INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Automatic Hate Speech Detection in Portuguese Social Media

Gil Antunes Silva Nogueira Ramos

Master in Data Science

Supervisors:
Doctor Fernando Manuel Marques Batista, Associate Professor, Iscte – Instituto Universitário de Lisboa

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

October, 2024

TECHNOLOGY
AND ARCHITECTURE

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Automatic Hate Speech Detection in Portuguese Social Media

Gil Antunes Silva Nogueira Ramos

Master in Data Science

Supervisors:
Doctor Fernando Manuel Marques Batista, Associate Professor, Iscte – Instituto Universitário de Lisboa

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

October, 2024

Acknowledgment

This work was funded in part by the European Union under Grant CERV-2021-EQUAL (101049306). However, the views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or Knowhate Project. Neither the European Union nor the Knowhate Project can be held responsible.

The work presented in this dissertation would not have been possible without the support I received along the way. From significant contributions to small, momentary gestures, each act of kindness and encouragement propelled me forward and motivated me to complete this dissertation to the best of my ability.

First and foremost, I would like to thank ISCTE, ISTAR-Iscte and the Knowhate Project for the opportunity to develop this work with their trust and support. I would also like to extend my gratitude to INESC-ID for granting me access to their virtual machines, that contributed to the completion of this work.

I would like to express my heartfelt thanks to my supervisors, Fernando and Ricardo, whose guidance, knowledge, and encouragement made this work possible. You always found time to offer your mentorship and help, especially during those late-night meetings on paper submission deadlines. Your support made these challenging moments feel manageable (and fun).

I would also like to express my gratitude to other professors who have influenced my master's journey, particularly Sérgio Moro and Diana Mendes. Thank you for your support and the opportunities you provided, which greatly enriched my academic experience.

And finally, I would like to express my wholehearted gratitude to my family and friends. To André and Mariana, for the camaraderie during our classes, and the moments we shared together in all our meetings "na Margem." To Considra, António, and Alberto, thank you for our enduring friendship, which has only grown stronger over these years. To Carolina, Nuno, Dulce and Salete, thank you for your support and encouragement, each in your own unique way. To Bia, thank you for your support in all aspects of my life, including this work. Even while we travelled the world together, you always reminded me to stay focused on finishing it. To my mother and father, thank you for always granting me the freedom to pursue my passions my entire life, regardless of the many fields I chose. Dad, thank you for introducing me to Data Science, with full confidence that I would succeed in it, it seems that, thanks to you, I finally found an area I will stick with. Mum, thank you for always showering me with love and support, making my life so much easier. I am truly grateful for everything you do for me.

Resumo

O rápido crescimento das redes sociais introduziu novas formas de comunicação digital, mas também contribuiu para um aumento preocupante do discurso de ódio online (DOO). Este fenómeno incentivou a investigação de métodos de processamento de linguagem natural para a deteção de discurso de ódio. Apesar dos avanços desta área, existe uma lacuna notável na investigação focada na variante europeia do português. Assim, investigou-se a eficácia de vários modelos de *transfer learning*, que estudos prévios demonstram ter neste contexto um desempenho superior em relação a abordagens tradicionais de aprendizagem profunda. Foram utilizados modelos do tipo *BERT*, como o BERTimbau e o mDeBERTa, pré-treinados em texto português, juntamente com modelos generativos como o GPT, o Gemini e o Mistral, para a deteção de DOO em conversações online no espaço português. Esta investigação recorreu a dois conjuntos de dados, anotados manualmente, compostos por comentários do YouTube e *tweets* do Twitter, ambos classificados como DOO ou não-DOO. Os resultados indicaram que uma variante do BERTimbau, re-treinada especificamente para deteção de DOO em *tweets* em português europeu, foi o modelo com melhor desempenho no conjunto de dados do YouTube. Este modelo obteve um *F-score* de 87,1% para a classe positiva, o que representa uma melhoria de 1,8% em relação ao BERTimbau original. No conjunto de dados do Twitter, o modelo GPT-3.5 foi o que apresentou melhor desempenho, com um *F-score* de 50,2% para a classe positiva, embora com resultados inferiores comparativamente ao conjunto de dados do YouTube. Adicionalmente, foram avaliados os efeitos do uso de conjuntos de treino *in-domain* versus *mixed-domain*, bem como o impacto da informação contextual nas *prompts* dos modelos generativos. Concluiu-se que os dados *mixed-domain* podem melhorar os resultados, desde que seja assegurada a sua qualidade, e que a informação contextual tem um impacto positivo tanto no DOO geral como no DOO indireto.

Palavras Chave: *Discurso de ódio*, *transfer learning*, *modelos transformer*, *modelos generativos*, *classificação de texto*

Abstract

The rapid rise of social media has brought about new ways of digital communication, along with a worrying increase in online hate speech (HS). This escalation has prompted researchers to develop various Natural Language Processing techniques for HS detection. Despite the progress made, there is a notable lack of research focused on the European Portuguese language, which is typical for many under-resourced languages. To fill this gap, we investigate the effectiveness of several transfer learning models that prior studies have indicated to outperform traditional Deep Learning approaches in this context. We utilize BERT-like models, including BERTimbau and mDeBERTa, pre-trained on Portuguese text, along with generative models such as GPT, Gemini, and Mistral, to identify HS in Portuguese online discourse. Our research is based on two annotated datasets comprised of YouTube comments and Twitter posts, both manually labelled as HS or non-HS. The results indicate that a retrained variant of BERTimbau, fine-tuned for the HS detection task using European Portuguese tweets, achieved the highest performance for the YouTube dataset, with an F-score of 87.1% for the positive class, showing an 1.8% improvement over the original BERTimbau. For the Twitter dataset, GPT-3.5 emerged as the top model, achieving an F-score of 50.2% for the positive class, with models having a far worse performance compared to when applied to the YouTube dataset. Additionally, we evaluate the effects of utilizing in-domain versus mixed-domain training sets and the role that contextual information in generative model prompts has on their overall performance, concluding that mixed-domain data has the potential to improve results, provided its quality is ensured, and that contextual information has a discernable impact in both general and covert HS.

KEYWORDS: *Hate speech*, *transfer learning*, *transformer models*, *generative models*, *text classification*

Contents

List of Figures

List of Tables

List of Acronyms

ML: Machine Learning

DL: Deep Learning

HS: Hate Speech

CHS: Covert Hate Speech

OHS: Overt Hate Speech

SLR: Systematic Literature Review

SVM: Support Vector Machine

LR: Logistic Regression

RF: Random Forest

DT: Decision Tree

BERT: Bidirectional Encoder Representations from Transformers

MLP: Multi-Layer Perceptron

CNN: Convolutional Neural Networks

LSTM: Long Short-Term Memory

BiGRU: Bidirectional Gated Recurrent Unit

GRU: Gated Recurrent Unit

RNN: Recurrent Neural Networks

BiLSTM: Bidirectional Long Short-Term Memory Network

NLP: Natural Language Processing

GAN: Generative Adversarial Network

GPT: Generative Pre-trained Transformer

IAA: Inter-Annotator Agreement

FP: False Postives

TP: True Postives

FN: False Negatives

TN: True Negatives

CHAPTER 1

Introduction

This dissertation focuses on the issue of detecting Hate Speech (HS) in online platforms, with a particular emphasis on the Portuguese language and on the Portuguese communities, and was developed in the context of the kNOwHATE: kNOwing online HATE speech project [1]. HS poses a significant threat to social harmony, and addressing it effectively requires advanced automated methods. This chapter introduces the problem, outlines the motivation behind the research, presents the goals and key research questions, and discusses the contributions of this work.

1.1. Context and Motivation

In recent years, the surge in social media usage has dramatically transformed the way individuals express themselves and engage with others [2]. With the widespread availability of smartphones and internet access, social media platforms have become accessible to a global audience, offering users the ability to share thoughts, opinions, and ideas freely. This democratization of expression has fostered empowerment and meaningful dialogue, but it has also brought significant challenges, most notably the proliferation of HS [3], which poses a severe threat to social cohesion, online communities, and society at large.

While there is no universally accepted definition of HS [4], the United Nations defines it as any form of communication that uses derogatory or discriminatory language directed at individuals or groups based on intrinsic identity factors [5]. Such communication can lead to severe emotional and psychological distress among its targets, manifesting as stress, anxiety, or even depression [6]. Moreover, prolonged exposure to HS erodes societal bonds, fostering an atmosphere of mistrust and polarization. This fragmentation often leads individuals to retreat into echo chambers, where biases and prejudices are amplified [7].

The scale of the problem has driven numerous organizations to implement policies aimed at limiting the spread of HS on digital platforms. However, given the vast volumes of data generated daily, manual moderation methods are impractical. Consequently, the need for automated systems capable of detecting and classifying HS has become increasingly urgent. In response, a wide range of Machine Learning (ML) techniques have been developed to address this challenge. From traditional ML models to more recent advances in Deep Learning (DL) and Transformer-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT) [8]. All these approaches have shown promising results in the detection of HS [9].

1.2. Background

As previously stated, defining HS is not an easy task, since this is a complex phenomenon that is heavily reliant on the subtleties of language. It is nonetheless necessary to understand how HS is defined and what constitutes it, in order to begin to detect and combat it. Many organizations, companies and countries have defined HS in their policies and below we can see some examples of these definitions. Since this work was developed in the scope of the kNOwHATE project, we also provide the definition used in it:

- United Nations: "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" [5].
- Meta hate speech policy: "a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation" [10].
- Twitter (now X) policy on hateful conduct: "attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" [11].
- YouTube hate speech policy: "content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy: 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" [12].
- Definition in the eBook *The Content and Context of Hate Speech*: "is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature... stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as undesirable... casts the target group as an undesirable presence and a legitimate object of hostility" [13].
- kNOwHATE project: building on scholar definitions [i.e., 14] and guidelines provided by the Council of Europe in its latest recommendation (CM/Rec/2022/16), the project defines online HS as "bias-motivated, derogatory language that spread, incite, promote, or justify hatred, exclusion, and/or violence/aggression against a person/group because of their group membership" [15].

When examining the various interpretations of HS used by multiple organizations and research initiatives, we can identify some similarities. Firstly, all definitions mention

that HS targets a specific group or individual based on group membership, and not concepts or institutions. Secondly, these groups are targeted with malicious intent, based on real or attributed characteristics, and some organizations consider these characteristics as protected. Depending on the characteristic that is being targeted, there are different categories of HS. The main characteristics mentioned in the aforementioned definitions include religion, ethnicity, nationality, race, colour, descent, gender, and sexual orientation.

As previously stated, this work was developed in the scope of the kNOwHATE project. This project aims to combat online HS by implementing an innovative, interdisciplinary, and participatory approach that integrates social and linguistic sciences with computational techniques. The primary goals of kNOwHATE include gaining a comprehensive understanding of the psychosocial and linguistic characteristics of HS, analysing its content and propagation within user-generated content in Portugal, and creating automatic detection models informed by scientific knowledge and the lived experiences of targeted communities. This work specifically contributes to the latter, by developing and evaluating ML models for detecting HS, thereby providing essential insights and tools for effectively addressing this pressing social issue.

1.3. Goals and Research Questions

The primary goal of this work is to advance the field of HS detection by exploring the effectiveness of modern ML techniques, particularly focusing on novel Transformer-based and generative models, in the context of the European Portuguese language. More specifically, this work aims to firstly provide a comprehensive overview of HS detection research, including the evolution of methods and datasets used. By systematically reviewing the existing literature, we aim to map out the current state of research and inform later stages of this work. Another key objective is to investigate the performance of Transformer-based models designed for Portuguese HS detection and comparing their effectiveness to traditional ML and DL approaches. This evaluation will assess whether these models, when applied to real-world social media data from platforms such as YouTube and Twitter, yield measurably better results in detecting HS. Additionally, this work explores the potential of mixed-domain learning and context-enhanced generative models in improving the detection of both overt and covert forms of HS. A central aspect of this goal is to evaluate how generative models like Generative Pre-trained Transformer (GPT) [16], Gemini [17], and Mistral [18] perform in comparison to BERT-based models. By incorporating contextual data into generative models, this work seeks to address the challenge of detecting Covert Hate Speech (CHS), which is often more nuanced and difficult to identify. Through these efforts, this research aspires to fill the identified gaps in the literature, particularly in the context of HS detection in the European Portuguese language.

To achieve the objectives outlined above, the following research questions will be explored in this work:

- RQ1: What is the landscape of HS detection research since the advent of Transformer-based models, and how has it evolved in terms of methods and datasets used?
- RQ2: How do Transformer-based models, specifically designed for HS detection in Portuguese, and generative models perform in comparison to traditional ML and DL approaches?
- RQ3: Can mixed-domain learning and the incorporation of context in generative models improve the detection of HS, and its covert forms, in Portuguese social media data?

1.4. Contributions

This work makes significant contributions to the field of HS detection, with a particular focus on the Portuguese language, through a series of publications that address various aspects of this problem.

The first major contribution is the following SLR of the topic: [19]. This review mapped the landscape of HS detection research, specifically analysing the evolution of the methods used for HS detection since the development of Transformer-based models. By systematically reviewing existing literature, this article filled a critical gap by offering a comprehensive overview of the state of the art in HS detection, focusing on how Transformer-based solutions compared to more traditional approaches. This review also highlighted the scarcity of research targeting HS detection in European Portuguese, thereby identifying a gap in the literature that laid the groundwork for the subsequent experimental studies.

The second contribution explored the effectiveness of generative models and Transformer-based models designed for Portuguese HS detection: [20]. In particular, a comparison between the performance of HS domain-specific models with more general models. We also experimented with mixed-domain learning to assess whether training on diverse data sources, such as YouTube and Twitter, improves model generalization. In addition, experiments were also made with generative models. The results of this study provided new insights into the relative effectiveness of different model architectures for HS detection in Portuguese, filling the gap in understanding how Transformer and generative models perform in comparison to traditional ML and DL approaches when applied to Portuguese social media data. Additionally, this work also contributes with two new datasets developed specifically for the HS detection task, that are one of a kind in terms of number of annotated messages and its annotation schema.

The third contribution focused on the detection of CHS, which is inherently more difficult to identify than Overt Hate Speech (OHS): [21]. This work utilized generative models (GPT-3.5-turbo and GPT-4-turbo), and explored how incorporating context could improve the detection of both overt and covert forms of HS. This study filled a key gap by addressing the challenge of detecting CHS with generative models, an area previously

underexplored in the HS detection literature. It provided a novel utilization of context-enhanced generative models for the task, contributing to the broader understanding of how context can influence model performance in detecting nuanced forms of HS.

Each of these publications not only contributed individually to advancing the field, but this dissertation, being the combination of the knowledge present in each work, addresses significant gaps in HS detection research, particularly in the European Portuguese language. By tackling issues related to the evolution of HS detection techniques, the performance of Transformer-based and generative models, and the detection of CHS, this work provides a comprehensive and multi-faceted contribution to the growing body of knowledge on automated HS detection.

Lastly, a significant contribution of this work is the development of a dedicated Hugging Face space for HS detection, where users can interact with the models developed in this research: `https://huggingface.co/knowhate`. This platform allows the public to perform real-time HS classification on any given text, providing an accessible and practical tool for HS detection.

1.5. Document Structure

This dissertation is structured in four chapters: In Chapter 1 (Introduction), the context and motivation behind the research are presented, followed by a detailed discussion of the background, goals, research questions, and contributions. Chapter 2 (Related Work) provides a comprehensive Systematic Literature Review (SLR), encompassing works from 2018 until 2024 and analysing the methods and data used for HS detection. In Chapter 3 (Automatic Classification of Hate Speech), the datasets used in this work are introduced and described in detail, in addition to the models and experiments conducted for HS detection, including the setup, the different experiments conducted, and the results obtained. Finally, Chapter 4 summarizes the key findings of this work and outlines potential directions for future research.

CHAPTER 2

Related Work

As mentioned previously, there is a growing need for automated systems capable of detecting and classifying HS, and as a consequence various techniques, ranging from approaches like traditional ML and DL models, have been applied for this task. With the development of Transformer-based models [22], we have seen a growing expansion in the HS detection landscape. The recent advances in Transformer-based models have introduced new possibilities, but a comprehensive synthesis of these efforts is lacking, particularly in terms of comparing them to other ML methods.

This SLR addresses this gap by exploring the current research landscape of HS detection on social media, with a specific focus on Transformer-based models, by answering the following questions:

- Q1: What is the landscape of HS detection literature since the development of Transformer-based models?
- Q2: How do Transformer-based models compare to other ML solutions in the context of HS detection?
- Q3: What are the characteristics of the data used for HS detection?

2.1. Prisma Methodology

This section presents an overview of the methodologies employed in this SLR. In developing our methodology, we drew inspiration from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [23]. PRISMA provides a widely recognized framework for conducting systematic reviews, ensuring transparency and methodological rigour in the review process. The structured approach outlined in PRISMA facilitated a comprehensive overview of the methodologies employed in our SLR, covering key aspects from search criteria delineation to data extraction. Our goal was to adhere to the principles of PRISMA to enhance the reliability and reproducibility of our review and to ensure a robust and exhaustive coverage of the literature under review.

Our primary goal is to provide an analysis focusing on key trends in performance across different methods employed in the field of HS detection within the context of social media. Specifically, our review seeks to address the following key objectives: First, we aim to examine the ML and Natural Language Processing (NLP) methods that have been utilized for the identification and classification of HS in social media platforms and how they have changed with the introduction of Transformer models, to better understand what are the current trends and future perspectives (Q1); Then, we compare the several methodologies employed with one another and with Transformer models, to identify which

ones achieve better results (Q2). Finally, we analyse the characteristics of the resources being used in the scope of this task, like languages and data sources, to identify which areas can be further developed (Q3).

In the end, we also delve into the current challenges and limitations that researchers face in this domain, with a focus on proposed strategies and potential solutions. By addressing these goals, we aim to contribute valuable insights into the state-of-the-art in HS detection, thus facilitating a better understanding of the field and its future directions.

To accomplish this, we first defined criteria to search and select studies to be examined in our SLR, relevant to our objectives. We selected two databases, Scopus and Web of Science, since they both have an extensive coverage of literature, across diverse academic fields. This is beneficial, since HS detection can be seen as a multidisciplinary problem ranging from linguistics and social sciences to computer science, so it is necessary to search in databases that index a wide range of journals, in a variety of disciplines.

The search query was designed to maximise the retrieval of studies pertinent to our subject, and for that the following keywords were established: 'hate speech', 'abusive', 'offensive', 'classification' and 'detection'. 'Hate speech' is the most common keyword used in this subject by the scientific community, since it is also a legal term in many countries. The terms 'offensive' and 'abusive' were also added since they convey a similar idea, in the sense that HS can be seen has an extreme of abusive text, and all of them share an offensive aspect [24]. These terms are also present in the literature as key terms to use when finding relevant studies [25]–[27]. These keywords were used in addition to Boolean operators to form our search query ("hate speech" OR "abusive" OR "offensive") AND ("classification" OR "detection"). Our query was applied to the following parts of the studies: title, abstract and keywords.

To define which articles should be included or omitted from our SLR some inclusion and exclusion criteria were set to keep only the studies that fulfilled our goals for this work.

The inclusion criteria where the following: firstly, to capture the most recent developments in the field, and since we want to focus on Transformer-based models, we limited our search to studies published from 2017 to the present day, since it was in 2017 that the Transformers' architecture was introduced [22], and with that came a growing interest in this area. Furthermore, to facilitate the comprehension and analysis of the research, we restricted our selection to studies written in the English language. To assure high-quality and peer-reviewed research, only journal articles were considered for inclusion, while conference papers, data papers, and similar publications were excluded. Additionally, we aimed to select studies that were published in journals with a high impact factor, specifically those ranked in Quartiles 1 and 2 based on Scimago[1] journal quality rankings. Given the emphasis of this review on HS classification, we prioritized articles whose

[1] www.scimagojr.com

primary focus centred on this specific area of research and that proposed or discussed solutions related to this classification task.

The exclusion criteria were the following: articles primarily focused on other forms of media, such as images, sound, memes, and non-textual content, articles that lacked a clear approach or technical content related to HS classification, and finally, studies that did not centre their main objectives on HS detection, but on another task, like the development of HS resources.

Although we decided to include only journal articles, we recognized that by excluding high impact peer-reviewed conferences we were limiting the inclusion of cutting-edge research, so in order to mitigate this side effect we decided to include the most relevant papers of two tasks held in the context of the SemEval international workshops of 2019 and 2020, published by the Association of Computational Linguistics (ACL). In these years' editions, the OffensEval task was held, that aimed at detecting offensive language. By including the most relevant studies of a competition with a high degree of participation, we believe we get a glimpse of that time's best techniques for the task. Additionally, to ensure comprehensive coverage of recent innovations, we extended our search to include ACL conference papers published between 2020 and 2024 that met our inclusion criteria, specifically selecting long papers from the main conference proceedings.

Figure 2.1 shows the number of records identified in the database search, and the filtering process that is applied afterwards, using a PRISMA flow diagram [28]. Our initial query resulted in 2876 studies, plus the 15 ACL studies selected. After the removal of duplicate entries, and the application of exclusion criteria, we were left with 105 articles for full-text analysis. After assessing the full text of the 105 articles selected from our inclusion/exclusion criteria, an additional three articles were discarded because the dataset used for HS detection was not manually annotated, but instead algorithms were used to automatically annotate the data used for building the classifiers [29]–[31]. Given the nuanced and context-dependent nature of HS, the reliance on automated processes for annotation introduces potential biases and inaccuracies that may compromise the robustness and reliability of the classifiers developed in these studies, leading to the final 102 articles considered for our SLR.

For the full-text analysis of our studies, data extraction is a critical component, since it helps to collect information in a methodological and comprehensive way, so we employed a rigorous and systematic approach that involved the identification and extraction of key elements from each study, that answered our initial objectives. The data collected was mainly about the datasets utilized in each study, the methods they used for the classification task (algorithms, pre-processing, feature representation, etc.), the metrics used to evaluate the performance (with the actual values obtained) and the principal findings and limitations. For this, an extraction form was used in order to ensure consistency.

The findings are divided into four distinctive categories: An overall analysis of the results of our search (Section 2.2), an analysis of the evolution of HS detection (Section 2.3),

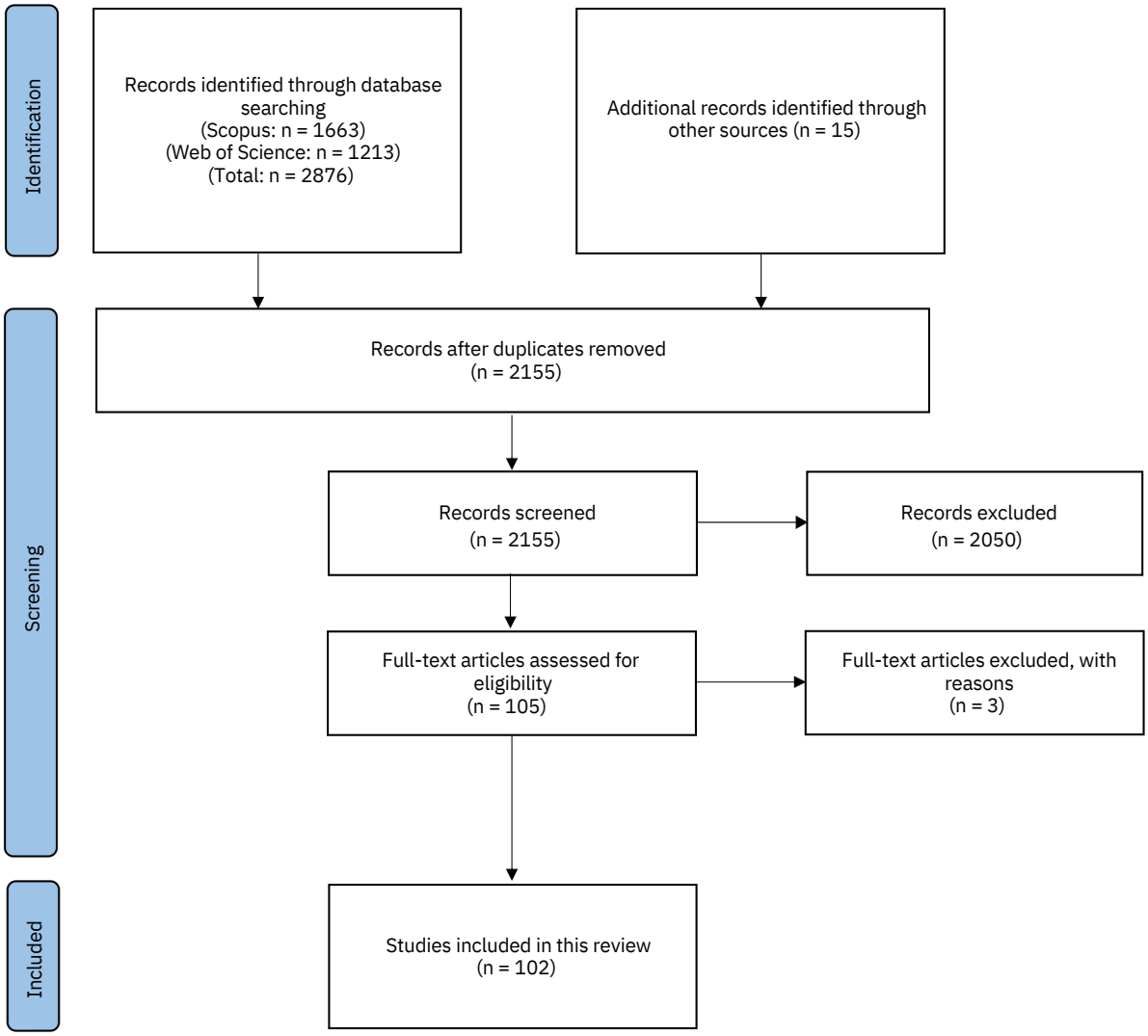

FIGURE 2.1. PRISMA flow diagram

Methods and Algorithms, where we will compare all different approaches employed for this task (Section 2.4), and Resources where both the languages and the types of data used for the detection will also be analysed (Section 2.5).

This SLR focuses on analyzing studies related to HS detection, especially those that define HS within comprehensive frameworks. It also includes studies addressing offensive and abusive speech, recognizing that these types of speech are frequently discussed alongside HS in the literature. Although offensive and abusive speech do not involve targeting individuals based on group membership (as is the case with HS) [15], the detection methods used for these types of speech are quite similar. So, in order to maintain clarity, the remainder of the SLR refers to these collective studies (HS, offensive, and abusive speech detection) as HS detection works. Nevertheless, Section 2.5 presents specific statistics about the number of studies addressing each type of speech, as this breakdown may be of interest to certain readers. This approach provides a clear and concise way to streamline

the discussion while still offering the detailed analysis and statistical information for those who may want to differentiate between the types of speech.

2.2. Overall Analysis

When analysing the initial results of the 2155 articles (after the removal of duplicates) resulting from our search query, we can see in Figure 2.2 a notable upsurge in the volume of studies related to HS detection, confirming the increasing significance of this topic within the research community. Over the years, we observed a considerable growth in publications, with the data indicating a substantial increase in the number of studies published annually. In 2017, 65 relevant studies were identified, which increased almost 10 times to the 588 results found in 2022. Since the search was conducted in September, and the current year (2023) has not come to an end at the time of writing, the lower number of publications found (369) is not surprising. We have also added the number of documents included in our SLR from each year. This graph confirms the growth of this research topic and the need for an updated review.

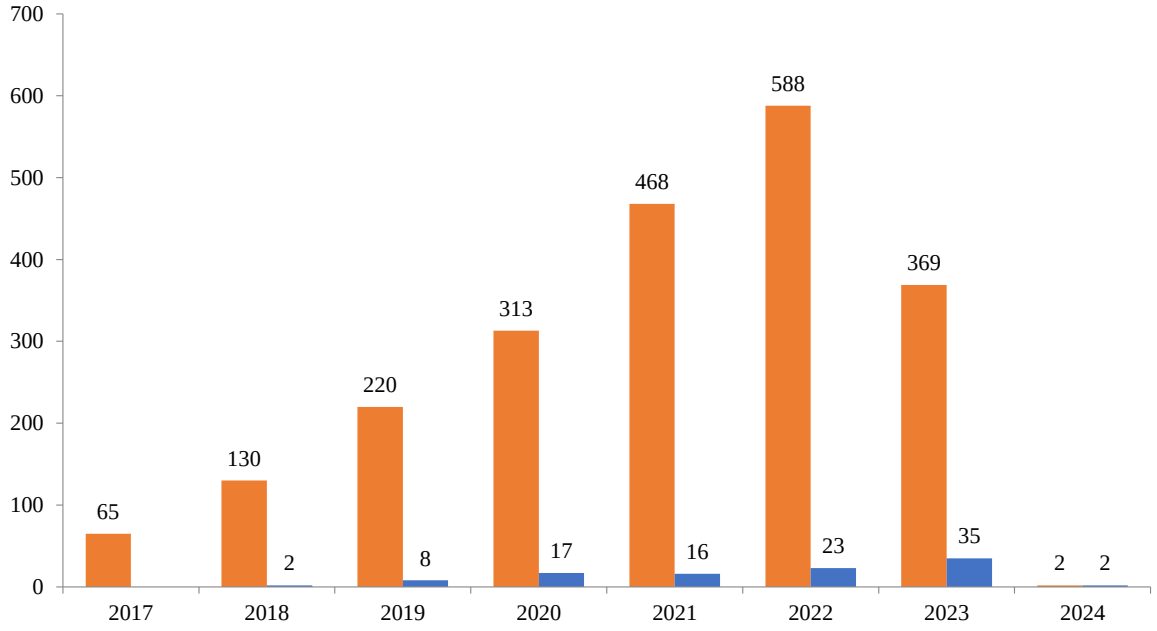

FIGURE 2.2. Number of documents by year.

Our search across the Scopus and Web of Science databases yielded a substantial number of results, with 1663 studies identified in Scopus and 1213 in Web of Science. The presence of these studies across both platforms emphasizes the widespread recognition and coverage of the topic within the academic community, while also reflecting the diversity of academic sources that contribute to this discourse.

Categorically, the types of studies were delineated into two main groups: conference papers and journal articles, has shown in Figure 2.3. The data demonstrated that conference papers constituted most of the studies, with 1645 identified. In contrast, 1025

studies were classified as journal articles. This can be explained in part by the number of competitions dedicated to the task of HS classification [32]–[34], from which many conference articles result, since each participant has their contribution in the form of a conference paper.

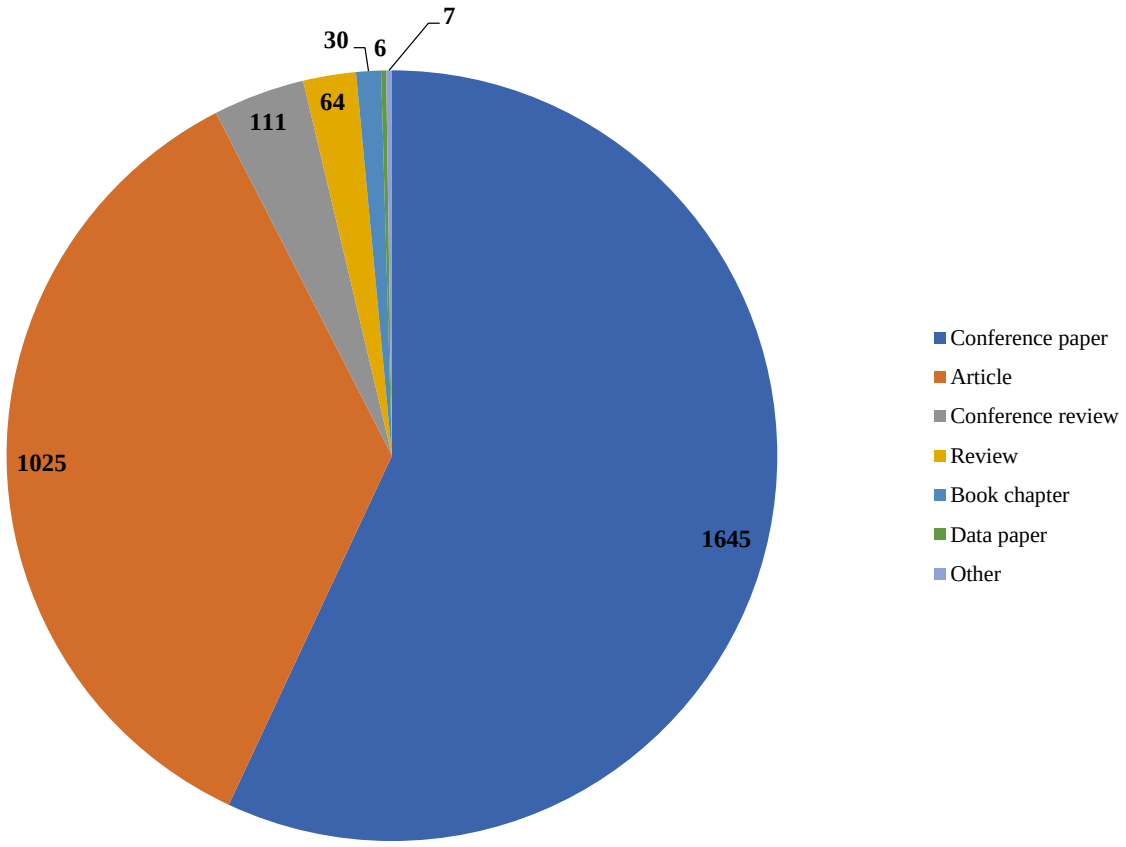

FIGURE 2.3. Number of documents by type.

Our initial search results show the growing prominence of HS classification as a research field, the substantial volume of studies dedicated to the topic, and the diverse types of publications contributing to this evolving discourse. This data forms a valuable foundation for the subsequent synthesis and filtering of the findings in our initial search. Moving forward, the results presented will be of the final 102 studies considered for this SLR.

2.3. Q1: Landscape of HS detection literature

Over the years, various approaches have been employed for HS detection, with notable evolution in the methods used. This section provides an overview of the five major approaches – Traditional ML, DL, Transformers, Generative Models, and Multi-Task Learning – and examines their progression and impact on HS detection throughout the years.

Figure 2.4 illustrates the evolving trends in the application of different approaches to HS detection, highlighting a clear shift in techniques over time. By 2019, DL techniques

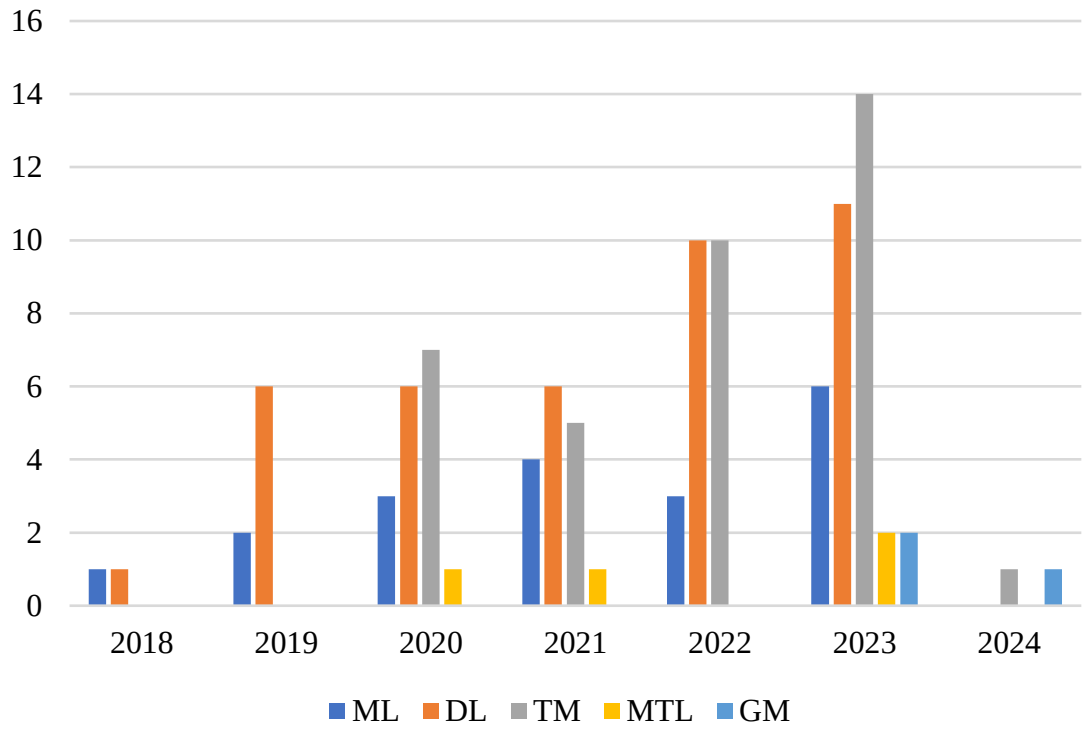

FIGURE 2.4. HS detection approaches by year (MTL: Multi-Task Learning).

became more prevalent, reflecting the growing interest in neural network-based methods for HS detection. This increase aligns with the first OffensEval task, where most participants employed DL models, marking them as the state-of-the-art approach at that time.

In 2020 and 2021, the landscape of HS detection continued to evolve. Transformer-based models began to gain significant traction, with seven studies in 2020 and five in 2021. This surge in popularity aligns with the introduction of the Transformer architecture by Vaswani *et al.* [22], which took about three years to be widely adopted for HS detection. The second OffensEval task further solidified this trend, as most competitors shifted to BERT-based models, confirming that Transformers had become the dominant approach during this period. Although traditional ML methods continued to be used, Multi-Task Learning emerged for the first time, with one study appearing in both 2020 and 2021.

In 2022 and 2023, we observed a more diverse set of approaches in HS detection. DL remained prominent, while Transformers continued to grow in popularity, becoming the go-to method with 10 studies in 2022 and 14 in 2023. Traditional ML techniques remained relevant, but with less usage compared to both methods mentioned before.

Generative and Multi-Task Learning models, newer approaches in the field, began to gain recognition in 2023, highlighting their potential for HS detection. In 2024, one study featuring Transformers was published, and another using generative models, both

coinciding with ACL papers extracted after the search, explaining their presence as the only studies from that year.

Fig 2.5 shows the total number of studies that employed each approach. DL and Transformers are the most frequently used methods, with 40 and 37 studies, respectively, accounting for over two-thirds of the research reviewed. Traditional ML follows with 19 studies, while Multi-Task Learning and generative models are represented by four and three studies, respectively. These findings underscore the significant impact of Transformers on the HS detection landscape, as they have become the preferred choice for many researchers in recent years.

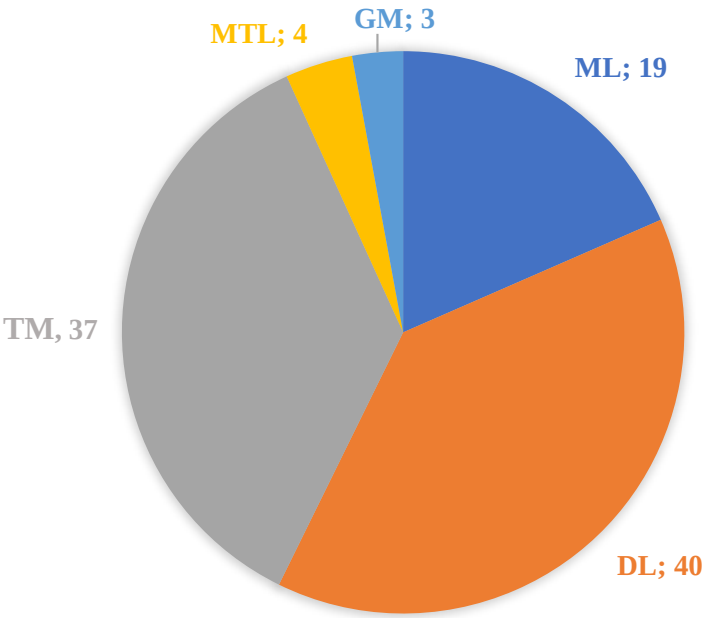

FIGURE 2.5. Number of HS studies, grouped by approach.

The results presented may be limited by the relatively small number of articles included in our analysis, potentially misrepresenting broader trends. To address this, we supplemented our review with conference papers from the top participants in OffensEval-2019 and OffensEval-2020, as well as other selected ACL papers, providing a more comprehensive representation of the state-of-the-art methods during that period. As shown in Table 2.1, the results from these conferences align with our findings, demonstrating a clear transition from ML and DL approaches in 2019 to the adoption of Transformer-based models in 2020. Furthermore, the authors of the OffensEval-2019 reported that over half of the participants explored DL models [32]. In contrast, OffensEval-2020 saw most teams utilizing pre-trained Transformer models, with all the top 10 teams employing either BERT, RoBERTa, or XLM-RoBERTa [33], which further confirms our findings.

In summary, the evolution of HS detection methods shows a clear shift from traditional, simpler ML techniques to more advanced DL and Transformer-based models. The field has also seen a growing diversity of approaches, with generative models and Multi-Task

TABLE 2.1. SemEval top papers.

Paper	Model	Method	Rank
OffensEval-2019			
[35]	SVM model with RBF kernel	ML	1st
[36]	stacked BiGRUs	**DL**	2nd
NA[a]	Multiple Choice CNN	**DL**	3rd
[37]	LSTM	**DL**	4th
[38]	linear-kernel SVM	ML	1st (Spanish task)
OffensEval-2020			
[39]	Ensemble of ALBERT models	**Transformer**	1st
[39]	RoBERTa-large	**Transformer**	2nd
[40]	XLM-R-base and XLMR-large	**Transformer**	3rd
[41]	XLM-R	**Transformer**	4th
[42]	BERT	**Transformer**	5th

[a]Participants did not publish paper

Learning gaining prominence in recent years. This progression highlights the dynamic nature of the research landscape and the continuous efforts to enhance HS detection in digital environments.

2.4. Q2: ML solutions for HS detection

As previously discussed, a wide range of approaches have been employed for HS detection, from traditional ML methods to more advanced DL and Transformer-based models. This section compares these approaches to determine which methods yield the most promising results and whether Transformers have consistently outperformed other models. To facilitate this comparison, we categorize the studies into five distinct approaches. Before examining each in detail, we provide a brief summary of each category to clarify their key differences.

ML focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit programming. The core idea is to allow machines to learn patterns and make decisions based on data. DL is a subset of ML that employs neural networks with many layers, that are more complex than traditional ML models. Multi-Task Learning is an approach where a single model is trained to perform multiple related tasks simultaneously. The goal is to enable the model to learn shared representations and features across tasks, potentially leading to improved performance compared to training separate models for each task. Generative Models are a class of ML models that aim to generate new data samples that resemble a given training dataset, increasing the amount of data available for training. Finally, Transformer-based approaches use transfer learning, by taking advantage of models pre-trained on large datasets for unsupervised tasks that capture general language patterns, and fine-tuning them with smaller labelled datasets on specific tasks, leveraging this pre-existing knowledge. This transfer of knowledge allows the model to generalize well to diverse tasks, enhancing performance and efficiency.

In the subsequent sections, we delve into the findings of studies adopting each of these approaches, assessing their effectiveness and making comparisons with one another.

2.4.1. Traditional Machine Learning

Starting with traditional ML techniques, we identified 19 studies that resorted to this type of method, and made comparisons with various algorithms. Support Vector Machine (SVM) and Logistic Regression (LR) were the algorithms that achieved better results, outperforming other ML algorithms in three different studies respectively. Pitropakis *et al.* [43], Shannaq *et al.* [44], and Mohapatra *et al.* [45] obtained better results with a combination of SVM with n-grams and pre-trained embeddings, when compared with other traditional ML models. Indurthi *et al.* [35] and Pérez and Luque [38] managed to obtain good results with an SVM model with a RBF and linear kernel respectively, topping the standings in the OffensEval-2019 task. Arcila-Calderón *et al.* [46], Vanetik and Mimoun [47], and Saeed *et al.* [48] employed an LR model with pre-trained embeddings and managed to outperform other traditional ML models. Other models that obtained good results were Random Forest (RF) with count vectorizer embeddings, that managed to outperform Bagging and Adaboost models [49], and the j48graft classifier, a type of Decision Tree (DT) model, combined with text features [3].

Recently, pre-trained Transformer embeddings have been used in combination with traditional ML models to improve performance. By using these embeddings as input features for traditional ML models, they benefit from their ability to capture intricate relationships and context in the text data, which can be challenging for traditional feature engineering methods. [47], [50] combined BERT embeddings with a Multi-Layer Perceptron (MLP) and LR models, respectively, and managed to outperform ML and ensemble models. In addition to this, [47], [51] showed that combining traditional ML models with BERT embeddings can even outperform DL and Transformers on its own.

Ensemble models have gained prominence in the realm of HS detection, as a strategic approach to overcome limitations associated with individual models. These models involve combining predictions from multiple models to enhance overall performance, making them a compelling alternative for addressing challenges posed by the use of single models in HS detection. Seven studies used an ensemble of ML models, and although these models did not outperform Transformers and DL models, they managed to outperform single ML models, showing that they can enhance the performance of these simpler models, by combining them. Four of these models used majority voting to get the predictions [52]–[55], two studies used a LR meta classifier [56], [57], and one study used a stacking approach [26].

Traditional ML models can be used effectively for the task of HS detection, and recent improvements show that this type of simpler model, when combined with a richer textual representation, or in an ensemble with other simple models, can even surpass

more complex models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional Gated Recurrent Unit (BiGRU) and BERT based models [47], [48], [51].

2.4.2. Deep Learning

Jumping to DL techniques, these have been extensively used for the task of HS detection, with 40 studies employing this approach. These studies have explored a variety of DL models, including CNNs, LSTMs, Gated Recurrent Unit (GRU)s, and hybrid or ensemble models that combine multiple DL architectures as we can see in Figure 2.6.

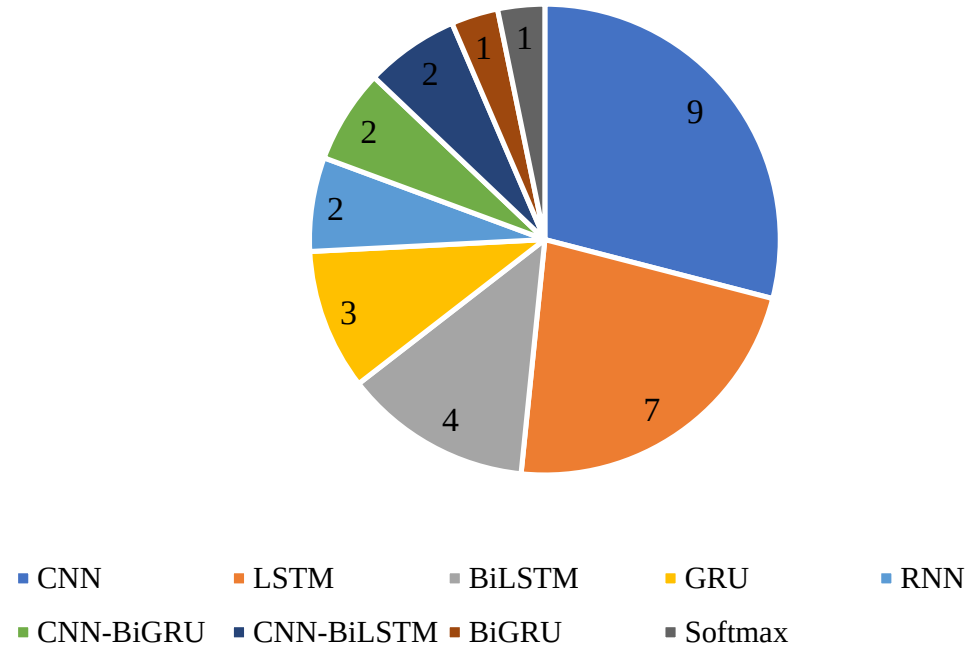

FIGURE 2.6. Different DL models for HS detection.

CNNs have been used to effectively capture the local patterns and features of text, making them well-suited for identifying HS. They have been applied in several HS detection studies [58]–[65] with promising results, even outperforming Transformers [64], and getting 3rd place in OffenseEval-2019.

LSTMs are another class of Recurrent Neural Networks (RNN)s that are capable of capturing long-range dependencies in text. This makes them well-suited for handling the sequential nature of language, which can be important for identifying HS. They have also been used in several HS detection studies [37], [66]–[71].

Both CNNs and LSTMs are two of the most widely used DL architectures for HS detection. CNNs can capture local patterns and features in text, while LSTMs are adept at handling long-range dependencies. The results of using CNNs and LSTMs for HS detection are somewhat mixed, with some studies have shown that CNNs outperform LSTMs [60], [61], while others have found the opposite [67], [68], [71].

Taking advantage of these mixed results, hybrid models that combine these two types of models have consistently shown strong performance. These models leverage the strengths of each architecture, leading to improved results and generalizability. For example, CNN-Bidirectional Long Short-Term Memory Network (BiLSTM) models have been shown to outperform even Transformers in some studies [72], [73]. This suggests that hybrid models may be able to more effectively capture the complexities and nuances of HS. In addition, CNN-BiGRU models have also shown promising results, by combining the local feature extraction ability of CNNs with the long-range dependency modelling ability of BiGRU's they managed to outperform all other single DL models [74], [75]. Nine other studies used an ensemble approach of DL models managing to outperform single DL and ML models, and in some cases even the state-of-the-art Transformers. A majority voting ensemble of several LSTM models with different features [76], a meta classifier of several combinations of models with different embeddings [77], a combination of a BERT, BiLSTM and BiGRU models [78] and finally a deep neural network with several text features [79] all managed to outperform ML and DL models with good results. In addition, five other studies managed to get better results than all other approaches (ML, DL and Transformer models). These studies employed an ensemble of CNN models [80], BERT models [81], a bagging of BiGRU, BiLSTM and CNN [82], a stacking of BiLSTM, LSTM, CNN and CNN-LSTM models [83] and a combination of a BERT, MuRIL and Deep Neural Network models [84].

Ensembles emerge as a compelling solution to HS detection, especially when individual models like CNNs or LSTM's do not perform well. By leveraging the strengths of diverse architectures and addressing limitations in generalization and imbalanced datasets, ensembles offer a robust and effective approach for enhancing the accuracy and reliability of HS detection systems even managing in some cases to outperform the state-of-the-art models.

Similarly to LSTM's, GRU's are also a type of RNN's that are capable of capturing short-term dependencies in text. They were used in three HS detection studies, even doe the comparisons were made with traditional ML models, that they outperformed [85]–[87]. Another study used a BiGRU model, managing to place top two in the OffensEval-2019 task [36]. Other DL models used were Bidirectional RNNs [88] and a Softmax classifier combined with text features [89].

These studies demonstrate the versatility and effectiveness of DL techniques for HS detection. DL models can capture complex patterns in text, making them well-suited for identifying subtle and nuanced forms of HS. Additionally, hybrid models can combine the strengths of different DL architectures to further improve performance.

2.4.3. Transformer-based Models

The Transformers were by far the ones that achieved the most promising results, surpassing the state-of-the-art models in almost all studies that employed them, outperforming all other approaches in most cases. It was also the most used approach, with 37 studies.

The success of the basic BERT model on a plethora of different NLP tasks lead to the widespread use of these models and many variants. This is mirrored on the large number of studies that employed this models for HS detection.

A fine-tuned version of the basic BERT model for the English language was used in nine studies [42], [90]–[97], outperforming all DL and ML models compared in the respective studies. Other variants of the BERT model that were retrained in other languages were also implemented, like BETO for Spanish [98]–[101], RuBERT for Russian [69], [102], RoBERTuito also for Spanish [103], UmBERTo for Italian [104], MARBERT for Arabic [25], HindiBERT for Hindi [105], Arabic BERT-mini also for Arabic [106], MuRIL for seventeen Indian languages [107] and NAIJAXLM-T for English and Nigerian [108]. It is also relevant to mention that this list goes beyond the set of articles found by our SLR and includes models such as BERTimbau widely used for Portuguese [109], [110] and Bertie for Dutch [111]. Besides these BERT models retrained for other languages, there are also multilingual models being developed like mBERT and XLM-RoBERTa that were trained with multilingual data and can be used in many languages. The mBERT model was used in four studies [107], [112]–[114] and the XLM-RoBERTa was used in five studies [40], [41], [115]–[117]. In addition to the models retrained on other languages, there have also been models with different architectures or hyperparameters than BERT, also used for HS detection like RoBERTa [39], [113], [118]–[121], ELECTRA [122] and AlBERT [39].

Transformers emerged as the most promising strategy for HS detection, consistently outperforming other methods across all studies. The versatility and adaptability of TM, coupled with the development of specialized variants and hybrid approaches, have significantly advanced the field of HS detection, paving the way for more comprehensive and effective measures to combat online HS.

2.4.4. Generative Models

As we have seen, there has been a recent surge in the use of generative models, with two studies employing this method in the year 2023 and one in 2024. Su *et al.* [123] utilized a Semi-Supervised Learning Generative Adversarial Network (GAN) architecture. The model incorporates RoBERTa sentence features as the backbone, combining them with a generator that introduces random noise and a discriminator for adversarial training. In this study, the authors also used vast amounts of unlabelled data from another related domain, and demonstrated that the generative model outperformed the baseline RoBERTa model without the additional data generation. In another study, Cohen *et al.* [124] combined multiple generative models for HS detection. This model utilizes DeBERTa Large as a foundational element and incorporates back-translation augmentation to enhance the diversity of the training dataset. Furthermore, the integration of GPT and Test-Time Augmentation demonstrated superior performance compared to baseline models, highlighting the effectiveness of generative models in achieving state-of-the-art results in HS detection. Finally, Zhang *et al.* [125] also used GPT-3.5 for this task.

The combination of pre-trained language representations, in this case RoBERTa and DeBERTa, and generative capabilities allows these models to capture intricate patterns and nuances present in HS texts. Generative techniques facilitate the augmentation of the training dataset, addressing issues related to limited labelled data in HS detection scenarios, like is the case with low-resource languages. This, in turn, enhances the generalization capabilities of the models, ensuring better performance on unseen HS text. In addition, adversarial training allows models to discern subtle differences between authentic and deceptive HS content, contributing to heightened discriminative power in HS detection. The utilization of generative models in HS detection has the potential to address one of the most common challenges in HS detection scenarios, being the lack of training data, that needs to be manually collected and annotated. With the introduction of these models, HS detection in low-resource languages can be done, without the need of extensive collection and annotation of data.

2.4.5. Multi-Task Learning

Previous studies have established the relevance of sentiment features in aiding HS detection tasks [3], [79], [89]. Recognizing the potential benefits of incorporating sentiment-related features, researchers have extended their exploration into Multi-Task Learning. The prevalent idea is that HS is a negative type of discourse, that has associated emotions like anger, rejection and criticism, so in the Multi-Task Learning framework, the model is designed to simultaneously learn and optimize multiple tasks during training, through shared representations. Specifically, in the context of HS detection, the model is tasked with emotion and sentiment classification in addition to HS detection. Shared representations are employed across these interconnected tasks, allowing the model to leverage common knowledge and patterns present in the data, aiming to enhance the overall performance of HS detection models.

Studies referenced earlier have highlighted the informative nature of sentiment features in HS detection. This recognition has spurred further investigation into Multi-Task Learning, where sentiment and emotion classification tasks are jointly addressed to bolster HS detection capabilities. Recently, four studies have employed Multi-Task Learning for HS detection task. Two studies leveraged Multi-Task Learning to concurrently address emotion and sentiment classification alongside HS detection [126], [127]. By sharing information across these related tasks, the model aimed to capture linguistic nuances associated with HS. This integrated approach demonstrated notable improvements over ML and DL models. Following this work, Min *et al.* [128] also developed a Multi-Task Learning model that tackled emotion classification in conjunction with HS detection, obtaining a better performance when compared with the baseline Single-Task Learning model. The last study that employed Multi-Task Learning diverged from the previous two, choosing to develop a model that addressed simultaneously post level and token level aggression [129].

Multi-Task Learning, specifically integrating emotion and sentiment classification with HS detection, emerges as a promising avenue for HS detection. The studies discussed underscore the effectiveness of Multi-Task Learning, leading to improved model performance. However, there is a downside to this approach, since the quality of corpora is important in a Multi-Task Learning environment, and having enough data with quality is not always possible, especially in low-resourced languages.

2.5. Q3: Data characteristics for HS detection

In this section, we look into the different languages where studies have been developed to detect HS, and also what are the different sources where researches look to gather data for the development of their models. This information will allow us to understand which languages researches have focused their work on, and which languages are less explored and may be more vulnerable to the negative effects HS. By looking at the data used, we will also be able to see if data has been collected from a vast plethora of places, or if studies have all converged to the same sources, thus making the models less likely to be able to perform well outside their scope.

2.5.1. Information Sources

The majority of studies use data collected from different social media platforms, as shown in Figure 2.7. They are a rich source of data for HS detection, given the extensive volume of user-generated content. Twitter[2], in particular, stands out as the dominant source in HS detection research, with a staggering 73 studies using Twitter data. The brevity and public nature of tweets make them highly accessible for research purposes. The Twitter platform has been a focus due to the ease of collecting and processing large datasets. While Twitter leads the way, other social media platforms also contribute to the HS detection landscape. Facebook[3], YouTube[4], Instagram[5] and Reddit[6] are also present with 10, 11, three and three studies respectively. These platforms, although less prevalent, offer insights into the multifaceted nature of HS across different online environments.

HS detection research also explores data outside social media, like news sites and alternative platforms that cater to specific communities. Sites like Fox News and others provide eight instances and niche platforms like GAB[7] and Stormfront[8], known for its association with far-right ideologies, contributes eight instances. The inclusion of such sources allows for a more comprehensive examination of HS across diverse online spaces.

It is important to note that not all data sources are created equal. Twitter, with its character limit, differs significantly from platforms like Facebook or YouTube, where users have more space to express their views. Furthermore, news websites and comments may

[2]www.twitter.com
[3]www.facebook.com
[4]www.youtube.com
[5]www.instagram.com
[6]www.reddit.com
[7]www.gab.com
[8]www.stormfront.org

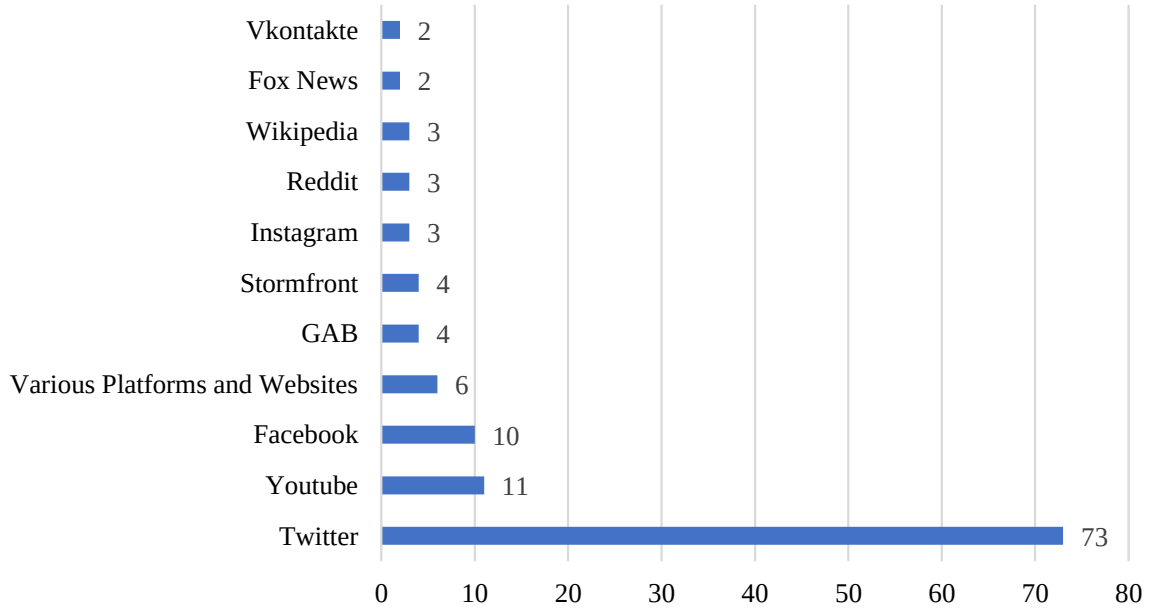

FIGURE 2.7. Data sources for HS detection.

not share the same characteristics as tweets, as they often involve more formal language and context. Researchers must consider these nuances when developing and evaluating HS detection models to ensure their applicability across various platforms.

HS detection research draws data from a wide range of sources, with Twitter being the primary contributor. The prevalence of Twitter data highlights its accessibility and suitability for large-scale studies. However, it is essential to recognize the distinctions among different sources in terms of content, context, and user behaviour. Future research in this field should continue to explore a diverse array of sources to gain a more comprehensive understanding of HS in the digital landscape.

2.5.2. Languages

HS is a pervasive problem that transcends geographic and linguistic boundaries. It is a global issue, and researchers have recognized the need to address it in various languages. However, the research landscape in the domain of HS detection has exhibited a notable focus on the English language, as evidenced by Figure 2.8. A significant portion of research efforts, resources, and datasets have been concentrated on English, with 60 studies focusing on this language. Nonetheless, other languages were explored, like Spanish, Arabic and Hindi, with 16, 11 and 8 studies respectively.

Recognizing the need to combat HS in various linguistic environments, researchers are increasingly turning their attention to low-resource languages. These languages often lack the extensive datasets and resources that are readily available for English, but has we can see, some work is beginning to be made in order to include these languages in this field.

One promising avenue for addressing HS in low-resource languages is the utilization of Transformer-based models, since they can leverage knowledge from languages with

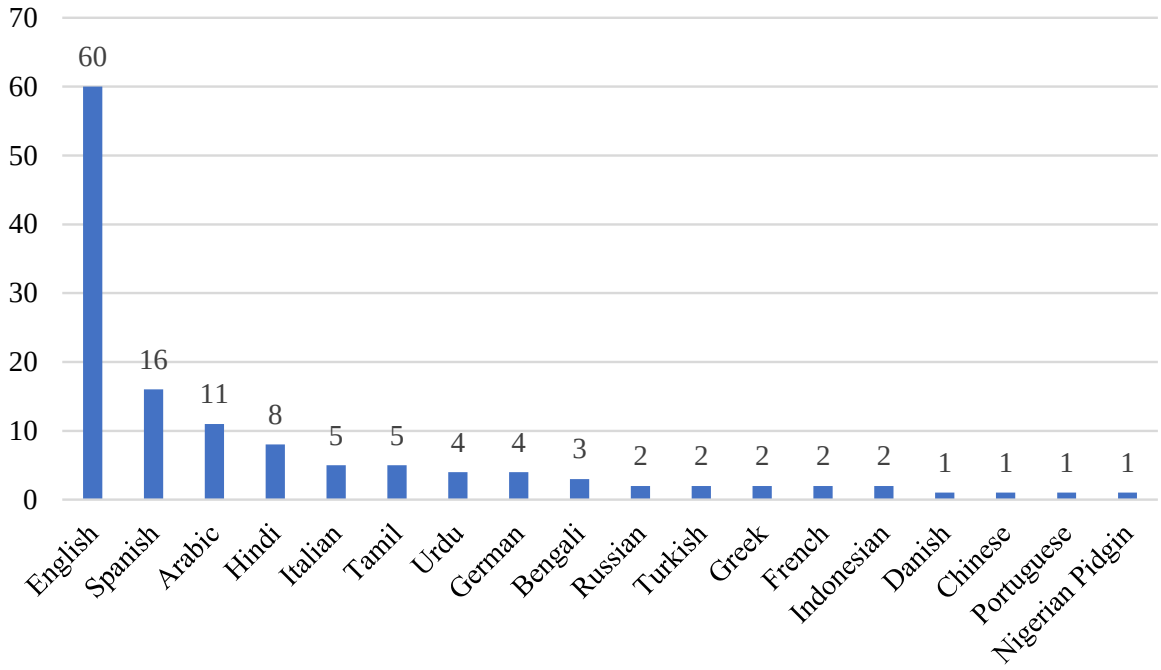

FIGURE 2.8. Languages where HS detection was conducted.

more extensive resources, like English, and fine-tune it on the limited available data for a specific language, bridging the resource gap to some extent. Transformer-based models, particularly those pre-trained on multilingual data, have shown promise in cross-lingual HS detection. These models can generalize across multiple languages, learning universal language features that enable them to detect HS irrespective of the language used. They can be effective on zero-shot cross-lingual HS detection, because, by using a high-resource language for training, like English, models can classify low-resource target languages with promising results [97], [114]. Additionally, by fine-tuning these models on a small dataset in the low-resource target language, researchers can effectively extend HS detection capabilities to languages with limited resources [115], [116].

2.5.3. Portuguese Language

For Portuguese, the literature on HS detection is relatively limited, with only a few studies focusing specifically on this language. In general, the existing work has primarily focused on Brazilian Portuguese, with few addressing European Portuguese.

For European Portuguese, initial work focused primarily on constructing a hierarchically labelled dataset for HS detection, but the authors also describe the development of an initial baseline classification for the dataset, using pre-trained word embeddings and an LSTM, they achieved a 71% micro F-score [130]. More recent studies, focused solely on the task of detecting HS, are based on BERT. [109] combine a GAN and a BERT-based model to obtain a 66.4% positive class F-score. [110] use a BERT-CNN architecture for the classification, and managed to obtain 72.1% F-score on the positive

class, by also considering the annotations that had a higher inter-annotator agreement score between them. Both of these studies used newly developed annotated datasets for European Portuguese with HS instances from YouTube and Twitter [131], [132]. Other recent works have addressed the European Portuguese variant by developing tools that can be used for the HS detection task, like foundation encoder models to expand the still very scarce ecosystem of large language models specifically developed for this language, such as the Albertina family of models [133].

Although there are several studies on HS for Brazilian Portuguese [134]–[136], this fact does not discard the need for further development in European Portuguese, since research has identified several differences between European and Brazilian Portuguese. First, variations exist in both contemporary language and technical vocabulary, as demonstrated by the differences in mood distribution. Brazilian Portuguese tends to prioritize the truth-value of a proposition, whereas European Portuguese are characterized by a more neutral tone [137]. Additionally, there are distinctions in the lexical, lexical-syntactic, and morphosyntactic usages of temporal adverbials [138]. More importantly, besides these technical differences, because HS is intrinsically dependent on both the target communities and social practice (i.e., the social and historical context), existing resources and models cannot be directly transferable or easily adapted to other linguistic and pragmatic contexts [139]–[141]. Therefore, in this case, models developed for Brazilian Portuguese are dependent on the context of the population that uses this variation of the language and are not suited for a different social and historical context like the European Portuguese one.

2.5.4. Types of Speech

As noted earlier, not all the included studies focus solely on HS, as our search criteria also encompassed offensive and abusive speech. As illustrated in Figure 2.9, the majority (83%) of studies included address either HS alone or a combination of HS with other types of speech. The remaining 17% were split between 11% of studies focused on offensive speech and 6% on abusive speech.

2.6. Summary

As we have seen, Transformers have had an impact on almost all areas of HS detection. Firstly, these models have been gaining traction in HS detection tasks, and since 2022 have been the most used models, which clearly indicates their popularity and success among researchers. These models, characterized by their ability to capture intricate linguistic patterns and contextual nuances, have consistently demonstrated superior performance compared to traditional ML techniques and other DL architectures. Studies highlighted in our review show that Transformers usually outperform other highly used models such as CNN's, LSTM's, SVM and Ensemble models. Moreover, besides Transformers having a better standalone performance, they have also been incorporated into other models to further enhance detection accuracy. Generative models have also recently started being

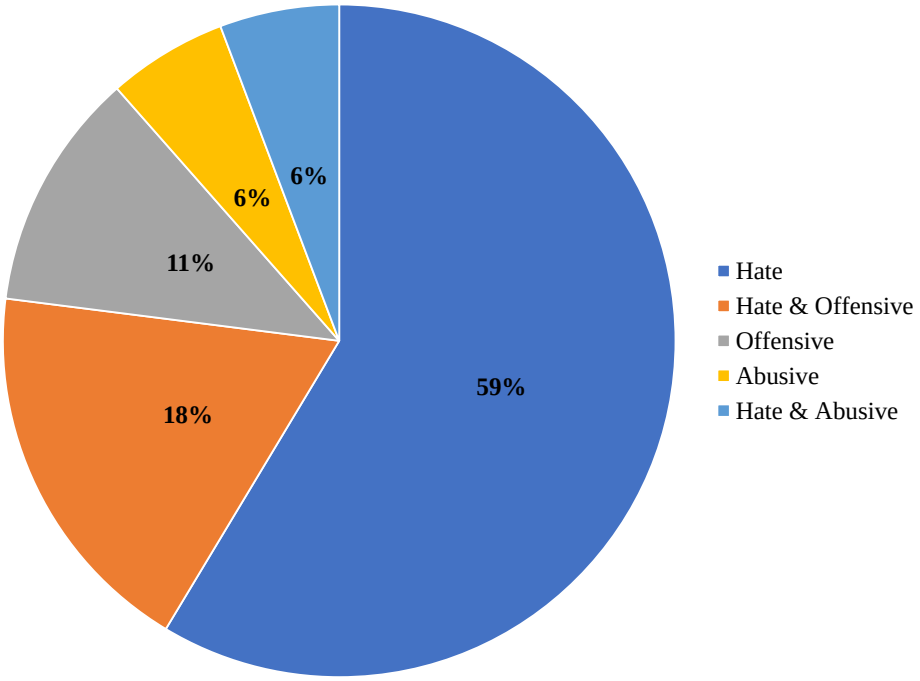

FIGURE 2.9. Percentage of works dealing with the different types of speech included in our SLR.

used in HS detection tasks with promising results, taking advantage of the recent interest and resources that these models have gained. Additionally, the review of the existing literature on HS detection in Portuguese has revealed a relevant gap in the detection of European Portuguese HS. Current research in this domain remains in its initial stages. Consequently, there is a need for further research aimed at addressing this gap, which could also provide valuable insights and methodologies applicable to other low-resourced languages. By developing tools for a language that, as it does not yet have many resources, may be more vulnerable to the risk of discrimination and abuse, we have the potential to make a significant difference in mitigating the harmful effects of HS on the Portuguese-speaking community.

CHAPTER 3

Automatic Classification of Hate Speech

This chapter presents the data, methodology and results of the automatic classification of HS using a variety of ML models. We begin by describing the data and models adopted for the classification task, with their configurations and the computational setup used for running the experiments. Next, we present the results of the experiments, covering both datasets and addressing different aspects of HS detection, including mixed-domain and context-rich experiments. The chapter then discusses the findings, providing insights into the performance of the models, followed by an error analysis to examine the misclassifications made by the models. Finally, we conclude with a discussion on the deployment of these models in a practical environment.

3.1. Data

This study uses two corpora containing annotated online HS messages, recently created in the scope of kNOwHATE project: a YouTube corpus consisting of YouTube comments, and a Twitter corpus containing tweets retrieved from Twitter. Table 3.1 presents both corpora distributions.

TABLE 3.1. Corpora distributions.

	#messages		HS proportion	
Corpus	Train	Test	Train	Test
YouTube	23912	825	64.90%	72.06%
Twitter	21546	805	11.48%	20.62%

3.1.1. YouTube

The YouTube corpus consists of 23912 comments collected from 88 distinct YouTube videos, covering topics and events targeting, directly or indirectly, four specific target groups: African descent, Roma, Migrants, and the LGBTQ+ communities. Initially, videos containing HS messages were selected. To broaden this selection, additional videos featured in the related section were also included, as illustrated in Figure 3.1.1. This decision was based on the hypothesis that frequently suggested videos when watching an already HS-flagged video were more likely to attract HS. To quantify the frequency of video suggestions and to identify those most likely to contain HS, a sorted list of suggested videos was generated. Videos that appeared more than 85 times on the list were added to the dataset. After obtaining the final group of videos that were potential candidates, videos were removed from selection if they did not have a minimum number of 1000 views

and 100 comments, resulting in the final 88 videos, distributed by target group as follows: Roma – 16, migrants – 19, LGBTQ+ – 24, and African descent – 29.

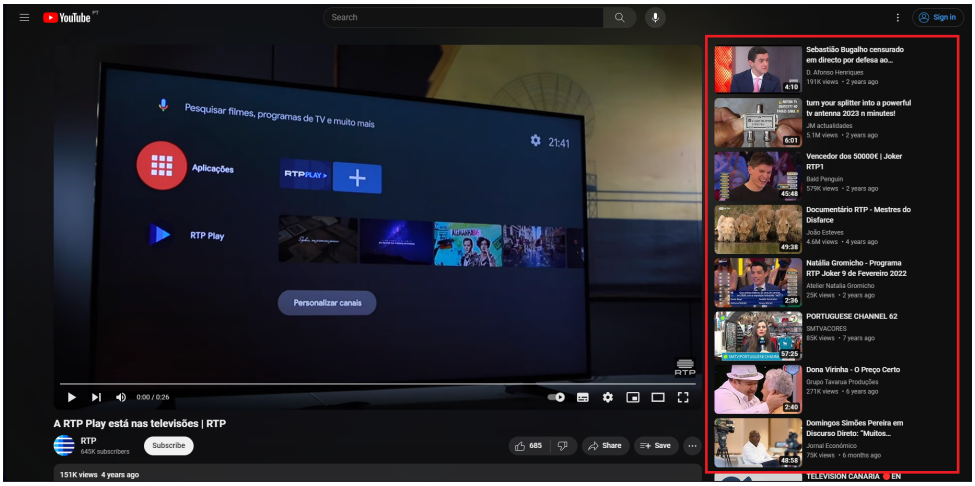

FIGURE 3.1. Related section (shown in red).

3.1.2. Twitter

The Twitter corpus consists of 21546 tweets retrieved using the Twitter API published between January 1, 2021, and December 31, 2022. For the collection of relevant tweets, a list of 259 keywords associated with the four specific target groups (African descent, Roma, migrants, and the LGBTQ+ communities) was compiled, and tweets containing those keywords were collected. From the collected tweets, only those written in Portuguese were selected, resulting in a dataset predominantly consisting of Brazilian Portuguese. Therefore, to ensure geographical relevance, the tweets were further narrowed to tweets only posted in Portugal. Additionally, the entire conversation to which the tweets belonged was also retrieved, ensuring that the parent tweet of all conversations was published in Portugal. In Table 3.2, we display some examples of messages with HS for the different target groups for both corpora.

3.1.3. Annotation

The corpora were manually annotated by interdisciplinary teams, consisting of four researchers for the YouTube corpus and three researchers for the Twitter corpus, all with backgrounds in language sciences and social psychology. Each annotator was tasked with annotating approximately 6000 comments on YouTube and, 7000 tweets on Twitter. Additionally, a subset of comments/tweets (825 for YouTube and 805 for Twitter), that served as our test sets, was assigned to all annotators to assess Inter-Annotator Agreement (IAA) and annotation reliability using Krippendorff's alpha [142]. The IAA for YouTube was moderate, at 0.546, whereas for Twitter was considerably lower, at 0.355, indicating variations in agreement levels between the annotators across the two datasets. As mentioned, this IAA subset also served as the test set for model evaluation, and given the task subjectivity, only the messages that were labelled as conveying HS by at least

TABLE 3.2. Hate Speech examples of both corpora for the different target groups.

Corpus	Target Group	Example Message
YouTube	Migrants	Isso pulhiticos merdosos, continuem a importar lixo, até Portugal deixar de ser Portugal. [*That is right shitty politicians, keep importing rubbish until Portugal stops being Portugal.*]
	Roma	Mais um bairro de ciganos onde eles é que fazem a lei.. Se o nosso belo governo lhes continuar a dar casas e dinheiro eles continuam a procriar e a encher bairros, onde depois o próprio governo não tem mão.. [*Another gypsy neighborhood where they make the law... If our beautiful government continues to give them houses and money they will continue to procreate and fill neighborhoods, where the government itself has no hand...*]
	LGBTQ+	as pessoas tem que perceber que ser "panasca"[a] não é deixar de ser homem, é deixar de ser humano kek [*People have to realize that being "panasca" doesn't mean stopping being a man, it means stopping being human kek*]
	African descent	Ao menos os branco de raça superior ainda criam biologia, já os pretos americanos nem sabem definir o que é biologicamente uma mulher.... Este mundo está perdido... alguém me sabe dizer se já aceitam voluntários para a missão a Marte? [*At least superior white people still create biology, while black Americans don't even know how to define what a woman is biologically.... This world is lost... can anyone tell me if they are already accepting volunteers for the mission to Mars?*]
Twitter	Migrants	Os zucas[b] podem ofender todos os portugueses mas se a gente riposta já somos uns filhos da puta Nós tamos é cansados de ser chacota desta escória de pessoas. Não são todos, aliás, a maioria são gente boa MAS tem muito cabrão aí [*The "zucas" may offend all Portuguese people, but if we fight back, we are already sons of bitches. We are tired of being made fun of by these scum of people. Not all of them, in fact, most of them are good people BUT there are a lot of bastards out there*]
	Roma	Ta tanto cigano no loureshopping foda se qual deles é q foi a julgamento [*There are so many gypsies in Loures Shopping, fuck which one went to trial*]
	LGBTQ+	Vai pá puta que te pariu seu paneleiro do caralho, virgem ofendida [*Go fuck yourself you fucking faggot, offended virgin.*]
	African descent	@UserID se calhar são os que cometem mais ilegalidades não??!? Esta questão do coitadinho que é preto já passou de moda. Resistiu às autoridades teve o que se encontra previsto na lei, sem pôr nem tirar! [*@UserID maybe they are the ones who commit the most illegalities, right??!? This question of the poor thing being black is out of fashion. He resisted the authorities and did what is stipulated by law, without putting in or taking away!*]

[a,b]Derogatory terms used to refer to homosexual men and people from Brazil respectively

two annotators were considered hatred content in the test sets. We did not consider the messages containing only one vote to discard unintentional errors introduced by the annotator, as the possibility that the majority of annotators made a mistake would be less likely.

The annotation scheme involved several dimensions describing: 1) different speech acts; 2) the different target groups/communities involved in the project; 3) different forms of discrimination; 4) a set of discursive strategies; 5) a selection of counter-speech strategies; 6) the type of intergroup contact between members of different social groups; 7) a set of rhetorical mechanisms that can be found in hate speech; and 8) a set of negative and positive emotions to characterize both online hate speech and counter speech, namely hate, anger, disgust, fear, guilt, shame, and hope.

3.2. Adopted Models

In this section, we present an overview of the different models used for HS detection, along with the experimental settings used to run these models and the metrics used to evaluate the performance of each model. Figure 3.2 presents the overall workflow of the experiments conducted. As we can see, we resorted to three different types of models: DL models, that served as baseline, Transformer-based models and generative models.

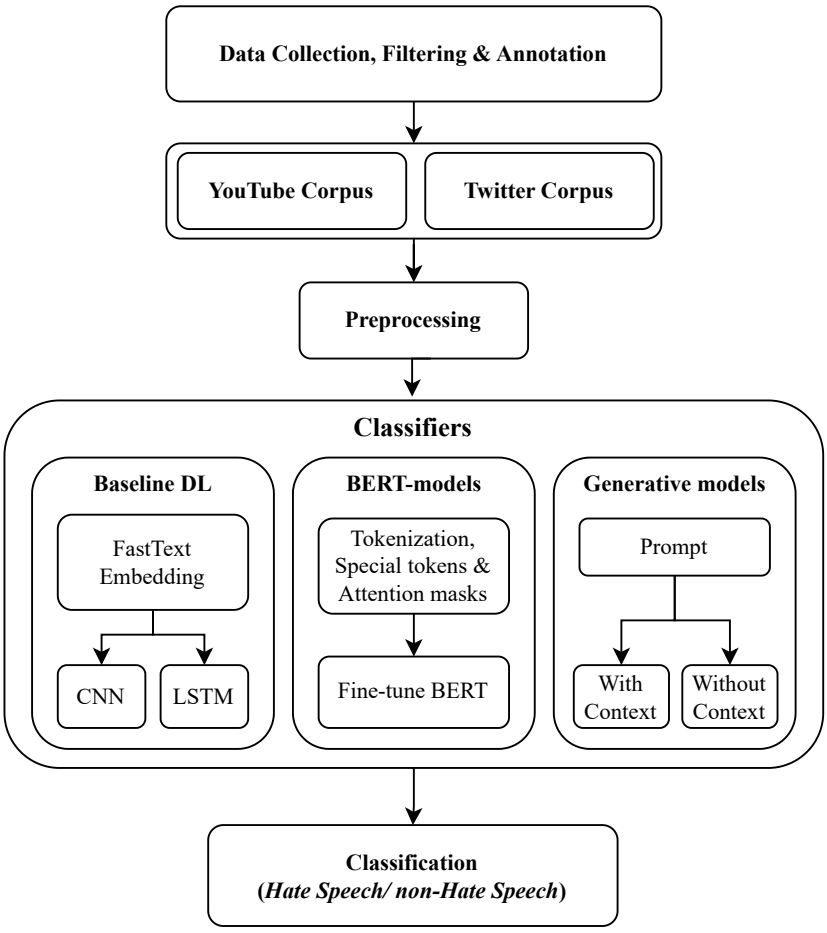

FIGURE 3.2. Workflow of our HS detection system.

3.2.1. Baseline

To serve as a baseline for comparison with the Transformer-based models, we employed a CNN model based on Safaya *et al.* [143] work with 160 convolutional filters of 5 different sizes (1, 2, 3, 4, and 5) and 32 filters for each size. We also employed a LSTM model with an initial layer comprising 128 units, followed by one dense layer with 64 units and an output layer with a softmax activation function. For both models, the embeddings used were FastText CBOW for Portuguese [144], with dimensions of 300.

3.2.2. Transformer-based models

For the Transformer-based models, we used four different models based on BERT. The BERT base model contains an encoder with 12 layers (transformer blocks), 12 self-attention heads, and 110 million parameters. Because BERT-based models are pre-trained on large general corpora, they were fine-tuned using the domain-specific one, and a linear layer was added on top of the BERT architecture for the classification. For this, the [CLS] token output of the 12^{th} transformer encoder, a vector of size 768, is given as an input to a fully connected network. Subsequently, the sigmoid activation function was applied to the hidden layer to make the predictions. During training, some of the BERT weights were also updated, allowing the model to adapt to the specific characteristics of our dataset. Four different BERT-based models were used:

- BERTimbau [145] – although developed for Brazilian Portuguese, since our work is also focused on a variant of Portuguese, we used BERTimbau, a pre-trained BERT model on the brWac corpus [146];
- BERTimbau-hatebr [147] – an already fine-tuned version of the BERTimbau model for HS with the HateBR corpus [148];
- mDeBERTa-hatebr [147] – a fine-tuned version of mDeBERTa [149], a multilingual version of DeBERTa, which is an improved version of BERT, for HS detection using the HateBR corpus;
- HateBERTimbau [150] – a retrained version of BERTimbau with 229103 tweets in European Portuguese associated with offensive conversations.

For the training hyperparameters of the BERT models we followed the original paper recommendations, with a batch size of 32, learning rate for Adam optimizer of $2^{\text{e-5}}$ and 3 epochs [8]. Other attempts were conducted with different parameters, also suggested by the original article, like a batch size of 16 and epochs between 1 and 5, but the used parameters proved to have better performance.

Although some of the models used were already fine-tuned on HS corpora, which was the case with the BERTimbau-hatebr and mDeBERTa-hatebr, we performed further fine-tuning of the models in our corpora, which led to better results. We did not use the previously mentioned Albertina models since at the time of our work only the large version was available, which is very resource intensive, and initial trials did not lead to a better performance.

3.2.3. Generative Models

In addition to the BERT-based models, we also explored three additional models for text generation: GPT, Gemini and Mistral. For GPT versions 3.5 and 4 were used, for Gemini version Gemini-Pro was used, and the Mistral version used was Mistral-7B-Instruct-v0.3. The inclusion of Mistral in this work was due to its static nature, which addresses the issue of varying performance over time, observed on other generative, that are updated over time in an opaque way [151]. Mistral ensures consistent characteristics for all users of the same version, which allows for a stable benchmark against which the dynamic nature of GPT and Gemini can be compared, enhancing the robustness of this work. All runs using the generative models were conducted on April 1st, 7th, 10th and June 27th, 2024, ensuring that the results align with the versions of the models current at the time of use.

3.3. Experimental Setup

All experiments were conducted using the computational resources of an NVIDIA RTX A6000 GPU with 48 GB of memory, housed within a dedicated machine accessed for the purposes of this work.

For all models, an initial pre-processing of the text was performed to replace all usernames with "@UserID". For the BERT models, the maximum sequence length of each text sample was set to 350 tokens to avoid overloading the GPU. Despite this limitation, a substantial number of messages did not exceed this length, with only 228 comments in the YouTube corpus and none in the Twitter corpus surpassing the threshold. This constraint did not adversely affect the model's performance. To obtain the evaluation metrics, an average of five runs was calculated, and the training data was split into 80% for the training set and 20% for the validation set.

In the Transformer-based models both corpora underwent in-domain and mixed-domain assessments. In the in-domain experiments, the model was trained exclusively on the corpus being tested. In the mixed-domain experiments, both corpora were used to train the model, to determine if additional information from a different source could further enhance performance. For example, the model was trained using data from both the YouTube and Twitter corpora and then tested on the Twitter corpus test set.

Concerning the generative models, the following prompts were used to obtain the labels for runs with and without context:

- **No context**: "Consider the comment/tweet delimited by <START> and <END> and indicate whether it is hate speech. Return only the number 1 if yes or the number 0 if not, as in a classification task <START>0<END>";
- **With context (YouTube)**: "Consider the following data about comments of a YouTube video, where <title> is the title of the video, <previous comments> corresponds to previous comments, and <comment> corresponds to the comment being analysed. Please classify the comment as being hate speech (1) or non hate-speech (0) based on the context provided. Return only the values

1 or 0, nothing more. <title>{title} <previous comments>{context} <comment>{current_comment}";

- **With context (Twitter)**: "Consider the following data about Twitter conversations, where <context> corresponds to previous tweets in the conversation, and <current_comment> corresponds to the tweet being analysed. Please classify the tweet as being hate speech (1) or non hate-speech (0) based on the context provided. Return only the value 1 or 0, nothing more. <context>{context} <comment>{current_comment}".

For the context runs, each YouTube comment was supplemented with the title of the video in which the comment appeared, as well as preceding comments in the conversation if they belonged to the same thread. For tweets, in the absence of video context, only preceding tweets in the thread were provided, if available. All prompts were composed in English, although the messages fed for classification were written in Portuguese, since in our prompt engineering experiments, detailed in Section 3.5.1, we verified that this configuration led to better results. We employed the OpenAI API to use the GPT models for our experiments – the temperature parameter was set to 0 to assure consistent results, and the Google API for the Gemini runs.

3.4. Evaluation Metrics

The performance of the models was evaluated using three standard metrics, namely, Precision, Recall, and F-score. These metrics are mathematically defined in Equations 1, 2, and 3, respectively, where True Postives (TP) refers to the total number of correctly classified HS instances, False Postives (FP) refers to the total number of non HS instances classified as HS, True Negatives (TN) refers to the total number of correctly classified non HS instances, and, finally, False Negatives (FN) refers to the total number of HS instances classified as non HS.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

$$\text{F-score} = \frac{2 * \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

We report the macro, weighted, and positive class scores, but when we assess the models we give more importance to the positive class F-score since it evaluates the class we want to detect and is a more balanced measure, taking into account both FP and FN. For the comparisons between both types of HS (CHS and OHS), we compared them based on the number of TP and FP that each model achieved, since we employed a binary classification where the models were prompted to classify text as either HS or non-HS. This approach does not allow us to determine whether the model distinguishes between

OHS or CHS. As a result, it is only possible to obtain the above-mentioned metrics for the general HS, and not for these finer types of HS.

3.5. Experiments and Results

This section is divided into four parts: the initial prompt engineering experiments conducted to assess optimal prompts for the HS detection task, the results of our HS detection experiments in both the YouTube and Twitter corpora (including the results of the BERT-based models for both in-domain and mixed-domain experiments, as well as the results of the generative models with and without contextual information), and finally, the results of the impact that context has on detecting CHS. To ensure the statistical significance of all results presented, we conducted McNemar's test.

3.5.1. Prompt Engineering

To identify the most effective prompt for our experiments, we conducted several trials using different prompt formulations and evaluated their performance based on the Positive F1 Score metric. For resource optimization, these experiments were primarily conducted on GPT-3.5-turbo. However, we also tested a smaller sample of data on GPT-4-turbo to confirm the results. The prompts tested, and their corresponding scores, are summarized in Table 3.3.

TABLE 3.3. Prompt engineering attempts.

Prompt	Description	Positive F1 Score
Task	Prompt with only the description of the task	70.6%
Definition	Prompt with the description of the task + our definition of HS	69.2%
Example	Prompt with the description of the task + an example message of HS	57.4%
Definition + Example	Prompt with the description of the task + our definition of HS + an example message of HS	59.5%

The results from our prompt engineering experiments indicate that the "Task" prompt, which includes only the description of the task, achieved the highest Positive F1 Score of 70.6%. This prompt formulation demonstrated better performance when compared to other variations that included additional components such as HS definitions or example messages of HS. It is worth noting that we also experimented with different prompt formulations and languages (Portuguese and English), to select the prompt, ultimately arriving at the best version, as presented in Section 3.2.

Based on these findings, we selected the "Task" prompt as the optimal formulation to be used for subsequent experiments, leveraging its effectiveness and simplicity in eliciting high-quality responses from our generative models.

3.5.2. YouTube

Table 3.4 summarises the results achieved for the in-domain and mixed-domain experiments of BERT-based models, as well as the results of the generative models with and without context. The results of the in-domain experiments reveal that all BERT-based models significantly outperformed the baseline DL models by more than 20 p.p. in regard to the positive class F-score (p-value < 0.01). The best model between the BERT-based models was HateBERTimbau, with an increase in F-score of 1.5 p.p. when compared to the next best model, with all BERT models having similar performance. No significant differences were observed between the BERT-based models with an in-domain setting, except HateBERTimbau and BERTimbau (p-value < 0.05).

For the mixed-domain experiments, the models were trained with the addition of the Twitter corpora for a total of 45458 messages. The mixed-domain section of Table 3.4 shows that, although BERTimbau and BERTimbau-hatebr models improved their performance by 0.5% and 0.2%, this difference was not significant, and that none of the models surpassed the overall best F-score obtained in the in-domain results, with the best model being again HateBERTimbau. Again, no significant differences were observed between BERT-based models in a mixed-domain setting, and there were no significant differences between in-domain and mixed-domain models, except for HateBERTimbau in-domain and mDeBERTa-hatebr mixed-domain (p-value < 0.05).

Finally, the generative models section of Table 3.4 presents the results of all generative models. Firstly, we can see that these models have a far worse performance than the BERT models, with a decrease of almost 10 p.p. in F-score between the best models. This was confirmed by the statistical test conducted, where all generative models were significantly worse than all BERT-based models (p-value < 0.01). Comparing the runs where context about the messages was provided versus the ones where no context was provided, we see that the best result was obtained in a context setting, with GPT-3.5 achieving a 0.796 F-score, significantly different from all other generative models (p-value < 0.01), excluding GPT-4 without context. The GPT-3.5 and GPT-4 models were the only ones that improved their performance with the addition of context by 4.6 p.p. and 1.4 p.p. respectively, with only the difference observed in GPT-3.5 being significant (p-value $<$ 0.01). Both Gemini-Pro and Mistral have better performance in a no-context setting, with only the differences observed in Mistral being significant (p-value < 0.05). Although the results obtained were consistent in multiple iterations on the same day, subsequent runs in different days with identical configurations revealed differences of approximately 25 p.p. in some models. For instance, on a previous run of the GPT-4 model, we got a positive class F1 of 0.554, which marks a difference of 17.4% to the F1 presented in Table 3.4 of 0.728. This goes in line with the literature that shows that the behaviour of the "same" model can change substantially in a relatively short amount of time, since these models are updated over time, in an opaque way [151], as mentioned before. This was also observed for the Twitter generative models.

TABLE 3.4. YouTube experiments for both BERT-based models and Generative models.

Model	Positive Class			Macro Avg			Weighted Avg		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Baseline Models									
CNN (Baseline)	0.840	0.590	0.690	0.610	0.640	0.590	0.720	0.610	0.640
LSTM (Baseline)	0.850	0.500	0.630	0.600	0.630	0.550	0.720	0.570	0.590
BERT-based models: In-domain									
BERTimbau	0.858	0.850	0.853	0.742	0.746	0.744	0.792	0.791	0.792
BERTimbau-hatebr	0.848	0.861	0.855	0.742	0.736	0.738	0.788	0.790	0.789
mDeBERTa-hatebr	0.866	0.847	0.856	0.749	0.758	0.753	0.800	0.796	0.798
HateBERTimbau	0.863	0.879	**0.871**	0.770	0.762	0.766	0.810	0.813	0.811
BERT-based models: Mixed-domain									
BERTimbau	0.867	0.848	0.858↑	0.751	0.760	0.756	0.802	0.798	0.800
BERTimbau-hatebr	0.861	0.853	0.857↑	0.749	0.752	0.750	0.797	0.796	0.796
mDeBERTa-hatebr	0.858	0.829	0.843↓	0.729	0.740	0.734	0.785	0.779	0.781
HATEBERTimbau	0.866	0.853	**0.860**↓	0.754	0.759	0.757	0.803	0.800	0.801
Generative Models: Without context									
Gemini-Pro	0.888	0.669	**0.763**	0.685	0.727	0.680	0.782	0.691	0.706
GPT-3.5	0.873	0.658	0.750	0.665	0.704	0.659	0.750	0.669	0.686
GPT-4	0.875	0.624	0.728	0.661	0.699	0.648	0.754	0.666	0.683
Mistral-7B-Instruct-v0.3	0.895	0.616	0.729	0.675	0.716	0.657	0.770	0.672	0.688
Generative Models: With context									
Gemini-Pro	0.924	0.555	0.693↓	0.681	0.719	0.639	0.779	0.634	0.650
GPT-3.5	0.775	0.817	**0.796**↑	0.611	0.601	0.604	0.729	0.660	0.676
GPT-4	0.862	0.651	0.742↑	0.658	0.693	0.651	0.746	0.675	0.691
Mistral-7B-Instruct-v0.3	0.881	0.464	0.608↓	0.632	0.653	0.567	0.740	0.571	0.585

3.5.3. Twitter

For the Twitter corpus, the results were far worse, when compared with the YouTube corpus, as shown in the in-domain section of Table 3.5. All BERT-based models had a positive class F-score bellow 50%, with the best being again HateBERTimbau with an F-score of 47.3% (more than 3.5 p.p. above all other BERT models), although without significant differences. Among the BERT models, all significantly outperformed the baseline CNN model (p-value < 0.01), but only HateBERTimbau significantly outperformed the LSTM model (p-value < 0.01), with a 3.3 p.p. increase.

In the Twitter corpus, the addition of information to the models, by incorporating the YouTube comments in the training phase, resulted in an increase in performance, as shown in the mixed-domain section of Table 3.5. There was in increase of 4 p.p., 5.2 p.p., and 5.1 p.p. in BERTimbau, BERTimbau-hatebr, and mDeBERTa-hatebr models, respectively, all being statistically significant (p-value < 0.05). The previously best performing model, HateBERTimbau, did not see an increase in performance, being significantly worse than its in-domain counterpart (p-value < 0.01). Contrary to the in-domain models, all mixed-domain models significantly outperformed both baseline models (p-value < 0.01).

Lastly, regarding the results of the generative models in the Twitter corpus, illustrated in the generative models section of Table 3.5, the inclusion of context did not prove to

be beneficial for enhancing the performance, with every model showing significant decline with context (p-value < 0.01), excluding GPT-3.5. However, it is noteworthy that the GPT-3.5 model without context achieved the highest performance out of any model, attaining a score of 50.2%. This model was significantly superior to all in-domain and mixed-domain BERT-based models, as well as all generative models without context (p-value < 0.01), being the only one to exceed the 50% threshold.

TABLE 3.5. Twitter experiments for both BERT-based models and Generative models.

Model	Positive Class			Macro Avg			Weighted Avg		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Baseline Models									
CNN	0.290	0.750	0.420	0.580	0.600	0.510	0.730	0.530	0.560
LSTM	0.300	0.800	0.440	0.590	0.620	0.520	0.750	0.530	0.560
Transformer Models: In-domain									
BERTimbau	0.511	0.371	0.429	0.679	0.639	0.652	0.778	0.796	0.784
BERTimbau-hatebr	0.494	0.395	0.438	0.672	0.645	0.655	0.777	0.792	0.782
mDeBERTa-hatebr	0.507	0.375	0.431	0.678	0.640	0.653	0.778	0796.	0.784
HATEBERTimbau	0.497	0.454	**0.473**	0.679	0.667	0.672	0.786	0.791	0.788
Transformer Models: Mixed-domain									
BERTimbau	0.422	0.528	0.469↑	0.645	0.670	0.654	0.777	0.754	0.763
BERTimbau-hatebr	0.440	0.542	**0.486**↑	0.657	0.682	0.666	0.784	0.754	0.763
mDeBERTa-hatebr	0.440	0.534	0.482↑	0.656	0.678	0.664	0.783	0.763	0.772
HATEBERTimbau	0.358	0.630	0.456↓	0.619	0.667	0.619	0.772	0.689	0.715
Generative Models: without context									
Gemini-Pro	0.388	0.614	0.476	0.635	0.681	0.643	0.774	0.737	0.751
GPT-3.5	0.400	0.675	**0.502**	0.649	0.706	0.656	0.799	0.720	0.743
GPT-4	0.389	0.705	0.501	0.646	0.708	0.649	0.797	0.711	0.735
Mistral-7b-v0.3	0.371	0.578	0.452	0.621	0.662	0.628	0.768	0.711	0.731
Generative Models: with context									
Gemini-Pro	0.468	0.398	0.430↓	0.659	0.640	0.648	0.758	0.779	0.766
GPT-3.5	0.293	0.729	0.418↓	0.589	0.636	0.546	0.806	0.597	0.633
GPT-4	0.350	0.645	**0.453**↓	0.616	0.667	0.613	0.772	0.680	0.707
Mistral-7b-v0.3	0.314	0.590	0.410↓	0.588	0.628	0.580	0.749	0.650	0.681

Regarding the time performance of the models, all BERT-based models had similar performance, which is to be expected since they are all BERT versions, sharing the same number of parameters and architecture. So BERTimbau, BERTimbau-hatebr, and HateBERTimbau had a testing time of 3.97, 3.97, and 3.95 seconds, respectively, for 825 sentences. mDeBERTa-hatebr was the slowest model, taking 5.79 seconds, probably because this model is based on DeBERTa-V3 which has 184 million parameters instead of the 110 million present on the other models. For the generative models, their testing time is dependent on the company that provides them, seeing that they control the number of requests allowed. For instance, the free version of Gemini is limited to 15 requests per minute (RPM) which accounts for a 55 minute testing time for the same 825 sentences.

GPT-3.5 and GPT-4 took 5m19s and 5m57s respectively, and Mistral took 33m17s, although Mistral was tested in a different machine, with the use of Ollama[1] for resource optimization, so it is not possible to draw direct comparisons.

3.5.4. Context role in Overt/Covert HS

The experiments conducted to assess the impact of context on the classification of OHS and CHS were done using both the GPT-3.5-turbo and GPT-4-turbo models. As we saw in the previous sections, both models showed improvement with the addition of context, increasing 4.6% and 1.4%, respectively, in general HS. Now focusing specifically on OHS and CHS, Table 3.6 presents the results for both types of HS classifications with and without context. Overall, the addition of context led to improvements in True Positives, particularly for CHS. This increase was more noticeable in GPT-3.5-turbo, which increased in 21.64% the number of True Positives for CHS, compared to 6.5% for OHS. GPT-4-turbo, although less noticeably, also increased the number of True Positives for CHS by 3.23% compared to 0% in OHS. Despite these improvements, it is noteworthy that the inclusion of context also resulted in a substantial increase in False Negatives, with an additional 84 and 11 messages misclassified by GPT-3.5-turbo and GPT-4-turbo, respectively.

TABLE 3.6. Results of OHS and CHS detection without and with addition of context to the prompt.

Metric	Type of HS	No context	Context	Gain
GPT-3.5-turbo				
True Positives	OHS	262	281	+19 (6.5%)
	CHS	212	299	**+88 (21.64%)**
False Positives	Both	57	141	+84 (10.18%)
GPT-4-turbo				
True Positives	OHS	255	255	+0 (0.0%)
	CHS	198	211	**+13 (3.23%)**
False Positives	Both	52	63	+11 (1.33%)

In analysing the performance across different target groups, it is crucial to acknowledge that CHS can vary significantly in expression based on the specific characteristics and social dynamics of each group. Therefore, understanding the detection capabilities for CHS within these groups is important, to verify the effectiveness of the models. Table 3.7 presents the Positive F1 Score obtained in the context run, as well as HS distribution and IAA for each group. Notably, both Roma and Migrants groups exhibit the best results, each achieving a Positive F1 Score exceeding 90%, in contrast to 73.7% for LGBTQ+ and 84.5% for African descent. Furthermore, our analysis reveals a correlation between performance and IAA, with the proportion of CHS messages within each group. Groups

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

with lower performance and IAA, such as LGBTQ+ and African descent, exhibit a higher percentage of CHS messages (more than 20%) compared to groups with higher performance.

TABLE 3.7. Results of HS detection by Target Group, with HS distributions and IAA.

Target Group	GPT-3.5 F1	GPT-4 F1	%HS (%CHS)	IAA
Roma	91.9%	85.6%	98.42% (48.6%)	0.691
LGBTQ+	73.7%	57.0%	86.93% (68.6%)	0.494
Migrants	90.8%	83.6%	90.48% (44.7%)	0.627
African descent	84.5%	80.7%	83.16% (65.6%)	0.464

3.6. Discussion

Firstly, the observed impact of prompt design on model performance raises questions about the role of information inclusion and rule specificity. Our findings suggest that the addition of explicit definitions or specific examples of HS within the prompt may inadvertently constrain the model's ability to generalize across diverse instances of HS. This phenomenon could result from the model becoming overly fixated on the provided definitions or examples, potentially missing out on nuanced or less typical instances of HS. Surprisingly, our results contrast with existing literature, which often indicates that the provision of examples and information enhances model performance [152]. This disagreement confirms the complexity of prompt engineering and the need for investigation into optimal prompt strategies for each case study.

Regarding the overall results of the models employed, we can conclude that BERT-based models are more effective for the HS detection task, when compared to generative models and other DL models. This finding aligns well with existing literature [153] and was, to some extent, anticipated, as BERT-based models underwent a fine-tuning process with the used datasets, whereas the generative models were not optimised for our data. Despite this, for the Twitter corpus, GPT-3.5 with the no-context prompt managed to obtain the best result out of all the models. A possible explanation for the surprising results in the Twitter corpus, where all models struggled to even break the 50% positive class F-score mark, could be the low IAA recorded, that showed the annotators had differing views on what constituted HS in this corpus. This divergence of annotations could have impacted the BERT models in the fine-tuning phase, which led to the poor performance. Although the performance of the generative models was not great on its own, they managed to match, and even outperform the BERT models with GPT-3.5. The generative models were also in more agreement between them, with an IAA of 0.542 in their predictions, greater than the 0.355 obtained by the annotators. The disparity observed between the performance obtained in the YouTube and Twitter corpora could also be explained by the differences in discourse style and linguistic characteristics inherent to each platform. Twitter, because of its character limit and fast-paced nature, often

has condensed and cryptic language, that can make interpreting and detecting HS more challenging compared to the relatively more verbose and explicit language typically found in YouTube comments. Finally, the prevalence of HS messages on each corpus can also be an explanation for the difference in results, since the Twitter corpus has a much lower prevalence of HS messages (11.48%) when compared to the YouTube one (64.9%), and although BERT-based models are generally not as susceptible to the quantity of data as traditional ML models, having more data for fine-tuning could still improve the performance of BERT models. This is even more relevant when the task requires domain-specific knowledge and when the dataset is highly diverse or complex, which is the case. Additionally, the standard deviations for the BERT models' results were around 0.002 to 0.012 for the YouTube results and 0.002 to 0.035 for the Twitter results. These higher standard deviations for the Twitter results indicate greater variability in model performance, which can be attributed to the low IAA and the challenging nature of the Twitter data, already discussed. In contrast, the lower standard deviations for the YouTube results suggest more consistent model performance in the YouTube corpus.

Upon examining the BERT-based models employed, we can see that HateBERTimbau exhibited the best overall performance for both corpora. This model was retrained with task-relevant data and further fine-tuned with our specific corpora. This model, alongside BERTimbau-hatebr, that was already fine-tuned for the downstream HS detection task and further fine-tuned on our corpora, had the best performances, outperforming both multilingual mDeBERTa and general BERTimbau. The achieved results were expected and are in line with the literature [98], [99]. These models are domain- and task-specific, making them more adept at capturing the intricacies of Portuguese HS language and context. This specialization results in improved performance compared to more general BERT models. In addition, we can see that by retraining the model on European Portuguese we have a better performance than using a model trained on Brazilian Portuguese, confirming the need for European Portuguese resources, even in the presence of Brazilian Portuguese ones. Regarding the mixed-domain tests, the results obtained do not conclusively demonstrate that incorporating information from a different context than the target domain enhances model performance. Although when training the models with both YouTube and Twitter data, we observed significant improved performance in the Twitter test set, this improvement was not observed in the YouTube test set. This discrepancy may be attributed to the Twitter data, which, as we already mentioned, may lack sufficient information for effective training due to its low IAA and unique language characteristic. Such specificities may have hindered the model's ability to learn discriminative features relevant to the YouTube domain, thereby limiting its performance. We believe that the inclusion of diverse data sources can enrich the model's understanding and generalization capabilities across domains, as shown in literature [68], however the quality of the data needs to be assured.

Finally, concerning the generative models, GPT-3.5 outperformed Gemini-Pro, GPT-4 and Mistral for the HS detection task, and although their results fell short of the results of BERT models, they still proved to be a viable alternative forHS detection, having fairly good results and outperforming both baseline DL models. In addition, these models were not fine-tuned with our datasets, so there is still room for improvement. When running the models, it was apparent that prompt design matters a lot in this type of setting, with different prompts leading to varying results, which is consistent with the literature [154]. Lastly, our experiments with adding context to the prompts of these models (as they were not fine-tuned) showed that there were improvements in GPT-3.5 and GPT-4 in the YouTube corpus. In all other runs, the addition of context decreased performance, which appears to contradict the literature [152], where context typically enhances performance. However, it has been demonstrated that while GPT-3.5 benefits from context, other models may not [153]. Our findings align with this observation. For the generative models, the standard deviations were between 0.003 and 0.025 for Gemini and between 0.001 and 0.007 for GPT models. These relatively low standard deviations for the GPT models indicate more consistent performance across runs, while the higher standard deviations for the Gemini model suggest more variability.

Still with regard to the generative models, the impact of the addition of context to the prompts on OHS and CHS classification, CHS, which relies heavily on contextual cues and subtleties, appears to benefit significantly more from the addition of context to the prompt, as is visible in Figure 3.3. This observation aligns with expectations, as context plays a pivotal role in deciphering the hidden or implicit nature of CHS messages.

However, the notable increase in False Positives with context suggests a potential drawback. The inclusion of contextual information may inadvertently lead the model to misclassify non-HS messages as HS, emphasizing the need for careful consideration when integrating contextual cues into HS detection systems. It is worth mentioning that the False Positives misclassified with the addition of context were not cases of counter speech, which can pose potential issues. Finally, our analysis reveals a correlation between the percentage of CHS and IAA within target groups and model performance. Target groups characterized by a higher proportion of CHS, such as LGBTQ+ and African descent, exhibit lower model performance and IAA compared to groups with a lower prevalence of CHS. This observation further highlights the challenges posed by CHS and the need for tailored approaches to address the specific dynamics of different target groups.

3.7. Error Analysis

To gain insights into the performance of our models, we conducted an error analysis, examining instances of FP and FN in the predictions. Notably, we can observe in Table 3.8 that many FP instances contained counter-speech instances, that commonly have words associated with HS, leading to misclassifications. For example, the comment “Shut up, wash your mouth.... white and black people do shit too” was classified as HS probably because of the inclusion of the term “black” and the negative connotation of the message,

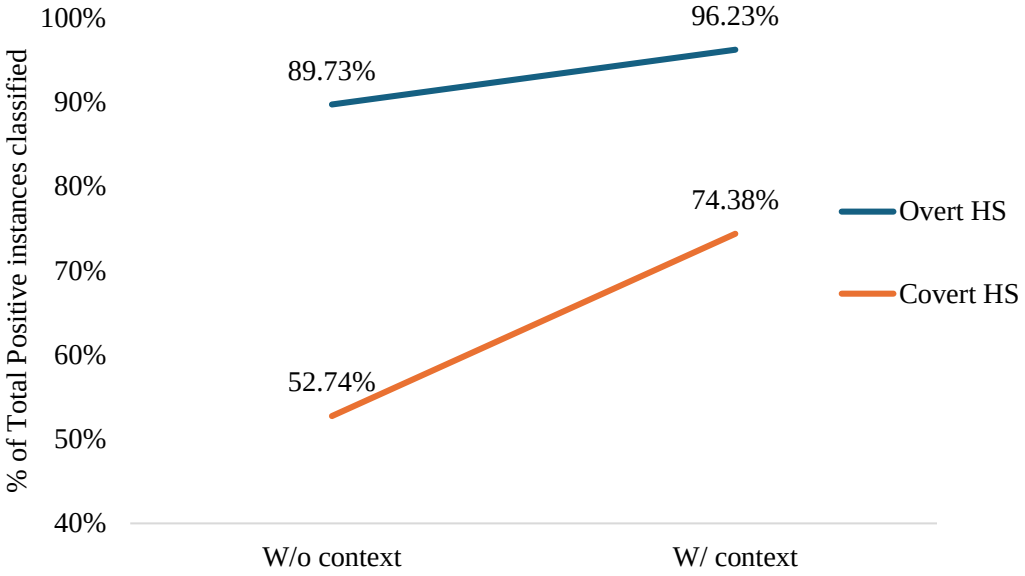

FIGURE 3.3. Impact of context in the prediction of Overt and Covert HS in GPT-3.5-turbo.

despite this being an instance of counter-speech, where the intention was to battle hate. Looking at the other examples, we see the same phenomena happening with other terms like "racism", "gypsies", "Angola", and "LGBT", that appear inserted in an aggressive message, where the intent is to combat HS. While these findings may suggest that the models rely heavily on lexical clues, it is important to note that the misclassified messages closely resemble HS messages in their structure and wording. Thus, while lexical cues play a role, the misclassifications may also stem from the nuanced similarity between these messages and actual instances of HS. Similarly, FN instances often required additional context to discern the presence of HS, particularly for covert forms. In these cases, the absence of explicit HS language made it challenging for the models to accurately identify the underlying harmful intent. For instance, the message "@UserID you are great." lacked overtly discriminatory language but probably implied support to a previously derogatory sentiment toward a specific group, illustrating the nuanced nature of CHS.

Observing Table 3.9 we have examples where the model without context misclassified a CHS instance as non-HS, and then the same model with the addition of context managed to classify the comment as HS, confirming that the addition of context is necessary in cases where the target comment is being implicitly hateful. In these examples, we can see that the comment on its own does not contain any of the characteristics of an HS comment, but when we integrate it in a thread of messages its intent becomes apparent.

TABLE 3.8. Examples of misclassifications of our models.

Type	**Message**
False Positive	Imagina a seguinte situação... Um angolano chega em Portugal e leva corona vírus... Agora pensem na reação de vcs. Sem mais... No cu dos outros é refresco ne [*Imagine the following situation... An Angolan arrives in Portugal and carries the coronavirus... Now think about your reaction. Nothing more... In other people's asses its refreshment, right?*]
	Cala te vai lavar boca.... brancos e pretos fazem merda tb [*Shut up, wash your mouth.... white and black people do shit too*]
	Os ciganos são muito bem vindes em Portugal. Eles tem cultura, educação e valorizam os povos nativos. Carrega ciganos! Saudações do Brasil. [*Gypsies are very welcome in Portugal. They have culture, education and value native people. Go gypsies! Greetings from Brazil.*]
	"Deles" quem meu?... Tás a falar de quem? Os portugueses são respeitados e bem vindos em Angola mas é claro que uma minoria residual de gente estúpida e mal formada existe em todo o lado. Isso não quer dizer com certeza que a maioria dos angolanos é contra o povo português. Longe disso. [*"Theirs" who man?... Who are you talking about? The Portuguese are respected and welcome in Angola but it is clear that a residual minority of stupid and poorly educated people exists everywhere. This does not mean for sure that the majority of Angolans are against the Portuguese people. Far from it.*]
False Negative	Força gonçalo [*Go gonçalo*]
	@UserID és grande. [*@UserID you are great.*]
	olá amiga Marta. obrigado eu pela paciência em comentar sempre as publicações e é um gosto perceber que concordo sempre contigo, e este não é exceção. cumprimentos e beijinhos [*Hello friend Marta. Thank you for your patience in always commenting on posts and it's a pleasure to know that I always agree with you, and this one is no exception. greetings and kisses*]
	Mais um comentário excelente. Parece tudo demasiado óbvio e até parece surreal alguém não entender ainda esta visão que é descrita pelo nosso amigo. Obrigado [*Another excellent comment. It all seems too obvious and it even seems surreal that someone still doesn't understand this vision described by our friend. Thanks*]

3.8. Model Deployment

In this section, we present the development and deployment of a functional prototype based on the research conducted throughout this work. The prototype serves as a practical application of the models created, offering a real-world solution for HS detection. This deployment includes both the organization and infrastructure built on Hugging Face, where the models are publicly available, as well as an interactive space that enables real-time HS classification. Through this prototype, we aim to demonstrate the practical utility and accessibility of the research outputs, allowing users to directly engage with the tools developed in this work.

TABLE 3.9. Effect of prompt context in predicted label.

Context	Prompt	Label
no	**Target:** Isso sim uma limpeza sem dor nem remorsos [*Yes a cleansing without pain or remorse*]	0
yes	**Context:** Está na hora de deixarmos de ser um povo de brandos costumes e limparmos a nossa casa. [*It's time for us to stop being a people of soft customs and clean our house.*] **Target:** Isso sim uma limpeza sem dor nem remorsos	1
no	**Target:** existem brasileiros do bem????????? [*are there good Brazilians?????????*]	0
yes	**Context:** Esses bastardos só causam desgosto e vergonha aos brasileiros de bem. [*These bastards only cause disgust and shame to good Brazilians.*] **Target:** existem brasileiros do bem?????????	1
no	**Target:** Tal e qual. [*Just like that.*]	0
yes	**Context:** Infelizmente quem imigrava respeitava o país que o recebia ficava grato por trabalho e estudo. Hoje dia eles levam o inferno com eles. [*Unfortunately, those who immigrated respected the country that received them and were grateful for work and study. Today, they take hell with them.*] **Target:** Tal e qual.	1

To ensure the accessibility of the research outputs, an organization was created on Hugging Face to centralize all resources related to this project,[1] as shown in Figure 3.4. This space serves as a repository for the models developed throughout the research, enabling other researchers, developers, and practitioners to easily find and access the tools needed for their own projects. The choice of Hugging Face as a platform was strategic due to its wide adoption within the NLP community, fostering an environment where the results of this work can be further validated, improved, and utilized by a larger audience. Additionally, the platform offers model cards and detailed documentation, helping to streamline the process for those unfamiliar with the specifics of the models developed in this work.

The primary outcome of this research is the set of models specifically fine-tuned to detect HS in Portuguese online content. Each model is publicly accessible on Hugging Face, complete with training details, evaluation metrics, and use-case documentation. Figure 3.5 shows the publicly available models, that researchers can easily integrate into their own systems through API calls or download them for offline use. This openness ensures their continued relevance and utility in various domains, from social media monitoring to academic research.

To demonstrate the real-world applicability of the research, a dedicated interactive prototype was developed for HS detection. This prototype is hosted on a Hugging Face space, also part of the organization created, allowing users to input their own text for analysis and receive real-time feedback on whether the content is classified as HS or non-HS.

[1] `https://huggingface.co/knowhate`

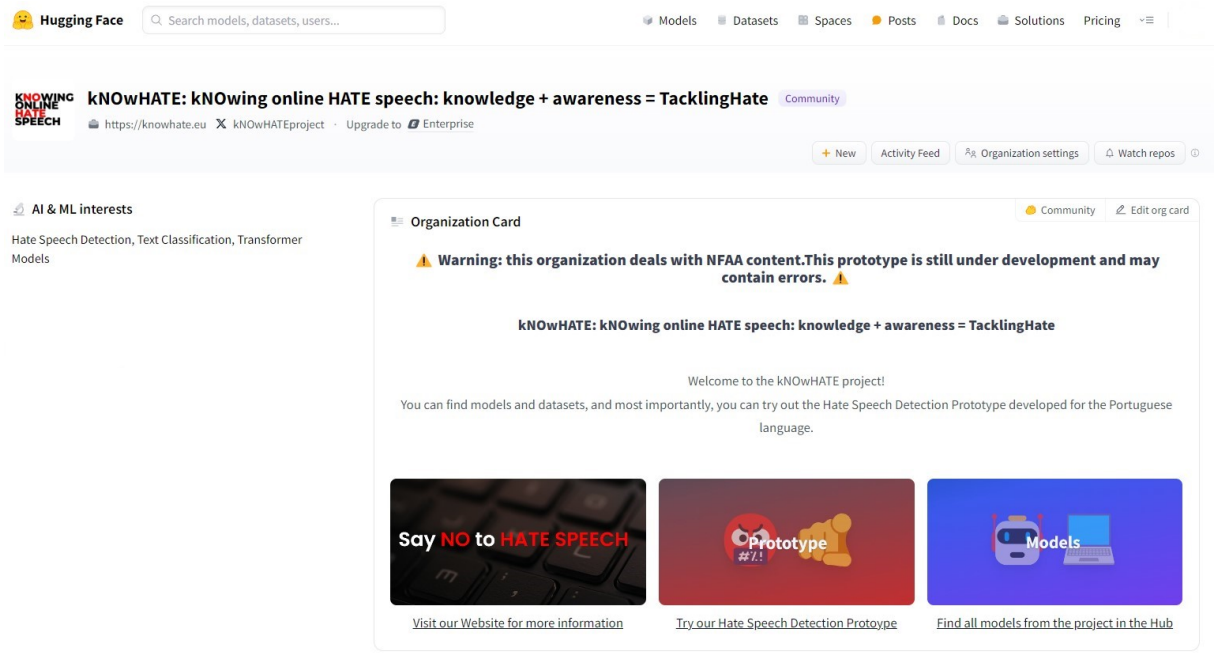

FIGURE 3.4. Hugging Face organization where the outputs of this work are housed.

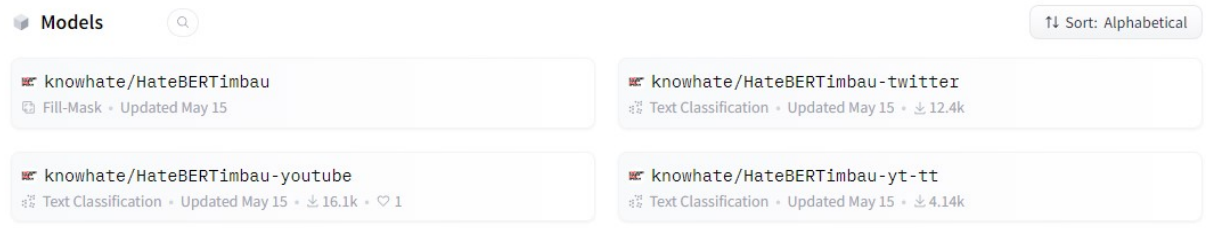

FIGURE 3.5. Models available for public use.

This space is particularly valuable for showcasing the capabilities of the models in a user-friendly and accessible manner, making the complex task of HS detection approachable for both experts and the public. Users can explore the models' performance first-hand, and test various input scenarios. The prototype also includes an intuitive interface, visualizing the results to help users better understand the predictions, as seen in Figure 3.6. This hands-on tool also raises awareness about the significance of HS detection in online environments, emphasizing the broader societal implications of the research.

As of the time of writing, the models developed have gained significant traction, ranking 4$^{\text{th}}$, 5$^{\text{th}}$, and 11$^{\text{th}}$ worldwide in the "hate" category based on the number of downloads, with over 30000 downloads combined. This recognition highlights the impact and practical significance of both the models and the developed prototype in advancing the field of HS detection, while also serving as a valuable tool for raising awareness about this topic.

3.9. Summary

In this chapter, we presented the results of applying various ML models for HS detection on two corpora: YouTube comments and Twitter data. The experiments included

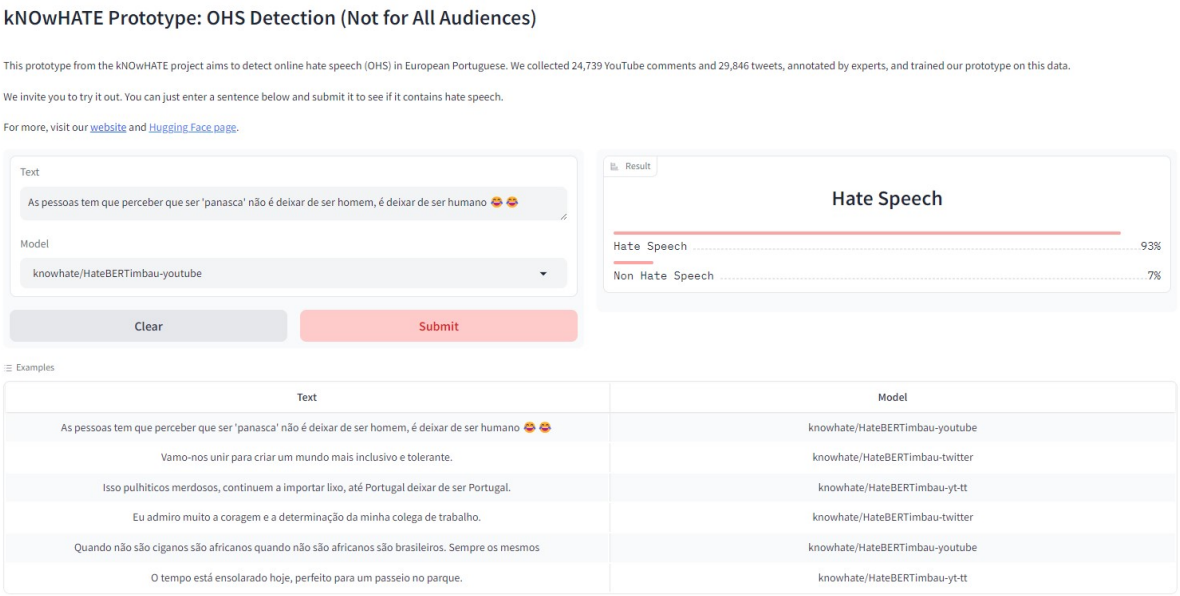

FIGURE 3.6. Prototype where users can perform HS detection on text using the developed models.

Transformer-based models such as BERTimbau, and HateBERTimbau, as well as generative models like GPT-3.5, Gemini, and Mistral. A summary of the results obtained is presented in Table 3.10, which shows the best-performing baseline, Transformer, and generative models for each dataset. The results demonstrated that Transformer-based models, particularly HateBERTimbau, consistently outperformed other models, achieving the highest F-scores for YouTube in-domain experiments. However, the addition of mixed-domain data did not yield significant performance improvements. Generative models demonstrated worse overall results, with noticeable variability in their outcomes over time. In contrast to the YouTube dataset, the Twitter corpus presented greater challenges, with all models struggling to exceed a 50% F-score. Despite this, the GPT-3.5 model without context performed the best, highlighting the difficulties posed by the shorter, cryptic language typical of Twitter, as well as lower IAA in this dataset. Lastly, the chapter explored the role of context in distinguishing between CHS and OHS. Results indicated that context significantly improved CHS detection, particularly for the GPT-3.5-turbo model, although this came at the cost of higher false negative rates. The chapter concluded with a discussion of the practical deployment of these models, including their integration into an accessible space for public use and contribution to the kNOwHATE project.

TABLE 3.10. Summary of results with best model by type for each dataset.

Type	**Model**	**Setup**	**Positive F-score**
YouTube			
Baseline	CNN	FastText	0.690
Transformer	HateBERTimbau	In-domain	0.871
Generative	GPT-3.5	With context	0.796
Twitter			
Baseline	LSTM	FastText	0.440
Transformer	BERTimbau-hatebr	Mixed-domain	0.486
Generative	GPT-3.5	No context	0.502

CHAPTER 4

Conclusion

In this work, we investigated the performance of various models in identifying HS in European Portuguese online discourse in a YouTube corpus and a Twitter corpus. Specifically, we compared different BERT-based models – BERTimbau, BERTimbau-hatebr, mDeBERTa-hatebr, and HateBERTimbau – along with four generative models – GPT-4, GPT-3.5, Gemini-Pro and Mistral-7B-Instruct-v0.3.

HateBERTimbau achieved the best positive class F-score with 87.1% for the YouTube corpus, surpassing the baseline scores by more than 20 p.p., and GPT-3.5 achieved the best performance for the Twitter corpus with a positive class F-score of 50.2%, with an increase of 6.2 p.p. compared to the baseline. We showed that the incorporation of mixed-domain data for the training of the models has the potential to improve performance, significantly increasing the performance of BERT models in the Twitter corpus, by training them with the Twitter and YouTube corpus simultaneously. In order to achieve this, it is necessary to ensure the quality of the data, since none of the models had an improvement in performance when the Twitter data was incorporated, which may be caused by the low IAA between annotators, potentially adding noise to the models.

For the generative models, they had a worse performance when compared with the BERT models in the YouTube corpus, but since there was no fine-tuning done, and the models did not learn from the annotations of the training data – they made predictions based on their representations of HS, that may not be aligned with our definition. This can also be the reason why in the Twitter corpus they managed to outperform the BERT models, because they were not exposed to the possible noisy data with low IAA. The addition of context had a discernible impact on the classification of OHS and CHS. Notably, CHS, which relies heavily on contextual cues, exhibited significant performance improvements with the inclusion of context. However, this enhancement was accompanied by an increase in False Positives. Furthermore, our analysis of target groups revealed distinct patterns of HS prevalence and model performance. Groups characterized by a higher proportion of CHS demonstrated lower model performance, highlighting the challenges associated with detecting subtle forms of HS within vulnerable communities

Overall, this work contributes to understanding the effectiveness of different models for HS detection, in general, and in European Portuguese online discourse, specifically. Our findings suggest that BERT-based models fine-tuned for the HS detection task have better performance than general BERT models not fine-tuned for a downstream task, and that models retrained on European Portuguese are more effective in identifying HS in European Portuguese than models trained on only Brazilian Portuguese.

Regarding the error analysis, we found that some of the messages mislabelled as non-HS did not have sufficient context to be able to be classified as HS. This underscores the necessity for additional context provided by preceding messages. Additionally, some of the messages mislabelled as HS were in fact counter-speech attempts or messages containing words that are often used in HS messages, which further confirms the need to provide some context to the models in order to accurately predict HS. To overcome this limitation, future work could focus on incorporating context alongside target messages to better inform the models, especially the BERT-based ones; distinguishing between OHS and CHS may also lead to better representations of the different types of HS and improve classification accuracy; and, finally, pre-fine-tuning generative models with training data to align with annotation criteria.

In conclusion, this research has thoroughly addressed the initial research questions by exploring the effectiveness of various models in identifying HS in European Portuguese online discourse. Regarding RQ1, this work has shown that the field of HS detection has undergone significant advancements with the integration of Transformer-based models, which have gradually superseded traditional ML and DL methods. RQ2 was explored through the comparative performance analysis of Transformer-based and generative models, specifically designed or adapted for Portuguese, against traditional DL models. Here, the Transformer-based models demonstrated superior performance over conventional methods for detecting HS in Portuguese, with notable results on the YouTube dataset. Lastly, RQ3 was addressed by examining the effects of mixed-domain learning and contextual information on model performance. The findings indicate that mixed-domain data can enhance model performance, while the inclusion of contextual cues in generative models meaningfully improves their detection capabilities for CHS.

While this work has provided valuable insights concerning the effectiveness of different transfer learning models for HS detection, it is important to acknowledge some limitations. Specifically, our corpora were annotated by a small number of annotators, ranging from three to four individuals, each with distinct backgrounds. This variability among annotators may introduce considerable data variance, and should be taken into account for future studies.

We believe that for future work, multi-class detection attempts should be made, especially in detecting HS directed at different target groups, such as those present in our datasets. Furthermore, recent studies have explored network immunization after detection, in various ways. Either by proactive approaches [155], tree-based approaches [156], community-based approaches [157], or real-time approaches [158], they aim to stop the propagation of problematic content in networks. We consider this a very promising avenue for application in the HS detection space. Future work should combine both tasks: HS detection and network immunization, to not only identify forms of HS but also to effectively mitigate their spread within online communities. This integrated approach could

enhance the overall effectiveness of HS management and contribute to creating safer online environments.

Reflecting on the overall conclusions of this dissertation, the advancements demonstrated in the performance of these models suggests that they are not only effective but also viable for real-world applications. Given their superior ability to identify HS, these technologies can be integrated into existing moderation systems on social media platforms. The findings indicate that with further refinement and implementation, such systems could enhance the safety of online communities by enabling timely detection and intervention. While there are challenges to address, like the mitigation of false positives and biases, the overall readiness of these technologies suggests a positive trajectory toward practical applications in combating online HS. As we move forward, it is crucial to explore the integration of these models into real-time monitoring tools, ensuring that these solutions are developed in a manner that respects free speech and does not cross the line into censorship, but rather supports open and safe dialogue.

References

[1] kNOwHATE, *Knowhate*, 2023. [Online]. Available: https://knowhate.eu/pt-pt/ (visited on 05/01/2024).

[2] Statista, *Number of social media users worldwide from 2017 to 2027*, Aug. 2023. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13 825–13 835, 2018, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2806394. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8292838 (visited on 10/11/2023).

[4] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," en, *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477–523, Jun. 2021, ISSN: 1574-0218. DOI: 10.1007/s10579-020-09502-8. [Online]. Available: https://doi.org/10.1007/s10579-020-09502-8 (visited on 10/12/2023).

[5] United Nations, "United Nations Strategy and Plan of Action on Hate Speech," United Nations, Tech. Rep., May 2019. [Online]. Available: https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/%7BAction%5C_plan%5C_on%5C_hate%5C_speech%5C_EN.pdf%7D.

[6] B. M. Tynes, M. T. Giang, D. R. Williams, and G. N. Thompson, "Online racial discrimination and psychological adjustment among adolescents," eng, *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, vol. 43, no. 6, pp. 565–569, Dec. 2008, ISSN: 1879-1972. DOI: 10.1016/j.jadohealth.2008.08.021.

[7] Media Smarts, *Impact of Online Hate*, 2021. [Online]. Available: https://mediasmarts.ca/online-hate/impact-online-hate (visited on 12/10/2023).

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[9] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, and Philip S. Yu, "A Survey on Text Classification: From Traditional to Deep Learning | ACM Transactions on Intelligent Systems and Technology," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022. DOI: https://doi.org/10.1145/3495162. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3495162 (visited on 10/12/2023).

[10] Meta, *Hate speech*, 2023. [Online]. Available: https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ (visited on 01/19/2024).

[11] Twitter, *Hateful conduct*, 2023. [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (visited on 01/19/2024).

[12] Google, *Hate speech policy*, 2019. [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en (visited on 01/19/2024).

[13] B. Parekh, "Is there a case for banning hate speech?" In *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge University Press, 2012, pp. 37–56. DOI: 10.1017/CBO9781139042871.006.

[14] A. A. Siegel, "Online hate speech," in *Social Media and Democracy* (SSRC Anxieties of Democracy), N. Persily and J. A. Tucker, Eds., SSRC Anxieties of Democracy. Cambridge University Press, 2020, pp. 56–88.

[15] P. Carvalho and R. Guerra, "D3.2/D3.3 Annotation Guidelines OHS & OCS," Iscte-Instituto Universitário de Lisboa, Tech. Rep., May 2023.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[17] Gemini Team, *Gemini: A family of highly capable multimodal models*, 2023. arXiv: 2312.11805 [cs.CL].

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[19] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, and C. Silva, "A comprehensive review on automatic hate speech detection in the age of the transformer," *Social Network Analysis and*

Mining, vol. 14, no. 1, p. 204, Oct. 2024, ISSN: 1869-5469. DOI: 10.1007/s13278-024-01361-3. [Online]. Available: https://doi.org/10.1007/s13278-024-01361-3.

[20] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, and C. Silva, "Leveraging transfer learning for hate speech detection in portuguese social media posts," *IEEE Access*, vol. 12, pp. 101 374–101 389, 2024. DOI: 10.1109/ACCESS.2024.3430848.

[21] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, R. Guerra, P. Carvalho, C. Marques, and C. Silva, "Bypassing the nuances of portuguese covert hate speech through contextual analysis," in *Progress in Artificial Intelligence (Volume 14969: Lecture Notes in Computer Science)*, forthcoming.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Long Beach, CA: Curran Associates, Inc., 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[23] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and J. E. McKenzie, "Prisma 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021. DOI: 10.1136/bmj.n160. eprint: https://www.bmj.com/content/372/bmj.n160.full.pdf. [Online]. Available: https://www.bmj.com/content/372/bmj.n160.

[24] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," en, *Information*, vol. 13, no. 6, p. 273, Jun. 2022, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2078-2489. DOI: 10.3390/info13060273. [Online]. Available: https://www.mdpi.com/2078-2489/13/6/273 (visited on 10/11/2023).

[25] B. Alrashidi, A. Jamal, and A. Alkhathlan, "Abusive Content Detection in Arabic Tweets Using Multi-Task Learning and Transformer-Based Models," *Applied Sciences (Switzerland)*, vol. 13, no. 10, 2023. DOI: 10.3390/app13105825.

[26] N. Mullah and W. Zainon, "Improving detection accuracy of politically motivated cyber-hate using heterogeneous stacked ensemble (HSE) approach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12 179–12 190, 2023. DOI: 10.1007/s12652-022-03763-7.

[27] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: A review on obstacles and solutions," *Queen Mary University of London*, vol. 7, 2021. DOI: https://doi.org/10.7717/peerj-cs.598.

[28] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," en, *BMJ*, vol. 372, n71, Mar. 2021, Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, ISSN: 1756-1833. DOI: 10.1136/bmj.n71. [Online]. Available: https://www.bmj.com/content/372/bmj.n71 (visited on 10/11/2023).

[29] F. Ayo, O. Folorunso, F. Ibharalu, I. Osinuga, and A. Abayomi-Alli, "A probabilistic clustering model for hate speech classification in twitter," *Expert Systems with Applications*, vol. 173, 2021. DOI: 10.1016/j.eswa.2021.114762.

[30] E. Lee, F. Rustam, P. Washington, F. Barakaz, W. Aljedaani, and I. Ashraf, "Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model," *IEEE Access*, vol. 10, pp. 9717–9728, 2022. DOI: 10.1109/ACCESS.2022.3144266.

[31] S. Roy, A. Roy, P. Samui, M. Gandomi, and A. Gandomi, "Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2023. DOI: 10.1109/TCSS.2023.3260217.

[32] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. [Online]. Available: https://aclanthology.org/S19-2007.

[33] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin, "SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020)," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1425–1447. DOI: 10.18653/v1/2020.semeval-1.188. [Online]. Available: https://aclanthology.org/2020.semeval-1.188.

[34] M. Wiegand, M. Siegel, and J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language," in *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, Viena (online): International Committee for Computational Linguistics, Sep. 2018, pp. 1–

10. [Online]. Available: https://www.lsv.uni-saarland.de/wp%20content/publications/2018/germeval2018%5C_wiegand.pdf.

[35] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma, "FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 70–74. DOI: 10.18653/v1/S19-2009. [Online]. Available: https://aclanthology.org/S19-2009.

[36] Y. Ding, X. Zhou, and X. Zhang, "YNU_DYX at SemEval-2019 task 5: A stacked BiGRU model based on capsule network in detection of hate," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 535–539. DOI: 10.18653/v1/S19-2096. [Online]. Available: https://aclanthology.org/S19-2096.

[37] A. Montejo-Ráez, S. M. Jiménez-Zafra, M. A. García-Cumbreras, and M. C. Díaz-Galiano, "SINAI-DL at SemEval-2019 task 5: Recurrent networks and data augmentation by paraphrasing," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 480–483. DOI: 10.18653/v1/S19-2085. [Online]. Available: https://aclanthology.org/S19-2085.

[38] J. M. Pérez and F. M. Luque, "Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 64–69. DOI: 10.18653/v1/S19-2008. [Online]. Available: https://aclanthology.org/S19-2008.

[39] G. Wiedemann, S. M. Yimam, and C. Biemann, "UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1638–1644. DOI: 10.18653/v1/2020.semeval-1.213. [Online]. Available: https://aclanthology.org/2020.semeval-1.213.

[40] S. Wang, J. Liu, X. Ouyang, and Y. Sun, "Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds.,

Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1448–1455. DOI: 10.18653/v1/2020.semeval-1.189. [Online]. Available: https://aclanthology.org/2020.semeval-1.189.

[41] T. Dadu and K. Pant, "Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2183–2189. DOI: 10.18653/v1/2020.semeval-1.290. [Online]. Available: https://aclanthology.org/2020.semeval-1.290.

[42] S. Sotudeh, T. Xiang, H.-R. Yao, S. MacAvaney, E. Yang, N. Goharian, and O. Frieder, "GUIR at SemEval-2020 task 12: Domain-tuned contextualized models for offensive language detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1555–1561. DOI: 10.18653/v1/2020.semeval-1.203. [Online]. Available: https://aclanthology.org/2020.semeval-1.203.

[43] N. Pitropakis, K. Kokot, D. Gkatzia, R. Ludwiniak, A. Mylonas, and M. Kandias, "Monitoring Users' Behavior: Anti-Immigration Speech Detection on Twitter," *Machine Learning and Knowledge Extraction*, vol. 2, no. 3, pp. 192–215, 2020, ISSN: 2504-4990. DOI: 10.3390/make2030011.

[44] F. Shannaq, B. Hammo, H. Faris, and P. Castillo-Valdivieso, "Offensive Language Detection in Arabic Social Networks Using Evolutionary-Based Classifiers Learned From Fine-Tuned Embeddings," *IEEE Access*, vol. 10, pp. 75 018–75 039, 2022. DOI: 10.1109/ACCESS.2022.3190960.

[45] S. Mohapatra, S. Prasad, D. Bebarta, T. Das, K. Srinivasan, and Y.-C. Hu, "Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques," *Applied Sciences (Switzerland)*, vol. 11, no. 18, 2021. DOI: 10.3390/app11188575.

[46] C. Arcila-Calderón, J. Amores, P. Sánchez-Holgado, and D. Blanco-Herrero, "Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish," *Multimodal Technologies and Interaction*, vol. 5, no. 10, 2021. DOI: 10.3390/mti5100063.

[47] N. Vanetik and E. Mimoun, "Detection of Racist Language in French Tweets," *Information (Switzerland)*, vol. 13, no. 7, 2022. DOI: 10.3390/info13070318.

[48] R. Saeed, H. Afzal, S. Rauf, and N. Iltaf, "Detection of Offensive Language and ITS Severity for Low Resource Language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, 2023. DOI: 10.1145/3580476.

[49] T. Turki and S. Roy, "Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer," *Applied Sciences (Switzerland)*, vol. 12, no. 13, 2022. DOI: 10.3390/app12136611.

[50] J. García-Díaz, S. Jiménez-Zafra, M. García-Cumbreras, and R. Valencia-García, "Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers," *Complex and Intelligent Systems*, vol. 9, no. 3, pp. 2893–2914, 2023. DOI: 10.1007/s40747-022-00693-x.

[51] R. Raut and F. Spezzano, "Enhancing hate speech detection with user characteristics," *International Journal of Data Science and Analytics*, 2023. DOI: 10.1007/s41060-023-00437-1.

[52] M. Khairy, T. Mahmoud, A. Omar, and T. Abd El-Hafeez, "Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection," *Language Resources and Evaluation*, 2023. DOI: 10.1007/s10579-023-09683-y.

[53] M. Aljero and N. Dimililer, "A novel stacked ensemble for hate speech recognition," *Applied Sciences (Switzerland)*, vol. 11, no. 24, 2021. DOI: 10.3390/app112411684.

[54] R. Rajalakshmi, S. Selvaraj, R. Faerie Mattins, P. Vasudevan, and M. Anand Kumar, "HOTTEST: Hate and Offensive content identification in Tamil using Transformers and Enhanced STemming," *Computer Speech and Language*, vol. 78, 2023. DOI: 10.1016/j.csl.2022.101464.

[55] F.-M. Plaza-Del-Arco, M. Molina-González, L. Ureña-López, and M. Martín-Valdivia, "Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies," *ACM Transactions on Internet Technology*, vol. 20, no. 2, 2020. DOI: 10.1145/3369869.

[56] S. Agarwal and C. Chowdary, "Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19," *Expert Systems with Applications*, vol. 185, 2021. DOI: 10.1016/j.eswa.2021.115632.

[57] O. Oriola and E. Kotze, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21 496–21 509, 2020. DOI: 10.1109/ACCESS.2020.2968173.

[58] H. Karayiğit, Ç. İ. Aci, and A. Akdağlı, "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods," *Expert Systems with Applications*, vol. 174, 2021. DOI: 10.1016/j.eswa.2021.114802.

[59] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimedia Systems*, vol. 28, no. 6, pp. 1925–1940, Dec. 2022, ISSN: 1432-1882. DOI: 10.1007/s00530-021-00784-8. [Online]. Available: https://doi.org/10.1007/s00530-021-00784-8.

[60] P. Roy, A. Tripathy, T. Das, and X.-Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204 951–204 962, 2020. DOI: 10.1109/ACCESS.2020.3037073.

[61] A. T. Kabakus, "Towards the Importance of the Type of Deep Neural Network and Employment of Pre-trained Word Vectors for Toxicity Detection: An Experimental Study," *Journal of Web Engineering*, vol. 20, no. 8, pp. 2243–2268, Nov. 2021, ISSN: 1544-5976. DOI: 10.13052/jwe1540-9589.2082.

[62] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019. DOI: 10.3233/SW-180338.

[63] R. Duwairi, A. Hayajneh, and M. Quwaider, "A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 4001–4014, 2021. DOI: 10.1007/s13369-021-05383-3.

[64] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," *Applied Sciences (Switzerland)*, vol. 10, no. 23, pp. 1–16, 2020. DOI: 10.3390/app10238614.

[65] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS ONE*, vol. 15, no. 8 August, 2020. DOI: 10.1371/journal.pone.0237861.

[66] I. Priyadarshini, S. Sahu, and R. Kumar, "A transfer learning approach for detecting offensive and hate speech on social media platforms," *Multimedia Tools and Applications*, vol. 82, no. 18, pp. 27 473–27 499, 2023. DOI: 10.1007/s11042-023-14481-3.

[67] F. Ayo, O. Folorunso, F. Ibharalu, and I. Osinuga, "Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks," *International Journal of Intelligent Computing and Cybernetics*, vol. 13, no. 4, pp. 485–525, 2020. DOI: 10.1108/IJICC-06-2020-0061.

[68] Ş. Dascălu and F. Hristea, "Towards a Benchmarking System for Comparing Automatic Hate Speech Detection with an Intelligent Baseline Proposal," *Mathematics*, vol. 10, no. 6, 2022. DOI: 10.3390/math10060945.

[69] E. Pronoza, P. Panicheva, O. Koltsova, and P. Rosso, "Detecting ethnicity-targeted hate speech in Russian social media texts," *Information Processing and Management*, vol. 58, no. 6, 2021. DOI: 10.1016/j.ipm.2021.102674.

[70] J. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors (Switzerland)*, vol. 19, no. 21, 2019. DOI: 10.3390/s19214654.

[71] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual

dataset and benchmark experiments," *Expert Systems with Applications*, vol. 215, 2023. DOI: 10.1016/j.eswa.2022.119342.

[72] S. Mundra and N. Mittal, "FA-Net: Fused attention-based network for Hindi English code-mixed offensive text classification," *Social Network Analysis and Mining*, vol. 12, no. 1, 2022. DOI: 10.1007/s13278-022-00929-1.

[73] M. Fazil, S. Khan, B. Albahlal, R. Alotaibi, T. Siddiqui, and M. Shah, "Attentional Multi-Channel Convolution With Bidirectional LSTM Cell Toward Hate Speech Prediction," *IEEE Access*, vol. 11, pp. 16 801–16 811, 2023. DOI: 10.1109/ACCESS.2023.3246388.

[74] A. Kamal, T. Anwar, V. Sejwal, and M. Fazil, "BiCapsHate: Attention to the Linguistic Context of Hate via Bidirectional Capsules and Hatebase," *IEEE Transactions on Computational Social Systems*, pp. 1–12, 2023. DOI: 10.1109/TCSS.2023.3236527.

[75] B. Aarthi and B. Chelliah, "HATDO: Hybrid Archimedes Tasmanian devil optimization CNN for classifying offensive comments and non-offensive comments," *Neural Computing and Applications*, vol. 35, no. 25, pp. 18 395–18 415, 2023. DOI: 10.1007/s00521-023-08657-z.

[76] G. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Applied Intelligence*, vol. 48, no. 12, pp. 4730–4742, 2018. DOI: 10.1007/s10489-018-1242-y.

[77] R. Cruz, W. de Sousa, and G. Cavalcanti, "Selecting and combining complementary feature representations and classifiers for hate speech detection," *Online Social Networks and Media*, vol. 28, 2022. DOI: 10.1016/j.osnem.2021.100194.

[78] A. Mazari, N. Boudoukhani, and A. Djeffal, "BERT-based ensemble learning for multi-aspect hate speech detection," *Cluster Computing*, 2023. DOI: 10.1007/s10586-022-03956-x.

[79] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," *Computing*, vol. 102, no. 2, pp. 501–522, 2020. DOI: 10.1007/s00607-019-00745-0.

[80] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," *IEEE Access*, vol. 8, pp. 128 923–128 929, 2020. DOI: 10.1109/ACCESS.2020.3009244.

[81] M. Mridha, M. Wadud, M. Hamid, M. Monowar, M. Abdullah-Al-Wadud, and A. Alamri, "L-Boost: Identifying Offensive Texts from Social Media Post in Bengali," *IEEE Access*, vol. 9, pp. 164 681–164 699, 2021. DOI: 10.1109/ACCESS.2021.3134154.

[82] E. Mahajan, H. Mahajan, and S. Kumar, "EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media," *Expert Systems with Applications*, vol. 236, 2024. DOI: 10.1016/j.eswa.2023.121228.

[83] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Information*, vol. 14, no. 8, 2023, ISSN: 2078-2489. DOI: 10.3390/info14080467.

[84] P. Roy, S. Bhawal, and C. Subalalitha, "Hate speech and offensive language detection in Dravidian languages using deep ensemble framework," *Computer Speech and Language*, vol. 75, 2022. DOI: 10.1016/j.csl.2022.101386.

[85] A. Keya, M. Kabir, N. Shammey, M. Mridha, M. Islam, and Y. Watanobe, "G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media," *IEEE Access*, vol. 11, pp. 79 697–79 709, 2023. DOI: 10.1109/ACCESS.2023.3299021.

[86] N. Albadi, M. Kurdi, and S. Mishra, "Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space," *Social Network Analysis and Mining*, vol. 9, no. 1, 2019. DOI: 10.1007/s13278-019-0587-5.

[87] P. Kar and S. Debbarma, "Sentimental analysis & Hate speech detection on English and German text collected from social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network," *Engineering Applications of Artificial Intelligence*, vol. 126, 2023. DOI: 10.1016/j.engappai.2023.107143.

[88] F. Anezi, "Arabic Hate Speech Detection Using Deep Recurrent Neural Networks," *Applied Sciences (Switzerland)*, vol. 12, no. 12, 2022. DOI: 10.3390/app12126010.

[89] P. Sharmila, K. Anbananthen, D. Chelliah, S. Parthasarathy, and S. Kannan, "PDHS: Pattern-Based Deep Hate Speech Detection with Improved Tweet Representation," *IEEE Access*, vol. 10, pp. 105 366–105 376, 2022. DOI: 10.1109/ACCESS.2022.3210177.

[90] Z. Boulouard, M. Ouaissa, M. Ouaissa, M. Krichen, M. Almutiq, and K. Gasmi, "Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning," *Applied Sciences (Switzerland)*, vol. 12, no. 24, 2022. DOI: 10.3390/app122412823.

[91] M. Casavantes, M. Aragón, L. González, and M. Montes-y-Gómez, "Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter," *Journal of Intelligent Information Systems*, vol. 61, no. 2, pp. 519–539, 2023. DOI: 10.1007/s10844-023-00779-z.

[92] C. Arcila-Calderón, J. Amores, P. Sánchez-Holgado, L. Vrysis, N. Vryzas, and M. Oller Alonso, "How to Detect Online Hate towards Migrants and Refugees? Developing and Evaluating a Classifier of Racist and Xenophobic Hate Speech Using Shallow and Deep Learning," *Sustainability (Switzerland)*, vol. 14, no. 20, 2022. DOI: 10.3390/su142013094.

[93] A. Toliyat, S. Levitan, Z. Peng, and R. Etemadpour, "Asian hate speech detection on Twitter during COVID-19," *Frontiers in Artificial Intelligence*, vol. 5, 2022. DOI: 10.3389/frai.2022.932381.

[94] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with hindi and english social media," *Information (Switzerland)*, vol. 12, no. 1, pp. 1–16, 2021. DOI: 10.3390/info12010005.

[95] H. Fan, W. Du, A. Dahou, A. Ewees, D. Yousri, M. Elaziz, A. Elsheikh, L. Abualigah, and M. Al-Qaness, "Social media toxicity classification using deep learning: Real-world application uk brexit," *Electronics (Switzerland)*, vol. 10, no. 11, 2021. DOI: 10.3390/electronics10111332.

[96] K. Shanmugavadivel, V. E. Sathishkumar, S. Raja, T. B. Lingaiah, S. Neelakandan, and M. Subramanian, "Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data," *Scientific Reports*, vol. 12, no. 1, p. 21 557, Dec. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-26092-3. [Online]. Available: https://doi.org/10.1038/s41598-022-26092-3.

[97] E. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Information Processing and Management*, vol. 58, no. 4, 2021. DOI: 10.1016/j.ipm.2021.102544.

[98] J. Benítez-Andrades, Á. González-Jiménez, Á. López-Brea, J. Aveleira-Mata, J.-M. Alija-Pérez, and M. García-Ordás, "Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT," *PeerJ Computer Science*, vol. 8, 2022. DOI: 10.7717/PEERJ-CS.906.

[99] F. Plaza-del-Arco, M. Molina-González, L. Ureña-López, and M. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Expert Systems with Applications*, vol. 166, 2021. DOI: 10.1016/j.eswa.2020.114120.

[100] J. Perez, F. Luque, D. Zayat, M. Kondratzky, A. Moro, P. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, and V. Cotik, "Assessing the Impact of Contextual Information in Hate Speech Detection," *IEEE Access*, vol. 11, pp. 30 575–30 590, 2023. DOI: 10.1109/ACCESS.2023.3258973.

[101] G. Valle-Cano, L. Quijano-Sánchez, F. Liberatore, and J. Gómez, "SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles," *Expert Systems with Applications*, vol. 216, 2023. DOI: 10.1016/j.eswa.2022.119446.

[102] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman Urdu Hate Speech Detection Using Transformer-Based Model for Cyber Security Applications," *Sensors*, vol. 23, no. 8, 2023, ISSN: 1424-8220. DOI: 10.3390/s23083909.

[103] J. Molero, J. Perez-Martin, A. Rodrigo, and A. Penas, "Offensive Language Detection in Spanish Social Media: Testing from Bag-of-Words to Transformers Models," *IEEE Access*, vol. 11, pp. 95 639–95 652, 2023. DOI: 10.1109/ACCESS.2023.3310244.

[104] A. Ramponi, B. Testa, S. Tonelli, and E. Jezek, "Addressing religious hate online: From taxonomy creation to automated detection," *PeerJ Computer Science*, vol. 8, 2022. DOI: 10.7717/PEERJ-CS.1128.

[105] M. Bhardwaj, M. Sundriyal, M. Bedi, M. Akhtar, and T. Chakraborty, "HostileNet: Multilabel Hostile Post Detection in Hindi," *IEEE Transactions on Computational Social Systems*, pp. 1–11, 2023. DOI: 10.1109/TCSS.2023.3244014.

[106] M. Almaliki, A. Almars, I. Gad, and E.-S. Atlam, "ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media," *Electronics (Switzerland)*, vol. 12, no. 4, 2023. DOI: 10.3390/electronics12041048.

[107] P. Kapil, G. Kumari, A. Ekbal, S. Pal, A. Chatterjee, and B. Vinutha, "HHSD: Hindi Hate Speech Detection Leveraging Multi-Task Learning," *IEEE Access*, vol. 11, pp. 101 460–101 473, 2023. DOI: 10.1109/ACCESS.2023.3312993.

[108] M. Tonneau, P. Quinta De Castro, K. Lasri, I. Farouq, L. Subramanian, V. Orozco-Olvera, and S. Fraiberger, "NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9020–9040. [Online]. Available: https://aclanthology.org/2024.acl-long.488.

[109] R. B. Santos, B. C. Matos, P. Carvalho, F. Batista, and R. Ribeiro, "Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains," in *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, J. Cordeiro, M. J. Pereira, N. F. Rodrigues, and S. Pais, Eds., ser. Open Access Series in Informatics (OASIcs), vol. 104, Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, 11:1–11:14, ISBN: 978-3-95977-245-7. DOI: 10.4230/OASIcs.SLATE.2022.11. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.11.

[110] B. C. Matos, R. B. Santos, P. Carvalho, R. Ribeiro, and F. Batista, "Comparing Different Approaches for Detecting Hate Speech in Online Portuguese Comments," in *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, J. Cordeiro, M. J. Pereira, N. F. Rodrigues, and S. Pais, Eds., ser. Open Access Series in Informatics (OASIcs), vol. 104, Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022, 10:1–10:12, ISBN: 978-3-95977-245-7. DOI: 10.4230/OASIcs.SLATE.2022.10. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.10.

[111] I. Markov, I. Gevers, and W. Daelemans, "An ensemble approach for dutch cross-domain hate speech detection," in *Natural Language Processing and Information Systems*, P. Rosso, V. Basile, R. Martínez, E. Métais, and F. Meziane, Eds., Cham: Springer International Publishing, 2022, pp. 3–15, ISBN: 978-3-031-08473-7.

[112] F. Rodriguez-Sanchez, J. Carrillo-De-Albornoz, and L. Plaza, "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data," *IEEE Access*, vol. 8, pp. 219 563–219 576, 2020. DOI: 10.1109/ACCESS.2020.3042604.

[113] S. Dowlagar and R. Mamidi, "Hate Speech Detection on Code-Mixed Dataset Using a Fusion of Custom and Pre-trained Models with Profanity Vector Augmentation," *SN Computer Science*, vol. 3, no. 4, 2022. DOI: 10.1007/s42979-022-01189-8.

[114] I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, "Label modification and bootstrapping for zero-shot cross-lingual hate speech detection," *Language Resources and Evaluation*, 2023. DOI: 10.1007/s10579-023-09637-4.

[115] L. Liu, D. Xu, P. Zhao, D. Zeng, P.-H. Hu, Q. Zhang, Y. Luo, and Z. Cao, "A cross-lingual transfer learning method for online COVID-19-related hate speech detection," *Expert Systems with Applications*, vol. 234, 2023. DOI: 10.1016/j.eswa.2023.121031.

[116] M. Awal, R. Lee, E. Tanwar, T. Garg, and T. Chakraborty, "Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2023. DOI: 10.1109/TCSS.2023.3252401.

[117] M. Subramanian, R. Ponnusamy, S. Benhur, K. Shanmugavadivel, A. Ganesan, D. Ravi, G. K. Shanmugasundaram, R. Priyadharshini, and B. R. Chakravarthi, "Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer," *Computer Speech & Language*, vol. 76, p. 101 404, Nov. 2022, ISSN: 0885-2308. DOI: 10.1016/j.csl.2022.101404. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230822000407.

[118] M. Arshad, R. Ali, M. Beg, and W. Shahzad, "UHated: Hate speech detection in Urdu language using transfer learning," *Language Resources and Evaluation*, vol. 57, no. 2, pp. 713–732, 2023. DOI: 10.1007/s10579-023-09642-7.

[119] O. Kaminska, C. Cornelis, and V. Hoste, "Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony," *Information Sciences*, vol. 625, pp. 521–535, 2023. DOI: 10.1016/j.ins.2023.01.054.

[120] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3309–3326. DOI: 10.18653/v1/2022.acl-long.234. [Online]. Available: https://aclanthology.org/2022.acl-long.234.

[121] S. Bansal, V. Garimella, A. Suhane, J. Patro, and A. Mukherjee, "Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for

Computational Linguistics, Jul. 2020, pp. 1018–1023. DOI: 10.18653/v1/2020.acl-main.96. [Online]. Available: https://aclanthology.org/2020.acl-main.96.

[122] T. T. Aurpa, R. Sadik, and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 24, Dec. 2021, ISSN: 1869-5469. DOI: 10.1007/s13278-021-00852-x. [Online]. Available: https://doi.org/10.1007/s13278-021-00852-x.

[123] X. Su, Y. Li, P. Branco, and D. Inkpen, "SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media," *Natural Language Engineering*, 2023. DOI: 10.1017/S1351324923000396.

[124] S. Cohen, D. Presil, O. Katz, O. Arbili, S. Messica, and L. Rokach, "Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time," *Information Fusion*, vol. 99, 2023. DOI: 10.1016/j.inffus.2023.101887.

[125] M. Zhang, J. He, T. Ji, and C.-T. Lu, *Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection*, 2024. arXiv: 2402.11406 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.11406.

[126] F. Plaza-Del-Arco, M. Molina-Gonzalez, L. Urena-Lopez, and M. Martin-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," *IEEE Access*, vol. 9, pp. 112 478–112 489, 2021. DOI: 10.1109/ACCESS.2021.3103697.

[127] X. Zhou, Y. Yong, X. Fan, G. Ren, Y. Song, Y. Diao, L. Yang, and H. Lin, "Hate speech detection based on sentiment knowledge sharing," in *Annual Meeting of the Association for Computational Linguistics*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236459847.

[128] C. Min, H. Lin, X. Li, H. Zhao, J. Lu, L. Yang, and B. Xu, "Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective," *Information Fusion*, vol. 96, pp. 214–223, 2023. DOI: 10.1016/j.inffus.2023.03.015.

[129] M. Zampieri, T. Ranasinghe, D. Sarkar, and A. Ororbia, "Offensive language identification with multi-task learning," *Journal of Intelligent Information Systems*, vol. 60, no. 3, pp. 613–630, Jun. 2023, ISSN: 1573-7675. DOI: 10.1007/s10844-023-00787-z. [Online]. Available: https://doi.org/10.1007/s10844-023-00787-z.

[130] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proceedings of the Third Workshop on Abusive Language Online*, S. T. Roberts, J. Tetreault, V. Prabhakaran, and Z. Waseem, Eds., Florence, Italy: Association for Computational

Linguistics, Aug. 2019, pp. 94–104. DOI: 10.18653/v1/W19-3510. [Online]. Available: https://aclanthology.org/W19-3510.

[131] P. Carvalho, D. Caled, C. Silva, F. Batista, and R. Ribeiro, "The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments," *Journal of Language Aggression and Conflict*, 2023, ISSN: 2213-1272. DOI: https://doi.org/10.1075/jlac.00085.car. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/jlac.00085.car.

[132] P. Carvalho, B. Matos, R. Santos, F. Batista, and R. Ribeiro, "Hate speech dynamics against African descent, Roma and LGBTQ+ communities in Portugal," eng, in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, European Language Resources Association (ELRA, 2022.

[133] R. Santos, J. Rodrigues, L. Gomes, J. Silva, A. Branco, H. L. Cardoso, T. F. Osório, and B. Leite, *Fostering the ecosystem of open neural encoders for portuguese with albertina pt* family*, 2024. arXiv: 2403.01897.

[134] A. A. Firmino, C. de Souza Baptista, and A. C. de Paiva, "Improving hate speech detection using cross-lingual learning," *Expert Systems with Applications*, vol. 235, p. 121 115, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.121115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423016172.

[135] A. Silva and N. Roman, "Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression," in *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, Evento Online: SBC, 2020, pp. 1–12. DOI: 10.5753/eniac.2020.12112. [Online]. Available: https://sol.sbc.org.br/index.php/eniac/article/view/12112.

[136] A. A. Firmino, C. S. de Baptista, and A. C. de Paiva, "Using cross lingual learning for detecting hate speech in portuguese," in *Database and Expert Systems Applications*, C. Strauss, G. Kotsis, A. M. Tjoa, and I. Khalil, Eds., Cham: Springer International Publishing, 2021, pp. 170–175, ISBN: 978-3-030-86475-0.

[137] R. Marques, "On the system of mood in european and brazilian portuguese," *Journal of Portuguese Linguistics*, vol. 3, pp. 89–109, 2004.

[138] T. Móia and A. T. Alves, "Differences between european and brazilian portuguese in the use of temporal adverbials," *Journal of Portuguese linguistics*, vol. 3, no. 1, 2004.

[139] F. Baider, "Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement," *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, vol. 35, pp. 1–25, Apr. 2022. DOI: 10.1007/s11196-022-09882-w.

[140] F. Baider, "Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech," *Pragmatics and Society*, vol. 11, no. 2, pp. 196–218, 2020, ISSN: 1878-9714. DOI: 10.1075/ps.20004.bai.

[141] M. Pohjonen and S. Udupa, "Extreme speech online: An anthropological critique of hate speech debates," *International Journal of Communication*, vol. 11, no. 0, 2017, ISSN: 1932-8036. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/5843.

[142] M. B. Giacomo Marzi and D. MArchiori, "K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient," *Methods X*, vol. 12, 2024.

[143] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds., Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2054–2059. DOI: 10.18653/v1/2020.semeval-1.271. [Online]. Available: https://aclanthology.org/2020.semeval-1.271.

[144] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[145] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[146] J. A. Wagner Filho, R. Wilkens, M. Idiart, and A. Villavicencio, "The brWaC corpus: A new open resource for Brazilian Portuguese," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1686.

[147] R. Chaves Rodrigues, M. Tanti, and R. Agerri, *Evaluation of Portuguese Language Models*, version 1.0.0, Mar. 2023. DOI: 10.5281/zenodo.7781848. [Online]. Available: https://github.com/ruanchaves/eplm.

[148] F. Vargas, I. Carvalho, F. Rodrigues de Góes, T. Pardo, and F. Benevenuto, "HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 7174–7183. [Online]. Available: https://aclanthology.org/2022.lrec-1.777.

[149] P. He, J. Gao, and W. Chen, *Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing*, 2021. arXiv: 2111.09543 [cs.CL].

[150] B. C. Matos, "Automatic Hate Speech Detection in Portuguese Social Media Text," eng, M.S. thesis, Instituto Superior Técnico, Nov. 2022.

[151] L. Chen, M. Zaharia, and J. Zou, *How is chatgpt's behavior changing over time?* 2023. arXiv: 2307.09009.

[152] K.-L. Chiu, A. Collins, and R. Alexander, *Detecting hate speech with gpt-3*, 2022. arXiv: 2103.12407 [cs.CL].

[153] G. Assis, A. Amorim, J. Carvalho, D. de Oliveira, D. Vianna, and A. Paes, "Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models?" In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, Eds., Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, Mar. 2024, pp. 301–311. [Online]. Available: https://aclanthology.org/2024.propor-1.31.

[154] L. Li, L. Fan, S. Atreja, and L. Hemphill, *"hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media*, 2023. arXiv: 2304.10619 [cs.CL].

[155] A. Petrescu, C.-O. Truică, E.-S. Apostol, and P. Karras, "Sparse shield: Social network immunization vs. harmful speech," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21, Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 1426–1436, ISBN: 9781450384469. DOI: 10.1145/3459637.3482481. [Online]. Available: https://doi.org/10.1145/3459637.3482481.

[156] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, and P. Karras, "Mcwdst: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media," *IEEE Access*, vol. 11, pp. 125 861–125 873, 2023. DOI: 10.1109/ACCESS.2023.3331220.

[157] E. S. Apostol, Ö. Coban, and C.-O. Truică, "Contain: A community-based algorithm for network immunization," *Engineering Science and Technology, an International Journal*, vol. 55, p. 101 728, 2024, ISSN: 2215-0986. DOI: https://doi.org/10.1016/j.jestch.2024.101728. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215098624001149.

[158] E.-S. Apostol, C.-O. Truică, and A. Paschke, *Contcommrtd: A distributed content-based misinformation-aware community detection system for real-time disaster reporting*, 2023. arXiv: 2301.12984 [cs.SI]. [Online]. Available: https://arxiv.org/abs/2301.12984.