020), while the undersampling technique reduces the majority class (MIOK et al., 2019). However, most of the works did not deal with class imbalance.

Feature extraction is an important task in text analysis. Several approaches are explored in hate speech detection and related subjects. Among these, dictionary or lexical resources (NOBATA et al., 2016; GITARI et al., 2015; BURNAP; WILLIAMS, 2015; MATHEW et al., 2019; TEH; CHENG; CHEE, 2018), distance metric (NANDHINI; SHEEBA, 2015; MOSSIE; WANG, 2020), bag-of-word (BURNAP; WILLIAMS, 2016; SENARATH; PUROHIT, 2020; WASEEM; THORNE; BINGEL, 2018), n-grams (CORAZZA et al., 2020; MOSSIE; WANG, 2020; WULCZYN; THAIN; DIXON, 2017; SENARATH; PUROHIT, 2020; SANTOSH; ARAVIND, 2019), term frequency (ALMATARNEH et al., 2019; MOSSIE; WANG, 2020; SALMINEN et al., 2020), text embedding and deep learning (CAO; LEE; HOANG, 2020; MIOK et al., 2019; SENARATH; PUROHIT, 2020; ZIMMERMAN; KRUSCHWITZ; FOX, 2018), meta-information (FOUNTA et al., 2019; PITSILIS; RAMAMPIARO; LANGSETH, 2018; WASEEM; HOVY, 2016), and so on. Different studies addressed hate speech detection on so-

cial media present better results when combining a set of features (SALMINEN et al., 2020; SENARATH; PUROHIT, 2020). In this study, we highlight several methods used to feature extraction and their advantages and limitations.

Although the feature engineering process's effective for text representation, the feature space can present a high dimensionality. However, in the context of hate speech detection, few studies (ROBINSON; ZHANG; TEPPER, 2018; ZHANG; ROBINSON; TEPPER, 2018) have evaluated the feature selection process's impact. The automatic feature selection algorithms can reduce the original feature space by 90% and improve machine learning algorithms' performance for hate speech detection (ROBINSON; ZHANG; TEPPER, 2018; ZHANG; ROBINSON; TEPPER, 2018).

Classic supervised machine learning methods have been explored for automated hate speech detection. Among these, Support Vector Machines (SVM) (BURNAP; WILLIAMS, 2015; SALMINEN et al., 2020), Logistic Regression (LR) (DAVIDSON et al., 2017; WASEEM; HOVY, 2016; KHAN; SHAHZAD; MALIK, 2021), Naive Bayes (NB) (SALMINEN et al., 2020; IBROHIM; BUDI, 2019), Random Forest (RF) (ALMATARNEH et al., 2019), C4.5 decision tree learning (WATANABE; BOUAZIZI; OHTSUKI, 2018). Although more expensive, ensemble approaches have presented robust results of the different classification tasks (BURNAP; WILLIAMS, 2015; MARKOV et al., 2021; NUGROHO et al., 2019; PASCHALIDES et al., 2020; ZIMMERMAN; KRUSCHWITZ; FOX, 2018). Another approach explored is the DNN, which has been used for feature extraction and classifiers' training. The most used approaches are CNN, LSTM, and GRU (PITSILIS; RAMAMPIARO; LANGSETH, 2018; DORRIS et al., 2020; ZHANG; LUO, 2019; RIZOS; HEMKER; SCHULLER, 2019; SANTOSH; ARAVIND, 2019; AL-MAKHADMEH; TOLBA, 2019; CAO; LEE; HOANG, 2020; ALSAFARI; SADAOUI; MOUHOUB, 2020; MOSSIE; WANG, 2020; MARPAUNG; RISMALA; NURRAHMI, 2021). This work discusses several methods used for hate speech detection on social media platforms in the following section.

The following sections present an overview of the datasets, feature extraction techniques, and classification methods employed for automatic hate speech detection.

2.5 DATASETS FOR HATE SPEECH CLASSIFICATION

Representative publicly available datasets are essential for developing automatic hate speech detection approaches. However, collecting and annotating data in the context of hateful messages is challenging, especially, as previously mentioned, no universal definition is adopted. The most common way of labelling this type of content is using social media platforms' definitions.

Besides, the number of hate speech texts compared to non-hate on social media platforms is significantly smaller. The studies adopted some strategies to collect the dataset, such as using terms and phrases related to hate content from dictionaries like HateBase, specific profiles, hashtags and keywords (WASEEM; HOVY, 2016; FOUNTA et al., 2018; DAVIDSON et al., 2017; FORTUNA et al., 2019).

Table 18 summarises the main information from several datasets proposed in the literature. These datasets vary considerably in their labels, number of instances, characteristics of hate speech, etc. The most popular data source is Twitter, which has attracted a significant part of the research due to the increasingly available data and free APIs (DAVIDSON et al., 2017; WASEEM; HOVY, 2016; WATANABE; BOUAZIZI; OHTSUKI, 2018). English has been the most popular language analysed, but we can also find works exploring other languages, such as Arabic, Spanish, Indonesian, Portuguese, German, French, and Greek.

Overall, the publicly available datasets for hate speech detection in different languages and social media platforms are scarce, with few studies publishing their datasets. In most cases, the datasets are not available for external researchers, such as a large annotated dataset of abusive language detection from the 'Yahoo! Finance and News' (NOBATA et al., 2016); Facebook, Italian language corpus of hate speech (VIGNA et al., 2017), Amharic language corpus for hate speech detection approach to vulnerable community identification (MOSSIE; WANG, 2020). (POLETTO et al., 2020) performed a further analysis in several datasets for hate speech detection, including methodology, topical focus, language, and other factors. The results presented different data sources and highlighted some issues and improvements.

Table 1 – Summary of datasets for hate speech classification.

Dataset	Year	Distribution	Number of instances	Labels (%)	Annotators	Origin source	Language
WH (WASEEM; HOVY, 2016)	2016	GitHub repository	16,914	sexism (20%), racism (12%) and none (68%)	authors	Twitter	English
WS (WASEEM, 2016)	2016	GitHub repository	6,909	sexism (13%), racism (11%), both (1%), neither(84%)	3 or more	Twitter	English
DV (DAVIDSON et al., 2017)	2017	GitHub repository	24,802	hate (5%) offensive(76%) neither(17%)	3 or more	Twitter	English
GB (GOLBECK et al., 2017)	2017	Need request access	35k	Harassing (15.7%) Non-Harassing (74.3%)	2-3	Twitter	English
FT (FOUNTA et al., 2018)	2018	Dataverse	80k	hateful(7.5%), abusive(11%), spam(22.5%), normal (59%)	5-20	Twitter	English
PR (PRATIWI; BUDI; ALFINA, 2018)	2018	GitHub repository	835	Hate speech 34.24% not hate speech 65.75%	3	Instagram	Indonesian
SE (BASILE et al., 2019)	2019	GitHub repository	19,600 (6,600 - Spanish; 13,000 - English)	Hate (43%)/Not Hate (57%)	3	Twitter	English, Spanish
FO (FORTUNA et al., 2019)	2019	GitHub repository	5,668	hate speech (22%), not hate speech(78%)	3	Twitter	Portuguese
IB (IBROHIM; BUDI, 2019)	2019	GitHub repository	13,169	hate speech (42.2%) not hate speech (57.8%)	3	Twitter	Indonesian
YT (PHILIPP; ROMAN, 2019)	2019	Zenodo platform	1k	hate speech (13.8%) not hate speech (86.2%)	–	YouTube	English
KU (KUMAR et al., 2019)	2019	Need request access	18k tweets (T) 21k Facebook (F)	Overtly Aggressive (T - 6.0% F 27.5%), Covertly Aggressive (T - 44.1% F 29.9%), Non-aggressive (T - 49.9% F 42.6%)	3	Facebook and Twitter	Code-mixed (Hindi-English)
AL (ALSAFARI; SADAOUI; MOUHOUB, 2020)	2020	GitHub repository	5,360	Hate 26.65% Offensive 8.18% Clean 65,17%	3	Twitter	Arabic
CH (CHARITIDIS et al., 2020)	2020	Zenodo platform	EN 92,022 DE 43,735 ES 37,688 FR 29,109 GR 61,481	Hate speech EN p 7.78% n 92.22% DE p 3.9% n 96.1% ES p 2.64 % n 97.36% FR p 9.3 % n 90.7% GR p 1.86% n 98.14%	1	Twitter	English, German, Spanish, French and Greek

Source: Prepared by the author.

2.6 FEATURE EXTRACTION APPROACHES

An essential task in text analysis is the meaningful feature extraction from data. The approaches selected often have a significant impact on the data analysis itself. However, extracting insights and patterns from a text can be challenging, especially in the context of social media, where there is the issue of unstructured text. Table 2 presents the advantages and limitations of the most widespread techniques for feature extraction used in the context of hate speech detection and related subjects. This section analyses features used in hate speech detection and related subjects.

Table 2 – Overview of the features used in the context of hate speech detection. n is the number of words/-tokens/strings in the document.

Method	**Advantages**	**Limitations**	**Average vector size**
Dictionaries or lexical resources	It is a simple method and effective to detect hate speech with derogatory terms.	The dependency of hateful keywords	$\sim 5 - 250$ words
Distance Metric	It captures the number of edit operations and semantic similarity.	It is few explored in the context of hate speech, and it is used as a complementary metric.	n strings
Bag-of-words (BoW)	The corpus are collected from the training data.	It ignores word sequences and their semantic and syntactic content, which may lead to misclassification of words used in various contexts.	n different words in the sentence
N-grams	Overcome the limitation of BoW. The subclass POS captures information about the syntactic structure of the text.	It can suffer from a high level of distance between related words. Besides, the POS technique can promote confusion between the classes due to the abundance of similar patterns.	items sequences (with n in range between $1 - 5$)
Term frequency	It provides good classification performance for hate speech detection, simple method.	It did not help the model generalise well across different dataset domains.	n tokens
Template-based Strategy	Structures predefined.	It can generate false positives. Besides, it is often useful to the specific context.	Template length
Typed Dependencies	It extracts a subset of dependency relationship labels.	It is often used as a complementary metric and can increase the number of false-negative instances.	number of sentences extracted
Text embedding and Deep learning approaches	Pre-trained word embeddings have proved useful for abusive text classification. Besides, it required fewer training samples to obtain a good performance. The DNN technique learns abstract feature representations for hate speech detection; It can be used for feature extraction as well as a machine learning classifier.	A problem faced with pre-trained word embedding is out-of-vocabulary (OOV) words. Moreover, a limitation of DNN techniques is the high cost of computational and explainability.	$25 - 300$ dimensions
Sentiment analysis	Usually, negative sentiment belongs to the hate speech message, besides several automatic tools available.	It needs to use other techniques to improve results.	Number of sentiment polarity (usually 'positive', 'negative', 'neutral', and 'compound')
Meta-information	It provides additional information about the context of the message.	It is scarce and often not readily available for external researchers; it might introduce bias in the model.	Amount of metadata

Source: Prepared by the author.

2.6.1 Dictionaries or lexical resources

A dictionary is a relevant approach used in natural language processing (NLP) based on keywords. This strategy lists potential keywords and counts the number of occurrences in the text or context.

These frequencies can be used as features or to compute scores. For hate speech context, different dictionaries have been available:

- Hatebase is a multilingual dataset of derogatory terms with data across 95+ languages and 175+ countries. This resource offers constant updates in the terminology and a broad vocabulary (<https://hatebase.org/>);
- Dictionary of general swear words and insults in English (<https://www.noswearing.com/>);
- Urban dictionary of colloquial language and slang words in English (<https://www.urbandictionary.com>).

Previous works used this approach, in general, considering negative or derogatory words (NOBATA et al., 2016; GITARI et al., 2015; BURNAP; WILLIAMS, 2015; MATHEW et al., 2019; TEH; CHENG; CHEE, 2018). (GITARI et al., 2015) built a lexicon of hate-related verbs which encourage violent acts (such as to discriminate, loot, riot, beat, kill, and evict). (MATHEW et al., 2019) created a lexicon with 45 hate words selected from the Hatebase and Urban Dictionary for further analysis of hateful and non-hateful users on Gab. (TEH; CHENG; CHEE, 2018) constructed a lexical of profane words frequently used in different types of hate speech from comments on YouTube, which showed that 35% of profane words are related to sexual orientation, based on 500 comments. (BURNAP; WILLIAMS, 2016) focused on specialised lists towards particular subtypes of hate, such as LGBT slang terms, ethnic slurs, and negative connotations against disabled people. (HAYATY; ADI; HARTANTO, 2020) focused on local languages in Indonesia for hate speech detection and created a dictionary of abusive words containing of 250 terms.

Despite their general effectiveness, a limitation of this approach is the dependency on hateful keywords (MACAVANEY et al., 2019). Thus, lexical features can be employed as an additional step of feature extraction (SCHMIDT; WIEGAND, 2017).

2.6.2 Distance Metric

The presence of noise and conjugations often makes it difficult to perform automatic detection of hateful content. Once derogatory words are intentionally used in text messages (NOBATA et al., 2016), it is possible to identify such words with character substitution such as '*ni99er*', '*@ss*', '*sh1t*' which can make the whole process even more challenging for automatic detection. Approaches to compute the minimum number of edit operations of individual characters like Levenshtein distance can also be used for this end (NANDHINI; SHEEBA, 2015).

There is no lexicon for hate speech detection in some languages, such as the Amharic language. Thus, one approach employed was translating the text into English using the Google translator tool. In this approach, the researchers used the cosine distance to evaluate the semantic similarity between each input word and the corresponding vectors in the model (MOSSIE; WANG, 2020).

2.6.3 Bag-of-Words (BoW)

Bag-of-Words (BoW) is another technique used to detect hateful speech (BURNAP; WILLIAMS, 2016; NOBATA et al., 2016; SENARATH; PUROHIT, 2020; WASEEM; THORNE; BINGEL, 2018). Similarly to the dictionary, this technique uses keywords, the main difference being that it creates a corpus from the collected training data, while the dictionary uses predefined words. After the data collection stage, word frequencies are used as a feature for training a classifier. A limitation of this approach is ignoring word sequence and its semantic and syntactic content. Hence, it may lead to the mistaken classification of words used in various contexts. Another technique that can be adopted to overcome this limitation is n-grams.

A statistical analysis conducted using BoW with all typed dependencies and with only hateful and derogatory terms to investigate its influence in the classification task is presented in (BURNAP; WILLIAMS, 2015), which follows the assumption that BoW can confuse the classification task when the same word is frequently in non-hateful and hateful scenarios. The study showed that using only hateful and derogatory terms can potentially increase the number of false negatives because the hateful content does not necessarily use derogatory or hateful terms.

2.6.4 N-grams

The n-grams is one of the most used techniques in automatic hate speech detection and related tasks (CORAZZA et al., 2020; MOSSIE; WANG, 2020; WULCZYN; THAIN; DIXON, 2017; SENARATH; PUROHIT, 2020; SANTOSH; ARAVIND, 2019; CHAKRABORTY; SEDDIQUI, 2019). It combines a sequence of n adjacent items into a list with size N, where the items can be words (most common), syllables, or characters (FORTUNA; NUNES, 2018). However, for the problem of hate speech detection, 'character n-grams' provided better performance than 'word n-grams', because it captures the changes in the words associated with hate (WASEEM; HOVY, 2016; VIGNA et al., 2017; UNSVÅG; GAMBÄCK, 2018).

Its main disadvantage is that it suffers from a high level of distance between related words (BURNAP; WILLIAMS, 2016), which is closely associated with the selection of the n value. Since n-grams may not be able to capture long-range dependencies between words, for example: '*Jews are lower class pigs*', the words '*Jews*' and '*pigs*', similarities would not be connected using only n-grams, depending on the n selected (NOBATA et al., 2016).

These features are often combined with other features to improve the hate speech classification. For instance, in (WATANABE; BOUAZIZI; OHTSUKI, 2018), the authors explored different features for hate speech detection, such as the most common word unigrams, pattern features, sentimental, and semantic features. They believed that unigram features could help identify explicit forms of hate speech. Overall, unigram features presented high accuracy, but all features combined performed better.

Part-of-speech (POS) is a subclass of the n-gram approach that detects the role of the word in the context of the sentence, which tags capture the syntactic function of the word, for instance, personal pronoun (PRP), verbs (VB), nouns (NN), adjectives (JJ). These approaches have been used for hate speech detection to capture information about the syntactic structure of the text to extract frequencies from unigrams, bigrams, and trigrams (DAVIDSON et al., 2017).

Furthermore, it was also used to collect unigrams with a specific syntactic function (e.g. noun, verb, adjective or adverb) from the training set to investigate occurrences in hateful and offensive tweets (WATANABE; BOUAZIZI; OHTSUKI, 2018). However, POS, when used as a feature, can promote confusion between the classes due to the abundance of similar patterns (BURNAP; WILLIAMS, 2015; FORTUNA; NUNES, 2018).

2.6.5 Term frequency

The word or term frequency indicates the relevance of the word in the document that contains it. The most common types of word frequency are Term Frequency (TF), Term Relative Frequency (TFR), Inverse Document Frequency (IDF), and Term Frequency-Inverse Document Frequency (TF-IDF) (LIU et al., 2019a). In (PLAZA-DEL-ARCO et al., 2020) used TF weighting to represent unigrams and bigrams as vectors of numerical features to misogyny and xenophobia detected in Spanish tweets.

Several works used TF-IDF weighting features for hate speech detection (ALMATARNEH et al., 2019; ELISABETH; BUDI; IBROHIM, 2020; MOSSIE; WANG, 2020; SALMINEN et al., 2020). The TF-IDF provided good classification performance for hate speech detection with the same dataset to train and test the models. However, it did not help the model generalise well when used across different dataset domains (SENARATH; PUROHIT, 2020).

2.6.6 Typed Dependencies

Typed dependencies have been widely used for hate speech detection (BURNAP; WILLIAMS, 2015; BURNAP; WILLIAMS, 2016; ALORAINY et al., 2019). The probabilistic parse trees, provided by Stanford Typed Dependency Parser (MARNEFFE; MANNING, 2008), can be used to extract a subset of dependency relationship labels and provide a description of the grammatical relationships in a sentence (ALORAINY et al., 2019). The introduction of typed dependency features for hate speech detection can reduce the false positive rate, but this can lead to an increase in false-negative instances. This approach performed better when combined with other features (BURNAP; WILLIAMS, 2016).

2.6.7 Template Based Strategy

In this strategy, the main idea is to build a corpus of structured sentences. (MONDAL; SILVA; BENEVENUTO, 2017) proposed the follow sentence structure "*I* <*intensity*> <*userintent* ><*hatetarget* >", to search hate speech post. Thus, they additionally designed two templates, focusing on exploring hate against groups of people. The first was simply "<one word> people" for scenarios when hate was directed towards a group, and the second template used words collected on Hatebase for <hate target> tokens.

2.6.8 Text embedding and Deep learning approaches

The embedding technique is aimed at training a model to provide a vector representation of sentence/word, which captures the semantic and the syntactic relationship between the words (INDURTHI et al., 2019). Word embedding methods have improved prediction accuracy for hate speech classification (LIU et al., 2019a), which can be illustrated by several studies using pre-trained word embedding approaches, such as Word2vec, GloVe, FastText, ELMo, LASER, XLM, BETO (CAO; LEE; HOANG, 2020; VITIUGIN; SENARATH; PUROHIT, 2021; MIOK et al., 2019; ARCO et al., 2021; SENARATH; PUROHIT, 2020; SREELAKSHMI; PREMJITH; SOMAN, 2020). The pre-trained word embedding had been proven effective for abusive text classification. Besides, it required fewer training samples to obtain a good performance (FOUNTA et al., 2019). Another approach is sentence embedding which represents sentences as vectors. (MIOK et al., 2019) proposed a model for hate speech detection in three datasets (from Twitter and YouTube) using word and sentence embedding. The approach used the LSTM model with Monte Carlo dropout obtained better performance by using pre-trained sentence embedding than word embedding and state-of-the-art features.

However, an issue faced with pre-trained word embedding is out-of-vocabulary (OOV) words. Particularly, present on social media data because of its colloquial nature, users often perform intentional obfuscation of words, which can be mitigated by performing pre-processing before feature extraction for noise reduction (ZHANG; LUO, 2019). (CORAZZA et al., 2020) investigated the impact of word embedding and emoji embedding on the specific domain and compared it with pre-trained embeddings, such as FastText. Specific embedding improved the results but needed a large amount of data. On the other hand, pre-trained embedding using binary models could mitigate the issue of OOV word, since this approach provided sub-words information.

Deep neural network (DNN) techniques have been recently explored to learn abstract feature representations for hate speech detection. The most popular approaches are the Convolutional Neural Network (CNN) and the Long Short-Term Memory network (LSTM). In the context of hate speech classification, CNN was applied as a feature extractor, and LSTM was used for modelling sequences of word or character dependencies (BOUAZIZI; NIIDA; OHTSUKI, 2021; KAPIL; EKBAL, 2020; SANTOSH; ARAVIND, 2019; SAJJAD et al., 2019; ZHANG; ROBINSON; TEPPER, 2018).

Even though very expensive, another approach explored was deep learning ensembles that

used CNN for feature extraction (ZIMMERMAN; KRUSCHWITZ; FOX, 2018; ZHOU et al., 2020). These techniques are robust and improve the results of the different classification tasks. In a study conducted in seven datasets from Twitter in the English language (ZHANG; LUO, 2019), CNN showed more effectiveness for specific types of hate (racism and sexism) than polarised data (hate and non-hate).

Other approaches have investigated the language model pre-training BERT (Bidirectional Encoder Representation from Transformers) (CALABRESE et al., 2021; WICH et al., 2021). BERT was designed to pre-train deep bidirectional representation. In (HENDRAWAN; ADIWIJAYA; FARABY, 2020), analysed the BiLSTM and the BiLSTM with BERT multilingual trained with Wikipedia from 104 languages. However, the BiLSTM with BERT was less effective than the BiLSTM and the Random Forest Decision Tree.

2.6.9 Sentiment analysis

Sentiment analysis is often considered synonymous with 'opinion mining', a field of study that aims to analyse a person's feelings, opinions, and emotions towards 'elements' (SERRANO-GUERRERO et al., 2015). The 'elements' in this context can represent individuals, events, services, products, and topics. Sentiment analysis and hate speech are related, and often negative sentiments are associated with hate speech messages (SCHMIDT; WIEGAND, 2017).

Several works have used the sentiment as a feature for hate speech detection (CAO; LEE; HOANG, 2020; CORAZZA et al., 2020; GITARI et al., 2015; RODRÍGUEZ; ARGUETA; CHEN, 2019). Features based on emotions and sentiments are relevant approaches and can improve classification tasks on hate speech detection (CORAZZA et al., 2020; MARKOV et al., 2021). However, supervised methods required labels for sentiment classification and hate speech datasets often did not have this information. Different automatic tools were explored to overcome this limitation of supervised methods for sentiment analysis, such as JAMMIN, an emotion analysis tool, and VADER (Valence Aware Dictionary for sEntiment Reasoning), a sentiment analysis tool (CAO; LEE; HOANG, 2020; RODRÍGUEZ; ARGUETA; CHEN, 2019).

Although related, it is arguable that hate speech detection is a different task requiring more sophisticated techniques (WATANABE; BOUAZIZI; OHTSUKI, 2018). In sentiment analysis, the presence of positive/negative words or expressions can be considered helpful in this process. The presence of negative words or expressions, even in such sentences using the word 'hate', depending on the context, does not make them related to hate speech. Thus, this approach for

feature extraction is usually used with other techniques to improve results (CAO; LEE; HOANG, 2020; CORAZZA et al., 2020; WATANABE; BOUAZIZI; OHTSUKI, 2018).

2.6.10 Meta-information

Additional information from social media can help better understand the characteristics of the post-context and provide valuable data for hate speech detection. Social media platforms offer a wide variety of information that can be collected through APIs, such as user gender, demographics data, timestamp, user profiles, and network structures (AYO et al., 2020; DESOUZA; DA-COSTA-ABREU, 2020).

Background information about the user can improve the predictably of hateful messages since hateful users are densely connected (RIBEIRO et al., 2018). In a study about the impact of information like user gender and demographic information in tweets (WASEEM; HOVY, 2016), these features brought slight improvement, but this could be because of the lack of coverage. Information about user gender was also explored in (UNSVÅG; GAMBÄCK, 2018), which used a similar approach performed in (WASEEM; HOVY, 2016), to identify the user gender based on username or profile names as well as the user description in messages. However, a limitation of this approach is names used for both females and males. Another approach investigated the metadata based on text content to analyse specific attributes in tweets, such as the number of hashtags and mentions of other users, emoticons in the tweet, words with only uppercase letters, URLs included, and frequency of punctuation marks (AL-MAKHADMEH; TOLBA, 2019; CHATZAKOU et al., 2017; FOUNTA et al., 2019; VIGNA et al., 2017).

Furthermore, another meta-information relevant is the user network, such as user friends and followers. These features are beneficial in classifying aggressive user behaviour (CHATZAKOU et al., 2017). Features about user behaviour are also useful for detecting racist and sexist messages (PITSILIS; RAMAMPIARO; LANGSETH, 2018). These features can help describe the user's tendency toward the class based on their tweet history, post content, and subsets of those tweets with labelled messages. This information is scarce and often not readily available for external research(CAO; LEE; HOANG, 2020; MACAVANEY et al., 2019). Since these data have sensitive information about users, publishing raises privacy issues. Moreover, user information can introduce bias in the model against particular users or groups (MACAVANEY et al., 2019).

2.6.11 Other techniques

Other features used in the classification task are based on Flesch-Kincaid Grade Level (**FKGL**) and Flesch Reading Ease (**FRE**) scores to measure the quality of a document (ŞAHI; KILIÇ; SAğLAM, 2018); **Pattern features** (WATANABE; BOUAZIZI; OHTSUKI, 2018); Latent Dirichlet Allocation (**LDA**), typically used for topic modelling. (CAO; LEE; HOANG, 2020) used LDA to determine the posts' topic distribution in each dataset, considering each post as a single document.

Texts extracted from social media platforms often contain URLs, punctuation, symbols, usernames, and tags such as '@', RT and $< >$. Some studies, before the feature extraction stage, have used **stemming** and removed special characters and **stop-words** (ZHANG; ROBINSON; TEPPER, 2018). However, using stemming, some words in the Indonesian language can be converted into words with different meanings, such as "*dadakan*" which means "*all of sudden*" to "*dada*" which means "*chest*". Besides, stop-word removal can reduce the information from the sentence (HENDRAWAN; ADIWIJAYA; FARABY, 2020).

Table 3 – Summary of studies for hate speech detection on social media

Ref.	Feature	Model	Social Media	dataset	Acc	AUC	R	P	F
(NOBATA et al., 2016)	Token and char N-grams (3-5), POS tags, word2vec, comment2vec, length of comment in tokens, number of punctuations, and so on.	Vowpal Wabbit's regression model	Yahoo	collected	–	–	–	–	Fin. 0.79 News 0.81
(CHATZAKOU et al., 2017)	Meta-information, Word embedding, sentiment, dictionary	RF	Twitter	collected	–	0.90	0.91	–	0.89
(VIGNA et al., 2017)	POS, sentiment, word2vec, char-lemma- and word- n-grams, repetition of n-grams char, punctuation	SVM, LSTM	Facebook	collected	0.80	–	∼0.79	∼0.83	∼0.78
(WULCZYN; THAIN; DIXON, 2017)	n-gram (word, char)	LR, MLP	Wikipedia	collected	–	0.96	–	–	–
(PITSILIS; RAMAMPIARO; LANGSETH, 2018)	User features	ensemble of LSTM	Twitter	WH	–	–	∼0.87	∼0.9	∼0.89
(WATANABE; BOUAZIZI; OHTSUKI, 2018)	sentiment, punctuation marks, all-capitalized words, POS, word unigram, pattern features	C4.5	Twitter	Crowdflower, DV, WH)	0.87	–	0.87	0.88	0.87
(ZHANG; LUO, 2019)	skipped CNN (sCNN), word2vec, CNN	CNN+GRU and CNN+sCNN	Twitter	WS (WS-S.amt, WS-S.exp, WS-S.gb, WS.pj), DV	–	–	–	–	0.83-0.94
(RIZOS; HEMKER; SCHULLER, 2019)	word2vec, GloVe, FastText, POS-tags	LSTM, CNN	Twitter	DV, WS, WH	–	–	DV 0.49	–	DV 0.74, WH 0.82, WS 0.83
(ALMATARNEH et al., 2019)	n-grams, TF-IDF and CountVectorizer	SVM, GNB, CNB, DT, K-NN, RF, and NN	Twitter	SE	–	–	–	–	EN 0.76 ES 0.77
(LIU et al., 2019a)	embedding; LDA	fuzzy ensemble	Twitter	(BURNAP; WILLIAMS, 2015)	0.93	–	–	–	–
(SANTOSH; ARAVIND, 2019)	char and word n-grams, negation words, punctuation marks	SVM, RF, Sub-word level LSTM, Hierarchical LSTM	Twitter	(BOHRA et al., 2018)	0.66	–	0.45	–	0.48
(ALMAKHADMEH; TOLBA, 2019)	sentiment, semantic,unigram and pattern features	ensemble deep learning	Twitter	collected	0.98	–	–	–	–
(SENARATH; PUROHIT, 2020)	BoW, tf-idf, n-grams, dictionary, FrameNet, word2vec	SVM	Twitter	DV, FT	DV 0.94, FT 0.94	DV 0.96, FT 0.78	DV 0.97 FT 0.90	–	DV 0.96, FT 0.70
(SALMINEN et al., 2020)	BoW, TF-IDF, GloVe, BERT, and all combined	LR, NB, SVM, XGBoost, and Neural Networks	YouTube, Reddit, Wikipedia, and Twitter	(SALMINEN et al., 2018), (ALMEREKHI et al., 2019), (WULCZYN; THAIN; DIXON, 2017), DV	–	–	–	–	D1 0.91, D2 0.77, D3 0.86, DV 0.98
(CAO; LEE; HOANG, 2020)	GloVe, word2vec, Paragram, sentiment and LDA.	LSTM, C-LSTM-Att	Twitter	(WH and WS combined), DV, FT, All combined	–	–	–	–	D1 0.78, DV 0.89, FT 0.79, D4 0.92
(ALSAFARI; SADAOUI; MOUHOUB, 2020)	Unigram, word and char n-grams, word embedding	NB, SVM, LR, CNN, LSTM, GRU	Twitter	collected	–	–	0.87	0.86	0.87
(MOSSIE; WANG, 2020)	word n-grams, TF-IDF and word2vec	GBT, RF, LSTM, GRU	Facebook	collected	0.92	0.97	–	–	–

Source: Prepared by the author.

2.7 CLASSIFICATION METHODS

Automated hate speech detection on social media is a complex problem. Several approaches have been explored to deal with this problem, such as classic supervised machine learning methods, ensemble, and DNN techniques. Table 3 summarises several studies with the results for the best model for each work.

Classic supervised machine learning methods have been explored for automated hate speech detection. (NOBATA et al., 2016) developed a machine learning method to detect hate speech from the 'Yahoo! Finance and News' dataset that outperformed a deep learning approach. The decision tree classifiers were also explored for hate speech detection and related subjects (CHATZAKOU et al., 2017; WATANABE; BOUAZIZI; OHTSUKI, 2018). In the study (CHATZAKOU et al., 2017), the Random Forest classifier presented a better performance in classifying bullying and aggressive behaviour from a Twitter dataset than other tree classifiers experimented (J48, LADTree, LMT, NBTree, and Functional Tree), with 90% AUC (Area Under Curve). The authors in (WATANABE; BOUAZIZI; OHTSUKI, 2018) also analysed datasets from Twitter. The data was collected and combined from three different datasets labelled as hateful, offensive, or clean. They selected the C4.5 decision tree to classify the data in two explored approaches: binary and ternary. The binary classification (polarised the tweets as offensive and clean) obtained an accuracy of 87.4%, and the ternary classification (polarised the tweets as hateful, offensive, and clean) had an accuracy of 78.4%.

(VIGNA et al., 2017) analysed the SVM classifier and a recurrent neural network LSTM on a dataset from Facebook in the Italian language. The classifiers presented a similar performance for hate speech detection. The LR and MLP are used in (WULCZYN; THAIN; DIXON, 2017) both classifiers obtained 96% AUC. Several classifiers are explored in (ALMATARNEH et al., 2019). In the study, the Complement Naive Bayes (CNB), SVM, and RF presented the best performances to identify specific hate speech against women and immigrants in English and Spanish languages. The SVM was also used in (SENARATH; PUROHIT, 2020) to evaluate semantic features of social media messages for hate speech detection.

(SALMINEN et al., 2020) analysed hate speech as a problem of multiple social media platforms (YouTube, Reddit, Wikipedia, and Twitter). They investigated multiple algorithms and individual features as well as combined features. The ensemble algorithm XGBoost (Extreme Gradient Boosted Decision Trees) presented a more significant performance than the other algorithms analysed (F1=0.92). In the analysis of the features, the models show the best

performance with BERT features.

Another approach explored is the Deep Neural Network (DNN), which has been used for feature extraction and classifier training. The most used classifiers are LSTM, CNN, and GRU (PITSILIS; RAMAMPIARO; LANGSETH, 2018; ZHANG; LUO, 2019; RIZOS; HEMKER; SCHULLER, 2019; SANTOSH; ARAVIND, 2019; AL-MAKHADMEH; TOLBA, 2019; CAO; LEE; HOANG, 2020; AL-SAFARI; SADAOUI; MOUHOUB, 2020; MOSSIE; WANG, 2020). (CAO; LEE; HOANG, 2020) proposed a framework for hate speech detection on social media, namely DeepHate. They evaluated the DeepHate using three public datasets and the combination of the three datasets. The DeepHate outperformed different CNN models.

An ensemble of recurrent neural networks is also investigated for hate speech detection (PITSILIS; RAMAMPIARO; LANGSETH, 2018). The authors proposed an ensemble of LSTM with the user's tendency towards each class as a feature method. Their model proposed has obtained more effective results than state-of-the-art with the detection of sexist messages (about F1-score=0.99), neutral (about F1-score=0.95), and racism (about F1-score=0.70).

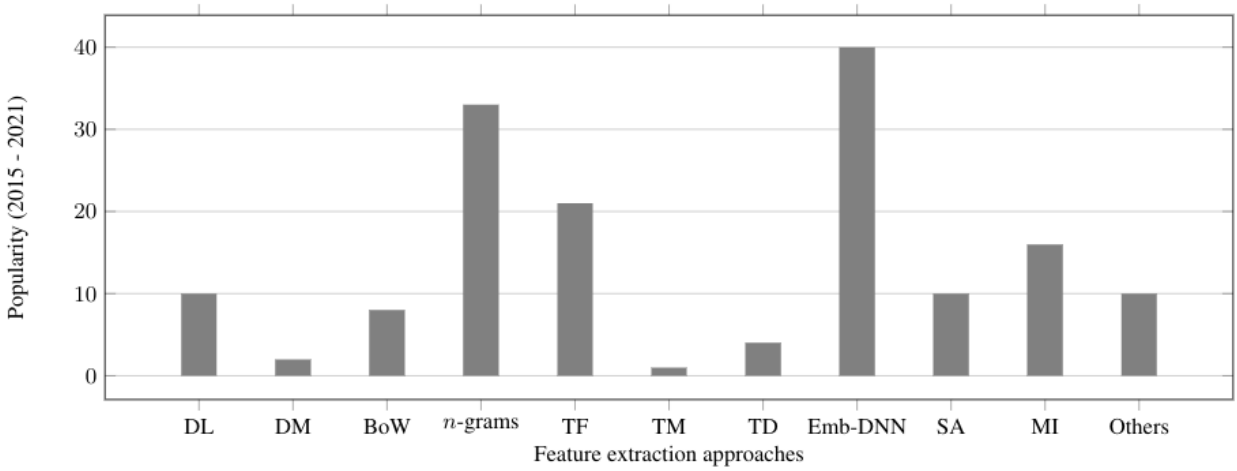
The ensemble deep learning method was also explored in (AL-MAKHADMEH; TOLBA, 2019). The authors proposed a hybrid approach, namely Killer Natural Language Processing Optimisation Ensemble Deep Learning (KNLPEDNN), which combines NLP and machine learning techniques. They used Stormfront (a neo-Nazi website) and CrowdFlower Twitter datasets. The ensemble method was used to minimise the weak features and to improve the prediction of hate. The system obtained 98.71% accuracy.

The models used different metrics to evaluate the performance of the models, such as Accuracy (Acc), AUC, Precision (P), Recall (R), and F-measure (F). Accuracy measures the number of correctly predicted samples among all predicted samples. The AUC computes the area under the ROC Curve. Precision measures the percentage of true positives among the true and false positives predicted. Recall measures the percentage of true positive cases that are correctly predicted positive. The F-measure calculates the harmonic average of precision and recall. Despite the results obtained in the studies evaluated, it needs to be clarified which model performed better. Furthermore, several works evaluate only the dataset collected by itself without evaluating whether the model generalises well to other domains.

2.8 RESEARCH DIRECTIONS AND GAPS FOR HATE SPEECH DETECTION ON SOCIAL MEDIA

This section aims to present challenges and points out automatic hate speech detection opportunities on social media platforms. As our previous sections suggested, the community has developed several resources to benefit from benchmark datasets for hate speech detection on social media platforms. Several feature extraction techniques and classification methods are employed on hate speech detection and related subjects. Figure 4 presents information about the popularity of the approaches used for feature extraction, and Figure 5 presents the classification method's popularity. The feature extraction techniques more used are embedding and DNN, and the n-grams. The classification method more used is SVM. Most works use more than one approach or a combination of them. In the following sections, we highlighted the challenges and opportunities.

Figure 4 – The frequency of feature extraction techniques from 2015 to July 2021. Dictionary or Lexical (DL); Distance Metrics (DM); Bag-of-Words (BoW); n-grams; Term Frequency (TF); Template Method (TM); Typed Dependencies (TD); Text Embedding and DNN (Emb-DNN); Sentiment Analysis (SA); Meta-information (MI).

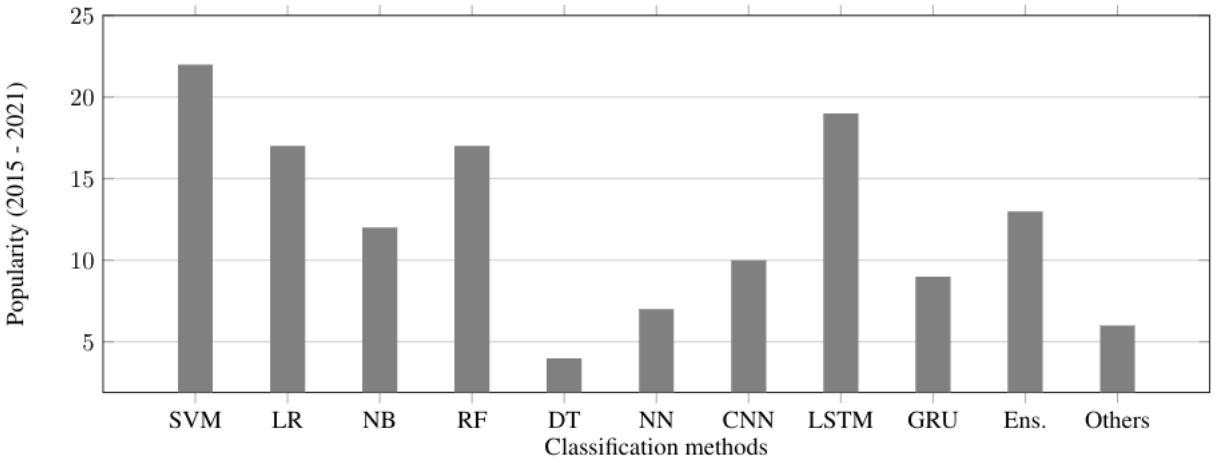


Source: Prepared by the author.

2.8.1 Challenges and opportunities

Hate speech detection is a complex phenomenon and difficult to recognise, both by humans and machines. Despite the efforts of the scientific community, different open challenges can be highlighted:

Figure 5 – The frequency of classification methods from 2015 to July 2021. Support Vector Machines (SVM); Logistic Regression (LR); Naive Bayes (NB); Random Forest (RF); Decision Tree (DT); classical Neural Network (NN); Convolutional neural network (CNN); Long Short-Term Memory (LSTM); Gated Recurrent Unit (GRU); Ensemble (Ens.). The 'Others' are techniques less used, such as K-Nearest Neighbors (K-NN), DeGroot's model, and so on.



Source: Prepared by the author.

- Issues with datasets: include bias because, in many cases, most data belong to the same user. Thus, dataset bias can overestimate the current state-of-the-art (ARANGO; PéREZ; POBLETE, 2019; CALABRESE et al., 2021). In particular, one of the most widely used datasets, proposed in (WASEEM; HOVY, 2016), most of the data are generated by a few users. The dataset has more than 16k tweets annotated as racist, sexist, and neither sexist nor racist, where only nine users sent the 1,972 for racist content;

- Context-dependent: transfers poorly across datasets, different approaches present high performance, however only within specific datasets, in which training and test sets were taken from the same dataset (ARANGO; PéREZ; POBLETE, 2019; GRöNDAHL et al., 2018; SENARATH; PUROHIT, 2020). This issue can be motivated by the influence of the social-demographic and cultural context of the dataset collection that can affect the data sampling and annotation methodology (WASEEM; THORNE; BINGEL, 2018);

- Polysemy words: when the word has many different meanings, hidden the actual text interpretation (SENARATH; PUROHIT, 2020);

- Imbalanced dataset: detection methods should not be vulnerable to imbalanced classes. Usually, hate speech datasets are highly imbalanced, with a small percentage of hate content, while most data are non-hate content. Practical resources often need to focus on the minority class (hate content). Therefore, the results evaluated using micro-average

metrics on the entire dataset can hide the real performance of minority classes (CHARITIDIS et al., 2020; ZHANG; LUO, 2019);

- Despite the efforts to automatically identify hate speech, a limitation is classifying messages without explicitly hateful words (ALORAINY et al., 2019; MACAVANEY et al., 2019);

Despite the challenges, we also can point out some opportunities in this field.

Feature selection: There is a clear lack of investigation on the impact of the feature selection process since text representation can deal with high dimensionality. In a study performed in (ROBINSON; ZHANG; TEPPER, 2018), the authors stated that automatic feature selection algorithms reduced about 90% of the feature space but only selected generic features. Therefore, to understand the contribution of distinct features to hate speech detection, there must be a focus on the existing feature selection techniques, which have proven to affect classification performance significantly.

Metadata: It is relevant that we can transpose our exploitative research into different languages. However, the study of features or indeed approaches for feature representation or metadata that works for more than one language is lacking since online social media platforms can offer a wide variety of information that improve the predictability of hateful (abusive) content (CHATZAKOU et al., 2017; FOUNTA et al., 2019; PITSILIS; RAMAMPIARO; LANGSETH, 2018), regardless of the text. Furthermore, in the study performed in (RIBEIRO et al., 2018), the authors have shown that users who produce hate speech are strongly linked. Therefore, metadata features can be helpful in this context.

Hate type: Better defining the specific characteristics of each type of hate speech (racism, gender hate, LGBT hate, religion, ethnicity, political view, etc.) can be potentially a significant advancement in this area.

Comparative studies: As we have pointed out, studies across datasets can help the analysis of the resulting generalisation models. In addition, different studies explored only the proposed dataset that often is not publicly available (NOBATA et al., 2016; VIGNA et al., 2017). Comparative studies using different models, features, and datasets are also necessary to understand better what is more effective for hate speech detection on online social media.

Multilingual research: Many researchers have explored datasets in only one language, the majority in English, which creates a lack of work focusing on cross-lingual scenarios. Few works use multilingual or bilingual content on social media platforms from the different Indian dialects (code-mixed language) (KUMAR et al., 2019; SANTOSH; ARAVIND, 2019). Different

particularities, such as distinct grammatical constructions and spelling variations, make the hate speech detection task in this context more difficult (SREELAKSHMI; PREMJITH; SOMAN, 2020). In order to deal with this, classification models and the datasets need to be more robust to lead to better classification performance on the code-mixed scenarios.

Ensemble learning: This approach has received relatively little attention in the context of hate speech detection. Moreover, ensemble methods have improved the results in different classification tasks.

'Memes' analysis: In certain cultures, there is heavy use of image-based with text dissemination of hate-related content. Such analysis has not yet been explored in this field, even though the distribution of such material is mainly done via social media sharing.

Free speech: There is a lack of comparative analysis of samples of free speech text and hate speech. For instance, in a study performed in (CASULA; ANUPAM; PARVIN, 2021), the authors discussed the effects of the moderation policies to avoid a toxic online environment in free speech. The researchers affirmed that even though online social media platforms state that they have developed a more inclusive online discourse environment, the moderation policies on online social media platforms can inhibit free speech and precipitate self-censorship. Since the preservation of free speech is essential in a democratic world, there is a need to create a mathematical analysis and definition of the main differences between those two models.

2.9 CONCLUSION

In this paper, we have presented a critical overview of automatic hate speech detection in text from the period between 2015-2021. So far, this task has been designed as a supervised learning problem and has used different techniques for feature extraction. Several works have applied simple features and feature extraction techniques, such as BOW, n-grams, or Term frequency, which provided a reasonable classification performance. Lexical resources are often used considering negative or derogatory words and have been employed as features or strategies for dataset collection. The pre-trained text embedding has been shown to be useful for abusive text classification. Features such as sentiment, meta-information, and extracted using DNN are relevant approaches and can improve the result when used to learn additional information. Other less frequently used features are FKGL and FRE scores, pattern features, LDA, and so on.

Judging which approaches are the best is a complex issue because several studies evaluate

only one dataset, and many are private. Hate speech detection is a recent subject, and different weaknesses still need to be explored.

3 UNINTENDED BIAS EVALUATION: AN ANALYSIS OF HATE SPEECH DETECTION AND GENDER BIAS MITIGATION ON SOCIAL MEDIA USING ENSEMBLE LEARNING

Francimaria RS Nascimento[1], George DC Cavalcanti[1], and Márjory Da Costa-Abreu[2]

[1]Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes s/n, Recife, Brazil

[2] Department of Computing, Sheffield Hallam University, Sheffield, UK

Article Published in ≪ ESWA ≫ submitted 2021, published 2022.

Abstract

Hate speech on online social media platforms is now at a level that has been considered a serious concern by governments, media outlets, and scientists, especially because it is easily spread, promoting harm to individuals and society and making it virtually impossible to tackle using just human analysis. Automatic approaches using machine learning and natural language processing are helpful for detection. For such applications, amongst several different approaches, it is essential to investigate the systems' robustness to deal with biases toward identity terms (gender, race, religion, for example). In this work, we analyse gender bias in different datasets and propose an ensemble learning approach based on different feature spaces for hate speech detection with the aim that the model can learn from different abstractions of the problem, namely **unintended bias evaluation metrics**. We have used nine different feature spaces to train the pool of classifiers and evaluated our approach on a publicly available corpus, and our results demonstrate its effectiveness compared to state-of-the-art solutions.

3.1 INTRODUCTION

The popularisation of social media platforms has driven the exponential growth of the textual content, making manual moderation of such content unsustainable (CAO; LEE; HOANG, 2020). In particular, social media platforms allow users to express themselves freely, giving them a false sense of 'no man's land' and promoting a fertile ground for hate speech cases and offensive language usage. Despite its scarcity compared to other contents, the easy dissemination of abusive content on these platforms can be potentially harmful to target individuals, society, governments, and social media (MIŠKOLCI; KOVÁČOVÁ; RIGOVÁ, 2020).

Hate speech is not a trivial phenomenon due to its subjective nature. (FORTUNA; NUNES, 2018) defined it as "*Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used*". It should be noted that hate speech is usually expressed against a group or a community and may cause potential harm to individuals and society.

In this context, sexist hate speech has a large space on online social media, usually used against women (CHIRIL et al., 2020). This type of speech discriminates or harms against a person or group based on a person's gender. Sexism often is based on a belief in the superiority of a specific sex or gender. Its dissemination can be potentially harmful, and we cannot underestimate its impact on online social media. As an example, widespread sexist hate speech on social media can disseminate gender stereotypes.

Several works have proposed methods to perform automatic hate speech detection on benchmark datasets using Natural Language Processing (NLP) with classic Machine Learning (ML) (SALMINEN et al., 2020; SENARATH; PUROHIT, 2020; WATANABE; BOUAZIZI; OHTSUKI, 2018) and Deep Learning techniques (ZHANG; LUO, 2019). So far, this task has been designed in the majority of cases using classic supervised machine learning approaches using metadata, user-based features, and text mining-based features, such as lexical approaches, n-grams, bag-of-words, text embedding, sentiment, etc., which require a previous definition of the feature extraction methods employed. Deep learning models have explored these approaches for feature extraction and classification (KAPIL; EKBAL, 2020; SANTOSH; ARAVIND, 2019). However, deep learning models require significant labelled data to perform well. Ensemble learning also has presented robust results, although few explored in the context of hate speech detection (AGARWAL; CHOWDARY, 2021; AL-MAKHADMEH; TOLBA, 2019; PITSILIS; RAMAMPIARO; LANGSETH, 2018). Even though different contributions have been dedicated to investigating these contents and presented high classification scores, the datasets and algorithms' potential biases did not receive attention in these researches.

The skewed distribution of specific terms in the training data can induce questionable trends for particular statements, and the representation learned by the model can not generalise well enough for practical use (BADJATIYA; GUPTA; VARMA, 2019; DIXON et al., 2018; PARK; SHIN; FUNG, 2018). Hence, the supervised model can give unreasonably high hateful scores to clearly non-hateful text, such as "*You are a great woman*". The source of this bias can

be associated with the highly frequent use of the word "*woman*" in hateful comments, which the model overgeneralised and associated with hateful comments. (DIXON et al., 2018) stated this phenomenon as *false positive bias* and defined this behaviour of recognition models as *unintended bias*. In particular, they said: *"a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others"*.

Despite previous efforts, recent studies have investigated concerns about systems' robustness and discuss the impact of unintended bias in the dataset (BADJATIYA; GUPTA; VARMA, 2019; DIXON et al., 2018; NOZZA; VOLPETTI; FERSINI, 2019). Some studies investigated bias regarding sensitive words (e.g., lesbian, gay, bisexual, transgender, trans, and so on) and tried to mitigate bias based on balancing the training dataset (DIXON et al., 2018) or using replacement strategies (BADJATIYA; GUPTA; VARMA, 2019). Moreover, some works presented evidence of racial and dialect biases in several corpora annotated for toxic content, based on the correlation between words related to African American English dialect (AAE) and toxicity ratings (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; SAP et al., 2019). Gender stereotypes in benchmark datasets are also a serious concern, in which a model can perform better with determinate identity terms than comments with others (PARK; SHIN; FUNG, 2018). Therefore, it is essential to consider the bias in the datasets and algorithms for hate speech detection. These biases in datasets or classifiers lead to unfairness against target groups, which the classifiers are usually designed to protect.

In this work, we proposed an ensemble learning method based on different feature spaces for unintended gender bias mitigation in the context of hate speech detection on online social media. The model combines base classifiers, each trained with a different feature representation. Each feature extraction method captures a different abstraction about the data and can present a different classification performance. Therefore, even though one method of feature extraction might fail due to inconsistencies in the data samples (SAJJAD et al., 2019) the system can still achieve a good performance as the system also considers other features. We analyse and mitigate gender bias in the datasets using bias-sensitive words and a replacement strategy for bias mitigation.

We believe that it will revolutionise the fight against gender-based hate speech if we can automatically detect messages of this nature and, therefore, deal with gender stereotypes present in the system. Thus, we analyse model biases, particularly gender identities (gender bias) in hate speech datasets. We also propose an approach based on ensemble learning to

classify hate speech on online social media and investigate the impact of gender bias in our ensemble method. Hence, this study aims to answer the following research questions: (1) Does the proposed multi-view stacked classifier combined with template-based mitigation outperform current techniques for hate speech detection in the context of unintended gender bias? (2) Can the bias mitigation method deal with gender biases in datasets without compromising the performance of the ensemble learning model?

In essence, the main contributions of this research are:

- Evaluation of a multi-view stacked classifier using nine different feature spaces combined with template-based mitigation for hate speech detection and gender bias mitigation.
- We perform our experiments in four real-world datasets in the context of gender bias mitigation.
- We explore the model's behaviour using three base classifiers while considering the unintended gender bias.

This work is organised as follows: Section 3.2 describes related work. Section 3.3 presents the problem statement, Section 3.4 describe the proposed methodology, Section 3.5 present the experimental setup, and Section 3.6 discusses the results. Section 3.7 concludes the work with the final remarks.

3.2 RELATED WORK

This section presents a comprehensive study of automatic hate speech detection and bias detection and mitigation in hate speech models and later specifically for gender-related hate speech.

3.2.1 Automatic hate speech detection

Hate speech is a complex problem that expresses the explicit intention to promote hatred or incite harm against a person or a targeted group. Several approaches have been proposed to hate speech detection on online social media platforms using classic machine learning methods, ensemble learning, and deep learning techniques. Twitter has attracted a significant part of the research due to the increasing number of available data and free tools for data collection

(DAVIDSON et al., 2017; WASEEM; HOVY, 2016; WASEEM, 2016; WATANABE; BOUAZIZI; OHTSUKI, 2018).

Classic supervised machine learning methods with different techniques for feature extraction have been frequently used in the literature for hate speech detection (ALMATARNEH et al., 2019; SANTOSH; ARAVIND, 2019). General feature representation methods of text mining have been successfully adapted to the problem of hate speech detection, such as Bag-of-Words (BoW) (BURNAP; WILLIAMS, 2016; NOBATA et al., 2016), n-grams (CORAZZA et al., 2020; SANTOSH; ARAVIND, 2019), dictionaries or lexical resources (GITARI et al., 2015; MATHEW et al., 2019), etc. Regarding classification perspective, different algorithms have been employed, such as Logistic Regression (DAVIDSON et al., 2017), Support Vector Machine (SVM) (SALMINEN et al., 2018), Random Forest (ELISABETH; BUDI; IBROHIM, 2020), Decision tree (PLAZA-DEL-ARCO et al., 2020).

(DAVIDSON et al., 2017) addressed the problem of hate speech detection on Twitter, focusing on distinguishing between hate speech and offensive language. They exhibited that the presence of offensive words does not necessarily represent hate speech. The researchers evaluated their own hate speech dataset with the Logistic Regression classifier that achieved an F1-score of 0.90. However, the classifier had difficulty differentiating tweets labelled as hate speech, mislabeling almost 40%.

The deep learning techniques learn abstract feature representations from the data, and different models can be used as feature extractors and classifiers for hate speech detection. Recently, several works have applied pre-trained word embedding approaches, such as Word2Vec, GloVe, and FastText, because of the semantic information extracted from the text (CAO; LEE; HOANG, 2020; FOUNTA et al., 2019; MIOK et al., 2019; SALMINEN et al., 2020). Regarding classification models, the most popular models are the Convolutional Neural Network (CNN) (ZHANG; LUO, 2019; VIGNA et al., 2017), Long Short-Term Memory Network (LSTM) (CAO; LEE; HOANG, 2020; ZHANG; ROBINSON; TEPPER, 2018), Gated Recurrent Unit (GRU) (CORAZZA et al., 2020), and Bidirectional Encoder Representations from Transformers (BERT) (MOZAFARI; FARAHBAKHSH; CRESPI, 2020).

Ensemble learning, or multiple classifier systems, have proven robust and improve the results of different classification tasks. In (AL-MAKHADMEH; TOLBA, 2019), (PITSILIS; RAMAMPIARO; LANGSETH, 2018), and (ZIMMERMAN; KRUSCHWITZ; FOX, 2018), the researchers explored the combination of deep neural networks. Even though the models achieve slightly higher classification results than the current state-of-art, these techniques are time-consuming compared to

the combination of other algorithms such as Logistic Regression and Decision Tree classifiers. In (RISCH; KRESTEL, 2020), the researchers proposed an ensemble of BERT models based on bootstrap aggregation (bagging) and used soft majority voting to combine the predictions. (LIU et al., 2019a) investigated the hate speech detection problem as multi-task learning. For the classification task, they proposed a fuzzy ensemble approach. The experimental results showed that the proposed method outperforms SVM and deep neural networks using embedding features.

3.2.2 Bias detection and mitigation in hate speech models

Recently, great efforts have been taken to detect and mitigate bias in hate speech detection models. (DIXON et al., 2018) investigated unintended bias in abusive detection models and evaluated the proposed method using a synthetic test set and an annotated dataset from Wikipedia Talk pages. The authors manually created a list of general identity terms (e.g., gay, transgender, feminist, and so on) to quantify the bias. Similarly, Nozza et al. (NOZZA; VOLPETTI; FERSINI, 2019) also used a list of terms to quantify and mitigate unintended bias.

In (BADJATIYA; GUPTA; VARMA, 2019), the researchers proposed a two-stage method for unintended stereotype bias detection and mitigation. Firstly, they design different heuristics to identify a set of bias-sensitive words. Further, in the second stage, the researchers proposed replacement strategies in training data to mitigate the bias. The results show that the proposed procedures can reduce the bias without compromising the model performance significantly.

(BOLUKBASI et al., 2016) demonstrated gender stereotypes in word2vec (MIKOLOV et al., 2013) and introduced an algorithm to reduce gender biases in word embeddings. (PARK; SHIN; FUNG, 2018) investigated gender bias on abusive language detection models. The authors used different methods to measure and debias gender bias, such as Debiased Word Embeddings, Gender Swap, and Bias fine-tuning strategies. Although the strategies for gender bias mitigation explored have reduced the performance of the classifiers, the authors stated that the method applied reduced the gender biases by 90-98%. In (KIRITCHENKO; MOHAMMAD, 2018), the researchers evaluated gender and race bias in 219 automatic sentiment analysis systems from SemEval-2018 Task 1. The study provided an Equity Evaluation Corpus (EEC) to evaluate those systems' gender and racial bias.

(SAP et al., 2019) investigated the unintended racial bias against speech produced by African Americans in two benchmark datasets widely used for hate speech detection. They used the

AAE dialect to quantify the toxicity rating and stated that AAE tweets are more likely to be associated with offensive classes than other tweets. In (MOZAFARI; FARAHBAKHSH; CRESPI, 2020), the researchers addressed the problem of racial bias in the trained classifier. They introduced a transfer learning approach based on the BERT using the fine-tuning of the algorithm to mitigate racial bias. The results achieved demonstrated evidence of racial bias in the trained classifier against tweets written in AAE.

In this study, we investigated a list of potential bias-sensitive words (available in Section 3.4.1.1) and looked for disproportionate representations, focusing on gender bias. We mitigated the gender bias based on a replacement strategy. Firstly, we evaluate the distribution of the bias-sensitive words in the hateful classes and overall. Then, we use a template strategy to replace the potential bias-sensitive words.

Despite different contributions for hate speech detection on online social media, it is relevant to highlight that a challenging task in hate detection is to select the best feature space for the classification. Furthermore, different feature spaces can capture different abstractions of the problem. However, classification models for detecting hate speech using multi-view learning are seldom explored. In this paper, we proposed an ensemble learning method based on several feature spaces and different classifiers using public datasets to fill this gap. Moreover, we address unintended gender bias in the training set.

3.3 PROBLEM FORMULATION

In this section, we formulate the problem statement and describe the datasets used. Furthermore, we discuss the strategy employed for gender bias mitigation and ensemble learning for hate speech detection.

3.3.1 Dataset description

We analyse public annotated datasets for hate speech detection. We limited our data source using four criteria: (a) Twitter as the data source because it is the third most popular online social media (ANTONAKAKI; FRAGOPOULOU; IOANNIDIS, 2021). Furthermore, Twitter is one of the most exploited sources for hate speech detection due to its policy on publicly available data and its free tools for data collection (POLETTO et al., 2020). (b) The dataset was available at the time of performing research. (c) Written in the English language. (d) Described in previous

studies. Thus, we obtained four datasets, described below and summarised in Table 4.

- **Waseem-Hovy (WH)** (WASEEM; HOVY, 2016): The corpus contains data collected from Twitter over the two months. The authors collected 130k tweets and performed an initial manual search with potential terms or phrases[1] they considered hateful. The authors then manually annotated a subset of these data based on guidelines inspired by critical race theory. The annotation was reviewed by "*a 25-year-old woman studying gender studies and a non-activist feminist*" to check annotator bias. The original dataset consists of 16,906 tweets annotated as sexism, racism, or neither.

- **Waseem (WS)** (WASEEM, 2016): This dataset explored an overlap of the dataset described in (WASEEM; HOVY, 2016) to investigate the influence of annotator in the labelling of data. Thus, the authors relabelled 2,876 tweets. The authors provide 6,909 labelled tweets by annotators, domain experts (feminist and anti-racist activists) and amateurs recruited on CrowdFlower. The authors also included a new label (racism and sexism) to identify tweets with both types of hate speech. However, we do not consider the new label (both) because it represents only 1% of the samples.

- **Davidson (DV)** (DAVIDSON et al., 2017): The authors used a hate speech lexicon from *Hatebase.org* to collect the corpus. The first sample was collected, resulting in 85.4 million tweets from the timeline of 33k Twitter users. Then, the authors selected a random sample of the 25k tweets using the lexicon. The CrowdFlower (CF) workers manually annotated the corpus as hate speech, offensive but not hate speech, or neither (neither offensive nor hate speech). In this process, the authors instructed the CF workers to think about the words and inferred context to avoid false positives. Thus, it has resulted in a dataset with 24,802 labelled tweets.

- **HatEval (HE)** (BASILE et al., 2019): collected the HatEval dataset for task 5 at SemEval-2019. They explored two categories of hate speech: misogyny and xenophobia. Different approaches were employed to compile potential hate speech and a lexicon of more frequent terms. The authors annotated the dataset from the crowdsourcing platform Figure Eight (F8) and two experts based on majority voting. The final dataset includes 19,600 tweets, 6,600 for Spanish and 13,000 for English. The data was annotated based

[1] Terms queried for: "MKR", "asian drive", "feminazi", "immigrant", "nigger", "sjw", "WomenAgainstFeminism", "blameonenotall", "islam terrorism", "notallmen", "victimcard", "victim card", "arab terror", "gamergate", "jsil", "racecard", "race card".

on three categories: Hate Speech (hateful or non-hateful); Target Range (individual or generic target); and Aggressiveness (aggressive or non-aggressive). However, we used only English tweets and the category Hate Speech.

Table 4 – Description of the datasets.

Dataset	Distribution	Number of instances	Label (%)	Target/Categories	Annotators
WH	GitHub repository	16,906	sexism (20%) racism (12%) neither (68%)	sexism, racism	1
WS	GitHub repository	6, 909	sexism (13%) racism (2%) neither(85%)	sexism, racism	4 or more
DV	GitHub repository	24,783	hateful (6%) offensive (77%) neither (17%)	general	3 or more
HE	GitHub repository	13,000	hateful (43%) non-hateful (57%)	misogyny, xenophobia	3

Source: Prepared by the author.

In the first phase, we pre-process the tweets for noise reduction. It includes removing the URLs (which start with "$http[s] : //$"), the mentions ("i.e.,$@user$"), numbers, punctuation and stopwords, and making all text lowercase and used stemming. Several works performed the pre-processing step before the feature extraction (DORRIS et al., 2020; ZHANG; ROBINSON; TEPPER, 2018; WATANABE; BOUAZIZI; OHTSUKI, 2018; DESOUZA; DA-COSTA-ABREU, 2020) because the informal language used on social media and a diversity of elements of the tweets (for example, user names, URLs) can introduce noise and confuse a text classifier. Furthermore, this data pre-processing reduces the feature dimensionality of different feature extraction methods.

3.3.2 Unintended gender bias mitigation

Text-related models can extract strong insights about the significant association between determinate terms and labels. In some cases, these associations can be positive and help the model improve performance. Nevertheless, it is not suitable for a hate speech detection model to depend on strong insight from individual word occurrences, but the combination of such words (BADJATIYA; GUPTA; VARMA, 2019). For example, "*Mary is a beautiful woman*". In this case, it might be beneficial for the classification model to use the knowledge extracted from

the significant association between the “woman" and the “female" label. However, it is not good to relate the word “woman" with a “hateful" label, which might have unintended learned from the training pattern.

Hate speech detection models tend to present gender biases toward specific identity terms (PARK; SHIN; FUNG, 2018). This issue can be motivated by the imbalanced nature of hate speech datasets and the disproportionate use of identity terms in hate speech sentences. For instance, some keywords such as “women" and “feminism" are highly associated with sexist comments in benchmark datasets (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; PARK; SHIN; FUNG, 2018). These factors can contribute to overfitting the original hate speech detection model. Consequently, the model can make generalisations such as associating the word “women" with a “hateful" label.

Different studies have investigated potential bias-sensitive words (BSWs). (DIXON et al., 2018) manually creates a list of 51 common identity terms and further analyses them from the training data. In (BADJATIYA; GUPTA; VARMA, 2019), the researchers also used the list of words proposed in (DIXON et al., 2018). Besides, the authors proposed two new approaches to selecting the words, called Skewed Occurrence Across Classes (SOAC), which select the word that is used significantly in a particular class ('Hateful'), and Skewed Predicted Class Probability Distribution (SPCPD), which select the word based on the probability distributions. In this study, we investigate a list of bias-sensitive words described in Section 3.4.1 based on the literature (NOZZA; VOLPETTI; FERSINI, 2019; KIRITCHENKO; MOHAMMAD, 2018), because we focus on a particular bias (gender bias), and we investigate disproportionate distribution among labelled classes.

Regarding bias correction, different strategies can be employed, such as statistical correction (DIXON et al., 2018), model correction or post-processing (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; PARK; SHIN; FUNG, 2018), and data correction (BADJATIYA; GUPTA; VARMA, 2019). The statistical correction includes techniques that distribute terms across the training set classes uniformly to balance the samples. In the model correction or post-processing, the mitigation of bias in the training set can either be made during the model fine-tuning in the post-processing or by modifying the word embeddings. The data correction strategy consists of generalising some attributes that the model should not use to classify the sentence as hateful, thus reducing the number of information in the training set available to the classifier.

In this paper, we employed the data correction strategy to mitigate unintended bias, focusing on gender bias. The data correction was employed because (i) selecting appropriate

samples for bias correction is challenging in the statistic correction. Furthermore, the balancing strategies used can introduce new skew distribution of terms in the training set; (ii) It does not require specific classifier models as the model correction strategy. Therefore, we can use it with any classification model; (iii) It has been successfully used as an unintended bias mitigation strategy for hate speech detection without compromising model performance (BADJATIYA; GUPTA; VARMA, 2019).

3.3.3 Ensemble learning

Hate speech datasets usually disproportionately use determinate terms (say, bias-sensitive words) highly correlated to minority class ('hateful'), enhancing bias stereotypes in the machine learning model. In this way, the classifier trained with biased data can deal with an increase in false-positive instances. Generally, different training data or feature spaces can emphasise different aspects of the problem. Even with the same method, each learning algorithm presents its own weaknesses and strengths (ZHOU et al., 2020).

Ensemble learning, or multiple classifier systems (MCS), is a machine learning technique that extracts the knowledge from the combination of several methods to increase the recognition accuracy in pattern recognition systems (KUNCHEVA, 2014). Bagging Algorithm (bootstrap aggregating) and Boosting are popular ensemble methods (WALMSLEY et al., 2018; RISCH; KRESTEL, 2020). These algorithms are based on a homogeneous set of weak learners and build diversity by sub-sampling or re-weighting the existing training examples. However, these methods used the same feature spaces for all classifiers, and a challenging issue in the hate speech detection task is to determine the right features for classification (FORTUNA; NUNES, 2018).

The hate speech detection on social media is a complex classification task in which different feature spaces can significantly change the performance. Moreover, a single classifier usually performs worse using a single feature space to handle inconsistencies and various data (SAJJAD et al., 2019). Several studies have argued that combining different feature spaces presents better results (BURNAP; WILLIAMS, 2016; WATANABE; BOUAZIZI; OHTSUKI, 2018), but combined spaces in a vector can deal with large dimensions.

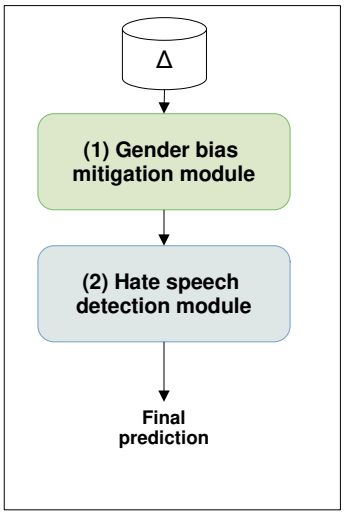
Therefore, we choose to combine the classifiers using different feature spaces. Each feature space represents a different view of the problem to capture different abstractions about the data. Thus, although one method might fail due to data inconsistencies, the system can still consider other feature spaces and perform well. The multi-view learning optimises the model

by learning one function based on different abstractions of the data that a single-view cannot comprehensively represent for all examples (CRUZ et al., 2013; ZHAO et al., 2017).

3.4 PROPOSED METHODOLOGY

This section introduces our methodology for hate speech detection and gender bias analysis and mitigation. The proposed model (Figure 6) consists of two main modules: **(1) Gender bias mitigation module** and **(2) Hate speech detection module**. These two modules are described in Sections 3.4.1 and 3.4.2, respectively.

Figure 6 – Overview of the proposed methodology. Δ is the training set.

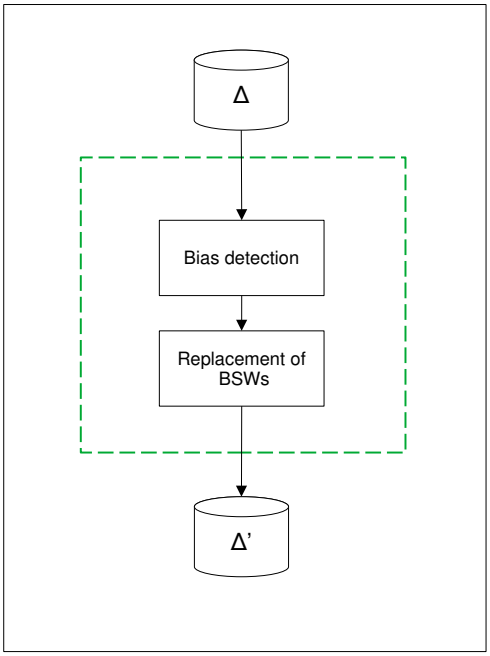


Source: Prepared by the author.

3.4.1 Gender bias mitigation

For gender bias mitigation, we investigate the disproportional distribution of specific terms on the datasets. Thus, we evaluate whether the model incorrectly predicted the sample's label based on specific words. The gender bias mitigation module is divided into two stages (see Figure 7): (1) Bias detection and (2) Replacement of bias-sensitive words (BSWs).

Figure 7 – Gender bias mitigation module. Δ and Δ' are the training set before and after the gender bias mitigation module, respectively. BSWs: bias-sensitive words.



Source: Prepared by the author.

3.4.1.1 Bias detection

In this stage, we evaluate the distribution of the bias-sensitive words in hateful tweets and the entire dataset to investigate disproportionate representations. In order to simplify our analysis, we only consider a binary gender. Table 5 presents the list of nouns used in our study representing females and males. Different nouns, such as 'she', 'her', 'he', and 'him', were disregarded because of the pre-processing step as we remove stop-words and, consequently, exclude these words. Besides, the word 'female' was written as 'femal' because of the pre-processing step.

Table 5 – Pairs of nouns representing a female or a male person used in this study.

Female	woman, women, girl, sister, daughter, wife, girlfriend, mother, aunt, mom, grandmother, femal
Male	man, men, boy, brother, son, husband, boyfriend, father, uncle, dad, grandfather, male

Source: Prepared by the author.

Bias in the training sets is a serious concern, and the high scores due to it can overestimate

the models (WIEGAND; RUPPENHOFER; KLEINBAUER, 2019). The significant occurrence of a word in a determinate class (say Hateful) can likely introduce unintended bias in the classifier model, which can probably learn this pattern and classify a sentence with that word into that class (BADJATIYA; GUPTA; VARMA, 2019). Therefore, we investigate the distribution of tweets with specific words. To do so, we compute $p(w|c)$ to measure the likelihood of the sentences in the class c containing the word w, and $p(w)$, which denote the likelihood of the sentences in the entire training/test set contain the word w. We analyse the disproportional distribution of determinate words in the hateful class and its overall distribution in the training and test sets. We used a cross-validation strategy to split three of the datasets in training, validation, and test sets (described in Section 3.5.1). Therefore, we present the average results for WH, WS, and DV datasets.

Figure 8 illustrates the top 10 average likelihood of words in the hateful comments and its overall likelihood with WH, WS, and DV datasets in the training and test sets, respectively. We used the partitions provided in (BASILE et al., 2019) for the HE dataset. The “class name", e.g. sexism, racism, hateful, and so on, in the columns, represents $p(w|c)$ and “overall" represents $p(w)$. We use a heat map with a grey colour bar where the legend indicates the likelihood values in colours.

Note that terms such as ‘women’ and ‘girl’ appear more frequently in ‘sexism’ and ‘hateful’ comments than overall comments with WH, WS, and HE datasets. On the other hand, even though terms such as ‘man’ and ‘girl’ have been more frequent in ‘hateful’ comments with the DV dataset, the amount of hateful comments containing these terms is not disproportional to the other classes. The term ‘man’ also occurred more frequently in ‘racism’ comments with the WS dataset. These behaviours occurred for both training and test sets’ samples for WH, WS, and DV datasets and only in the training set in the HE dataset.

It is relevant to observe that the high disproportional distribution of the term ‘women’ usually occurs in datasets composed of tweets related to sexism or misogyny categories. Furthermore, even though the term ‘man’ occurred with a higher frequency in the DV dataset in ‘hateful’ comments, the distribution of the term is much smaller than the word ‘women’ in other datasets.

3.4.1.2 Replacement of BSWs

In the replacement stage, we use a template strategy based on (BADJATIYA; GUPTA; VARMA, 2019) to replace the potential bias-sensitive words (listed in Table 5) for the $< identity >$ tag and reduce gender bias introduced by these terms without compromise the model accuracy. This process masks some of the information available in the training set, inhibiting bias through these BSWs in the classification model. Different examples are illustrated in Table 6.

Table 6 – Examples of sentences using the replacement strategy.

Tweet	**After replacement strategy**
RT @user: I'm not sexist, but girl fights just plain s*ck.	RT @user: I'm not sexist, but $< identity >$ fights just plain s*ck.
I'm not sexist but I hate serving women!	I'm not sexist but I hate serving $< identity >$!
This boy is an idiot followed by a bunch of idiots, this is a lack of leadership and direction.	This $< identity >$ is an idiot followed by a bunch of idiots, this is a lack of leadership and direction.

Source: Prepared by the author.

The idea is to reduce the differentiation of similar terms related to gender, such as 'women' and 'men'. In the hate speech domain, the term 'women' is usually more frequently used than 'men', although they represent a similar group. Hence, the significantly high use of a term in a specific class (say, Hateful) can likely introduce bias in the model.

3.4.2 Hate speech detection

The hate speech detection module consists of two phases (Figure 9):

1. *Pool generation phase*, where the pool of classifiers P is generated using the training instances based on the combination of the classifier c_i with each feature of the feature space $F : \{f_1, f_2, ..., f_n\}$, composed of n feature extraction methods; Then, $P : \{f_1c_1, f_2c_2, ..., f_nc_n\}$.

2. *Combination phase*, where the predictions are combined using the stacked generalization to give the final prediction.

Table 7 – Feature extraction methods. The N is the number of different sequences of words/characters across the dataset.

Name	**Feature**	**Description**	**Vector dimension**
f_1	GloVe	Global Vectors for Word Representation. Pre-trained word embedding.	200
f_2	FastText	Pre-trained word embedding.	300
f_3	BERT	Bidirectional Encoder Representations from Transformers (BERT). Pre-trained embedding method.	768
f_4	TF	Term Frequency.	vocabulary size
f_5	TF-IDF	Term Frequency-Inverse Document Frequency.	vocabulary size
f_6	Word bi-grams	Count vector of word bigrams.	N sequences of two adjacent words
f_7	Word tri-grams	Count vector of word trigrams.	N sequences of three adjacent words
f_8	Char bi-grams	Count vector of character bigrams.	N sequences of two adjacent characters
f_9	Char tri-grams	Count vector of character trigrams.	N sequences of three adjacent characters

Source: Prepared by the author.

3.4.2.1 Pool generation

The pool generation phase is performed using a heterogeneous approach since each model is trained with different feature spaces. We investigated three different base classifiers: Logistic Regression (LR) (DAVIDSON et al., 2017; UNSVÅG; GAMBÄCK, 2018), Decision Tree (DT) (PLAZA-DEL-ARCO et al., 2020; SALMINEN et al., 2018), and Support Vector Machine (SVM) (MACAVANEY et al., 2019; SENARATH; PUROHIT, 2020). We selected these models because they are frequently used for the classification of hate speech. Although recent works have addressed the use of Deep Learning models, these techniques are data-hungry and time-consuming compared to algorithms such as LR and DT.

Each feature set f_i captures a different representation of the instances and can capture different properties of the dataset. Thus, using distinct sets of features, even though one instance representation might fail due to feature space, the model can consider the other data representation. We selected nine feature representations currently used in the literature (WATANABE; BOUAZIZI; OHTSUKI, 2018; MOZAFARI; FARAHBAKHSH; CRESPI, 2020; CORAZZA et al., 2020; CAO; LEE; HOANG, 2020; FORTUNA; NUNES, 2018). Table 7 presents a summary of

the feature methods used.

We selected three popular embedding methods, GloVe (PENNINGTON; SOCHER; MANNING, 2014), FastText (BOJANOWSKI et al., 2017), and BERT (DEVLIN et al., 2019), for f_1, f_2, and f_3 representations, respectively. These embedding methods had been used in several studies and proven effective for hate speech classification (FOUNTA et al., 2019; MOZAFARI; FARAHBAKHSH; CRESPI, 2020; RIZOS; HEMKER; SCHULLER, 2019; SAJJAD et al., 2019).

We choose the highest dimension (200) available for GloVe embedding trained over the Twitter data, as it produced the best results in the literature (FOUNTA et al., 2019). For FastText embedding, the dimension is 300. The BERT has two models, and both have uncased (only lowercase letters) and cased versions, named BERT$_{BASE}$ and BERT$_{LARGE}$. The BERT$_{BASE}$ model contains 12 layers, 12 self-attention heads, and 110 million parameters, and the BERT$_{LARGE}$ model has 24 layers, 16 attention heads, and 340 million parameters. This work uses the uncased version of the pre-trained BERT$_{BASE}$ model because training BERT is computationally expensive.

Furthermore, the BERT$_{BASE}$ effectively detected hate speech in (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; RISCH; KRESTEL, 2020; SALMINEN et al., 2020). We used the implementation from Transformers library (WOLF et al., 2020) for the BERT model and Zeugma library[2] for the other word embedding methods.

For the representations f_4 to f_9, we selected traditional feature extraction methods used for hate speech detection (ALMATARNEH et al., 2019; CORAZZA et al., 2020; ELISABETH; BUDI; IBROHIM, 2020; SALMINEN et al., 2020; SENARATH; PUROHIT, 2020; SANTOSH; ARAVIND, 2019). These methods are based on the Bag-of-Words (BoW) technique. Thus, for the TF and TF-IDF, the feature vector's size used depends on the dataset vocabulary size. The n-grams technique combines the n adjacent items (words, characters, syllables, etc.) into a list of size N. We selected two approaches 'word n-grams' and 'character n-grams', with n equal 2 and 3. We used the implementations from scikit-learn (PEDREGOSA et al., 2011) for the extraction of these features.

It is relevant to highlight that the proposed framework is general to work with different feature extraction methods and classifiers. The new techniques added to the system only need to be careful with the feature representation standard required by the classifier selected. Therefore, the proposed methodology can be continuously refined and improve the classification results with new feature extraction methods and new classification models.

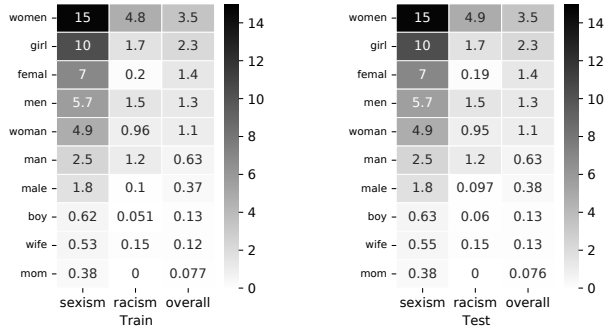
[2] https://zeugma.readthedocs.io/en/latest/

3.4.2.2 Combination phase

In the combination phase, the outputs of the classifiers are aggregated to obtain the final decision. The aggregation of the models can be performed based on different strategies, such as non-trainable, trainable and dynamic weighting (CRUZ; SABOURIN; CAVALCANTI, 2018). In this work, we used a trainable aggregation strategy (WOLPERT, 1992). The Stacked Generalization (or "stacking") consists of two levels of learning (ORIOLA, 2020). At the first level, different base learning algorithms learn from the training dataset. Each trained algorithm is then used to create a new dataset from the predictions collected using the validation dataset. Then, at the second level, another learning algorithm, also called meta-learner, is fitted based on the new dataset, which learns the aggregation function to provide the final prediction.

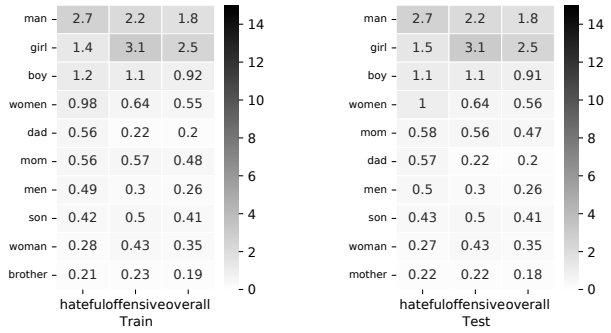
This architecture presents more robust than non-trainable ones as it does not require assumptions about the base model. Furthermore, the stacked generalisation does not use fixed rules and can be adjusted to the characteristics of the problem (CRUZ; SABOURIN; CAVALCANTI, 2018). It has also been successfully used as a fusion rule in different classification problems, for instance, sentiment analysis (AL-AZANI; EL-ALFY, 2017) and hate speech classification (MACAVANEY et al., 2019; MONTANI; SCHÜLLER, 2018; PASCHALIDES et al., 2020). We use the StackedClassifier implementation provided by the Deslib Python library (CRUZ et al., 2020).

Figure 8 – The top 10 likelihood of tweets with the terms related to the gender terms in each dataset (WH, WS, DV, and SE). Sorted by the first column in descending order. Average results of cross-validation for WH, WS, and DV datasets.

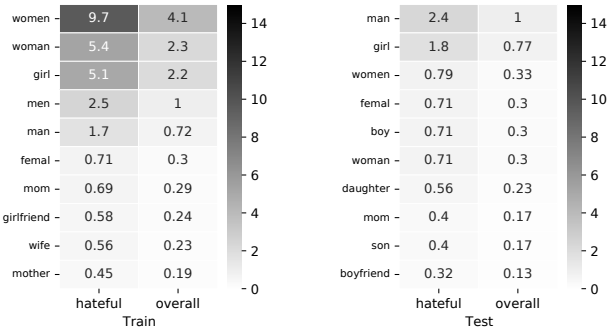


(a) WH dataset.

(b) WS dataset.



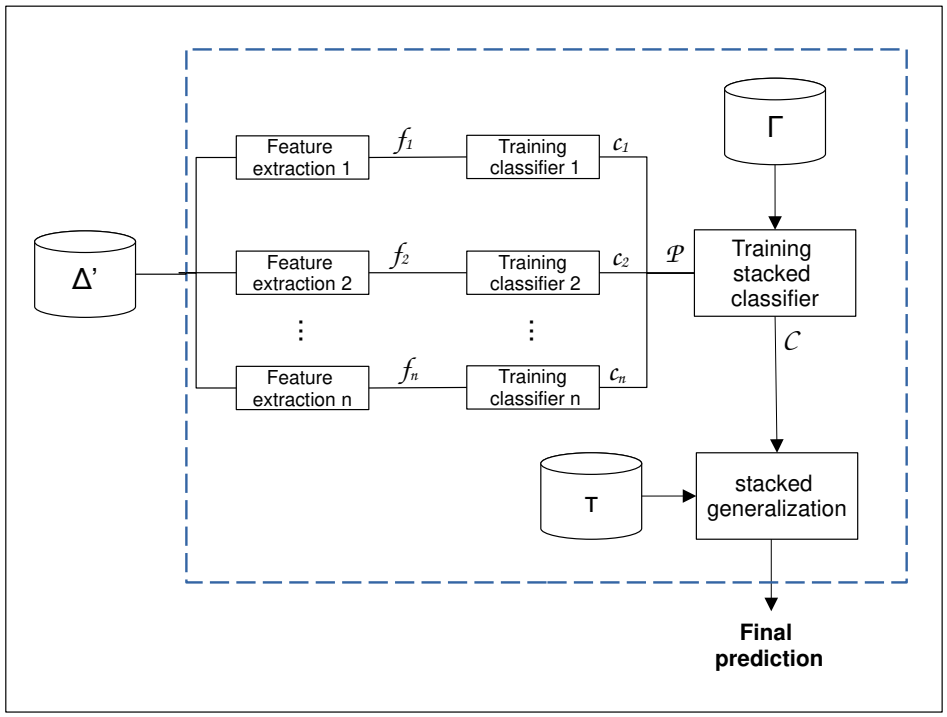
(c) DV dataset.



(d) SE dataset.

Source: Prepared by the author.

Figure 9 – Hate speech detection module. Δ', Γ, and τ are the training after the bias mitigation module, validation, and test sets, respectively.



Source: Prepared by the author.

3.5 EXPERIMENTAL SETUP

3.5.1 Datasets

The experiments were conducted using four public datasets for hate speech detection described in Section 3.3.1. We used stratified 5-fold cross-validation to evaluate the model. However, we need to partition the data into training, validation, and test because we used the validation set predictions to fit the stacked model. The cross-validation scheme partitioned the dataset into 5 disjoint subsets (1 fold for test and 4 folds for training/validation). Then, we applied a stratified 4-fold cross-validation in the training/validation folds divided into 3 folds for training and 1 for validation. Resulting in 20 executions with 3 folds for training, 1 for validation, and 1 for test. We used a stratified division because this strategy preserves the prior percentage of samples for each class. For the HE dataset, we used the partition provided in (BASILE et al., 2019) to compare the results with the literature easily.

3.5.2 Parameters setting

As stated in Section 3.4.2.1, we consider three base classifiers in this study: LR, SVM, and DT. These models were selected because they are the most used for hate speech detection. We trained each classifier with a different feature space resulting in nine models using each classifier.

Table 8 – Hyperparameters of the models evaluated for all datasets.

Classifier	**Hyperparameter**
LR	'penalty': ['l1', 'l2']
DT	'criterion': ['gini', 'entropy'], 'splitter': ['best','random']
SVM	'kernel': ['linear', 'sigmoid', 'rbf', 'poly']

Source: Prepared by the author.

We used a grid search to select the best hyperparameters of the models for all datasets. Table 8 shows the hyperparameters evaluated for each classifier. We fitted the algorithms using the training set and evaluated their performance using the validation set. Then, we selected the hyperparameters setup with the best performance for each model based on the macro F1-score metric.

3.5.3 Evaluation Metrics

We evaluated the overall **performance** of the classification with the macro F1-score. The F1-score measures the harmonic mean of the Precision and Recall. The precision is computed by the number of samples correctly classified as positives divided by the total of samples predicted as positives. The recall is the number of samples correctly classified as positives divided by the total samples identified as positives, including the false negatives. In multi-class problems, the F1-score is often applied to each class and aggregated using micro-average or macro-average to give a final result. In this work, we use the macro-average due to the imbalanced nature of the datasets. Furthermore, the micro-averaging can mask the real performance of minority classes (CHARITIDIS et al., 2020).

We look at divergences between the terms to measure the performance of the bias mitigation module, which we are calling as **Unintended bias evaluation metrics**. However, the disproportional distribution in the original dataset can be followed by the test set and influence the de-bias evaluation (PARK; SHIN; FUNG, 2018; NOZZA; VOLPETTI; FERSINI, 2019). Therefore, we generated a synthetic unbiased test set, with similar comments distribution for each term, based on the identity term template method proposed in (DIXON et al., 2018; PARK; SHIN; FUNG, 2018; NOZZA; VOLPETTI; FERSINI, 2019).

The idea is to build sentences changing only the identity term, for instance, "*Women* should be protected" and "*Men* should be protected". We define several templates filled with the terms described previously (see Table 5). Thus, each identity term occurs in the same context. Table 9 presents examples of templates. Several studies have also used this approach to evaluate unintended gender bias in different abusive language datasets (NOZZA; VOLPETTI; FERSINI, 2019; PARK; SHIN; FUNG, 2018). The synthetic test set[3] created comprises $1,248$ samples, of which 648 are non-hateful, and 600 are hateful, and all identity terms appear in the same contexts.

For evaluation of the unintended bias, we use metrics introduced in a recent state-of-the-art work (DIXON et al., 2018). The *Error Rate Equality Differences* compute the aggregation of the difference between the false positive rate (FPR) or false negative rate (FNR) on the entire test set and the per-term values, FPR_t and FNR_t. False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) are defined in Equations 4.2 and 4.3, respectively, where $T = \{female, male, girl, boy, ...\}$.

[3] <https://github.com/Francimaria/Hate_speech_gender_bias>

Table 9 – Template examples used to generate the synthetic test set.

Example template	**Label**
I am $< identity >$	non-hateful
$< identity >$ should be protected	non-hateful
You are a good $< identity >$	non-hateful
Filthy $< identity >$	hateful
I hate all $< identity >$	hateful
$< identity >$ should be killed	hateful

Source: Prepared by the author.

$$FPED = \sum_{t \epsilon T} |FPR - FPR_t| \tag{3.1}$$

$$FNED = \sum_{t \epsilon T} |FNR - FNR_t| \tag{3.2}$$

The error rate equality differences measure the model's fairness based on the hypothesis that a model without unintended bias has a similar error rate across all identity terms. Therefore, for these metrics, the ideal result is zero. It is relevant to mention that these metrics aim to evaluate bias. Thus, the punctual values of these metrics are not necessary, but rather whether they have similar values across all terms. Hence, we want to evaluate whether a specific term influences the error rates and, consequently, is subjected to unintended bias.

The *Pinned AUC Equality Difference* ($pAUC$) metric is also investigated in the literature to measure unintended bias (DIXON et al., 2018). However, we decided not to apply the $pAUC$ metric because it suffer from several limitations (BORKAN et al., 2019). Moreover, its competence to measure unintended bias depends on the sampling procedure used (BADJATIYA; GUPTA; VARMA, 2019).

3.5.4 Statistical analysis

For statistic analysis of the classifiers, we used the non-parametric Friedman test to compare the classification performance of all classifiers over the datasets as recommended in (DEMŠAR, 2006). The Friedman test ranks each algorithm for each dataset. The best performing algorithm gets the rank 1, the second-best rank 2, and so on. In the case of ties, average ranks are used. Then, the average rank is computed using all datasets. We performed the tests with 95% confidence, i.e., the significance level $\alpha = 0.05$.

We also performed a post-hoc Bonferroni-Dunn test for pairwise comparison between the average ranks for each classifier over the datasets. The critical difference is measured to evaluate whether the performance of the two classifiers is significantly different. The performance of the two classifiers is considered significantly different when the average rank is higher than the critical difference. The critical difference (CD) is defined in Equation 3.3. The critical value q_α is based on the Studentized range statistic divided by $\sqrt{2}$.

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{3.3}$$

We used the critical difference diagram proposed in (DEMŠAR, 2006) to describe post-hoc test results projected onto the average rank axis. The thick horizontal line connects classifiers that are not significantly different on the CD diagram.

Furthermore, we also investigated a second pairwise statistical analysis test to examine whether there is a significant difference between the classification methods. We use the Wilcoxon non-parametric signed-rank test with the level of significance $\alpha = 0.05$. (DEMŠAR, 2006) stated that this method is robust for pairwise comparison between classification algorithms.

3.6 RESULTS AND DISCUSSION

In this work, we divided the experimental study into four parts. In Section 3.6.1, the base classifiers are evaluated for each dataset using the test set. In Section 3.6.2, evaluate the proposed model performance with the test set and the unintended bias mitigation using the synthetic test set. In Section 3.6.3, we compare the proposed methodology against the state-of-art in order to evaluate the classification performance using the test set and the bias toward identity terms using the synthetic test set. Then, in Section 3.6.4, we analyse the case of studies to evaluate the effectiveness of the proposed methodology for gender bias mitigation using the synthetic test set.

3.6.1 Base classifiers evaluation

Firstly, we analysed the behaviour of each base classifier (LR, DT, and SVM) across different feature extractors. We used the macro F-measure scores to compare the general

model performance. Table 10 presents the average and standard deviation results for the datasets evaluated. The best results are highlighted in bold, and the second-best results are underlined for each dataset.

As seen in Table 10a, the LR classifier obtained the highest results with the TF feature extractor in WH, WS, and DV datasets. The pair LR and TF-IDF presented the best scores for the DV dataset. For the HE dataset, the monolithic classifier analysed presented a different behaviour. The LR classifier achieved the highest macro F-score with FastText word embedding and word 2-grams.

The DT classifier achieved slightly lower results than the other classifiers analysed for the four datasets evaluated (see Table 10b). This classifier performed better for the WH dataset with the TF-IDF feature extractor. Different feature extractors presented better results for the WS dataset as TF, TF-IDF, and character 3-grams. For the HE dataset, this classifier obtained the highest macro F-score with Glove word embedding and word 2-grams.

For the SVM classifier (Table 10c), the TF feature extractor obtained the highest classification performance for WH, WS, and DV datasets. On the other hand, the SVM with FastText word embedding and word 2-grams feature extractors presented the best classification performance for the HE dataset, similarly to the LR classifier.

The Friedman statistic test shows that there is a significant difference between the classification performance of each algorithm with the nine feature extraction techniques. Then, we evaluated a pairwise comparison using a post-hoc test. Figure 10 shows the Critical Difference (CD) diagram of the statistical test. The TF, TF-IDF, and character 3-grams feature extraction algorithms presented the highest rank values with the three classifiers. These results demonstrated that the vocabulary used is similar in these datasets, and the BoW and character n-grams approaches are still relevant in this context. Moreover, the experiments showed that the performance of the classifiers is highly dependent on the selected feature space and the dataset under analysis.

3.6.2 Proposed model evaluation

In this section, we will analyse the results obtained with the proposed methodology using three different base classifiers (LR, DT, and SVM) and our focus is to evaluate whether the bias mitigation model compromises the performance of the ensemble learning model. For the stacking generalisation (WOLPERT, 1992), we have used the Logistic Regression algorithm as

Table 10 – Performance of the base classifiers varying the feature spaces. Average and standard deviation results of the macro F-score. The best results are highlighted in bold, and the second-best results are underlined for each dataset. We present the results of the average rank in the column named 'Avg rank' of the tables.

LR					
Feature	**WH**	**WS**	**DV**	**HE**	**Avg. rank**
BERT	0.72 (0.01)	0.61 (0.03)	0.59 (0.02)	0.49	5.00
Glove	0.66 (0.01)	0.48 (0.02)	0.65 (0.02)	0.54	4.75
FastText	0.65 (0.01)	0.44 (0.01)	0.59 (0.02)	**0.56**	5.63
TF	**0.76 (0.01)**	**0.70 (0.02)**	**0.71 (0.02)**	0.46	2.50
TF-IDF	0.73 (0.01)	0.65 (0.05)	**0.71 (0.02)**	0.45	3.63
w2grams	0.57 (0.01)	0.44 (0.03)	0.44 (0.02)	**0.56**	6.25
w3grams	0.38 (0.01)	0.41 (0.02)	0.30 (0.01)	0.42	9.00
c2grams	0.72 (0.01)	0.64 (0.03)	0.64 (0.02)	0.43	5.38
c3grams	0.75 (0.01)	0.70 (0.04)	0.70 (0.02)	0.47	2.88

(a) Results obtained with Logistic Regression classifier (LR).

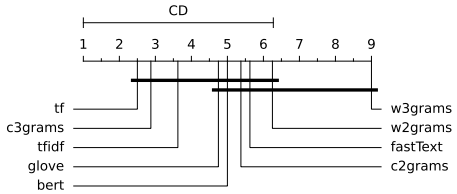
DT					
Feature	**WH**	**WS**	**DV**	**HE**	**Avg. rank**
BERT	0.52 (0.01)	0.46 (0.02)	0.46 (0.01)	0.52	6.5
Glove	0.55 (0.01)	0.44 (0.02)	0.54 (0.01)	**0.54**	5.50
FastText	0.56 (0.01)	0.44 (0.01)	0.52 (0.01)	0.53	5.875
TF	0.71 (0.01)	**0.69 (0.04)**	**0.70 (0.01)**	0.41	3.50
TF-IDF	**0.72 (0.01)**	**0.69 (0.04)**	0.69 (0.01)	0.43	3.25
w2grams	0.60 (0.01)	0.52 (0.04)	0.49 (0.02)	**0.54**	4.625
w3grams	0.42 (0.01)	0.45 (0.03)	0.33 (0.01)	0.50	7.50
c2grams	0.65 (0.01)	0.66 (0.03)	0.62 (0.01)	0.46	4.50
c3grams	0.70 (0.01)	**0.69 (0.04)**	0.67 (0.01)	0.44	3.75

(b) Results obtained with Decision Tree classifier (DT).

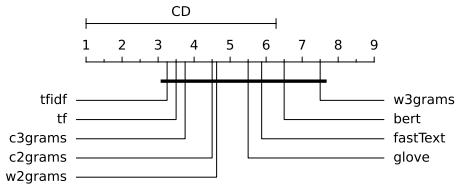
SVM					
Feature	**WH**	**WS**	**DV**	**HE**	**Avg. rank**
BERT	0.71 (0.01)	0.63 (0.03)	0.57 (0.02)	0.48	5.625
Glove	0.70 (0.01)	0.48 (0.02)	0.61 (0.02)	0.55	5.00
FastText	0.71 (0.01)	0.47 (0.01)	0.59 (0.02)	**0.56**	5.125
TF	**0.75 (0.01)**	**0.70 (0.03)**	**0.71 (0.02)**	0.42	2.625
TF-IDF	**0.75 (0.01)**	0.68 (0.03)	0.68 (0.02)	0.44	3.125
w2grams	0.56 (0.01)	0.47 (0.03)	0.45 (0.02)	**0.56**	6.25
w3grams	0.38 (0.01)	0.44 (0.02)	0.31 (0.01)	0.49	7.75
c2grams	0.72 (0.01)	0.65 (0.03)	0.60 (0.02)	0.39	5.50
c3grams	0.74 (0.01)	0.69 (0.02)	0.67 (0.01)	0.40	4.00

(c) Results obtained with the Support Vector Machine classifier (SVM).

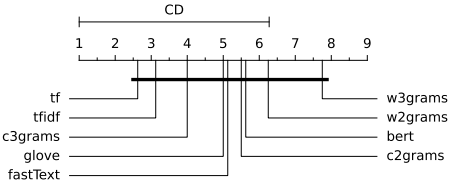
Figure 10 – Graphical representation of the average rank for each classifier over all datasets. For each classifier, we evaluated the performance with nine different feature extraction techniques. We used Bonferroni-Dunn post-hoc test to compute the critical difference (CD). Techniques with no statistical difference are connected by horizontal lines.



(a) LR classifier results.



(b) DT classifier results.



(c) SVM classifier results.

Source: Prepared by the author.

the meta-classifier. We selected this classifier because it is simpler and quicker than SVM and obtained better performance than the DT classifier (over the WH, WS, and DV datasets). Moreover, the Wilcoxon signed-rank test results demonstrate no significant difference between the performance of the LR and SVM classifiers for three of the datasets analysed (WH, DV, and HE).

Table 11 describes the results obtained with the proposed methodology using the original training set and the bias mitigation module. We presented the average and standard deviation results for WH, WS, and DV datasets. For the HE dataset, we used the partitions proposed in (BASILE et al., 2019). The best results are highlighted in bold for each metric. For each dataset, we performed a pairwise comparison of the proposed methodology with the original training set and after using the bias mitigation module. We used the Wilcoxon statistical test to compare the models, and significantly better results are marked with a $*$.

For the WH dataset (Table 11a), our proposed methodology obtained the best macro F-score, FPDE, and FNDE results. These results suggest that the proposed methodology reduces the unintended gender bias without compromise the model performance in this dataset. However, the bias mitigation scores tended to have a slightly increased with the SVM classifier.

Table 11b presents the results of the WS dataset. The proposed classifier obtained similar results with the original training set and with the bias mitigation module. On the other hand, for the DV dataset (Table 11c), the proposed classifier achieved macro F-score slightly inferior

Table 11 – Results obtained by the proposed method. Before and after applying the bias mitigation module. The best results are highlighted in bold for each metric. Results that are significantly better are marked with $*$

	Original training set			Bias Mitigation module		
Model	**F-score**	**FPED**	**FNED**	**F-score**	**FPED**	**FNED**
proposed (LR)	0.77 (0.009)	1.07 (0.486)	1.49 (0.499)	$*$ **0.79 (0.011)**	$*$**0.20 (0.237)**	$*$**0.47 (0.318)**
proposed (DT)	0.74 (0.010)	0.91 (0.628)	1.00 (0.377)	$*$ **0.77 (0.011)**	$*$ **0.29 (0.343)**	$*$**0.37 (0.329)**
proposed (SVM)	0.77 (0.008)	$*$**0.84 (0.313)**	$*$**1.43 (0.445)**	$*$ **0.79 (0.012)**	2.94 (1.183)	3.03 (1.091)

(a) Results obtained by the proposed method with the WH dataset.

	Original training set			Bias Mitigation module		
Model	**F-score**	**FPED**	**FNED**	**F-score**	**FPED**	**FNED**
proposed (LR)	**0.71 (0.026)**	**0.07 (0.129)**	$*$**0.17 (0.339)**	0.71 (0.034)	0.08 (0.159)	0.47 (0.592)
proposed (DT)	0.68 (0.043)	0.01 (0.031)	**0.00 (0.017)**	**0.69 (0.039)**	**0.00 (0.000)**	**0.00 (0.000)**
proposed (SVM)	0.70 (0.024)	$*$**0.10 (0.193)**	$*$**0.53 (0.418)**	0.70 (0.034)	0.24 (0.391)	0.90 (1.112)

(b) Results obtained by the proposed method with the WS dataset.

	Original training set			Bias Mitigation module		
Model	**F-score**	**FPED**	**FNED**	**F-score**	**FPED**	**FNED**
proposed (LR)	$*$**0.72 (0.022)**	5.72 (1.584)	3.99 (1.320)	0.71 (0.023)	$*$**4.39 (1.495)**	**3.60 (0.730)**
proposed (DT)	**0.67 (0.017)**	2.37 (1.561)	2.40 (1.572)	0.66 (0.014)	$*$ **0.80 (0.868)**	$*$**0.84 (0.597)**
proposed (SVM)	$*$**0.71 (0.025)**	$*$**5.62 (0.873)**	$*$**3.82 (0.604)**	0.70 (0.024)	6.03 (0.894)	4.29 (0.612)

(c) Results obtained by the proposed method with the DV dataset.

	Original training set			Bias Mitigation module		
Model	**F-score**	**FPED**	**FNED**	**F-score**	**FPED**	**FNED**
proposed (LR)	**0.46**	0.27	3.82	0.45	**0.00**	**1.54**
proposed (DT)	**0.44**	4.02	4.69	0.42	**0.00**	**0.15**
proposed (SVM)	0.42	0.14	3.00	0.42	**0.00**	**1.92**

(d) Results obtained by the proposed method with the HE dataset.

Source: Prepared by the author.

with the bias mitigation module. The HE dataset (see Table 11d)) also was evaluated in task 5-A at SemEval-2019, the mean of the baseline results with the dataset in English were 0.451 and 0.367, with the SVM and MFC (this assigns the most frequent labels), respectively, and the proposed model obtained similar results with LR classifier. Moreover, for this dataset, the proposed methodology reduces the unintended gender bias.

In order to improve the general classification performance of the proposed method for the HE dataset, we also evaluated the proposed model with different feature extractor combinations. After conducting empirical tests, we found a better trade-off between macro F-score and gender bias mitigation using three feature extraction methods: the word embedding FastText, Glove, and word 2-grams. The performance of the monolithic models with the other feature extractors can have influenced the results obtained. The results are described in Table 12. Al-

though the better classification performance, the proposed method using all features obtained better bias mitigation with the LR and SVM classifiers than using a subset of the features.

Table 12 – Results obtained by the proposed method adapted for HE dataset. Before and after applying the bias mitigation module. *In the pool of classifiers, we used only three feature extractors (FastText, Glove, and word 2-grams). The best results are highlighted in bold for each metric.

	Original training set			**Bias Mitigation module**		
Model*	**F-score**	**FPED**	**FNED**	**F-score**	**FPED**	**FNED**
proposed (LR)	0.55	4.25	7.43	0.55	**2.85**	**5.76**
proposed (DT)	0.54	0.41	0.28	**0.55**	**0.00**	**0.00**
proposed (SVM)	0.55	0.76	3.87	0.55	**0.45**	**3.31**

Source: Prepared by the author.

It has been shown in the literature the use of monolithic classifiers for hate speech classification task, such as (WASEEM; HOVY, 2016) and (DAVIDSON et al., 2017) used LR; (SALMINEN et al., 2018) used LR, DT, SVM and also used ensemble models; and (SENARATH; PUROHIT, 2020) used SVM. We have evaluated these classifiers with different feature extraction methods (see Section 3.6.1) and the proposed method achieved better performance than these methods for WH, WS, and DV datasets (Table 11). Even though we employed a simple strategy for bias mitigation and classical machine learning classifiers, the proposed methodology proved robust to unintended gender bias mitigation without compromising the model performance.

For the HE dataset, the method proposed using only three feature extractors (see Table 12) would be placed at the third position out of 69 submissions to the English Subtask A[4]. It is relevant to highlight that even though the team in the second position obtained a 0.571 macro f-score, they did not provide the system descriptions for a fair comparison. Moreover, our method also deals with unintended bias mitigation, which is in addition to classification performance.

Despite the ensemble learning techniques being time-consuming compared to monolithic models, the base models of the proposed stacked classifier can be executed simultaneously, reducing the processing time of the proposed model. Furthermore, once trained, its prediction is faster.

[4] <https://docs.google.com/spreadsheets/d/1wSFKh1hvwwQIoY8_XBVkhjxacDmwXFpkshYzLx4bw-0/edit#gid=0>

3.6.3 Proposed methodology versus the best base classifier

This section compares the results obtained by the proposed methodology against the best results obtained by the LR base classifier (Table 13). We evaluated the classification performance of the models using the macro F-score, while the FPED and FNED metrics are employed for bias evaluation.

Table 13 – Performance of the proposed method and the LR classifier. *In the pool of classifiers, we used only three feature extractors (FastText, Glove, and word 2-grams) for the HE dataset marked with $*$.

Dataset	Model	F-score	FPED	FNED
WH	TF + LR	0.76 (0.01)	0.61 (0.68)	1.41 (0.64)
	proposed (LR)	**0.79 (0.01)**	**0.20 (0.24)**	**0.47 (0.32)**
WS	TF + LR	0.70 (0.02)	**0.00 (0.00)**	**0.00 (0.00)**
	proposed (LR)	**0.71 (0.03)**	0.08 (0.16)	0.47 (0.59)
DV	TF + LR	0.71 (0.02)	**2.36 (1.83)**	**2.43 (1.48)**
	TF-IDF + LR	0.71 (0.02)	2.91 (1.47)	2.50 (0.77)
	proposed (LR)	0.71 (0.02)	4.39 (1.50)	3.60 (0.73)
HE	FastText + LR	**0.56**	6.24	6.62
	w2grams + LR	**0.56**	0.19	**0.22**
	proposed (LR)	0.45	**0.00**	1.54
	proposed (LR)$*$	0.55	2.85	5.76

Source: Prepared by the author.

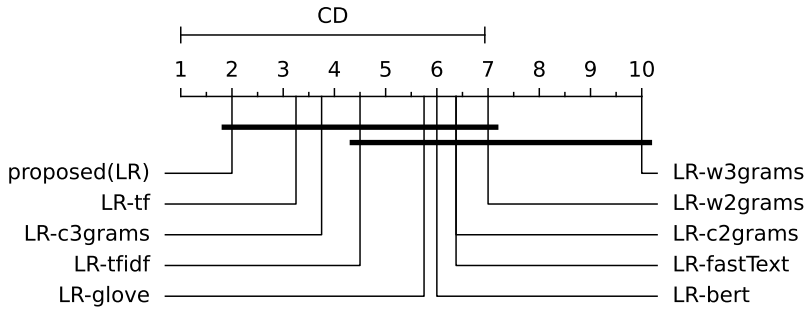
The proposed method obtained better classification performance and bias mitigation results for the WH dataset than with the best monolithic classifier evaluated. Even though the proposed method obtained higher macro F-score results than the LR classifier, it presented a slightly higher gender bias for the WS dataset. For the DV dataset, the proposed method presented the same macro F-score result as the LR classifier. Although the proposed method presented a slightly inferior macro F-score for the HE dataset, it achieved better bias mitigation results. Furthermore, even though the pre-trained word embedding FastText had better classification performance for the HE dataset (see Table 10), the FPED and FNED metrics obtained higher values.' This behaviour of word embedding confirms the results in (BOLUKBASI et al., 2016), stating that even word embeddings trained with millions of data can present bias.

The WH and HE datasets presented the highest frequency of specific identity terms in the "sexism" and "hateful" labelled classes (see Section 3.4.1.1), respectively. We can infer that the disproportionate representation of identity terms in the training set influenced the performance of the monolithic models for particular identity terms in these datasets.

For statistical analysis of the proposed methodology performance and the monolithic clas-

sifiers, we used the Friedman statistic test, which shows a significant difference in the performance of the classifiers. Then, we performed a pairwise comparison using a post-hoc test. Figure 11 presents the CD diagram of the statistical test. The pairwise comparison between the models presented that the proposed model presents a better average rank than different monolithic classifiers. However, its performance there is not a significant difference of some classifiers. Even though the proposed methodology only presented a minor classification improvement in contrast with some classifiers, the main objective of the proposed model is to reduce the unintended gender bias without compromising the classification model performance.

Figure 11 – Graphical representation of the average rank for each model using the LR classifier over all datasets. For the HE dataset was used the proposed methodology results in only three features. The Bonferroni-Dunn post-hoc test computed the critical difference (CD). Horizontal lines connect techniques with no statistical difference. The best classifier is the one presenting the lowest average rank.



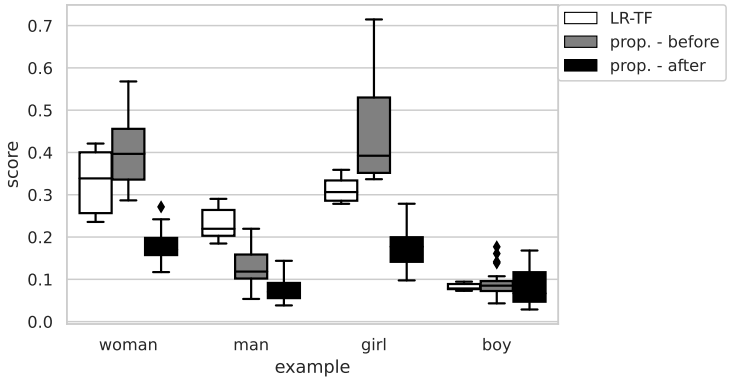
Source: Prepared by the author.

The unintended gender bias also has been investigated in the literature. (PARK; SHIN; FUNG, 2018) analysed three different strategies for gender bias mitigation (Debiased Word Embeddings, Gender Swap, and Bias fine-tuning) for the WS dataset. The methods analysed presented values between 0.006 and 0.333 for the FNED metric, and between 0.027 and 0.337 for the FPED, with different models and bias mitigation method combinations. However, the methods used have affected the classification performance evaluated with the AUC metric. Although our proposed model has presented FNED higher than the presented in (PARK; SHIN; FUNG, 2018), it has reduced the unintended gender bias without compromising the classification model performance.

3.6.4 Case Studies

This section evaluates the effectiveness of the proposed methodology using different pairs of examples from the proposed synthetic dataset. Table 14 presents the hateful score predicted by the classifiers for each pair of samples using the LR classifier trained on the WH dataset. This dataset was selected because it presented a higher disproportionate representation of identity terms (see Section 3.4.1.1). The examples presented are clearly non-hateful tweets. For instance, the first sample "*You are a great woman*", the Logistic Regression classifier with Term Frequency feature extractor (LR + TF) predicted the hateful label (score) of 0.33 while the proposed model after the bias mitigation gave the probability score of 0.18. We can infer that the significant frequency of particular identity terms in hateful comments and the imbalance nature of the training data used for hate speech detection can contribute to the increase of *false positive bias*, in which the model can give unreasonable high hateful score to the clearly non-hateful sentence due to the use of particular identity terms, similar to the reports in (DIXON et al., 2018).

Figure 12 – Case studies sentences predictions across the k-fold cross-validation. Logistic Regression classifier with Term Frequency feature extractor (LR-TF), proposed model before bias removal with original data (prop. - before), and proposed model after bias removal (prop. - after).



Source: Prepared by the author.

We also showed a boxplot (Figure 12) of these examples' hateful score for better visualisation because we collected the score across the k-fold cross-validation. Thus, we used the scores from the 20 executions. Each example is identified by the BSW used. The obtained results show the effectiveness of the proposed methodology. Even though using a simple method for bias mitigation, it performed well. Moreover, the proposed classifier demonstrated be less sensitivity to unintended gender bias than monolithic models.

Table 14 – Sentence predictions obtained by a monolithic classifier and the proposed method before and after the bias mitigation stage. The bias sensitive words are highlighted in bold. All examples are non-hateful. Bias Sensitive Words (BSWs).

		LR + TF	**Before bias removal**	**After bias removal**
BSW	**Examples**	***sexism***	***sexism***	***sexism***
woman	You are a great **woman**	0.33(0.067)	0.40(0.079)	0.18(0.038)
man	You are a great **man**	0.23(0.037)	0.13(0.044)	0.08(0.032)
girl	I am **girl**	0.31(0.027)	0.44(0.115)	0.17(0.044)
boy	I am **boy**	0.08(0.007)	0.09(0.036)	0.09(0.045)

Source: Prepared by the author.

3.7 CONCLUSIONS AND FUTURE WORK

In this paper, we have discussed how to identify and analyse bias mitigation, particularly toward gender identity terms, in the hate speech detection task, namely **unintended gender bias**. We have proposed a methodology based on two different modules to address the problem. In the first module, we proposed a gender bias mitigation strategy. Then, in the second module, a multi-view stacked classifier for hate speech detection. We selected nine different feature extraction methods, and we evaluated the proposed methodology with three base classifiers (LR, DT, and SVM).

Overall, the proposed classifier outperforms different models using several feature extractors using the WH, WS, and DV datasets. Furthermore, the proposed methodology reduced the unintended gender bias without compromising the performance of the WH dataset. The dataset presented the highest disproportionately in the representation of identity terms. Although some results are slightly inferior, the proposed methodology demonstrates effectiveness compared to state-of-the-art solutions.

It is relevant to highlight that the proposed multi-view stacked classifier is general enough to work with different feature extractors and classification models. Therefore, the proposed classifier can be extended and improve the classification results continuously.

In future work, we intend to explore complementary feature extraction techniques that better fitting for each dataset and newer ensemble learning strategies as dynamic selection methods (CRUZ; SABOURIN; CAVALCANTI, 2018). Furthermore, we also propose to investigate other strategies to select potential bias-sensitive words related to gender stereotypes. Although the proposed methodology focuses on gender terms, the method proposed can be expanded to work with other identity problems, such as racial stereotypes.

4 GENDER BIAS DETECTION ON HATE SPEECH CLASSIFICATION: AN ANALYSIS AT FEATURE-LEVEL

Francimaria RS Nascimento[1], George DC Cavalcanti[1], and Márjory Da Costa-Abreu[2]

[1]Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes s/n, Recife, Brazil

[2] Department of Computing, Sheffield Hallam University, Sheffield, UK

Article Published in ≪ SSRN ≫ submitted 2023.

Abstract

Hate speech is a growing problem on social media due to the larger volume of content being shared. Recent works demonstrated the usefulness of distinct machine learning algorithms combined with natural language processing techniques to detect hateful content. However, when not constructed with the necessary care, learning models can magnify discriminatory behaviour and lead the model to incorrectly associate comments with specific identity terms (e.g., woman, black, and gay) with a particular class, such as hate speech. Moreover, some specific characteristics should be considered in the test set when evaluating the presence of bias, considering that the test set can follow the same biased distribution of the training set and compromise the results obtained by the bias metrics. This work argues that considering the potential bias in hate speech detection is needed and focuses on developing an intelligent system to address these limitations. Firstly, we proposed a comprehensive, **unbiased dataset** to unintended gender bias evaluation. Secondly, we propose a framework to help analyse bias from feature extraction techniques. Then, we evaluate several state-of-the-art feature extraction techniques, specifically focusing on the bias toward identity terms. We consider six feature extraction techniques, including TF, TF-IDF, FastText, GloVe, BERT, and RoBERTa, and six classifiers, LR, DT, SVM, XGB, MLP, and RF. The experimental study across hate speech datasets and a range of classification and unintended bias metrics demonstrates that the choice of the feature extraction technique can impact the bias on predictions, and its effectiveness can depend on the dataset analysed. For instance, combining TF and TF-IDF with DT and MLP resulted in higher bias, while BERT and RoBERTa showed lower bias with the same classifier for the HE and WH datasets. The proposed dataset and source code will be publicly available when the paper is published[1].

[1] <https://github.com/Francimaria/hate_speech_bias_feature>

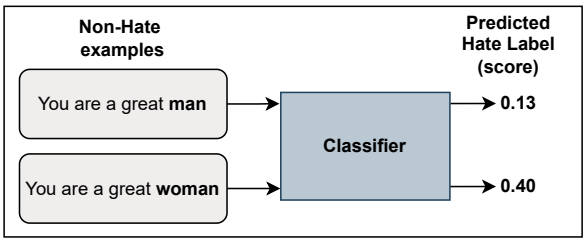
4.1 INTRODUCTION

Social media platforms power user-generated content about various subjects to spread quickly and easily. As a result, the easy dissemination of content and anonymity on social media platforms has facilitated online hate speech to proliferate. In (FORTUNA; NUNES, 2018), hate speech is defined as "*Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used*". The dissemination of hate speech on these platforms is potentially harmful and causes serious impacts on the victims. However, the enormous amount of content generated makes human moderation slow, expensive, and ineffective.

Recent studies have proposed several methods using distinct Machine Learning (ML) models, such as Deep Learning (DL) algorithms combined with Natural Language Processing (NLP) techniques to detect hate speech content automatically (BALOUCHZAHI et al., 2022; CRUZ; SOUSA; CAVALCANTI, 2022; KAPIL; EKBAL, 2020; SALMINEN et al., 2020; SENGUPTA et al., 2022). However, when badly designed, learning models can exhibit unintended unfair behaviours and lead the model to make decisions based on identity terms, such as woman, gay and black (ZHAO; ZHANG; HOPFGARTNER, 2022). As (DIXON et al., 2018) pointed out, "*a model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others*". Figure 13 shows an example of unintended bias from the model evaluated in (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022) before the bias mitigation procedures. This example illustrates the model's behaviour when it overgeneralises the association of a specific term ("woman") and the hate label. It results in a high probability of the model classifying as hate a non-hate sample. In this example, the classifier predicted the samples using the word "man" with a hate label score of 0.13, while the same example with the word "woman" with a higher score of 0.40.

The potential bias in learning models raises concerns regarding the robustness of the systems and the impact of this unintended bias on the generalisation of the systems in practical applications (BADJATIYA; GUPTA; VARMA, 2019; JAHAN; OUSSALAH, 2023). Different studies have exhibited bias associated with identity terms (e.g., lesbian, gay, transgender, and so on) in benchmark datasets (DIXON et al., 2018; BADJATIYA; GUPTA; VARMA, 2019). Moreover, racial and dialectic biases have been proven in trained classifiers for hate speech detection, as ev-

Figure 13 – Example of unintended bias in non-hateful tweets.



Source: Prepared by the author.

idenced by the correlation between words associated with African American English dialect (AAE) and the hate speech label (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; SAP et al., 2019). Studies also demonstrated the presence of gender bias in trained classifiers for hate speech detection (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022; PARK; SHIN; FUNG, 2018). Therefore, it is essential to consider the potential bias in the model development process since it can cause unfairness towards specific groups that these classifiers are usually designed to protect.

The development of machine learning models can lead to unintended bias at different stages (LEE; SINGH, 2021). An essential aspect of developing a machine learning solution for text classification is extracting meaningful features from data (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2023a). The feature representation used as input is a relevant factor contributing to a machine learning model's effectiveness. Several feature extraction techniques have been applied in the hate speech detection context, including methods based on Bag-of-Words (SENARATH; PUROHIT, 2020), lexical resources (NOBATA et al., 2016), and text embedding and deep learning approaches (CAO; LEE; HOANG, 2020). The embedding techniques have improved the classification performance for abusive and hate speech detection (CAO; LEE; HOANG, 2020; FOUNTA et al., 2019). However, a comparative study analysing the impact of feature extraction techniques for unintended bias in the classification of hate speech is still an open research question.

This problem's understanding is essential because it can help mitigate stereotypes in hate speech detection systems and related domains. For instance, there is an increasing number of applications based on textual content analysis, such as machine translation systems (WU et al., 2016), recommendation systems (KARN et al., 2023), and large language models like ChatGPT[2](Chat Generative Pre-trained Transformer), that can be influenced by biases as well. The bias can manifest in various forms and negatively impact the effectiveness of these systems.

[2] <https://openai.com/>

Therefore, it is essential to consider the potential bias when designing and implementing text-based technologies.

In this study, we investigate unintended bias, specifically related to gender identity (gender bias). Gender bias can result in the model showing preference or prejudice towards a particular gender. Its dissemination can reinforce harmful stereotypes in the systems, resulting in real-world consequences (SUN et al., 2019). For instance, concerns have been raised about sexist behaviours of Artificial Intelligence (AI) tools for resume filtering systems penalising women in the hiring process based solely on their gender[3] (DASTIN, 2018; DESHPANDE; PAN; FOULDS, 2020). Although these, few studies have addressed this issue related to the feature extraction technique in the hate speech detection context.

We performed a comprehensive analysis, considering six feature extraction methods TF (Term Frequency), TF-IDF (Term Frequency-Inverse Document Frequency), RoBERTa (Robustly Optimized BERT Pre-training Approach), BERT(Bidirectional Encoder Representations from Transformers), FastText, and GloVe(Global Vectors for Word Representation) used for feature extraction and different classification algorithms. To understand whether the feature extraction method impacts the unintended gender bias learned by the model and if this behaviour is followed in different datasets. Hence, this study aims to answer the following research questions: (1) Does the choice of the feature extraction technique impact the presence of unintended gender bias on the model's prediction? (2) Do feature extraction techniques tend to present bias when dealing with different datasets? (3) Can the bias affect the performance of the models? Experiments on three real-world English datasets for hate speech detection demonstrate that feature extraction techniques can impact the bias on predictions. Moreover, we explored the behaviour of the feature extraction techniques using several classifiers with various metrics. It allows us to explore different nuances of the bias problem.

The main contributions of this paper are:

1. The design of a framework to help analyse the biased behaviour of feature extraction techniques.

2. The proposal of an unbiased dataset for assessing unintended gender bias in the context of hate speech detection, while existing studies mainly focused on debiasing datasets. This dataset comprises all identity terms in the same context to ensure a fair and unbiased evaluation.

[3] https://www.bbc.co.uk/news/technology-45809919

3. Our experiments show that feature extraction techniques can impact unintended gender bias in predictions. For instance: TF and TF-IDF presented more biased behaviour than FastText, GloVe, BERT, and RoBERTa for the FPED and FNED metrics.

Thus, we aim to achieve these contributions by presenting our work which is organised as follows: Section 4.2 presents related work. Section 4.3 discusses the proposed methodology and the proposed unbiased dataset. The experimental setups are described in Section 4.4. Section 4.5 presents the results. Section 4.6 provides a comprehensive discussion. Section 4.7 concludes the work with the final remarks.

4.2 RELATED WORK

Several approaches for hate speech detection have been proposed based on classic machine learning models (SALMINEN et al., 2020; SENARATH; PUROHIT, 2020), ensemble learning (MAZARI; BOUDOUKHANI; DJEFFAL, 2023), and deep learning techniques (KAPIL; EKBAL, 2020) combined with different techniques for feature extraction. General feature extraction techniques for text mining have been applied to the hate speech detection problem. The word embedding methods have been more frequently used than classical methods, such as bag-of-words (BoW) and n-grams. These techniques can capture semantic information from the text and the syntactic relationship between the words (INDURTHI et al., 2019).

Cruz, SOUSA and Cavalcanti (2022) proposed a framework that evaluates and selects multiple feature extraction techniques and classification models that complement each other to design a robust multiple classifier system. The authors demonstrated the effectiveness of the proposed methodology in four hate speech datasets.

Even though these contributions have improved the performance of hate speech detection models, they did not explicitly consider the potential bias in the models. Dixon et al. (2018) introduced the concept of unintended bias in the toxicity language datasets. The authors investigate unintended bias regarding identity terms (atheist, gay, transgender, etc.) and try to mitigate the bias using statistic correction to balance the data with the most disproportional distribution. In (ZHAO; ZHANG; HOPFGARTNER, 2022), the authors investigated the subjectivity level of a comment and the presence of identity terms to mitigate its bias.

Gender stereotypes hosted in hate speech datasets are also a serious concern, in which a model can perform better with determinate gender identity terms than comments with others

(PARK; SHIN; FUNG, 2018). In (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022), the authors proposed a multi-view ensemble learning approach to learn distinct abstractions of the problem and found effective results compared to the literature. In (ŞAHINUÇ et al., 2022), the authors analysed the effect of debiased embedding for mitigating gender bias in English and Turkish tweets. They concluded that the classification performance of hate speech detection models based on neural embeddings could be improved by removing the gender-related bias.

Table 15 summarises the related works that address hate speech detection and investigate concepts related to bias. This table shows the reference of the paper and its publication year, the datasets, the feature extraction technique, and the classifiers evaluated. The column “gender bias" denotes whether the work considered the unintended gender bias in the proposed model, and the column “unbiased test set" indicates whether the dataset used to evaluate the bias follows an impartial data distribution in the context applied.

Despite the previous contributions to hate speech detection, the potential biases did not receive attention in most works (INDURTHI et al., 2019; SALMINEN et al., 2020; SENARATH; PUROHIT, 2020; KAPIL; EKBAL, 2020; CRUZ; SOUSA; CAVALCANTI, 2022; MAZARI; BOUDOUKHANI; DJEFFAL, 2023). Some efforts have investigated related bias concepts in hate speech detection, addressing the analysis only considering one input vector or features from the same families (DIXON et al., 2018; ZHAO; ZHANG; HOPFGARTNER, 2022). Although the analysis of different features has been performed in some studies (PARK; SHIN; FUNG, 2018; ŞAHINUÇ et al., 2022; NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022), these studies usually investigate the impact of the proposed bias mitigation model without considering the potential bias introduced by the original feature.

Moreover, none of these works presents a clear methodology for analysing the impact of unintended gender bias from multiple feature representations and how it affected the performance of classifiers. To fill this gap, this paper proposed a methodology for analysing the relationship between the feature extraction technique and the unintended bias in different hate speech datasets. The proposed methodology is presented in the following section.

Table 15 – Related works summary.

Year	Ref.	Dataset	Feature	Classifier	Gender Bias	Unbiased test set
2018	(DIXON et al., 2018)	Wikipedia Talk Pages	Convolutional Neural Network (CNN)	CNN	×	✓
2018	(PARK; SHIN; FUNG, 2018)	Twitter WH (WASEEM; HOVY, 2016) , FN (FOUNTA et al., 2018)	Word2Vec, FastText, randomly initialised embeddings (random)	CNN,Gated Recurrent Units (GRU), α-GRU	✓	✓
2019	(INDURTHI et al., 2019)	Twitter HE (BASILE et al., 2019)	InferSent, Concatenated Power Mean Word Embedding, Lexical Vectors, Universal Sentence Encoder,Embeddings from Language Model (ELMo)	Logistic Regression(LR), Random Forest(RF), Support Vector Machines - Radial Basis Function(SVM-RBF), Extreme Gradient Boosted Decision Trees (XGBoost)	×	×
2020	(SALMINEN et al., 2020)	YouTube (SALMINEN et al., 2018), Reddit (ALMEREKHI et al., 2019), Wikipedia (WULCZYN; THAIN; DIXON, 2017), Twitter dataset (DAVIDSON et al., 2017)	BoW,Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, BERT, and all combined	LR, Naïve Bayes, , XGBoost, and Neural Networks	×	×
2020	(SENARATH; PUROHIT, 2020)	WHO, 2021 Twitter DV (DAVIDSON et al., 2017), FN (FOUNTA et al., 2018)	BoW, TF-IDF, n-grams, dictionary (Hatebase), FrameNet, Word2Vec	SVM	×	×
2020	(KAPIL; EKBAL, 2020)	Twitter DV (DAVIDSON et al., 2017), WH (WASEEM; HOVY, 2016), Hindi–English, OLID (ZAMPIERI et al., 2019), Harassment (GOLBECK et al., 2017)	word and char embeddings, CNN	Deep Multi-task Learning (MTL), CNN, LSTM, stacking of CNN+GRU, and CNN$_\alpha$+GRU	×	×
2022	(CRUZ; SOUSA; CAVALCANTI, 2022)	Twitter DV (DAVIDSON et al., 2017), WH (WASEEM; HOVY, 2016), HE (BASILE et al., 2019), DV + WH	GLoVe, Word2Vec, FastTex, Term-Frequency (TF), TF-IDF	ensemble learning	×	×
2022	(ZHAO; ZHANG; HOPFGARTNER, 2022)	Stormfront (GIBERT et al., 2018), Twitter WH (WASEEM; HOVY, 2016), FN (FOUNTA et al., 2018) , Kaggle-Wikipedia[4]	BERT	Subdentity-BERT (SSBERT)	×	×
2022	(NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022)	Twitter WH (WASEEM; HOVY, 2016), WS (WASEEM, 2016), DV (DAVIDSON et al., 2017), HE (BASILE et al., 2019)	GloVe, FastText, BERT, TF, TF-IDF, char and word n-grams	ensemble learning	✓	✓
2022	(ŞAHINUÇ et al., 2022)	Twitter (TORAMAN; ŞAHINUÇ; YILMAZ, 2022)	BoW, FastText, BERT	SVM, BiLSTM	✓	✓
2023	(MAZARI; BOUDOUKHANI; DJEFFAL, 2023)	Kaggle-Wikipedia[5]	GloVe, FastText, BERT	BERT-based ensemble learning	×	×
2023	**Our**	Twitter DV (DAVIDSON et al., 2017), WH (WASEEM; HOVY, 2016), HE (BASILE et al., 2019)	TF, TF-IDF, RoBERTa, GloVe, FastText, BERT	SVM, LR,Decision Tree (DT), XGBoost, Multi-Layer Perceptron (MLP), RF	✓	✓

Source: Prepared by the author.

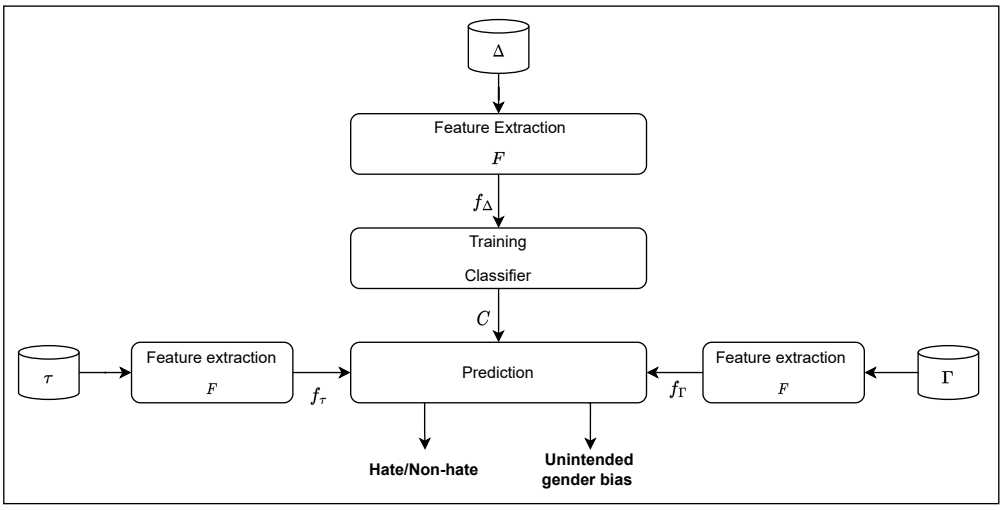
4.3 PROPOSED METHODOLOGY AND UNBIASED DATASET

This work investigates the relationship between the feature extraction technique and the unintended gender bias measured in the predictions of state-of-the-art machine learning techniques. Hate speech detection models are usually designed to classify the data in binary labels as Hate/Non-hate or multi-classes, and the model performance is calculated using the predictions from a test set. However, it is essential to consider the potential bias in the trained model against identity terms.

It is crucial to remark that while the original test set can be used to assess traditional metrics such as accuracy, it should not be evaluated to assess the bias since it may have the same biased distribution of identity terms as the training set, making bias identification challenging.

Therefore, a dataset with all identity terms in a similar context, the **unbiased dataset**, is necessary to properly evaluate the bias metrics, as these metrics depend on the identity term information. Considering the relevance and necessity of this dataset with these specific characteristics, we proposed a new unbiased dataset described in Section 4.3.1.

Figure 14 – Proposed methodology. Δ, τ, and Γ are the training, test, and unbiased test sets. F is the feature extractor. f_Δ, f_τ, and f_Γ are the matrices generated by F using the training, test, and unbiased test sets, respectively. C is the trained classifier.



Source: Prepared by the author.

Figure 14 presents the proposed framework. The proposed comprises three main stages: Feature extraction, Training, and Prediction, which receive three datasets: training (Δ), test (τ), and unbiased test (Γ) as input. Thus, given a training set Δ with text in natural language, the feature extractor F transforms the text in numeric feature spaces f_Δ. The training set's

data representation (f_Δ) is used to train a classifier. The trained classifier C predicts the classes from the unbiased test set numeric feature spaces f_Γ generated using F for bias evaluation. Then, C predicts the classes from the test set numeric feature spaces f_τ also generated using F for classification performance evaluation. The outputs are the hate/non-hate prediction accuracy computed using τ and the unintended bias assessment using Γ.

Feature extraction. In the context of hate speech, datasets are usually available as raw text for analysis. Therefore, feature extraction aims to transform the natural language text into a numerical vector space suitable as model inputs. Several feature extraction techniques can be applied, such as Bag-of-Words (BoW) techniques (ALMATARNEH et al., 2019; CORAZZA et al., 2020), lexical resources (GITARI et al., 2015), and embedding methods (MAZARI; BOUDOUKHANI; DJEFFAL, 2023; MOZAFARI; FARAHBAKHSH; CRESPI, 2020). The feature extractor F transforms the raw text in numerical feature spaces, and each dataset is represented as a set of feature vectors denoted by f_Δ, f_τ, and f_Γ.

Training. The training process is responsible for the learning task, where the input vector (f_Δ) and its respective labels are used to fit a classification model, resulting in the trained classifier C. Several classifiers, such as classical machine learning and deep learning algorithms, can be applied to this task. We investigate a diverse set of state-of-art classifiers.

Prediction. As shown in Figure 14, the prediction module receives as input the trained classifier C and the feature vectors f_τ and f_Γ, which represent the datasets τ and Γ, respectively. Thus, the same model C is used to predict the instances from two datasets; C predicts the hateful labels from f_τ and the predictions from f_Γ are used to calculate the unintended gender bias metrics since the test set τ may exhibit a biased distribution similar to the training set, which can impact the bias evaluation, as highlighted in (DIXON et al., 2018).

4.3.1 Unbiased dataset

In the field of machine learning, it is common to split the dataset into at least two parts: training and testing. The evaluation of the models is performed using the test dataset. However, this dataset will likely follow the biased distribution of identity terms from the training set, so it should not be considered to evaluate the unintended bias accurately (DIXON et al., 2018; NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022). To overcome this issue, a relevant strategy is to create a dataset with equal comment distribution for each term using the identity term template. Moreover, the bias metrics that are based on the difference in error rate equality

need all identity terms in the same context. Therefore, we proposed a dataset with equal comment distribution for each term using the identity term template.

Table 16 shows examples of the identity term templates used. The idea is to generate sentences that change only the identity term, for example “you are a good woman" and “you are a good man". Therefore, a model without bias is expected to generalise well and generate similar predictions for the same sentence, modifying only the identity term. The availability of reliable labels across a range of terms is crucial in assessing bias effectively. Therefore, to generate a more extensive data sample and cover different expressions of hate, we investigated the combination of the dataset proposed in (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022) and the HateCheck dataset (RÖTTGER et al., 2021).

Table 16 – Examples of templates. $< identity >$ denotes an identity term.

Example template	**Label**
$< identity >$ think they're special. I hate them.	hateful
$< identity >$ should all be killed.	hateful
I hate all $< identity >$	hateful
$< identity >$ should be protect.	non-hateful
We should not let $< identity >$ suffer.	non-hateful
I am $< identity >$	non-hateful

Source: Prepared by the author.

In (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022), the authors proposed templates filled with 24 identity terms (Table 17) to generate 1,248 samples, of which 600 are hateful, and 648 are non-hateful. The HateCheck is a functional hate speech test dataset, meaning that this dataset is designed to test different functionalities of a model with targeted test cases, which was validated through a structured annotation process. The dataset comprises different cases of hate speech that can be particularly challenging cases for classification models. It is composed of 421 templates using identity terms. We filled the templates using a list of identity terms (Table 17) to ensure all identity terms appear in the same context.

Table 17 – Identity terms. The word 'female' was spelled as 'femal' due to the pre-processing step.

Identity terms
woman, women, girl, sister, daughter, wife, girlfriend, mother, aunt, mom, grandmother, femal, man, men, boy, brother, son, husband, boyfriend, father, uncle, dad, grandfather, male

Source: Prepared by the author.

However, we used only 400 templates because some changed only the identity term style since we did not distinguish the templates by the identity term style, e.g. plural (<

$identity\_p >$), singular ($< identity\_s >$), plural with spaces ($< identity\_p\_space\_add >$), and so on. For example, the templates "$< identity\_p >$ are disgusting." and "$< identity\_p\_space\_add >$ are disgusting." are considered as "$< identity >$ are disgusting.". As the idea is to build sentences by changing only the identity term, these templates can generate repeated samples. It resulted in 9,600 samples, of which 7,296 are hateful and 2,304 non-hateful. The combination of the datasets resulted in a dataset with 10,848 examples. However, some examples presented the same text content after the pre-processing step (described in Section 4.4.2). Therefore, we remove these examples resulting in 10,728 instances, of which 7,776 are hateful and 2,952 non-hateful. The proposed dataset will be publicly available when the paper is published[6].

4.4 EXPERIMENTAL METHODOLOGY

This section describes the experimental setup used in this study. We describe the datasets, pre-processing steps, feature extraction, training classifier, evaluation metrics, and parameter settings.

The Python programming language was used to conduct the experiments on a computer with an processor Intel Core i7-10510U CPU @ 1.80 GHz x 4, 15.3 GiB of memory, and an Intel Corporation UHD Graphics card.

4.4.1 Datasets

Table 18 summarises the datasets selected. We analyse three (WH, DV, and HE) widely-used English Twitter datasets to evaluate the proposed methodology. Furthermore, considering that the test set from the original dataset can hold the same biased distribution as the training set and affect the bias evaluation (DIXON et al., 2018), we use an unbiased dataset (**UB**) (described in Section 4.3.1) for bias evaluation because this dataset includes all identity terms in the same context, ensuring non-bias towards identity terms.

Waseem-Hovy (WH)(WASEEM; HOVY, 2016): The corpus has more than 16k samples collected from Twitter. The initial search used a list of potential hateful terms and phrases[7]. The authors manually annotated the dataset based on guidelines inspired by critical race theory. The annotation was reviewed by "*a 25-year-old woman studying gender studies and a*

[6] <https://github.com/Francimaria/hate_speech_bias_feature/tree/main/dataset/UB>

non-activist feminist". The dataset consists of tweets labelled as sexist, racist or neither.

Davidson (DV) (DAVIDSON et al., 2017): The authors used a lexicon from *Hatebase.org* to search the tweet and extracted the timeline for each user. They then selected random samples, and the CrowdFlower (CF) workers manually annotated. They labelled the corpus as hate speech, offensive or neither (neither offensive nor hate speech). The authors instructed the CF workers to consider the words and the inferred context to avoid false positives in this process. The final dataset has resulted in 24,802 labelled tweets.

HatEval (HE) (BASILE et al., 2019): The HatEval dataset is a multilingual corpus for hate speech detection against women and immigrants. The authors used different gathering strategies based on previous studies proposed in the literature to collect the dataset. Figure Eight (F8) workers and two experts annotated the dataset labelled based on majority voting. The final dataset comprises 19,600 tweets, 6,600 for Spanish and 13,000 for English. However, we used only English tweets. The data was annotated based on three categories: first, Hate Speech (hateful and non-hateful); second, Target Range (individual target and generic target); and Aggressiveness (aggressive or non-aggressive).

We selected those datasets because they address different nuances of hate speech problems, such as sexism, racism, and xenophobia. Moreover, they have distinct data collection and annotation processes. Therefore, we can use a diverse set of datasets to evaluate the proposed methodology under different hate-speech detection scenarios.

We conducted our experiments using stratified 5-fold cross-validation to divide the WH and DV datasets in 4 folds for training (Δ) and 1 fold for testing (τ). So, we used 15% of the training set as the validation set for the classifiers' parameter tuning. This strategy is used to compute the mean and standard deviation of the results achieved and thus help us find more precise estimators of the model performance (CRUZ; SOUSA; CAVALCANTI, 2022). Moreover, we used the stratified version of cross-validation to ensure the proportion of each class is represented as in the original dataset across each fold to avoid class bias.

For the HE dataset, we used the original training (Δ), validation, and testing (τ) division used in the competition (BASILE et al., 2019). For the unbiased dataset (UB), we used the complete dataset as a test set (Γ) to evaluate the bias on predictions.

[7] Terms queried for: "MKR", "asian drive", "feminazi", "immigrant", "nigger", "sjw", "WomenAgainstFeminism", "blameonenotall", "islam terrorism", "notallmen", "victimcard", "victim card", "arab terror", "gamergate", "jsil", "racecard", "race card".

Table 18 – Summary of datasets.

Name	Available	Tweets	Label (%)	Target	Annotator
WH	GitHub repository	16,906	sexism (20%) racism (12%) neither (68%)	sexism, racism	1
DV	GitHub repository	24,783	hate (6%) offensive (77%) neither (17%)	general	3 or more
HE	GitHub repository	13,000	hate (43%) non-hate (57%)	misogyny, xenophobia	3
UB	GitHub repository	10,728	hate (76%) non-hate (24%)	general	-

Source: Prepared by the author.

4.4.2 Pre-processing

In the context of Twitter, the text often contains elements such as URLs, hashtags, slang, mentions, RT, etc. This content can raise noise in the classification task (ASIRI et al., 2022; DESOUZA; DA-COSTA-ABREU, 2020; MOZAFARI; FARAHBAKHSH; CRESPI, 2020). Therefore, we performed different pre-processing criteria to clean our model's input for clarity and generality. It includes: converting all text to lowercase, remove the mentions ("i.e.,$@user$"), URLs (which start with "$http[s] : //$"), RT symbols(Retweet), numbers, punctuation marks, stopwords, and redundant white spaces, and stemming the text to reduce word flexion.

4.4.3 Feature extraction

For feature extraction, we considered six methods in this study:

- **TF**: Term Frequency (TF) (FARHANGIAN; CRUZ; CAVALCANTI, 2024; PLAZA-DEL-ARCO et al., 2020), also called of count vectorizer, represents the textual features based on the occurrence and frequency of words in a document. This feature extraction method is relatively simple. However, in the case of large textual datasets, the representation matrix can become exceedingly sparse, necessitating a significant computational effort. Our work used the maximum features equal to 2000, as used in the literature (KUMARI; JAMATIA, 2022). We used the implementation from the Scikit-learn Python library (PEDREGOSA et al., 2011).
- **TF-IDF**: Term Frequency-Inverse Document Frequency (TF-IDF) (SHMUELI et al., 2017)

is a widely used statistical feature representation technique from textual data. This method scores and weighs words that occur in a document. The primary objective is to identify the relevant and significant words that impact the document most meaning and relevance. The length of the feature vector depends on the document vocabulary size, and the representation matrix can become sparse as with the TF method. Our work used the maximum features equal to 2000, as used in the literature (KUMARI; JAMATIA, 2022). We used the implementation from the Scikit-learn Python library (PEDREGOSA et al., 2011).

- **GloVe**: GloVe, an acronym for Global Vectors for Word Representation (PENNINGTON; SOCHER; MANNING, 2014), learns word representations by incorporating global statistics (word-word co-occurrence counts). In essence, it is a global log-bilinear model with a weighted least-squares objective for the unsupervised learning of word representations. Our work used GloVe embeddings trained on Twitter data (2B tweets, 27B tokens, 1.2M vocab, uncased) with a feature dimension of 200. We used the implementation from the Zeugma library[8].

- **FastText**: FastText model (BOJANOWSKI et al., 2017) learns word representations based on character n-grams, which assumes that each word is the sum of the n-grams vectors. Thus, considering subword information that helps the model build word vectors for out-of-vocabulary words. For the current work, we use the FastText embedding with a feature dimension of 300 (implementation from the Zeugma library) pre-trained with 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).

- **BERT**: BERT, an abbreviation for Bidirectional Encoder Representations from Transformers (DEVLIN et al., 2019). The BERT is a pre-trained embedding method defined in two models, BERT$_{BASE}$ and BERT$_{LARGE}$, both with uncased (only lowercase letters) and cased versions. The BERT$_{LARGE}$ model consist of 24 layers, 16 attention heads, and 340 million parameters and the BERT$_{BASE}$ model consists of 12 layers, 12 self-attention heads, and 110 million parameters. We selected the pre-trained BERT$_{BASE}$ uncased because the training process of a BERT model is computationally expensive. Furthermore, it has presented promising results for hate speech detection in (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; RISCH; KRESTEL, 2020; SALMINEN et al., 2020). We

[8] https://zeugma.readthedocs.io/en/latest/

used the implementation from Transformers library (WOLF et al., 2020) with a feature dimension of 768.

- **RoBERTa**: RoBERTa, an acronym to Robustly Optimized BERT Pre-training Approach (LIU et al., 2019b). RoBERTa is a language model developed based on BERT architecture. This model was designed to improve its results by adjusting key hyperparameters of the BERT model, such as longer sequences, changes in the length of batch size, and removal of the next sentence prediction objective. We selected the pre-trained RoBERTa$_{BASE}$ for this study. We used the implementation from Transformers library (WOLF et al., 2020) with a feature dimension of 768.

4.4.4 Training classifier

We selected various classification algorithms to evaluate the different feature extraction techniques. Our objective is to analyse different classifiers to investigate if the techniques' biased behaviour is generalised for a wide range of classification algorithms.

This study includes the following algorithms in the experiments: Support Vector Machine (SVM) (CORTES; VAPNIK, 1995), Logistic Regression Classifier (LR), Decision Tree Classifier (DT), Extreme Gradient Boosted Decision Trees (XGBoost) (CHEN; GUESTRIN, 2016), Multi-Layer Perceptron Neural Network (MLP) (AGGARWAL et al., 2018), and Random Forest (RF).

4.4.5 Evaluation

We assessed all methods using distinct evaluation metrics to provide different performance perspectives. The objective is to analyse the possible advantages and limitations of each technique. Table 19 summarises the selected metrics for bias and classification evaluation.

Regarding the unintended bias evaluation, we investigated different metrics widely used in the literature (NASCIMENTO; CAVALCANTI; COSTA-ABREU, 2022; PARK; SHIN; FUNG, 2018). These metrics measure the bias based on the outputs of the algorithms. We selected a threshold agnostic metric derived from the ROC-AUC (or AUC) metric (BORKAN et al., 2019), called, **subgroup AUC**. To facilitate the assessment of the bias in the context of our analysis, we measure the average across all identity terms. The equation for the subgroup AUC is defined in Equation 4.1.

$$subgroup\ AUC = \frac{1}{|T|} \sum_{t \epsilon T} AUC(D_t^- + D_t^+) \tag{4.1}$$

where D_t^- denotes the negative examples (non-hate speech) and D_t^+ the positive one (hate speech) that mention the identity term $t\ \epsilon\ T$, where $T = [woman, ..., male]$ (complete list in Table 17) and $|T|$ denotes the number of identity terms in T.

The subgroup AUC measures the model performance of each subset that mentions a specific identity term, so we compute the average value of these results. Therefore, low results indicate that the model had difficulty distinguishing the labels of the samples in the context of identity terms.

In (BORKAN et al., 2019), the authors also proposed other metrics based on AUC with different objectives. But we decided to use only the subgroup AUC since this paper focuses on investigating the feature extraction biased behaviour against identity terms. Moreover, we measured the average value for the subgroup AUC across all identity terms.

In addition, we also used two metrics based on the **Error Rate Equality Difference** introduced in (DIXON et al., 2018). The **False Positive Equality Difference (FPED**) and **False Negative Equality Difference (FNED**) defined in Equations 4.2 and 4.3, respectively.

$$FPED = \frac{1}{|T|} \sum_{t \epsilon T} |FPR - FPR_t| \tag{4.2}$$

$$FNED = \frac{1}{|T|} \sum_{t \epsilon T} |FNR - FNR_t| \tag{4.3}$$

The FPED (or FNED) computes the sum of the difference between the False Positive Rate (FPR) or False Negative Rate (FNR) on the complete dataset and each subset containing a specific identity term, FPR_t and FNR_t. As for the AUC subgroup, we also calculate the average value to normalise the metric values between 0 and 1. To facilitate the understanding and contrast of different metrics.

The FPED and FNED measure the bias based on the error rate equality differences. Therefore, a model without unintended bias is expected to present similar values across all terms, where $FPR = FPR_t$ and $FNR = FNR_t$ for all identity terms. The wide divergence in these values across the identity terms suggests a high unintended bias, so the best result is zero.

On the other hand, a partial objective of this experiment is to evaluate the classification performance. Therefore, we evaluated the general performance of the models using the macro F-score and Area under the ROC curve (AUC).

F1, or F-score, is measured based on the Precision and Recall harmonic mean, which are defined as in Equations 4.4 and 4.5. The number of instances correctly classified is defined as TP (True Positives) and TN (True Negatives). In contrast, FP (False Positives) and FN (False Negatives) represent the number of those incorrectly classified. Then, F1 can be defined as in Equation 4.6.

$$Precision = \frac{TP}{TP + FP} \tag{4.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.5}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.6}$$

The F1, usually in multi-class problems, can be aggregated using micro or macro averages. In this paper, we selected the macro-average due to the imbalance nature observed in the hate speech datasets evaluated. In imbalanced datasets, the micro-averaging can mask the model performance for minority classes (CHARITIDIS et al., 2020).

The AUC is computed as the area underneath the receiver operating characteristic curve. This probability curve plots the True Positive Rate (synonym for recall) against the False Positive Rate (FPR), defined in Equation 4.7, at various threshold values from 0 to 1. The AUC is designed for binary classification problems. However, we can use it for multiclass problems using the One-vs-Rest technique. It computes the AUC of each class against the rest.

$$FPR = \frac{FP}{FP + TN} \tag{4.7}$$

Table 19 – Summary of the selected metrics.

Evaluation	**Metric**	**Meaning**
Bias	subgroup AUC	Compute the AUC from examples with identity terms
	FPED	False Positive Equality Difference
	FNED	False Negative Equality Difference
Classification	F1	Harmonic mean of the Precision and Recall
	AUC	Area under the ROC curve

Source: Prepared by the author.

4.4.6 Parameters setting

Table 20 presents the parameters considered in this study for each classification model. We selected the best set of hyperparameters using grid-search and the macro F1-score as evaluation metric. The classifier was trained with the training set, and its performance was measured with the validation set. The column Library defines the library of each model and its version. The parameters for each classifier are in the Github repository [9].

Table 20 – Enumeration of parameters used throughout the experiments.

Method	**Hyperparameters**	**Library**
SVM	kernel = [linear, rbf]	sklearn v1.2.2[10]
LR	penalty = [l1, l2]	sklearn v1.2.2
MLP	activation = [relu, logistic]	sklearn v1.2.2
DT	criterion = [gini, entropy]	sklearn v1.2.2
XGBoost	n_estimators = [50,100]	xgboost 1.7.5[11]
RF	n_estimators = [50,100]	sklearn v1.2.2

Source: Prepared by the author.

Regarding the MLP, it requires defining the network architecture. Therefore, we use a standard architecture with a single hidden layer containing 100 neurons as in (CRUZ; SOUSA; CAVALCANTI, 2022).

4.5 EXPERIMENTAL RESULTS

This section presents the experimental results aiming to answer the research questions defined in the Introduction section. The experiments evaluate the feature extraction techniques to analyse the unintended gender bias on the predictions using an unbiased test set (Section 4.5.1) and to investigate the bias impact on the classification performance using the test set (Section 4.5.2). In addition, the results with the standard deviation for all metrics are available in the supplementary material.

[9] <https://github.com/Francimaria/hate_speech_bias_feature>
[11] Scikit-learn Python library (PEDREGOSA et al., 2011)
[11] <https://xgboost.readthedocs.io/en/stable/install.html>

4.5.1 Unintended gender bias

To answer the research question **RQ1** – Does the choice of the feature extraction technique impact the presence of unintended gender bias on the model predictions? – for each dataset, we compared the results of the feature extraction techniques using the unbiased test dataset. As mentioned previously, the unbiased test dataset uses the strategy of identity term templates to generate a data sample where all identity terms appear in the same context to evaluate the unintended bias from identity terms.

The unbiased test dataset was labelled as hate and non-hate. Therefore, to analyse the unintended bias metrics, we consider the predictions of “racism" and “sexism" as “hate" and “neither" as “non-hate" for the WH dataset. For the HE dataset, we did not perform modifications. For the DV dataset, we assume “hate" and “offensive" as “hate" and “neither" as “non-hate", as in related work (SALMINEN et al., 2020)

In all tables, we abbreviated the name of the classifier as XGB – XGBoost. We highlighted the best results in bold for each classifier and underlined ties. For each dataset, we compare the feature extractors per classification model using the Wilcoxon statistical test, and the significantly better result is marked with $*$. The significance level adopted was 0.05. We selected the Wilcoxon statistical test because, as stated in (DEMŠAR, 2006), this test is robust for pairwise comparison between models.

Table 21 presents the results obtained with the FNED metric, which measures the bias based on the false negative rate, in which the closer the result is to zero, the lower the bias. For the HE dataset, it is important to note that the FastText demonstrated more biased behaviour when combined with MLP, which obtained 0.214, and the TF with DT achieved 0.223. For the WH dataset, the classifiers presented more biased behaviour with GloVe and FastText, especially the LR classifier, which found results bigger than 0.20 when combined with these feature extractors. Moreover, for WH dataset, presented more bias on prediction with five of the six classifiers analysed. For the DV dataset, TF and TF-IDF presented more biased results for the DV dataset when combined with LR, SVM, and MLP, finding results bigger than 0.20. These results evidenced that specific hate speech samples were considered non-hate speech when the samples contained some identity terms but not others.

Table 22 presents the results obtained with the FPED metric, which measures the bias based on the false positive rate. For this metric, as for FNED, results closer to zero present less bias. The BERT and RoBERTa presented less biased behaviour for all datasets evaluated

Table 21 – Results obtained using FNED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	**0.041**	0.223	**0.025**	0.000	0.204	0.111
TF-IDF	0.142	0.227	0.154	0.000	0.231	0.122
GloVe	0.175	**0.034**	0.139	0.105	0.158	0.052
FastText	0.132	0.065	0.137	0.082	0.214	0.062
BERT	0.131	0.037	0.069	0.042	0.125	0.026
RoBERTa	0.053	0.037	0.069	0.031	**0.115**	**0.016**

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.084	0.215	0.098	0.000	0.167	0.181
TF-IDF	0.193	0.209	0.128	0.000	0.238	0.204
GloVe	0.250	0.069	0.198	0.128	0.138	0.067
FastText	0.200	0.084	0.159	0.142	0.174	0.085
BERT	0.131	0.051	**0.057***	0.028	0.113	**0.017**
RoBERTa	**0.083**	0.051	0.078	0.036	**0.098**	0.020

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.237	0.107	0.130	0.000	0.245	0.081
TF-IDF	0.132	0.075	0.215	0.000	0.217	0.079
GloVe	0.152	0.058	0.133	0.084	0.129	0.106
FastText	**0.067**	0.057	0.078	0.060	0.097	0.074
BERT	0.105	**0.036**	**0.065***	0.037	**0.081**	0.050
RoBERTa	0.101	0.051	0.103	0.068	0.124	**0.041***

(c) DV dataset

Source: Prepared by the author.

for most classifiers. These results were statistically better for the WH and DV datasets. As for the metric FNED, the DT classifier presented more bias in predictions for the HE dataset with the TF and TF-IDF (0.228 and 0.237) and the MLP with TF-IDF (0.190). Contrasting these results with those obtained through BERT and RoBERTa with DT, MLP, and RF, it is possible to note that these feature extractors present almost twice the result for HE and WH datasets. These results evidenced that TF and TF-IDF, combined with DT, MLP, and RF, consider more non-hate samples as hate when the samples contain some identity terms but not others.

Table 23 presents the results obtained with the subgroup AUC metric that measures the

Table 22 – Results obtained using FPED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.028	0.228	**0.009**	0.000	0.173	0.103
TF-IDF	0.095	0.237	0.096	0.000	0.190	0.118
GloVe	0.182	0.053	0.103	0.072	0.108	0.036
FastText	0.125	0.052	0.097	0.068	0.174	0.049
BERT	0.094	**0.035**	0.028	0.018	0.067	0.015
RoBERTa	**0.021**	0.047	0.031	0.025	**0.064**	**0.012**

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.092	0.212	0.103	0.000	0.167	0.166
TF-IDF	0.184	0.209	0.126	0.000	0.217	0.194
GloVe	0.251	0.078	0.215	0.117	0.138	0.065
FastText	0.232	0.094	0.162	0.151	0.171	0.083
BERT	0.112	**0.043**	**0.024***	0.016	0.095	**0.013***
RoBERTa	**0.070***	0.055	0.070	0.039	**0.088**	0.023

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.243	0.115	0.123	0.000	0.263	0.077
TF-IDF	0.138	0.076	0.212	0.000	0.232	0.077
GloVe	0.206	0.072	0.160	0.106	0.151	0.115
FastText	**0.084**	0.062	0.105	0.065	0.120	0.078
BERT	0.117	**0.045**	**0.081***	0.048	**0.102**	0.064
RoBERTa	0.128	0.064	0.127	0.077	0.163	**0.047***

(c) DV dataset

Source: Prepared by the author.

classifier performance in the context of identity terms. For the HE dataset, FastText presented the best results for SVM, XGB, and RF. Moreover, it is relevant highlight for this dataset GloVe with MLP (0.572) found better results than more complex more models, such as BERT and RoBERTa. For the WH dataset, BERT presented less biased behaviour for DT, SVM, and RF classifiers. However, it was statistically better only for the SVM classifier. The best results for the DV dataset were found with GloVe for SVM, XGB, and RF and with BERT for LR and MLP. In contrast with FNED and FPED, which analyse whether the model presents different performance among the identity terms, this metric measures the average model effectiveness in the samples evaluated with all identity terms.

Table 23 – Results obtained using Subgroup AUC bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.508	0.518	0.511	0.502	0.519	0.499
TF-IDF	0.529	0.517	0.541	0.502	0.531	0.517
GloVe	**0.554**	0.498	0.547	0.520	**0.572**	0.518
FastText	0.552	0.532	**0.553**	**0.538**	0.565	**0.528**
BERT	0.515	**0.533**	0.530	0.533	0.533	0.516
RoBERTa	0.523	0.516	0.528	0.515	0.546	0.508

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.497	0.508	0.491	0.500	0.493	0.492
TF-IDF	0.505	0.507	0.502	0.500	0.499	0.498
GloVe	0.498	0.502	0.490	**0.513**	0.480	0.501
FastText	**0.518**	0.507	0.504	0.503	**0.517**	0.501
BERT	0.510	**0.510**	**0.515***	0.508	0.513	**0.504**
RoBERTa	0.508	0.504	0.501	0.499	0.505	0.497

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.533	**0.530**	0.530	0.508	0.535	0.518
TF-IDF	0.516	0.508	0.531	0.508	0.534	0.500
GloVe	0.524	0.515	**0.538**	**0.525**	0.529	**0.531**
FastText	0.500	0.505	0.511	0.495	0.515	0.507
BERT	**0.534**	0.525	0.526	0.523	**0.551**	0.521
RoBERTa	0.495	0.509	0.497	0.507	0.526	0.500

(c) DV dataset

Source: Prepared by the author.

In addition, it is relevant to notice that most of the results from the Subgroup AUC metric (Table 23) are close to 0.5, meaning that the algorithms had difficulty classifying the examples with identity terms. However, as previously mentioned (see Section 4.4.1), the unbiased test dataset used to evaluate the unintended bias comprises different cases of hate speech, the majority challenging cases for classification models. We obtained the best results with the models trained with the HE dataset. These results also can evidence a context-dependence of these models.

Based on all the evidence presented above, we can answer the research question **RQ1: Yes, the choice of the feature extraction technique impacts the presence of unintended**

gender bias on the model predictions. We could verify that some classifiers presented more bias on predictions with some feature extraction techniques. For instance, the DT using the TF and TF-IDF as a feature extractor found a result higher than 0.20 with the bias metric FNED and using BERT, the same classifier found results lower than 0.06 for the HE and WH dataset (see Table 21). For this metric, the ideal value is zero, so the higher the value, the more bias. TF and TF-IDF are textual features that score and weight words based on their occurrence and frequency in a document. Thus, it may lead to a bias in relation to terms that are merely common in the dataset rather than truly informative for classification.

In addition, we can also answer the research question **RQ2 – Do feature extraction techniques tend to present bias when dealing with different datasets?** – The results in Tables 21, 22, and 23 endorse the need to properly and wisely select the feature extraction technique for each dataset matters for the effectiveness of the unbiased behaviour on the model predictions. The BERT and RoBERTa as input vectors achieved the best results for the FPED and FNED metrics with most classification models. However, it presented the best result only for some classifiers for the Subgroup AUC and in most cases, the results were not significantly better. Moreover, the analysis with the metric subgroup AUC showed different performances of the feature extractors in distinct datasets.

4.5.2 Classification performance

To answer the research question **RQ3 – Can the bias affect the performance of the models?** – for each dataset, we compare the results of each feature extraction technique with different classifiers and contrast them with the results in Section 4.5.1. We then aim to answer if the bias in the model predictions impacts the classification performance.

Tables 24 and 25 present the AUC and macro F1 metric results. For the HE dataset, the classification models presented the best AUC performance with GloVe for the DT and RF, FastText for the SVM and MLP, RoBERTa for the LR, and BERT for the XGB. With F1, FastText presented the best results for the LR, SVM, XGB, and RF, while GloVe with DT and MLP. In contrast with the results obtained with the bias metrics evaluated in Table 23, GloVe with DT presented more bias on prediction than the other feature extractors for the subgroup AUC metric, finding results under 0.5. The TF-IDF presented the best classification performance with both metrics for the WH dataset with the DT, SVM, and RF, while FastText with XGB. We achieved the best classification performance for the DV dataset using TF with

all classifiers for the macro F1 metric. This feature extractor also presented more bias on predictions than FastText, BERT, and RoBERTa for different classifiers with the metrics FPED and FNED, as shown in Section 4.5.1.

Table 24 – Results obtained using AUC for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.584	0.536	0.616	0.502	0.588	0.555
TF-IDF	0.626	0.538	0.638	0.518	0.591	0.548
GloVe	0.599	**0.562**	0.626	0.599	0.611	**0.647**
FastText	0.622	0.555	**0.648**	0.623	**0.647**	0.646
BERT	0.626	0.559	0.638	**0.624**	0.605	0.624
RoBERTa	**0.632**	0.529	0.631	0.590	0.617	0.618

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	**0.901**	0.761	0.887	0.832	0.878	0.891
TF-IDF	0.899	**0.770**	**0.898***	0.809	**0.887**	**0.901***
GloVe	0.862	0.659	0.885	0.840	0.871	0.833
FastText	0.864	0.653	0.885	**0.841**	0.884	0.829
BERT	0.867	0.630	0.870	0.828	0.864	0.813
RoBERTa	0.871	0.652	0.873	0.840	0.875	0.834

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.924	**0.800***	0.906	0.899	0.903	0.915
TF-IDF	**0.933***	0.783	**0.917***	0.899	0.900	**0.916**
GloVe	0.913	0.672	0.908	0.859	0.911	0.856
FastText	0.903	0.655	0.899	0.863	**0.915**	0.850
BERT	0.875	0.610	0.862	0.806	0.870	0.785
RoBERTa	0.890	0.630	0.877	0.833	0.900	0.827

(c) DV dataset

Source: Prepared by the author.

Based on all the above evidence, we can answer the research question **RQ3: It depends on the analysed dataset**. For the HE, the feature extraction techniques that present more bias on predictions also present the best classification performance. We can infer that the test dataset can follow the same biased behaviour as the training set and influence these results, similar to the conclusions in (DIXON et al., 2018). Therefore, evaluating the model with an unbiased test is relevant and can help investigate different insights into the problem.

Table 25 – Results obtained using macro F1 for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with *.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.489	0.435	0.446	0.421	0.487	0.409
TF-IDF	0.495	0.462	0.475	0.420	0.504	0.420
GloVe	0.525	**0.541**	0.539	0.544	**0.527**	0.579
FastText	**0.566**	0.538	**0.555**	**0.571**	0.517	**0.589**
BERT	0.500	0.535	0.500	0.532	0.512	0.541
RoBERTa	0.502	0.500	0.493	0.515	0.460	0.520

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	**0.749***	0.698	0.741	**0.701**	0.721	0.742
TF-IDF	0.730	**0.709**	**0.747**	0.700	0.725	**0.762***
GloVe	0.661	0.556	0.703	0.622	0.707	0.615
FastText	0.640	0.551	0.704	0.623	**0.726**	0.606
BERT	0.702	0.515	0.679	0.604	0.692	0.577
RoBERTa	0.684	0.545	0.695	0.629	0.688	0.598

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
TF	**0.707**	**0.693**	**0.706***	**0.703**	**0.698**	**0.715***
TF-IDF	0.702	0.681	0.681	0.696	0.691	0.682
GloVe	0.642	0.536	0.606	0.569	0.690	0.579
FastText	0.574	0.515	0.581	0.546	0.682	0.547
BERT	0.593	0.458	0.512	0.478	0.611	0.483
RoBERTa	0.571	0.487	0.543	0.499	0.629	0.489

(c) DV dataset

Source: Prepared by the author.

4.6 DISCUSSION

We evaluated the unintended gender bias from TF, TF-IDF, BERT, RoBERTa, GloVe, and FastText in the predictions of six state-of-the-art machine learning classifiers in hate speech datasets. Based on the proposed analysis, we identified three main aspects: (1) feature extractor choice matters from a biased perspective, (2) training and testing based on the same dataset cannot properly estimate the bias in the predictions, and (3) the bias influence is dataset-dependent in the classification performance. Section 4.6.1 presents the models execution time evaluation, Section 4.6.2 discusses the overall relationship between bias and

classification performance metrics, and a more profound analysis using the classifier's AUC and Subgroup AUC is addressed in Section 4.6.3.

4.6.1 Models execution time evaluation

This section analyses the average training time for each feature extraction method, including the training time and the representation step. The analysis was performed on the HE dataset, and we executed each experiment five times. Table 26 presents the execution time for each feature extraction method in each set. The BERT and RoBERTa presented a higher execution time for the train, test, and validation (val) sets. These results are due to the higher complexity of these models in contrast with TF, TF-IDF, GloVe, and FastText. Furthermore, the execution time of BERT and RoBERTa were similar, as expected, considering that they have similar architectures.

Table 26 – Models execution time evaluation for the representation step. The feature extraction with the lowest execution time for each classifier is highlighted in bold.

Feature	Train	Test	Val
TF	0.238	**0.132**	0.172
TF-IDF	**0.198**	0.263	0.158
GloVe	0.420	0.156	0.063
FastText	0.370	0.139	**0.056**
BERT	398.348	140.224	46.328
RoBERTa	386.538	136.511	44.620

Source: Prepared by the author.

We also calculate the classifiers' training time without considering the representation step to make the models comparable. Table 27 presents the model's execution time evaluation for the classification step. The classifiers presented the most cost-effective with GloVe as a feature extractor for five of the six classifiers analysed.

4.6.2 Classification performance metrics versus unintended bias metrics

This analysis was performed based on five metrics: AUC and macro F1 for performance evaluation and FNED, FPED, and Subgroup AUC for bias evaluation. As previously mentioned, the feature extractors that performed well regarding FNED and FPED had similar false negative and positive rates for different identity terms. On the other hand, the models with higher

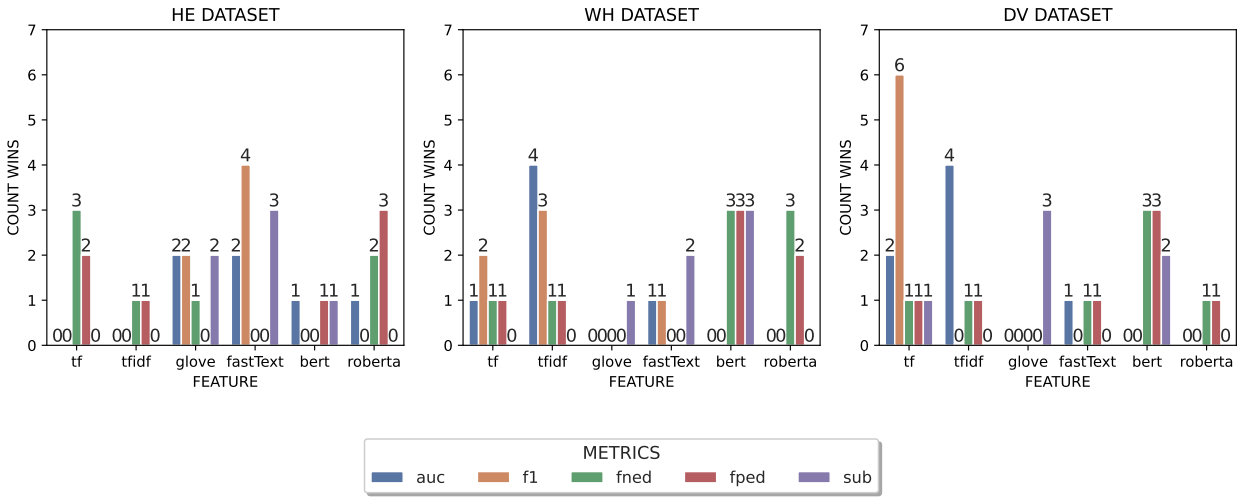
Table 27 – Models execution time evaluation for the classification step. The feature extraction with the lowest execution time for each classifier is highlighted in bold.

Feature	LR	DT	SVM	XGB	MLP	RF
TF	0.465	5.712	410.910	18.700	51.795	4.288
TF-IDF	**0.347**	7.076	430.032	18.790	47.957	4.500
GloVe	0.348	**2.084**	**60.432**	**7.557**	**8.353**	**2.741**
FastText	0.386	2.800	83.708	13.282	19.754	3.303
BERT	4.256	7.354	158.321	37.468	18.666	5.407
RoBERTa	1.472	9.141	191.830	36.169	19.168	5.502

Source: Prepared by the author.

Subgroup AUC scores found less difficulty in classifying samples containing identity terms. Figure 15 shows the results for all metrics for each dataset. This graphic represents the number of times each feature extractor wins for each metric independent of the six classifiers evaluated; so, the maximum number is six. In the case of ties, all who tie are considered winners.

Figure 15 – Classification performance metrics versus unintended bias metrics. f1_score is macro F1 score, and subgroup denotes Subgroup AUC.



Source: Prepared by the author.

As we can observe from the results reported in Figure 15, none of the feature extractors achieved the best results for all metrics. In addition, in some cases, the feature extractor that found the best overall classification performance also presented more bias on prediction.

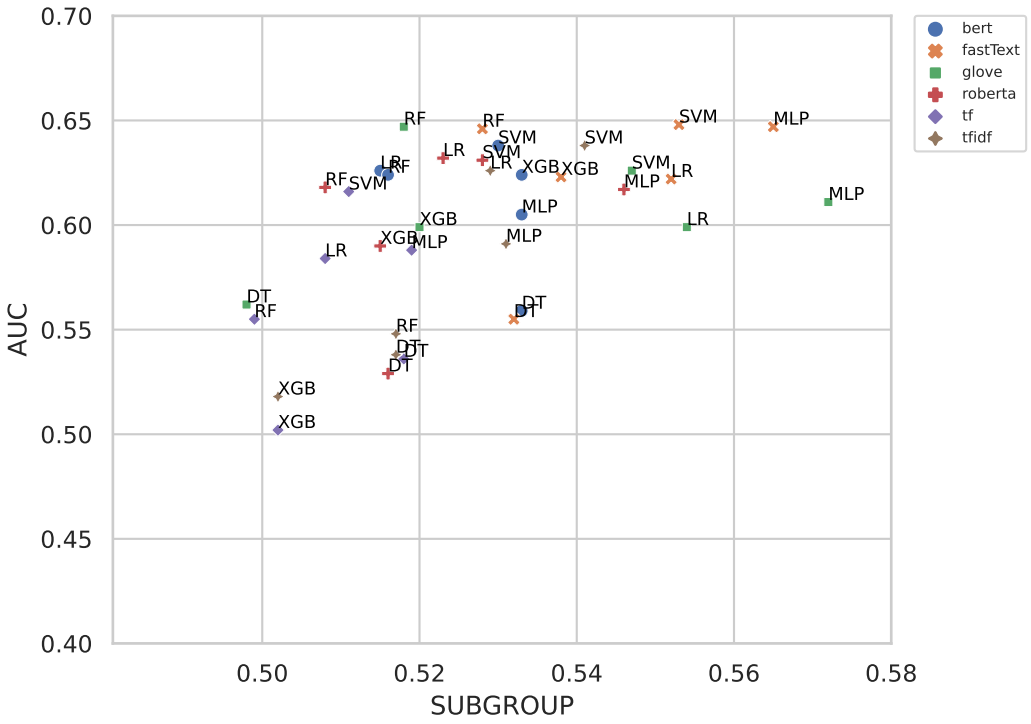
For instance, FastText presented more bias on prediction for the HE dataset than the other feature extractors for the FNED and FPED metrics, even though it had achieved better results than the other feature extractor techniques for the AUC and macro F1 metrics. These results suggest that when the model is trained using this feature extractor, it presents different

performances for examples that mention distinct identity terms. As expected, these results also suggest that the test set may have followed the same biased distribution as the training set.

For the WH dataset, the GloVe, in contrast with the other feature extractor analysed, presented the worst results for the bias metrics. In addition, it is relevant to note that BERT achieved the best result for the bias metrics and TF-IDF for the classification performance metrics.

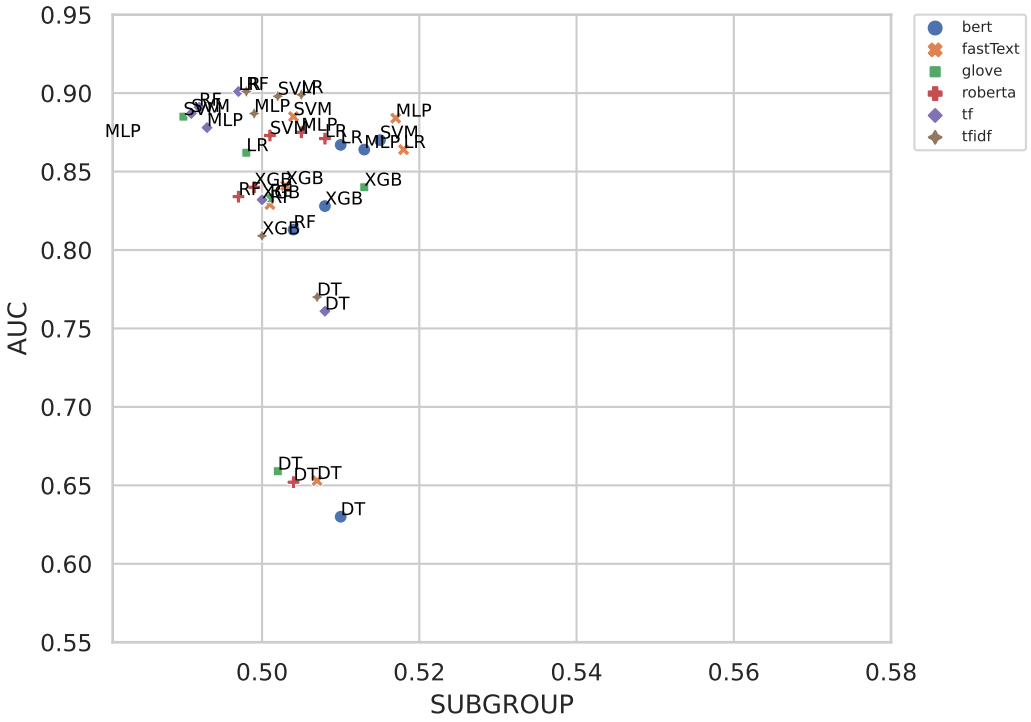
Moreover, an interesting behaviour can be observed for the DV dataset. For the classification performance metrics, BERT, RoBERTa, and GloVe achieved presented worst results with the AUC and macro F1. In addition, for the bias metrics, RoBERTa also presents poor results, especially with the metric subgroup AUC. These results can evidence the classifiers' poor performance across all identity terms when combined with RoBERTa. On the other hand, GloVe presented better results only with subgroup AUC bias metric.

Figure 16 – AUC versus Subgroup AUC for the HE dataset.



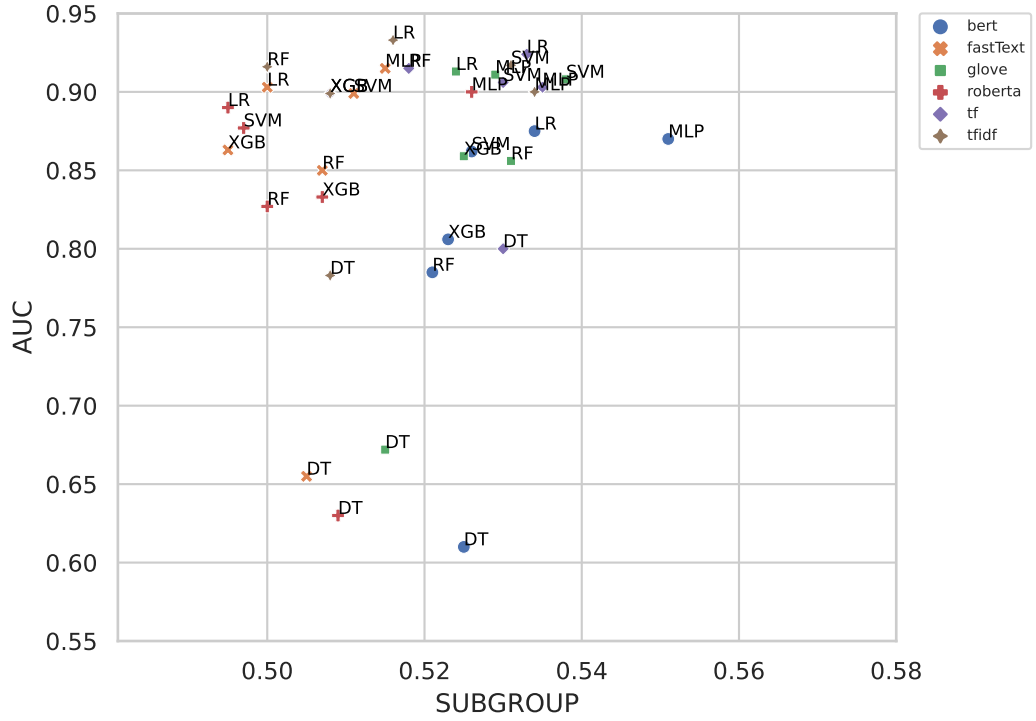
Source: Prepared by the author.

Figure 17 – AUC versus Subgroup AUC for the WH dataset.



Source: Prepared by the author.

Figure 18 – AUC versus Subgroup AUC for the DV dataset.



Source: Prepared by the author.

4.6.3 Case studies

This section evaluates the relationship between classification performance and the unintended gender bias metric. For this analysis, we consider two metrics, AUC and Subgroup AUC. However, the results with all combinations of metrics are available in the GitHub repository available in supplementary information.

Figures 16, 17 and 18 present the results using the pair of metrics for the HE, WH, and DV datasets, respectively. For all datasets, the DT classifier achieved the worst results related to bias and classification performance. In contrast with the other datasets, for the HE dataset, the combination of classifiers and feature extractors presented less biased behaviour (for more details, see Section 4.5). Considering the trade-off between the bias metric (subgroup AUC) and the performance metric (AUC), FastText presented the best results for the HE and WH datasets when combined with SVM, RF, and MLP, while GloVe and BERT presented the best results for the DV dataset when combined with LR, MLP, and SVM.

4.7 CONCLUSION

This study performed a comprehensive analysis to understand the impact of unintended gender bias from different feature extractors and how it can affect classification performance. We performed a broad experiment with six feature extractors, six classification methods, and three hate speech datasets. The results were evaluated based on several metrics to investigate different nuances of the problem.

The outcomes of our analysis reveal that the feature extraction method plays a crucial role in determining the occurrence of unintended gender bias in model predictions. Additionally, selecting the appropriate feature extraction technique depends highly on the dataset. Consequently, selecting the best feature extraction technique for each dataset is essential to ensure that the model predictions remain unbiased. Our findings emphasise the significance of such analyses as a critical tool in model selection.

Researchers face complex difficulties due to the growing incidence of hate speech in modern media. Although significant progress has been made in automatic hate speech detection, hate speech detection methods face challenges and limitations. The results obtained in this paper demonstrate that some feature extractors can lead to gender bias. Moreover, there are numerous biases within the context of hate speech, including racial, cross-geographic, and

political biases (GARG et al., 2023). To address this issue, conducting a comprehensive analysis of various biases may provide valuable insights into developing procedures that improve the model's generalisation power.

The data collection and annotation can also impact the dataset's characteristics and lead to bias in the model. In the context of hate speech detection, the real-world distribution of non-hate is tiny, which makes collecting hate speech comments hard. The researchers usually use specific topics, hashtags, or users to increase the hate speech content (DAVIDSON et al., 2017). Consequently, it introduces unintended biases into the dataset and the modelling pipeline. Therefore, cross-dataset analysis is crucial to identify and address dataset biases.

In future works, we intend to investigate further the dataset annotation process to understand its influence on the bias. In addition, we intend to extend this study to deep learning classifiers, such as CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), and so on. This analysis also can be extended to work with other identity problems, such as racial, religious, and xenophobic stereotypes.

In addition, this study showed that the unintended bias depends on the feature extractor and the classifier applied. Therefore, we can find an automatic way to indicate the best feature extractor and classifier for each dataset in future work.

5 GENERAL CONCLUSION

This chapter presents the final remarks about the main content discussed in this work and future work.

5.1 FINAL REMARKS

In this thesis, we addressed the issue of identifying and mitigating unintended bias in the context of hate speech detection, with a specific focus on identifying potential bias towards gender identity terms. We proposed a framework for mitigating unintended gender bias and hate speech detection composed of two modules: bias mitigation and hate speech detection. The proposed framework aims to reduce potential gender bias in hate speech detection without compromising the algorithms' classification performance. Moreover, unintended bias can occur at different stages of the ML model development. A fundamental stage in text classification is feature extraction, as discussed in Chapter 4. To address this issue, we proposed a framework to analyse unintended bias at the feature level and a comprehensive, unbiased dataset. This framework evaluates the presence of bias and its impact on the classification performance of machine learning algorithms.

In Chapter 2, we performed an extensive analysis of automatic hate speech detection in textual content from online social media, including feature extraction techniques, classifiers and datasets. The detection of hate speech has been performed as a supervised learning task using various feature extraction techniques, such as Bag-of-Words (BoW), n-grams, and pre-trained text embedding. The study findings showed that pre-trained text embedding, meta-information, and Deep Learning models are relevant approaches for enhancing classification performance. However, judging which approaches are the best is complex due to the lack of publicly available datasets and different weaknesses that still need to be explored.

In Chapter 3, a multi-view stacked framework using a strategy for bias mitigation is proposed. The framework is based on two modules: gender bias mitigation, in which the bias is detected and mitigated, and hate speech detection, in which is proposed a multi-view stacked classifier. The proposed framework reduced the bias and outperformed different models using several feature extractors for three of the four datasets evaluated. Although some results are slightly inferior, the proposed methodology demonstrates to be effective compared to state-of-

the-art solutions. Moreover, the proposed framework is general enough to work with different feature extractors and classification models. This makes it possible to extend and continuously improve the classification results.

Lastly, in Chapter 4, we proposed a framework and an unbiased dataset to evaluate the impact of unintended gender bias from different feature extractors on classification performance. The experiment included GloVe, FastText and BERT feature extractors, several classification methods, and three hate speech datasets. The results demonstrated that feature extraction technique selection could affect unintended gender bias in predictions, and it varied depending on the dataset. Therefore, the appropriate selection of the feature extraction technique for each dataset is crucial to avoid biased behaviour in model predictions.

5.2 CHALLENGES AND FUTURE DIRECTIONS

Detecting hate speech is a complex task, even for humans, due to its subjective nature. Therefore, hate speech detection methods have faced challenges and limitations. This thesis addresses the unintended gender bias problem in this context. However, this subject presents other difficulties and further studies can be performed. The findings of this study indicate some potential perspectives for future work on this topic.

5.2.1 Hate speech detection using multiple feature representations

Extracting relevant features from data is crucial for text classification using ML. Several methods have been proposed for feature extraction and significant progress has been made, as discussed in Chapter 2, including Bag-of-Words techniques, Large Language Models (LLMs), Deep Neural Network (DNN). However, properly selecting the adequate method can be a complex task.

According to the experimental study conducted in Chapter 3, the combination of different methods for feature extraction can improve the performance of classification models. Moreover, the study performed in Chapter 4 evidence that the feature selection matters in the context of unintended gender bias. Therefore, multiples features can extract different abstraction of the data and introduce in the model complementary for the model deal with inconsistencies.

5.2.2 Exploring Large Language Models

The advent of the LLMs has resulted in the emergence of a promising avenue for language processing by machines. The analysis in Chapter 4 shows that the bias can be incorporated in different stages of the algorithm development, and even state-of-art methods can present biased behaviours. Therefore, investigating new technologies developed in this context can be a promising research direction.

In future works, we intend to investigate cutting-edge language model, such as Falcon (PENEDO et al., 2023), LLaMA (TOUVRON et al., 2023) and Generative Pre-trained Transformer (GPT) (OUYANG et al., 2022). In particular, these language models have presented significant results (FARHANGIAN; CRUZ; CAVALCANTI, 2024).

5.2.3 Unintended bias mitigation

The hate speech detection models can present different types of unintended bias, such as racial, annotation, cross-geographic, political, and so on (DAVANI et al., 2023; GARG et al., 2023). Moreover, some models can present intersectional bias when two or more biases are related, for instance, racial and gender bias for AAE (KIM et al., 2020). Bias can develop due to limited perspective and repeated exposure to similar behavior. Therefore, designing a robust model to address various types of bias can be a promising avenue for research.

Although the proposed solutions focus on unintended gender bias, the methods proposed can be expanded to address other identity bias, such as racial stereotypes. In the literature, the racial bias identification and mitigation has been performed based on the AAE (MOZAFARI; FARAHBAKHSH; CRESPI, 2020; GARG et al., 2023).

REFERENCES

AGARWAL, S.; CHOWDARY, C. R. Combating hate speech using an adaptive ensemble learning model with a case study on covid-19. *Expert Systems with Applications*, v. 185, p. 115632, 2021. ISSN 0957-4174.

AGGARWAL, C. C. et al. Neural networks and deep learning. *Springer*, Springer, v. 10, n. 978, p. 3, 2018.

AL-AZANI, S.; EL-ALFY, E.-S. M. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, Elsevier, v. 109, p. 359–366, 2017.

AL-HASSAN, A.; AL-DOSSARI, H. Detection of hate speech in social networks: A survey on multilingual corpus. *6th International Conference on Computer Science and Information Technology*, v. 9, n. 2, p. 83–100, 2019.

AL-MAKHADMEH, Z.; TOLBA, A. Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, Springer, v. 102, n. 2, p. 501–522, 2019.

ALMATARNEH, S.; GAMALLO, P.; PENA, F. J. R.; ALEXEEV, A. Supervised classifiers to identify hate speech on english and spanish tweets. In: SPRINGER. *International Conference on Asian Digital Libraries*. Cham, 2019. p. 23–30.

ALMEREKHI, H.; KWAK, H.; JANSEN, B. J.; SALMINEN, J. Detecting toxicity triggers in online discussions. In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, 2019. p. 291–292.

ALORAINY, W.; BURNAP, P.; LIU, H.; WILLIAMS, M. L. "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, ACM New York, NY, USA, v. 13, n. 3, p. 1–26, 2019.

ALSAFARI, S.; SADAOUI, S.; MOUHOUB, M. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, Elsevier, v. 19, p. 100096, 2020.

ANTONAKAKI, D.; FRAGOPOULOU, P.; IOANNIDIS, S. A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, Elsevier, v. 164, p. 114006, 2021.

ARANGO, A.; PéREZ, J.; POBLETE, B. Hate speech detection is not as easy as you may think: A closer look at model validation. In: . New York, NY, USA: ACM, 2019. (SIGIR'19), p. 45–54. ISBN 9781450361729.

ARCO, F. M. P. del; MOLINA-GONZáLEZ, M. D.; UREñA-LóPEZ, L. A.; MARTíN-VALDIVIA, M. T. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, Elsevier, v. 166, p. 114120, 2021. ISSN 0957-4174.

ASIRI, Y.; HALAWANI, H. T.; ALGHAMDI, H. M.; HAMZA, S. H. A.; ABDEL-KHALEK, S.; MANSOUR, R. F. Enhanced seagull optimization with natural language processing based hate speech detection and classification. *Applied Sciences*, MDPI, v. 12, n. 16, p. 8000, 2022.

ASWANI, R.; KAR, A. K.; ILAVARASAN, P. V. Experience: managing misinformation in social media—insights for policymakers from twitter analytics. *Journal of Data and Information Quality (JDIQ)*, ACM New York, NY, USA, v. 12, n. 1, p. 1–18, 2019.

AYO, F. E.; FOLORUNSO, O.; IBHARALU, F. T.; OSINUGA, I. A. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, Elsevier, v. 38, p. 100311, 2020.

BADJATIYA, P.; GUPTA, M.; VARMA, V. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: *The World Wide Web Conference*. New York, NY, USA: ACM, 2019. p. 49–59. ISBN 9781450366748.

BALOUCHZAHI, F.; SHASHIREKHA, H. L.; SIDOROV, G.; GELBUKH, A. A comparative study of syllables and character level n-grams for dravidian multi-script and code-mixed offensive language identification. *Journal of Intelligent & Fuzzy Systems*, IOS Press, v. 43, n. 6, p. 6995–7005, 2022.

BASILE, V.; BOSCO, C.; FERSINI, E.; NOZZA, D.; PATTI, V.; PARDO, F. M. R.; ROSSO, P.; SANGUINETTI, M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis: ACL, 2019. p. 54–63.

BOHRA, A.; VIJAY, D.; SINGH, V.; AKHTAR, S. S.; SHRIVASTAVA, M. A dataset of hindi-english code-mixed social media text for hate speech detection. In: *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*. New Orleans, Louisiana, USA: ACL, 2018. p. 36–41.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 5, p. 135–146, 2017.

BOLUKBASI, T.; CHANG, K.-W.; ZOU, J. Y.; SALIGRAMA, V.; KALAI, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: *30th Conference on Neural Information Processing Systems*. Barcelona, Spain: Advances in Neural Information Processing Systems, 2016. p. 4349–4357.

BORKAN, D.; DIXON, L.; SORENSEN, J.; THAIN, N.; VASSERMAN, L. Nuanced metrics for measuring unintended bias with real data for text classification. In: *Companion proceedings of the 2019 world wide web conference*. New York, NY, USA: ACM, 2019. p. 491–500.

BOUAZIZI, M.; NIIDA, N.; OHTSUKI, T. All-in-one hate speech detectors may not be what you want. In: *2021 The 4th International Conference on Software Engineering and Information Management*. New York, NY, USA: ACM, 2021. (ICSIM 2021), p. 165–170. ISBN 9781450388955.

BURNAP, P.; WILLIAMS, M. L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, Wiley Online Library, v. 7, n. 2, p. 223–242, 2015.

BURNAP, P.; WILLIAMS, M. L. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, Springer, v. 5, n. 1, p. 11, 2016.

CALABRESE, A.; BEVILACQUA, M.; ROSS, B.; TRIPODI, R.; NAVIGLI, R. Aaa: Fair evaluation for abuse detection systems wanted. In: *13th ACM Web Science Conference 2021*. New York, NY, USA: ACM, 2021. (WebSci '21), p. 243–252. ISBN 9781450383301. Available at: <https://doi.org/10.1145/3447535.3462484>.

CAO, R.; LEE, R. K.-W.; HOANG, T.-A. Deephate: Hate speech detection via multi-faceted text representations. In: *12th ACM Conference on Web Science*. New York, NY, USA: ACM, 2020. (WebSci '20), p. 11–20. ISBN 9781450379892.

CASULA, P.; ANUPAM, A.; PARVIN, N. "we found no violation!": Twitter's violent threats policy and toxicity in online discourse. In: *C&T '21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech*. New York, NY, USA: ACM, 2021. (C&T '21), p. 151–159. ISBN 9781450390569. Available at: <https://doi.org/10.1145/3461564.3461589>.

CHAKRABORTY, P.; SEDDIQUI, M. H. Threat and abusive language detection on social media in bengali language. In: IEEE. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. Dhaka, Bangladesh, 2019. p. 1–6.

CHARITIDIS, P.; DOROPOULOS, S.; VOLOGIANNIDIS, S.; PAPASTERGIOU, I.; KARAKEVA, S. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, Elsevier, v. 17, p. 100071, 2020.

CHATZAKOU, D.; KOURTELLIS, N.; BLACKBURN, J.; CRISTOFARO, E. D.; STRINGHINI, G.; VAKALI, A. Mean birds: Detecting aggression and bullying on twitter. In: *Proceedings of the 2017 ACM on web science conference*. New York, NY, USA: ACM, 2017. p. 13–22.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Available at: <https://doi.org/10.1145/2939672.2939785>.

CHIRIL, P.; MORICEAU, V.; BENAMARA, F.; MARI, A.; ORIGGI, G.; COULOMB-GULLY, M. He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. online: ACL, 2020. p. 4055–4066.

COHEN-ALMAGOR, R. Freedom of expression v. social responsibility: Holocaust denial in canada. *Journal of Mass Media Ethics*, Taylor & Francis, v. 28, n. 1, p. 42–56, 2013.

CORAZZA, M.; MENINI, S.; CABRIO, E.; TONELLI, S.; VILLATA, S. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, ACM New York, NY, USA, v. 20, n. 2, p. 1–22, 2020.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, p. 273–297, 1995.

CRUZ, R. M.; CAVALCANTI, G. D.; TSANG, I. R.; SABOURIN, R. Feature representation selection based on classifier projection space and oracle analysis. *Expert Systems with Applications*, v. 40, n. 9, p. 3813–3827, 2013.

CRUZ, R. M.; HAFEMANN, L. G.; SABOURIN, R.; CAVALCANTI, G. D. Deslib: A dynamic ensemble selection library in python. *Journal of Machine Learning Research*, v. 21, n. 8, p. 1–5, 2020.

CRUZ, R. M.; SABOURIN, R.; CAVALCANTI, G. D. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, Elsevier, v. 41, p. 195–216, 2018.

CRUZ, R. M.; SOUSA, W. V.; CAVALCANTI, G. D. Selecting and combining complementary feature representations and classifiers for hate speech detection. *Online Social Networks and Media*, v. 28, p. 100194, 2022. ISSN 2468-6964.

DASTIN, J. Amazon scraps secret ai recruiting tool that showed bias against women. In: *Ethics of data and analytics*. Online: Auerbach Publications, 2018. p. 296–299.

DAVANI, A. M.; ATARI, M.; KENNEDY, B.; DEHGHANI, M. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA . . . , v. 11, p. 300–319, 2023.

DAVIDSON, T.; WARMSLEY, D.; MACY, M.; WEBER, I. Automated hate speech detection and the problem of offensive language. In: *Eleventh international AAAI conference on web and social media*. Montréal, Canada: AAAI Press, 2017. v. 11, p. 512–515.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, JMLR. org, v. 7, p. 1–30, 2006.

DESHPANDE, K. V.; PAN, S.; FOULDS, J. R. Mitigating demographic bias in ai-based resume filtering. In: *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*. New York, NY, USA: ACM, 2020. p. 268–275. Available at: <https://doi.org/10.1145/3386392.3399569>.

DESOUZA, G.; DA-COSTA-ABREU, M. Automatic offensive language detection from twitter data using machine learning and feature selection of metadata. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, UK: IEEE, 2020. p. 1–6.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: ACL, 2019. p. 4171–4186.

DIXON, L.; LI, J.; SORENSEN, J.; THAIN, N.; VASSERMAN, L. Measuring and mitigating unintended bias in text classification. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, 2018. (AIES '18), p. 67–73.

DORRIS, W.; HU, R. R.; VISHWAMITRA, N.; LUO, F.; COSTELLO, M. Towards automatic detection and explanation of hate speech and offensive language. In: *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*. New York, NY, USA: ACM, 2020. (IWSPA '20), p. 23–29. ISBN 9781450371155.

ELISABETH, D.; BUDI, I.; IBROHIM, M. O. Hate code detection in indonesian tweets using machine learning approach: A dataset and preliminary study. In: *2020 8th International Conference on Information and Communication Technology (ICoICT)*. Yogyakarta, Indonesia: IEEE, 2020. p. 1–6.

FACEBOOK, C. S. *Hate Speech*. 2020. Available: <https://www.facebook.com/communitystandards/hate_speech>. Accessed: 2020-09-09.

FARHANGIAN, F.; CRUZ, R. M.; CAVALCANTI, G. D. Fake news detection: Taxonomy and comparative study. *Information Fusion*, Elsevier, v. 103, p. 102140, 2024.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 4, p. 1–30, 2018.

FORTUNA, P.; SILVA, J. R. da; WANNER, L.; NUNES, S. et al. A hierarchically-labeled portuguese hate speech dataset. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: ACL, 2019. p. 94–104.

FOUNTA, A. M.; CHATZAKOU, D.; KOURTELLIS, N.; BLACKBURN, J.; VAKALI, A.; LEONTIADIS, I. A unified deep learning architecture for abuse detection. In: *Proceedings of the 10th ACM Conference on Web Science*. Florence, Italy: ACL, 2019. p. 105–114.

FOUNTA, A. M.; DJOUVAS, C.; CHATZAKOU, D.; LEONTIADIS, I.; BLACKBURN, J.; STRINGHINI, G.; VAKALI, A.; SIRIVIANOS, M.; KOURTELLIS, N. Large scale crowdsourcing and characterization of twitter abusive behavior. In: *Twelfth International AAAI Conference on Web and Social Media*. Stanford, California, USA: AAAI, 2018.

GARG, T.; MASUD, S.; SURESH, T.; CHAKRABORTY, T. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 13s, p. 1–32, 2023.

GIACHANOU, A.; ROSSO, P. The battle against online harmful information: The cases of fake news and hate speech. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York, NY, USA: ACM, 2020. p. 3503–3504.

GIBERT, O. de; PEREZ, N.; GARCIA-PABLOS, A.; CUADROS, M. Hate speech dataset from a white supremacy forum. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: ACL, 2018. p. 11–20.

GITARI, N. D.; ZUPING, Z.; DAMIEN, H.; LONG, J. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, v. 10, n. 4, p. 215–230, 2015.

GOLBECK, J.; ASHKTORAB, Z.; BANJO, R. O.; BERLINGER, A.; BHAGWAN, S.; BUNTAIN, C.; CHEAKALOS, P.; GELLER, A. A.; GNANASEKARAN, R. K.; GUNASEKARAN, R. R. et al. A large labeled corpus for online harassment research. In: *Proceedings of the 2017 ACM on web science conference*. New York, NY, USA: ACM, 2017. p. 229–233.

GRöNDAHL, T.; PAJOLA, L.; JUUTI, M.; CONTI, M.; ASOKAN, N. All you need is "love": Evading hate speech detection. In: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. New York, NY, USA: ACM, 2018. (AISec '18), p. 2–12. ISBN 9781450360043.

HAYATY, M.; ADI, S.; HARTANTO, A. D. Lexicon-based indonesian local language abusive words dictionary to detect hate speech in social media. *Journal of Information Systems Engineering and Business Intelligence*, v. 6, n. 1, p. 9–17, 2020.

HENDRAWAN, R.; ADIWIJAYA; FARABY, S. A. Multilabel classification of hate speech and abusive words on indonesian twitter social media. In: *2020 International Conference on Data Science and Its Applications (ICoDSA)*. Bandung, Indonesia: IEEE, 2020. p. 1–7.

IBROHIM, M. O.; BUDI, I. Multi-label hate speech and abusive language detection in indonesian twitter. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: ACL, 2019. p. 46–57.

INDURTHI, V.; SYED, B.; SHRIVASTAVA, M.; CHAKRAVARTULA, N.; GUPTA, M.; VARMA, V. Fermi at semeval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women on twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: ACL, 2019. p. 70–74.

JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, v. 546, p. 126232, 2023. ISSN 0925-2312. Available at: <https://www.sciencedirect.com/science/article/pii/S0925231223003557>.

KAPIL, P.; EKBAL, A. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, v. 210, p. 106458, 2020. ISSN 0950-7051.

KAR, A. K.; ASWANI, R. How to differentiate propagators of information and misinformation–insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences*, Taylor & Francis, v. 42, n. 6, p. 1307–1335, 2021.

KARN, A. L.; KARNA, R. K.; KONDAMUDI, B. R.; BAGALE, G.; PUSTOKHIN, D. A.; PUSTOKHINA, I. V.; SENGAN, S. Customer centric hybrid recommendation system for e-commerce applications by integrating hybrid sentiment analysis. *Electronic Commerce Research*, Springer, v. 23, n. 1, p. 279–314, 2023.

KHAN, M. M.; SHAHZAD, K.; MALIK, M. K. Hate speech detection in roman urdu. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, ACM, New York, NY, USA, v. 20, n. 1, Mar. 2021. ISSN 2375-4699. Available at: <https://doi.org/10.1145/3414524>.

KIM, J. Y.; ORTIZ, C.; NAM, S.; SANTIAGO, S.; DATTA, V. Intersectional bias in hate speech and abusive language datasets. In: AAAI ORGANIZATION. *Proceedings of the Fourteenth International Conference on Web and Social Media (ICWSM), Data Challenge Workshop*. [S.l.], 2020.

KIRITCHENKO, S.; MOHAMMAD, S. M. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018*, p. 43, 2018.

KUMAR, R.; REGANTI, A.; BHATIA, A.; MAHESHWARI, T. Aggression-annotated corpus of hindi-english code-mixed data. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2019. p. 1425–1431.

KUMARI, K.; JAMATIA, A. An approach of hate speech identification on twitter corpus. In: SPRINGER. *International Conference on Frontiers of Intelligent Computing: Theory and Applications*. [S.l.], 2022. p. 115–125.

KUNCHEVA, L. I. *Combining pattern classifiers: methods and algorithms*. 2. ed. Hoboken, New Jersey: John Wiley & Sons, 2014.

LEE, M. S. A.; SINGH, J. Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, 2021. (AIES '21), p. 704–714. ISBN 9781450384735. Available at: <https://doi.org/10.1145/3461702.3462572>.

LIU, H.; BURNAP, P.; ALORAINY, W.; WILLIAMS, M. L. Fuzzy multi-task learning for hate speech type identification. In: *The World Wide Web Conference*. New York, NY, USA: ACM, 2019. p. 3006–3012.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

MACAVANEY, S.; YAO, H.-R.; YANG, E.; RUSSELL, K.; GOHARIAN, N.; FRIEDER, O. Hate speech detection: Challenges and solutions. *PloS one*, Public Library of Science San Francisco, CA USA, v. 14, n. 8, p. e0221152, 2019.

MARKOV, I.; LJUBEŠIĆ, N.; FIŠER, D.; DAELEMANS, W. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: ACL, 2021. p. 149–159. Available at: <https://aclanthology.org/2021.wassa-1.16>.

MARNEFFE, M.-C. D.; MANNING, C. D. *Stanford typed dependencies manual*. Stanford University, 2008.

MARPAUNG, A.; RISMALA, R.; NURRAHMI, H. Hate speech detection in indonesian twitter texts using bidirectional gated recurrent unit. In: *2021 13th International Conference on Knowledge and Smart Technology (KST)*. Bangsaen, Chonburi, Thailand: IEEE, 2021. p. 186–190.

MATHEW, B.; DUTT, R.; GOYAL, P.; MUKHERJEE, A. Spread of hate speech in online social media. In: *Proceedings of the 10th ACM conference on web science*. New York, NY, USA: ACM, 2019. p. 173–182.

MAZARI, A. C.; BOUDOUKHANI, N.; DJEFFAL, A. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, Springer, p. 1–15, 2023.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. *Efficient Estimation of Word Representations in Vector Space*. 2013.

MIOK, K.; NGUYEN-DOAN, D.; ŠKRLJ, B.; ZAHARIE, D.; ROBNIK-ŠIKONJA, M. Prediction uncertainty estimation for hate speech classification. In: *International Conference on Statistical Language and Speech Processing*. Cham: Springer, 2019. p. 286–298.

MIŠKOLCI, J.; KOVÁČOVÁ, L.; RIGOVÁ, E. Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, SAGE Publications, Los Angeles, CA, v. 38, n. 2, p. 128–146, 2020.

MONDAL, M.; SILVA, L. A.; BENEVENUTO, F. A measurement study of hate speech in social media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, 2017. (HT '17), p. 85–94. ISBN 9781450347082.

MONTANI, J. P.; SCHÜLLER, P. Tuwienkbs at germeval 2018: German abusive tweet detection. In: *14th Conference on Natural Language Processing KONVENS*. Vienna, Austria: Austrian Academy of Sciences, 2018. p. 45.

MOSSIE, Z.; WANG, J.-H. Vulnerable community identification using hate speech detection on social media. *Inf. Process. Manage.*, Pergamon Press, Inc., USA, v. 57, n. 3, May 2020. ISSN 0306-4573.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE*, Public Library of Science, San Francisco, CA, USA, v. 15, n. 8, p. 1–26, 2020.

NANDHINI, B. S.; SHEEBA, J. I. 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)*. New York, NY, USA: ACM, 2015. (ICARCSET '15). ISBN 9781450334419.

NASCIMENTO, F. R.; CAVALCANTI, G. D.; COSTA-ABREU, M. D. Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, p. 117032, 2022. ISSN 0957-4174. Available at: <https://www.sciencedirect.com/science/article/pii/S095741742200447X>.

NASCIMENTO, F. R. S.; CAVALCANTI, G. D. C.; COSTA-ABREU, M. D. Exploring automatic hate speech detection on social media: A focus on content-based analysis. *SAGE Open*, v. 13, n. 2, p. 21582440231181311, 2023. Available at: <https://doi.org/10.1177/21582440231181311>.

NASCIMENTO, F. R. S.; CAVALCANTI, G. D. C.; COSTA-ABREU, M. D. Gender bias detection on hate speech classification: an analysis at feature-level. *pre-print SSRN*, 2023.

NASEEM, U.; RAZZAK, I.; EKLUND, P. W. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, Springer, p. 1–28, 2020.

NASIR, J. A.; KHAN, O. S.; VARLAMIS, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, Elsevier, v. 1, n. 1, p. 100007, 2021.

NOBATA, C.; TETREAULT, J.; THOMAS, A.; MEHDAD, Y.; CHANG, Y. Abusive language detection in online user content. In: *Proceedings of the 25th international conference on world wide web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016. p. 145–153.

NOZZA, D.; VOLPETTI, C.; FERSINI, E. Unintended bias in misogyny detection. In: *IEEE/WIC/ACM International Conference on Web Intelligence*. New York, NY, USA: ACM, 2019. p. 149–155.

NUGROHO, K.; NOERSASONGKO, E.; FANANI, A. Z.; BASUKI, R. S. et al. Improving random forest method to detect hatespeech and offensive word. In: *2019 International Conference on Information and Communications Technology (ICOIACT)*. Yogyakarta, Indonesia: IEEE, 2019. p. 514–518.

OBERMEYER, Z.; POWERS, B.; VOGELI, C.; MULLAINATHAN, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, p. 447–453, 2019.

ORIOLA, O. A stacked generalization ensemble approach for improved intrusion detection. *International Journal of Computer Science and Information Security*, v. 18, n. 5, p. 62–67, 2020.

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C.; MISHKIN, P.; ZHANG, C.; AGARWAL, S.; SLAMA, K.; RAY, A. et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, v. 35, p. 27730–27744, 2022.

PARK, J. H.; SHIN, J.; FUNG, P. Reducing gender bias in abusive language detection. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: ACL, 2018. p. 2799–2804.

PASCHALIDES, D.; STEPHANIDIS, D.; ANDREOU, A.; ORPHANOU, K.; PALLIS, G.; DIKAIAKOS, M. D.; MARKATOS, E. Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.*, ACM, New York, NY, USA, v. 20, n. 2, 2020. ISSN 1533-5399.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENEDO, G.; MALARTIC, Q.; HESSLOW, D.; COJOCARU, R.; CAPPELLI, A.; ALOBEIDLI, H.; PANNIER, B.; ALMAZROUEI, E.; LAUNAY, J. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. GloVe: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Doha, Qatar: ACL, 2014. p. 1532–1543.

PHILIPP, K.; ROMAN, K. *YouToxic English (Version 1.0.0)*. 2019. <https://zenodo.org/record/2586669#.X4l053VKjeR>. Accessed in October 16, 2020.

PITSILIS, G. K.; RAMAMPIARO, H.; LANGSETH, H. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, Springer, v. 48, n. 12, p. 4730–4742, 2018.

PLAZA-DEL-ARCO, F.-M.; MOLINA-GONZÁLEZ, M. D.; UREÑA-LÓPEZ, L. A.; MARTÍN-VALDIVIA, M. T. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, ACM New York, NY, USA, v. 20, n. 2, p. 1–19, 2020.

POLETTO, F.; BASILE, V.; SANGUINETTI, M.; BOSCO, C.; PATTI, V. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, Springer, p. 1–47, 2020.

PRATIWI, N. I.; BUDI, I.; ALFINA, I. Hate speech detection on indonesian instagram comments using fasttext approach. In: *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. Yogyakarta, Indonesia: IEEE, 2018. p. 447–450.

RIBEIRO, M. H.; CALAIS, P. H.; SANTOS, Y. A.; ALMEIDA, V. A.; JR, W. M. Characterizing and detecting hateful users on twitter. In: *Twelfth international AAAI conference on web and social media*. [S.l.: s.n.], 2018.

RISCH, J.; KRESTEL, R. Bagging bert models for robust aggression identification. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Marseille, France: ELRA, 2020. p. 55–61.

RIZOS, G.; HEMKER, K.; SCHULLER, B. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2019. p. 991–1000.

ROBINSON, D.; ZHANG, Z.; TEPPER, J. Hate speech detection on twitter: feature engineering vs feature selection. In: *European Semantic Web Conference*. Berlin, Heidelberg: Springer, 2018. p. 46–49.

RODRÍGUEZ, A.; ARGUETA, C.; CHEN, Y.-L. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In: *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. Okinawa, Japan: IEEE, 2019. p. 169–174.

ROSS, B.; RIST, M.; CARBONELL, G.; CABRERA, B.; KUROWSKY, N.; WOJATZKI, M. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *Bochumer Linguistische Arbeitsberichte*, p. 6–9, 2016.

RÖTTGER, P.; VIDGEN, B.; NGUYEN, D.; WASEEM, Z.; MARGETTS, H.; PIERREHUMBERT, J. et al. Hatecheck: Functional tests for hate speech detection models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *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)*. online, 2021. p. 41.

ŞAHI, H.; KILIÇ, Y.; SAğLAM, R. B. Automated detection of hate speech towards woman on twitter. In: *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. Sarajevo, Bosnia and Herzegovina: IEEE, 2018. p. 533–536.

ŞAHINUÇ, F.; YILMAZ, E. H.; TORAMAN, C.; KOÇ, A. The effect of gender bias on hate speech detection. *Signal, Image and Video Processing*, Springer, p. 1–7, 2022.

SAJJAD, M.; ZULIFQAR, F.; KHAN, M. U. G.; AZEEM, M. Hate speech detection using fusion approach. In: *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*. Taxila, Pakistan: IEEE, 2019. p. 251–255.

SALMINEN, J.; ALMEREKHI, H.; MILENKOVIC, M.; JUNG, S.-g.; AN, J.; KWAK, H.; JANSEN, B. J. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In: *Twelfth International AAAI Conference on Web and Social Media*. Stanford, California, USA: AAAI, 2018. p. 330–339.

SALMINEN, J.; HOPF, M.; CHOWDHURY, S. A.; JUNG, S.-g.; ALMEREKHI, H.; JANSEN, B. J. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, Springer, v. 10, n. 1, p. 1, 2020.

SANTOSH, T.; ARAVIND, K. Hate speech detection in hindi-english code-mixed social media text. In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. New York, NY, USA: ACM, 2019. p. 310–313.

SAP, M.; CARD, D.; GABRIEL, S.; CHOI, Y.; SMITH, N. A. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: ACL, 2019. p. 1668–1678.

SCHMIDT, A.; WIEGAND, M. A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: ACL, 2017. p. 1–10.

SENARATH, Y.; PUROHIT, H. Evaluating semantic feature representations to efficiently detect hate intent on social media. In: *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. San Diego, CA, USA: IEEE, 2020. p. 199–202. ISSN 2325-6516.

SENGUPTA, A.; BHATTACHARJEE, S. K.; AKHTAR, M. S.; CHAKRABORTY, T. Does aggression lead to hate? detecting and reasoning offensive traits in hinglish code-mixed texts. *Neurocomputing*, v. 488, p. 598–617, 2022. ISSN 0925-2312. Available at: <https://www.sciencedirect.com/science/article/pii/S0925231221017306>.

SERRANO-GUERRERO, J.; OLIVAS, J. A.; ROMERO, F. P.; HERRERA-VIEDMA, E. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, Elsevier, v. 311, p. 18–38, 2015.

SHMUELI, G.; BRUCE, P. C.; YAHAV, I.; PATEL, N. R.; JR, K. C. L. *Data mining for business analytics: concepts, techniques, and applications in R*. [S.l.]: John Wiley & Sons, 2017.

SOHN, H.; LEE, H. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In: *2019 International Conference on Data Mining Workshops (ICDMW)*. Beijing, China: IEEE, 2019. p. 551–559.

SREELAKSHMI, K.; PREMJITH, B.; SOMAN, K. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, Elsevier, v. 171, p. 737–744, 2020.

SUN, T.; GAUT, A.; TANG, S.; HUANG, Y.; ELSHERIEF, M.; ZHAO, J.; MIRZA, D.; BELDING, E.; CHANG, K.-W.; WANG, W. Y. Mitigating gender bias in natural language processing: Literature review. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: ACL, 2019. p. 1630–1640.

TEH, P. L.; CHENG, C.-B.; CHEE, W. M. Identifying and categorising profane words in hate speech. In: *Proceedings of the 2nd International Conference on Compute and Data Analysis*. New York, NY, USA: ACM, 2018. (ICCDA 2018), p. 65–69. ISBN 9781450363594.

TORAMAN, C.; ŞAHINUÇ, F.; YILMAZ, E. Large-scale hate speech detection with cross-domain transfer. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022. p. 2215–2225. Available at: <https://aclanthology.org/2022.lrec-1.238>.

TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A.; LACROIX, T.; ROZIÈRE, B.; GOYAL, N.; HAMBRO, E.; AZHAR, F. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

TWITTER. *Hateful conduct policy*. 2020. Available: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>. Accessed: 2020-09-09.

UNSVÅG, E. F.; GAMBÄCK, B. The effects of user features on twitter hate speech detection. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: ACL, 2018. p. 75–85.

VIGNA, F. D.; CIMINO, A.; DELL'ORLETTA, F.; PETROCCHI, M.; TESCONI, M. Hate me, hate me not: Hate speech detection on facebook. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. Venice, Italy: CEUR-WS.org, 2017. p. 86–95.

VITIUGIN, F.; SENARATH, Y.; PUROHIT, H. Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In: *13th ACM Web Science Conference 2021*. New York, NY, USA: ACM, 2021. (WebSci '21), p. 130–138. ISBN 9781450383301.

WALMSLEY, F. N.; CAVALCANTI, G. D.; OLIVEIRA, D. V.; CRUZ, R. M.; SABOURIN, R. An ensemble generation method based on instance hardness. In: IEEE. *2018 International Joint Conference on Neural Networks*. Rio de Janeiro, Brazil, 2018. p. 1–8.

WASEEM, Z. 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*. Austin, Texas: ACL, 2016. p. 138–142. Available at: <http://aclweb.org/anthology/W16-5618>.

WASEEM, Z.; HOVY, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: *Proceedings of the NAACL Student Research Workshop*. San Diego, California: ACL, 2016. p. 88–93. Available at: <http://www.aclweb.org/anthology/N16-2013>.

WASEEM, Z.; THORNE, J.; BINGEL, J. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In: *Online harassment*. Cham: Springer, 2018. p. 29–55.

WATANABE, H.; BOUAZIZI, M.; OHTSUKI, T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, IEEE, v. 6, p. 13825–13835, 2018.

WICH, M.; BREITINGER, M.; STRATHERN, W.; NAIMAREVIC, M.; GROH, G.; PFEFFER, J. Are your friends also haters? identification of hater networks on social media: Data paper. In: _____. *Companion Proceedings of the Web Conference 2021*. New York, NY, USA: ACM, 2021. p. 481–485. ISBN 9781450383134. Available at: <https://doi.org/10.1145/3442442.3452310>.

WIEGAND, M.; RUPPENHOFER, J.; KLEINBAUER, T. Detection of abusive language: the problem of biased datasets. In: *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, 2019. p. 602–608.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. M. Transformers: State-of-the-art natural language processing. In:

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: ACL, 2020. p. 38–45.

WOLPERT, D. H. Stacked generalization. *Neural networks*, Elsevier, v. 5, n. 2, p. 241–259, 1992.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

WULCZYN, E.; THAIN, N.; DIXON, L. Ex machina: Personal attacks seen at scale. In: *Proceedings of the 26th International Conference on World Wide Web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017. p. 1391–1399.

YOUTUBE. *Hate speech policy*. 2020. Available: <https://support.google.com/youtube/answer/2801939?hl=en>. Accessed: 2020-09-09.

ZAMPIERI, M.; MALMASI, S.; NAKOV, P.; ROSENTHAL, S.; FARRA, N.; KUMAR, R. Predicting the type and target of offensive posts in social media. In: *Proceedings of NAACL-HLT*. Minneapolis, Minnesota: ACL, 2019. p. 1415–1420.

ZHANG, Z.; LUO, L. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, IOS Press, v. 10, n. 5, p. 925–945, 2019.

ZHANG, Z.; ROBINSON, D.; TEPPER, J. Hate speech detection using a convolution-lstm based deep neural network. *ESWC 2018: The semantic web*, Lyon, 2018.

ZHAO, J.; WANG, T.; YATSKAR, M.; ORDONEZ, V.; CHANG, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: ACL, 2018. p. 15–20.

ZHAO, J.; XIE, X.; XU, X.; SUN, S. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, Elsevier, v. 38, p. 43–54, 2017.

ZHAO, Z.; ZHANG, Z.; HOPFGARTNER, F. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media*, Elsevier, v. 29, p. 100205, 2022.

ZHOU, Y.; YANG, Y.; LIU, H.; LIU, X.; SAVAGE, N. Deep learning based fusion approach for hate speech detection. *IEEE Access*, IEEE, v. 8, p. 128923–128929, 2020.

ZIMMERMAN, S.; KRUSCHWITZ, U.; FOX, C. Improving hate speech detection with deep learning ensembles. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. ISBN 979-10-95546-00-9.

APPENDIX A – SUPPLEMENTARY INFORMATION

The unbiased dataset and the source code are available in <https://github.com/Francimaria/hate_speech_bias_feature>.

A.1 SUPPLEMENTARY RESULTS

This Section presents the mean and the standard deviation of all results described in Section Experimental Results.

Table 28 – Results obtained using FNED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.175 ± 0.000	**0.034 ± 0.000**	0.139 ± 0.000	0.105 ± 0.000	0.158 ± 0.000	0.052 ± 0.000
Fast-Text	0.132 ± 0.000	0.065 ± 0.000	0.137 ± 0.000	0.082 ± 0.000	0.214 ± 0.000	0.062 ± 0.000
BERT	**0.131 ± 0.000**	0.037 ± 0.000	**0.069 ± 0.000**	**0.042 ± 0.000**	**0.125 ± 0.000**	**0.026 ± 0.000**

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.250 ± 0.007	0.069 ± 0.009	0.198 ± 0.009	0.128 ± 0.006	0.138 ± 0.008	0.067 ± 0.005
Fast-Text	0.200 ± 0.012	0.084 ± 0.011	0.159 ± 0.008	0.142 ± 0.007	0.174 ± 0.015	0.085 ± 0.012
BERT	**0.131 ± 0.007***	**0.051 ± 0.010**	**0.057 ± 0.007***	**0.028 ± 0.004***	**0.113 ± 0.031**	**0.017 ± 0.002***

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.152 ± 0.010	0.058 ± 0.010	0.133 ± 0.006	0.084 ± 0.008	0.129 ± 0.010	0.106 ± 0.010
Fast-Text	**0.067 ± 0.002***	0.057 ± 0.007	0.078 ± 0.007	0.060 ± 0.004	0.097 ± 0.014	0.074 ± 0.007
BERT	0.105 ± 0.003	**0.036 ± 0.006***	**0.065 ± 0.005***	**0.037 ± 0.002***	**0.081 ± 0.016**	**0.050 ± 0.004***

(c) DV dataset

Source: Prepared by the author.

Table 29 – Results obtained using FPED bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.182 ± 0.000	0.053 ± 0.000	0.103 ± 0.000	0.072 ± 0.000	0.108 ± 0.000	0.036 ± 0.000
Fast-Text	0.125 ± 0.000	0.052 ± 0.000	0.097 ± 0.000	0.068 ± 0.000	0.174 ± 0.000	0.049 ± 0.000
BERT	**0.094 ± 0.000**	**0.035 ± 0.000**	**0.028 ± 0.000**	**0.018 ± 0.000**	**0.067 ± 0.000**	**0.015 ± 0.000**

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.251 ± 0.005	0.078 ± 0.006	0.215 ± 0.009	0.117 ± 0.006	0.138 ± 0.015	0.065 ± 0.007
Fast-Text	0.232 ± 0.012	0.094 ± 0.009	0.162 ± 0.012	0.151 ± 0.011	0.171 ± 0.027	0.083 ± 0.009
BERT	**0.112 ± 0.011***	**0.043 ± 0.004***	**0.024 ± 0.003***	**0.016 ± 0.002***	**0.095 ± 0.018***	**0.013 ± 0.001***

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.206 ± 0.009	0.072 ± 0.014	0.160 ± 0.017	0.106 ± 0.004	0.151 ± 0.016	0.115 ± 0.007
Fast-Text	**0.084 ± 0.005***	0.062 ± 0.007	0.105 ± 0.010	0.065 ± 0.006	0.120 ± 0.006	0.078 ± 0.009
BERT	0.117 ± 0.004	**0.045 ± 0.006***	**0.081 ± 0.008***	**0.048 ± 0.006***	**0.102 ± 0.018**	**0.064 ± 0.005***

(c) DV dataset

Source: Prepared by the author.

Table 30 – Results obtained using Subgroup AUC bias metrics for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	**0.554 ± 0.000**	0.498 ± 0.000	0.547 ± 0.000	0.520 ± 0.000	**0.572 ± 0.000**	0.518 ± 0.000
Fast-Text	0.552 ± 0.000	0.532 ± 0.000	**0.553 ± 0.000**	**0.538 ± 0.000**	0.565 ± 0.000	**0.528 ± 0.000**
BERT	0.515 ± 0.000	**0.533 ± 0.000**	0.530 ± 0.000	0.533 ± 0.000	0.533 ± 0.000	0.516 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.498 ± 0.007	0.502 ± 0.008	0.490 ± 0.001	**0.513 ± 0.006**	0.480 ± 0.017	0.501 ± 0.002
Fast-Text	**0.518 ± 0.006**	0.507 ± 0.009	0.504 ± 0.002	0.503 ± 0.005	**0.517 ± 0.014**	0.501 ± 0.002
BERT	0.510 ± 0.006	**0.510 ± 0.008**	**0.515 ± 0.002**	0.508 ± 0.002	0.513 ± 0.016	**0.504 ± 0.002**

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.524 ± 0.007	0.515 ± 0.012	**0.538 ± 0.004***	**0.525 ± 0.004**	0.529 ± 0.009	**0.531 ± 0.006***
Fast-Text	0.500 ± 0.003	0.505 ± 0.008	0.511 ± 0.007	0.495 ± 0.004	0.515 ± 0.007	0.507 ± 0.005
BERT	**0.534 ± 0.003**	**0.525 ± 0.007***	0.526 ± 0.005	0.523 ± 0.005	**0.551 ± 0.020**	0.521 ± 0.003

(c) DV dataset

Source: Prepared by the author.

Table 31 – Results obtained using AUC for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.599 ± 0.000	**0.562 ± 0.000**	0.626 ± 0.000	0.599 ± 0.000	0.611 ± 0.000	**0.647 ± 0.000**
Fast-Text	0.622 ± 0.000	0.555 ± 0.000	**0.648 ± 0.000**	0.623 ± 0.000	**0.647 ± 0.000**	0.646 ± 0.000
BERT	**0.626 ± 0.000**	0.559 ± 0.000	0.638 ± 0.000	**0.624 ± 0.000**	0.605 ± 0.000	0.624 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.862 ± 0.007	**0.659 ± 0.009**	**0.885 ± 0.006**	0.840 ± 0.010	0.871 ± 0.008	**0.833 ± 0.010**
Fast-Text	0.864 ± 0.007	0.653 ± 0.012	0.885 ± 0.007	**0.841 ± 0.005**	**0.884 ± 0.006***	0.829 ± 0.010
BERT	**0.867 ± 0.005**	0.630 ± 0.017	0.870 ± 0.007	0.828 ± 0.009	0.864 ± 0.006	0.813 ± 0.011

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	**0.913 ± 0.005***	**0.672 ± 0.007***	**0.908 ± 0.004***	0.859 ± 0.005	0.911 ± 0.009	**0.856 ± 0.008**
Fast-Text	0.903 ± 0.005	0.655 ± 0.004	0.899 ± 0.007	**0.863 ± 0.003**	**0.915 ± 0.003**	0.850 ± 0.005
BERT	0.875 ± 0.007	0.610 ± 0.005	0.862 ± 0.005	0.806 ± 0.012	0.870 ± 0.008	0.785 ± 0.009

(c) DV dataset

Source: Prepared by the author.

Table 32 – Results obtained using macro F1-score for all datasets. The table shows the average obtained from the k-fold for each feature extractor combined with each classifier. The best feature extractor result for each classifier is boldfaced, and ties are underlined. Significantly better results are marked with $*$.

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.525 ± 0.000	**0.541 ± 0.000**	0.539 ± 0.000	0.544 ± 0.000	**0.527 ± 0.000**	0.579 ± 0.000
Fast-Text	**0.566 ± 0.000**	0.538 ± 0.000	**0.555 ± 0.000**	**0.571 ± 0.000**	0.517 ± 0.000	**0.589 ± 0.000**
BERT	0.500 ± 0.000	0.535 ± 0.000	0.500 ± 0.000	0.532 ± 0.000	0.512 ± 0.000	0.541 ± 0.000

(a) HE dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	0.661 ± 0.019	**0.556 ± 0.013**	0.703 ± 0.012	0.622 ± 0.014	0.707 ± 0.016	**0.615 ± 0.016**
Fast-Text	0.640 ± 0.019	0.551 ± 0.012	**0.704 ± 0.011**	**0.623 ± 0.010**	**0.726 ± 0.017***	0.606 ± 0.017
BERT	**0.702 ± 0.012***	0.515 ± 0.020	0.679 ± 0.014	0.604 ± 0.015	0.692 ± 0.011	0.577 ± 0.014

(b) WH dataset

Feature	LR	DT	SVM	XGB	MLP	RF
GloVe	**0.642 ± 0.005***	**0.536 ± 0.010***	**0.606 ± 0.005***	**0.569 ± 0.005***	**0.690 ± 0.013**	**0.579 ± 0.006***
Fast-Text	0.574 ± 0.016	0.515 ± 0.007	0.581 ± 0.008	0.546 ± 0.012	0.682 ± 0.016	0.547 ± 0.008
BERT	0.593 ± 0.010	0.458 ± 0.006	0.512 ± 0.011	0.478 ± 0.010	0.611 ± 0.035	0.483 ± 0.008

(c) DV dataset

Source: Prepared by the author.