Social Network Analysis and Mining (2024) 14:204
https://doi.org/10.1007/s13278-024-01361-3

REVIEW PAPER

A comprehensive review on automatic hate speech detection in the age of the transformer

Gil Ramos[1] · **Fernando Batista**[2,3] · **Ricardo Ribeiro**[2,3] · **Pedro Fialho**[1] · **Sérgio Moro**[1,4] · **António Fonseca**[1] · **Rita Guerra**[5] · **Paula Carvalho**[3,6] · **Catarina Marques**[7] · **Cláudia Silva**[8]

Received: 27 February 2024 / Revised: 7 September 2024 / Accepted: 22 September 2024

Abstract

The rapid proliferation of hate speech on social media poses significant challenges to maintaining a safe and inclusive digital environment. This paper presents a comprehensive review of automatic hate speech detection methods, with a particular focus on the evolution of approaches from traditional machine learning and deep learning models to the more advanced Transformer-based architectures. We systematically analyze over 100 studies, comparing the effectiveness, computational requirements, and applicability of various techniques, including Support Vector Machines, Long Short-Term Memory networks, Convolutional Neural Networks, and Transformer models like BERT and its multilingual variants. The review also explores the datasets, languages, and sources used for hate speech detection, noting the predominance of English-focused research while highlighting emerging efforts in low-resource languages and cross-lingual detection using multilingual Transformers. Additionally, we discuss the role of generative and multi-task learning models as promising avenues for future development. While Transformer-based models consistently achieve state-of-the-art performance, this review underscores the trade-offs between performance and computational cost, emphasizing the need for context-specific solutions. Key challenges such as algorithmic bias, data scarcity, and the need for more standardized benchmarks are also identified. This review provides crucial insights for advancing the field of hate speech detection and shaping future research directions.

Keywords Hate speech detection · Machine learning · Deep learning · Transfer learning · Transformers · Literature review

✉ Gil Ramos
gasnr@iscte-iul.pt

✉ Fernando Batista
fernando.batista@iscte-iul.pt

✉ Ricardo Ribeiro
ricardo.ribeiro@iscte-iul.pt

1 Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal

2 Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal

3 INESC-ID, Lisbon, Portugal

4 University of Jordan, Amman, Jordan

5 ISCTE-Instituto Universitário de Lisboa and Center for Psychological Research and Social Intervention (CIS-ISCTE), Lisbon, Portugal

6 Universidade de Aveiro, Aveiro, Portugal

7 ISCTE-Instituto Universitário de Lisboa and Business Research Unit (BRU-ISCTE), Lisbon, Portugal

8 ITI-LARSyS and IST, Lisbon, Portugal

1 Introduction

In recent years, the surge in social media usage has transformed the landscape of digital communication, fundamentally altering how individuals express themselves and connect with others (Statista 2023). With the widespread availability of smartphones and Internet access, social media platforms have become easily accessible to a global audience, providing a seamless channel for individuals to share their thoughts and ideas. This democratization of expression, while empowering people to voice their opinions and engage in meaningful conversations, has also brought to the forefront a pressing issue: the widespread proliferation of Hate Speech (HS) (Watanabe et al. 2018), which poses a critical threat to online communities and society, in general.

There are no universally accepted and precise definitions of HS (Poletto et al. 2021), but, according to the United Nations (2019), HS is defined as any form of communication that targets and employs derogatory or discriminatory language concerning individuals or groups based on intrinsic

Published online: 09 October 2024

attributes such as religion, ethnicity, nationality, race, color, descent, gender, or other identity factors. This type of discourse can lead to significant psychological and emotional distress among recipients, such as stress, anxiety and depression (Tynes et al. 2008). Beyond the immediate emotional impact, prolonged exposure to HS can also erode social cohesion, fostering an atmosphere of mistrust and polarization. This divisiveness can further perpetuate the cycle of hate and individuals may become increasingly isolated within their own echo chambers, reinforcing existing biases and prejudices (MediaSmarts 2021).

Many organizations, recognizing the urgency of addressing the proliferation of HS on social media, have initiated the release of guidelines and policies designed to mitigate this issue. However, the sheer scale of the problem, characterized by the continuous generation of vast volumes of data on these platforms, presents an inherent challenge to manual classification methods. Manual intervention is ultimately rendered impractical due to its time-consuming nature, underscoring the need to employ Machine Learning (ML) techniques to automate and streamline the classification process, thereby producing more dependable and efficient results (Qian Li et al. 2022). As a consequence of this technological shift, a dynamic landscape of research and development has emerged, aimed at harnessing the power of ML for HS detection.

Various techniques, ranging from approaches like traditional ML and Deep Learning (DL) models, have been applied with promising results, and recently, with the development of Transformer-based models (Vaswani et al. 2017), we have seen a growing expansion in the HS detection landscape.

The recent advances in Transformer-based models have introduced new possibilities in HS detection, but a comprehensive synthesis of these efforts is lacking, particularly in terms of comparing them to other ML methods.

This Systematic Literature Review (SLR) addresses this gap by exploring the current research landscape of HS detection on social media, with a specific focus on Transformer-based models. We aim to answer the following research 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?

This article makes three key contributions: (1) it offers a comprehensive review of HS detection methods with a focus on Transformer-based models, (2) it compares these models with other ML techniques in terms of performance and applicability, and (3) it identifies key datasets and challenges in the field to inform future research.

This document is organised as follows: Sect. 2 gives some background on what is HS and how it is defined across several organizations and research initiatives; Sect. 3 delves into the methodological aspects of the SLR, outlining the search strategy, inclusion criteria, and data extraction processes; Sect. 4 presents a comprehensive analysis of the selected studies, highlighting the key findings and principal results; finally, Sect. 5 presents the major conclusions and pinpoints current limitations and future directions.

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 bellow we can see some examples of this definitions. Since this SLR was developed in the scope of the kNOwHATE: kNOwing online HATE speech project (kNOwHATE 2023), we also provide the definition used in the project:

- 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" (United Nations 2019).
- 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" (Meta 2023).
- Twitter 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" (Twitter 2023).
- 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" Google (2019).

- 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" (Parekh 2012).
- kNOwHATE project: building on scholar definitions (i.e., Siegel 2020) 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" (Carvalho and Guerra 2023).

When examining the various interpretations of Hate Speech 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 his/her 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 this 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.

This work 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) (Carvalho and Guerra 2023), the detection methods used for these types of speech are quite similar.

In order to maintain clarity, the remainder of the article refers to these collective studies (HS, offensive, and abusive speech detection) as HS detection works. Nevertheless, Sect. 4 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.

3 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 (Page et al. 2021). PRISMA provides a widely recognized framework for conducting systematic reviews, ensuring transparency and methodological rigor 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 offer valuable insights into the state-of-the-art in HS detection in social media, 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 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 as previously mentioned 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 (Alkomah and Ma 2022). This terms are also present in the literature as key terms to use when finding relevant studies (Alrashidi et al. 2023; Mullah and Zainon 2023; Yin and Zubiaga 2021). 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 were 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 (Vaswani et al. 2017), 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 primary focus centred on this specific area of research and that proposed or discussed solutions related to this classification task.

The exclusion criteria were: articles primarily focused on other forms of media, such as images, sound, memes, and non-textual content, articles that lack a clear approach or technical content related to HS classification, and finally, studies that do 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 recognize that by excluding high impact peer-reviewed conferences we are 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 were held, that aimed at detecting offensive language. By including the most relevant studies papers 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.

Fig. 1 shows the number of records identified in the database search, and the filtering process that is applied afterwards, using a PRISMA flow diagram (Page et al. 2021). 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 (Ayo et al. 2021; Lee et al. 2022; Roy et al. 2023). 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.

[1] www.scimagojr.com.

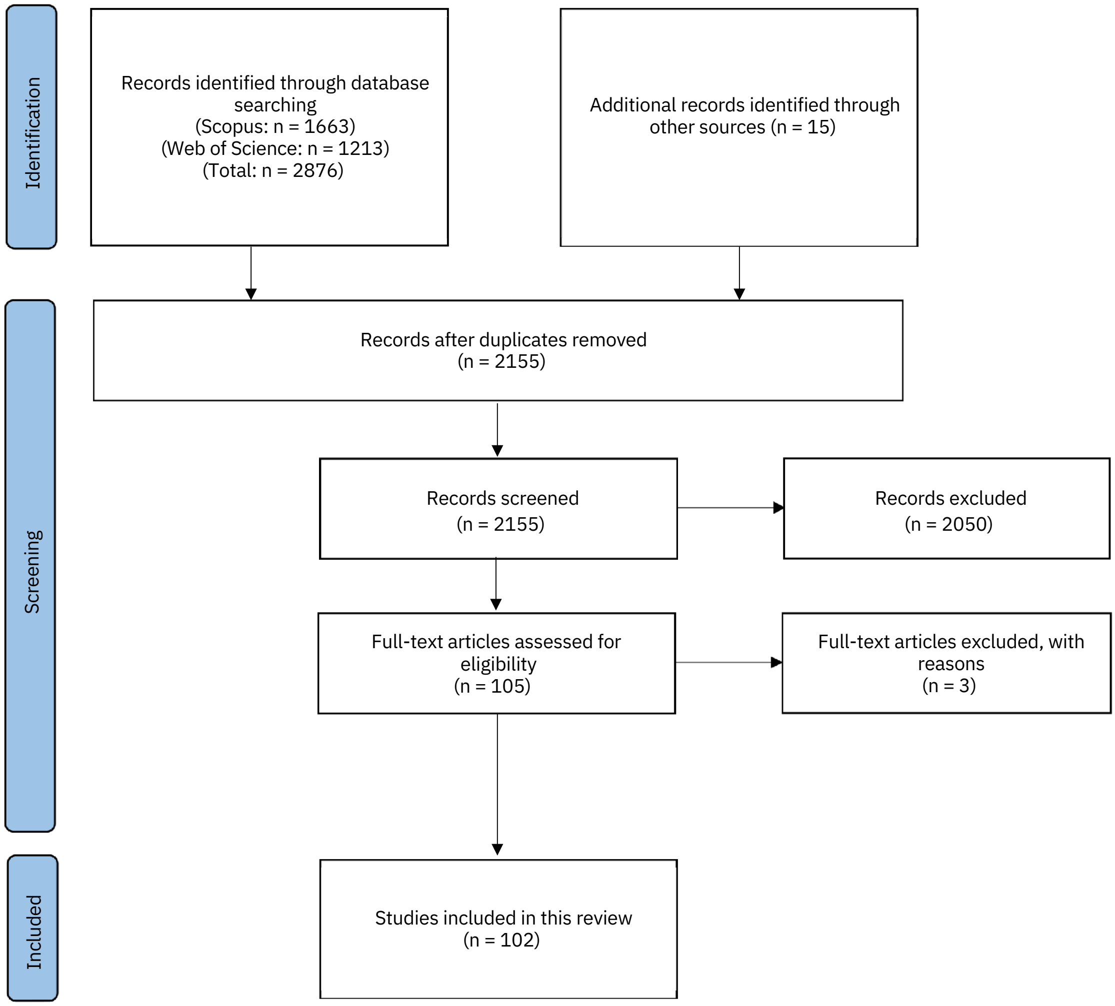

Fig. 1 PRISMA flow diagram

4 Results

The findings of this SLR are presented in this section, divided into four distinctive categories: An overall analysis of the results of our search (Sect. 4.1), an analysis of the evolution of HS detection (Sect. 4.2), Methods and Algorithms where we will compare all different approaches employed for this task (Sect. 4.3), and Resources where both the languages and the types of data used for the detection will also be analyzed (Sect. 4.4). Through a meticulous synthesis of empirical evidence and critical evaluation, this section aims to provide a comprehensive overview of the state-of-the-art in HS detection.

4.1 Overall analysis

When analysing the initial results of the 2155 (2140 plus 15) articles not duplicated, resulting from our search query we can see in Fig. 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.

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

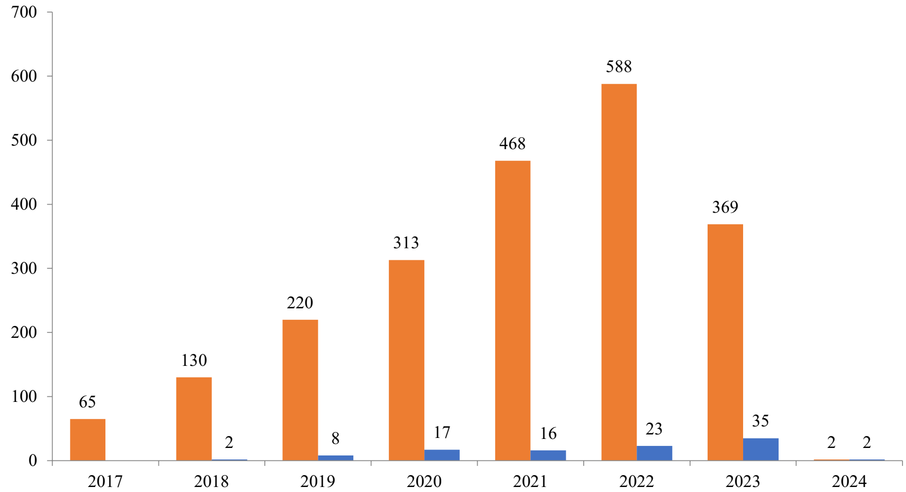

Fig. 2 Number of search results (orange) and included documents (blue) by year

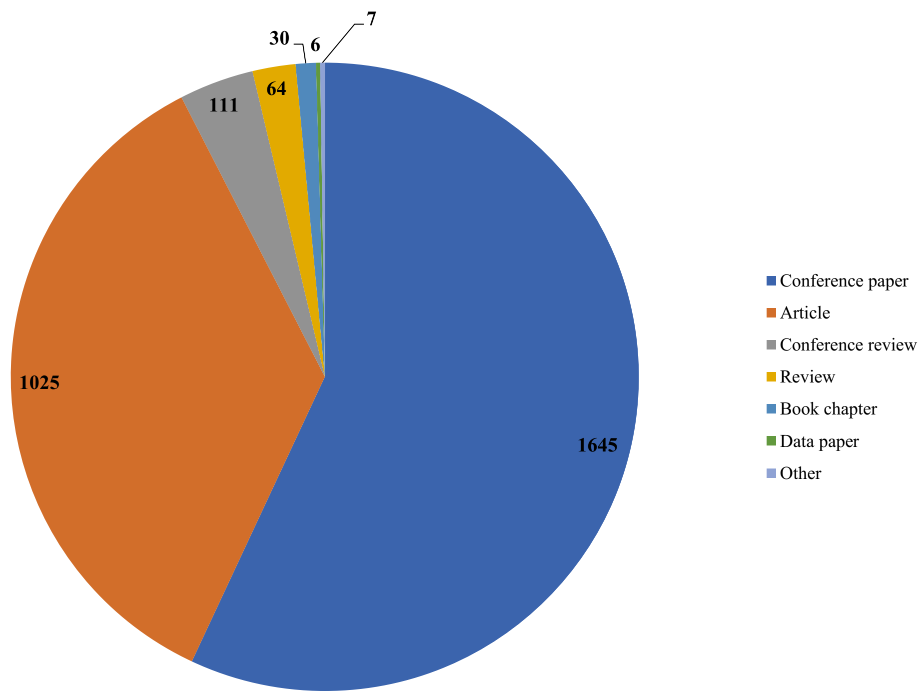

Fig. 3 Number of documents by type

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 Fig. 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 (Basile et al. 2019; Zampieri et al. 2020; Wiegand et al. 2018), from which a large number of conference articles result, since each participant has their contribution in the form of a conference paper.

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 87 studies considered for this SLR.

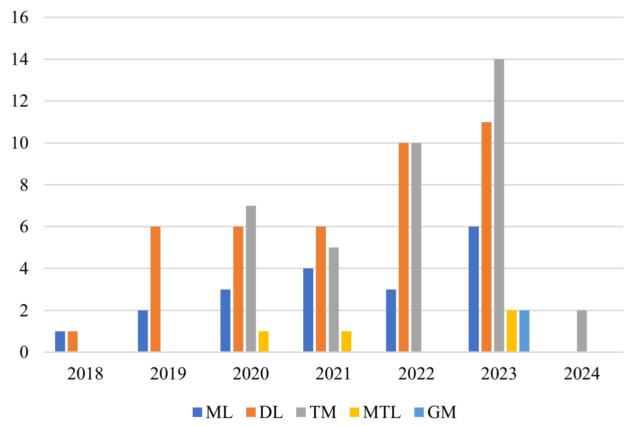

Fig. 4 HS detection approaches by year (MTL: Multi-Task Learning)

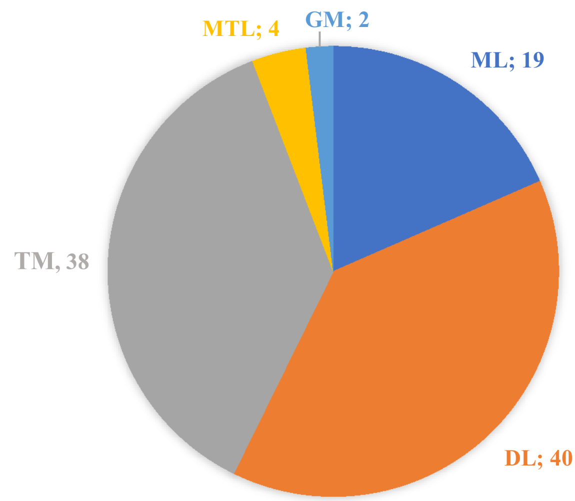

Fig. 5 Number of HS studies, grouped by approach

4.2 Q1: Landscape of HS detection literature

Over the years, various approaches have been employed for hate speech (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.

Fig. 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 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. (2017), 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 (MTL) 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. Although traditional ML techniques remained relevant, their usage declined. 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, two studies featuring Transformer models were published, both coinciding with ACL papers extracted after the search, explaining their presence as the only studies from that year.

Fig 5 shows the total number of studies that employed each approach. DL and Transformers are the most frequently used methods, with 40 and 38 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 two 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.

The authors of the OffensEval-2019 reported that over half of the participants explored Deep Learning models (Basile et al. 2019). In contrast, OffensEval-2020 saw most teams utilizing pre-trained Transformer models, with all of the top 10 teams employing either BERT, RoBERTa, or XLM-RoBERTa (Zampieri et al. 2020)..

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, to provide a more comprehensive representation of the state-of-the-art solutions during that period. As shown in Table 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.

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

Table 1 SemEval top papers

Paper	Model	Method	Rank
OffensEval-2019			
Indurthi et al. (2019)	SVM model with RBF kernel	ML	1st
Ding et al. (2019)	stacked BiGRUs	DL	2nd
Alonzorz	Multiple Choice CNN	DL	3rd
Montejo-Ráez et al. (2019)	LSTM	DL	4th
Pérez and Luque (2019)	linear-kernel SVM	ML	1st (Spanish Task)
OffensEval-2020			
Wiedemann et al. (2020)	Ensemble of ALBERT models	TM	1st
Wiedemann et al. (2020)	RoBERTa-large	TM	2nd
Wang et al. (2020)	XLM-R-base and XLMR-large	TM	3rd
Dadu and Pant (2020)	XLM-R	TM	4th
Sotudeh et al. (2020)	BERT	TM	5th

also seen a growing diversity of approaches, with Generative Models (GM) and Multi-Task Learning (MTL) 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.

4.3 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, to analyze and learn from data.

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, Transformers 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 labeled 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.

4.3.1 Traditional machine learning

Starting with traditional ML techniques, we identified 16 studies that resorted to this type of method, and made comparisons with various algorithms. Support Vector Machines (SVM) and Logistic Regression (LR) where the algorithms that achieved better results, outperforming other ML algorithms in three different studies respectively. Pitropakis et al. (2020); Shannaq et al. (2022); Mohapatra et al. (2021) 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. (2019) and Pérez and Luque (2019) 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. (2021); Vanetik and Mimoun (2022); Saeed et al. (2023) employed a 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 (Turki and Roy 2022), and the j48graft classifier, a type of Decision Tree (DT) model, combined with text features (Watanabe et al. 2018).

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

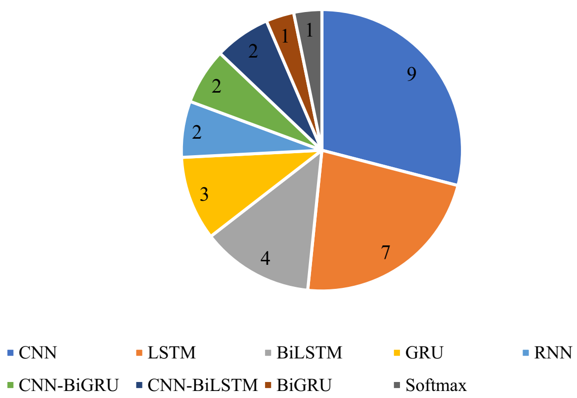

Fig. 6 Different DL models for HS detection

capture intricate relationships and context in the text data, which can be challenging for traditional feature engineering methods. (García-Díaz et al. 2023; Vanetik and Mimoun 2022) combined Bidirectional Encoder Representations from Transformers (BERT) embeddings with a Multi-Layer Perceptron (MLP) and LR models respectively, and managed to outperform ML and EM. In addition to this, (Raut and Spezzano 2023; Vanetik and Mimoun 2022) 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. This 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 this models used majority voting to get the predictions (Khairy et al. 2023; Aljero and Dimililer 2021; Rajalakshmi et al. 2023; Plaza-Del-Arco et al. 2020), two studies used a LR meta classifier (Agarwal and Chowdary 2021; Oriola and Kotze 2020), and one study used a stacking approach (Mullah and Zainon 2023).

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 (Saeed et al. 2023; Raut and Spezzano 2023; Vanetik and Mimoun 2022).

4.3.2 Deep learning

Jumping to DL techniques, these have been extensively used for the task of HS detection, with 37 studies employing this method. These studies have explored a variety of DL models, including CNNs, LSTMs, GRUs, and hybrid or ensemble models that combine multiple DL architectures as we can seen in Fig. 6.

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 (Karayiğit et al. 2021; Akhter et al. 2022; Roy et al. 2020; A. T. Kabakus 2021; Zhang and Luo 2019; Duwairi et al. 2021; Alshalan and Al-Khalifa 2020; Mozafari et al. 2020) with promising results, even outperforming Transformers (Alshalan and Al-Khalifa 2020), and getting 3rd place in OffenseEval-2019.

LSTMs are another class of recurrent neural networks (RNNs) 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 (Priyadarshini et al. 2023; Ayo et al. 2020; Dascălu and Hristea 2022; Pronoza et al. 2021; Pereira-Kohatsu et al. 2019; Madhu et al. 2023; Montejo-Ráez et al. 2019).

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 (Roy et al. 2020; A. T. Kabakus 2021), while others have found the opposite (Madhu et al. 2023; Dascălu and Hristea 2022; Ayo et al. 2020).

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-BiLSTM models have been shown to outperform even Transformers in some studies (Mundra and Mittal 2022; Fazil et al. 2023). 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 modeling ability of BiGRU's they managed to outperform all other single DL models (Kamal et al. 2023; Aarthi and Chelliah 2023). 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 (Pitsilis et al. 2018), a meta classifier of several combinations of models with different embeddings (Cruz et al. 2022), a combination of a BERT, BiLSTM and BiGRU models (Mazari et al. 2023) and finally a deep neural network with several text features (Al-Makhadmeh and Tolba 2020) 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 TL). These studies employed an ensemble of CNN models (Zhou et al. 2020), BERT models (Mridha et al. 2021), bagging of BiGRU, BiLSTM, CNN (Mahajan et al. 2024), a stacking of BiLSTM, LSTM, CNN and CNN-LSTM models (Muneer et al. 2023) and a combination of a BERT, MuRIL and DNN models (Roy et al. 2022).

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 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 (Keya et al. 2023; Albadi et al. 2019; Kar and Debbarma 2023). Another study used a BiGRU model managing to place top two in the OffensEval-2019 task (Ding et al. 2019). Other DL models used where Bidirectional RNNs (BiRNNs) (Anezi 2022) and a Softmax clasifier combined with text features (Sharmila et al. 2022).

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.

4.3.3 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 29 studies. The success of the basic BERT model on a plethora of different NLP tasks lead to the widespread use of this models and a large number of 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 (Boulouard et al. 2022; Casavantes et al. 2023; Arcila-Calderón et al. 2022; Toliyat et al. 2022; Vashistha and Zubiaga 2021; Fan et al. 2021; Shanmugavadivel et al. 2022; Pamungkas et al. 2021; Sotudeh et al. 2020), 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 (Benítez-Andrades et al. 2022; Plaza-del Arco et al. 2021; Perez et al. 2023; Valle-Cano et al. 2023), RuBERT for Russian (Bilal et al. 2023; Pronoza et al. 2021), RoBERTuito also for Spanish (Molero et al. 2023), UmBERTo for Italian (Ramponi et al. 2022), MARBERT for Arabic (Alrashidi et al. 2023), HindiBERT for Hindi (Bhardwaj et al. 2023), Arabic BERT-mini also for Arabic (Almaliki et al. 2023), MuRIL for seventeen indian languages (Kapil et al. 2023) and NAIJAXLM-T for English and Nigerian (Tonneau et al. 2024). 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 (Santos et al. 2022; Matos et al. 2022) and BERTje for Dutch (Markov et al. 2022). Besides this 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 (Rodriguez-Sanchez et al. 2020; Dowlagar and Mamidi 2022; Kapil et al. 2023; Bigoulaeva et al. 2023) and the XLM-RoBERTa was used in five studies (Liu et al. 2023; Awal et al. 2023; Subramanian et al. 2022; Wang et al. 2020; Dadu and Pant 2020). 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 (Dowlagar and Mamidi 2022; Arshad et al. 2023; Kaminska et al. 2023; Hartvigsen et al. 2022; Wiedemann et al. 2020; Bansal et al. 2020), ELECTRA (Aurpa et al. 2021) and AlBERT (Wiedemann et al. 2020). More recently, models like GPT−3.5 are also being used for this task, like the case of Zhang et al. (2024).

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.

4.3.4 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. Su et al. (2023) 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. (2023) 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 Generative Pre-trained Transformer (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.

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 labeled 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 this models, HS detection in low-resource languages can be done, without the need of extensive collection and annotation of data.

4.3.5 Multi-task learning

Previous studies have established the relevance of sentiment features in aiding HS detection tasks (Al-Makhadmeh and Tolba 2020; Sharmila et al. 2022; Watanabe et al. 2018). 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 (Plaza-Del-Arco et al. 2021; Zhou et al. 2021). 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. (2023) 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 (Zampieri et al. 2023).

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's 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-resource languages.

4.4 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.

4.4.1 Information sources

The majority of studies use data collected from different social media platforms as seen by Fig. 7. They are a rich source of data for HS detection, given the extensive volume

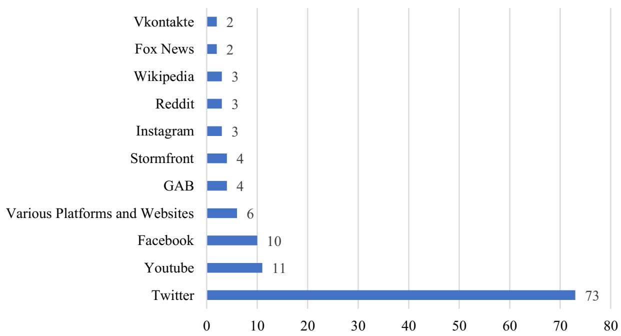

Fig. 7 Data sources for HS detection

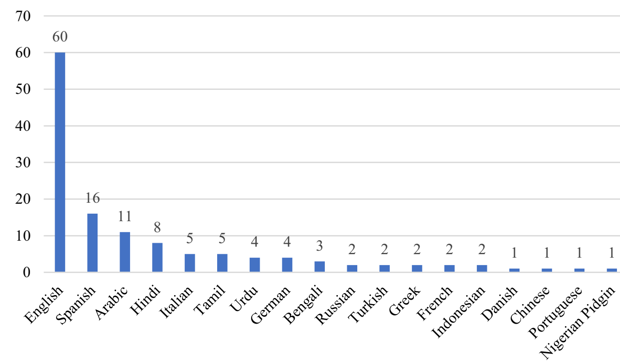

Fig. 8 Languages where HS detection was conducted

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 of 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 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 behavior. 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.

4.4.2 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 Fig. 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 eight 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 this languages in this field. For instance, even though Portuguese has only one study in our research, there have been recent attempts to create curated datasets for HS detection (Carvalho et al. 2022, 2023).

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 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

[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.

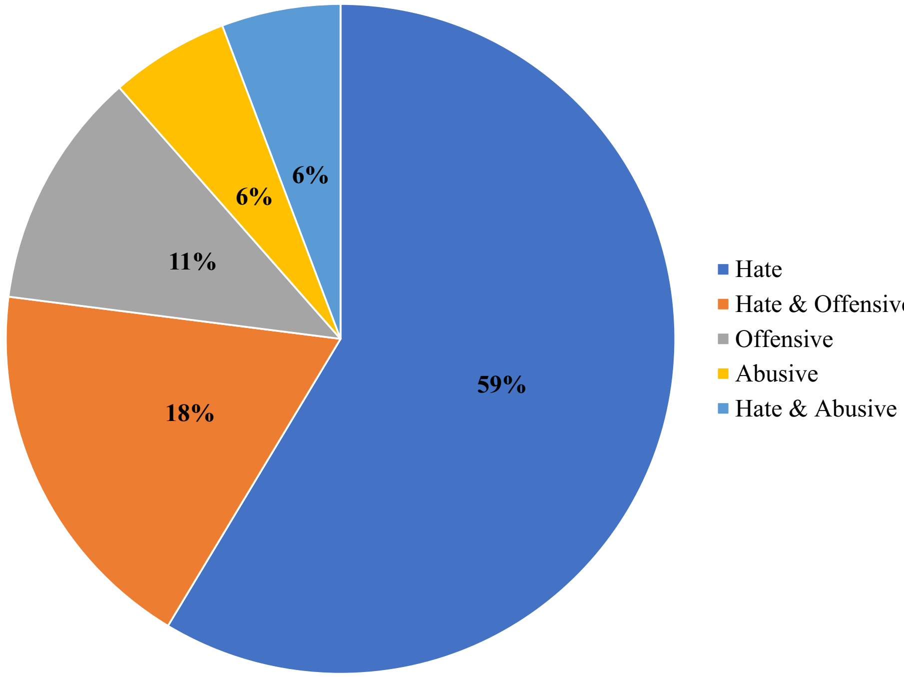

Fig. 9 Percentage of works dealing with the different types of speech included in our SLR

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 where by using a high resource source language for training, like is the case with English has we've seen, models can classify low-resource target languages with promising results (Bigoulaeva et al. 2023; Pamungkas et al. 2021). 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 (Awal et al. 2023; Liu et al. 2023).

4.4.3 Types of speech

As noted earlier, not all of the included studies focus solely on HS, as our search criteria also encompassed offensive and abusive speech. As illustrated in Fig. 9, the majority (83%) 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.

Tables 2, 3, 4, 5, 6, 7, 8, 9, 10 in Appendix A, present all the studies that were included in this SLR with the most important information about the approach each work followed for the task of HS detection.

5 Impact of transformer-based models

As we have seen throughout this SLR, Transformers have had an impact on almost all areas of HS detection. Firstly, theses 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 consistently show that Transformers outperform other highly used models such as CNN's, LSTM's, SVM's and Ensemble models. Moreover, besides Transformers having a better standalone performance they have also been incorporated into other models to further enhance detection accuracy. They have been used to enhance the performances of other models, or by taking advantage of their rich text representation, has features, or by combining them into hybrid or ensemble models. Furthermore, the advent of Transformers has catalyzed the development of HS literature and research, particularly in addressing challenges posed by low-resource languages. By leveraging pre-trained multilingual representations and fine-tuning on target languages with limited resources, Transformers have significantly expanded the scope of HS detection to encompass a broader array of linguistic contexts. In summary, the impact of Transformers on HS detection cannot be overstated. Their superior performance, integration into hybrid models, and facilitation of research in low-resource languages underscore their significance as the cornerstone of modern HS detection methodologies. Moving forward, continued advancements in Transformers hold immense promise in furthering our understanding of online HS dynamics and fostering safer digital environments for all users.

6 Conclusion

This work provides a comprehensive review of the evolution of hate speech (HS) detection, particularly focusing on the shift from traditional machine learning (ML) approaches to the dominance of Transformer-based models. Our review has shed light on several key lessons that will shape future efforts in this field. First, while Transformer models consistently outperform traditional ML and deep learning (DL) approaches in terms of performance, the trade-offs in computational demands highlight the need for context-specific solutions. Transformers excel in large-scale, multilingual applications, but DL models may offer faster, resource-efficient alternatives for specific tasks.

Second, our review reveals a growing yet underexplored interest in generative models and multi-task learning for HS detection. These approaches, while still in their infancy, show promise for handling more complex linguistic features of hate speech and addressing cross-platform variations in data. Moreover, the multilingual and cross-lingual capabilities of Transformer models present a significant advance, particularly for low-resource languages, suggesting a positive shift toward a more inclusive and globally applicable HS detection framework.

This review unveils several relevant insights: (1) Transformer models consistently outperform other methods, but their high computational requirements suggest that hybrid approaches, combining deep learning with traditional machine learning, may be more appropriate in certain contexts; (2) Although significant strides have been made in addressing low-resource languages, there is still a need for further work to improve inclusivity across a wider range of linguistic and cultural contexts; and (3) transparency and reproducibility remain critical challenges in the field, as the lack of publicly available code and datasets in many studies limits progress, hindering replication efforts and the development of generalizable models.

Looking ahead, we identify several key directions for future research. First, addressing algorithmic bias is imperative. Our review shows that despite advances in HS detection, bias mitigation remains under explored, especially for low-resource languages and marginalized communities. Future research should prioritize the development of fair and ethical models that avoid reinforcing societal inequalities. Second, there is a clear need for more standardized benchmarks and open-access resources. The difficulty of comparing results across studies due to inconsistencies in code and dataset availability is a major barrier to progress. Establishing common benchmarks, promoting data sharing, and ensuring transparency in methodology will be crucial in driving the field forward. Third, further exploration of emerging technologies such as multi-task learning and generative models could unlock new possibilities in HS detection. These techniques, which allow models to learn from multiple tasks simultaneously or generate more contextualized responses, may offer solutions to the inherent challenges of capturing the subtle and evolving nature of hate speech.

Our vision for the future of HS detection is one of interdisciplinary collaboration. As the scope of hate speech expands across different platforms and cultures, contributions from linguistics, computer science, ethics, and social sciences are essential to create holistic, reliable, and ethically sound solutions. We envision a future where HS detection systems are not only highly accurate but also transparent, fair, and adaptable to the needs of diverse online communities. By fostering such interdisciplinary efforts, we can ensure that HS detection tools contribute meaningfully to creating safer, more inclusive digital spaces.

One limitation of our review is its primary focus on journal articles, along with recent ACL papers and selected contributions from the OffensEval task at SemEval. This approach may have overlooked some cutting-edge research typically presented at conferences. While peer-reviewed journals provide a rigorous evaluation process, conferences are often hubs for the dissemination of innovative ideas and emerging trends. Consequently, the exclusion of a broader range of conference papers may have resulted in certain dimensions of the topic being underrepresented.

Although we decided to include only journal articles, we recognize that by excluding high impact peer-reviewed conferences we are 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 SemEval international workshops of 2019 and 2020 published in ACL. In these years' editions the OffensEval task was held, aimed at detecting offensive language. By including the most relevant studies papers 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

In summary, this work provides a comprehensive overview of the current research landscape in HS detection, with a particular focus on the increasing impact of Transformer-based models. It highlights key insights, identifies gaps in the existing literature, and suggests directions for future research. We aim for this review to serve as a foundation for further progress in the field, equipping researchers to tackle the complex and evolving challenges of detecting online hate speech.

Appendix: List of works analyzed in this SLR

See Tables 2, 3, 4, 5, 6, 7, 8, 9, 10, 11.

Table 2 Studies that employed traditional ML for HS detection.

References	Data source	Language	Features	Model	Outperformed
Raut and Spezzano (2023)	Twitter	English	BERT, user features and word count	XGB	CatBoost, BiGRU and BERT
García-Díaz et al. (2023)	Twitter	Spanish	Fine-tuned BETO embeddings	MLP	-
Watanabe et al. (2018)	Twitter	English	Sentiment, semantic, unigrams and pattern	J48graft-DT	RF & SVM
Shannaq et al. (2022)	Twitter	Arabic	Skip-grams	GA-SVM	KNN, NB, LR, DT, SVM, RF and XGB
Arcila-Calderón et al. (2021)	Twitter	Spanish	BOW	LR	NB, MNB, BNB, SGD, LSVC and RNN
Saeed et al. (2023)	Twitter	Urdu	Word n-grams	SVM	CNN, LSTM and BERT
Pitropakis et al. (2020)	Twitter	English	Word n-grams	SVM	LR, NB and n-grams
Turki and Roy (2022)	Twitter	English	Count vectorizer	RF	Bagging & AdaBoost
Vanetik and Mimoun (2022)	Twitter	French	mBERT embeddings	LR	RF, LR and XGB
Mohapatra et al. (2021)	Facebook	English-Odia[1]	Word2vec	SVM	NB & RF

[1] Code-mixed

Table 3 Studies that employed ensembles for HS detection.

References	Data source	Language	Features	Model	Outperformed
Khairy et al. (2023)	Twitter & Facebook	Arabic	TF-IDF	Hard Voting: LR+KNN+LSVC	LR, KNN and LSVC
Aljero and Dimililer (2021)	Twitter	English	Word2vec & USE sentence embeddings	Meta classifier: SVM+LR+XGB	KNN, LR, SVM, NB, RF and XGB
Mullah and Zainon (2023)	Twitter	English	TF-IDF	Stacking: RF+SVM+MNB +DT+LR+GBC +XGB+AdaB	RF, SVM, MNB, DT, LR, GNB, KNN, GBC, XGB and AdaB
Agarwal and Chowdary (2021)	Twitter	English	Word embeddings	Meta Classifier: SVM+GBDT +MLP+KNN +ELM	-
Rajalakshmi et al. (2023)	YouTube	Tamil	MuRIL embeddings	Majority Voting: RF+DT+NB	LR, SVM, SGD, RF, DT and NB
Plaza-Del-Arco et al. (2020)	Twitter	Spanish	TF unigrams and bigrams	Hard Voting: NB+LR	DT, SVM, MNB, LR and LSTM
Oriola and Kotze (2020)	Twitter	English	word n-grams and character n-grams	Meta Classifier: SVM+RF+GB	LR, SVM, RF and GB
Al-Makhadmeh and Tolba (2020)	Twitter & Stormfront	English	Semantic, sentiment, unigram and pattern features	DNN with a layer for each feature	TWEN-MLP, NLP-SVM, CGDNN and CANLNN

Table 4 Studies that employed Ensembles for HS detection.

References	Data source	Language	Features	Model	Outperformed
Muneer et al. (2023)	Twitter	English	CBOW	Stacking: LSTM+CNN +BiLSTM +Conv1DLSTM	LSTM, CNN, BiLSTM and BERT
Pitsilis et al. (2018)	Twitter	English	racism, sexism and neutral tendency	Majority Voting: LSTM for each combination of features	LR, LSTM-GBDT and Hybrid CNN
Mridha et al. (2021)	Websites and Platforms	Bengali	BERT embeddings	BERT-LSTM +BERT-AdaBoost	SVM, DT, RF,LR, LSTM, CNN, BiLSTM, mBERT and Bangla BERT
Cruz et al. (2022)	Twitter	English	Word2vec & TF-IDF	Meta Classifier: CNN+RF+NB +MLP	SVM, LR, RF, NB, KNN, MLP and CNN
Zhou et al. (2020)	Twitter	English	Character embeddings	Max fusion: 3xCNN	ELMo, BERT, BERT+ELMo +CNN
Mazari et al. (2023)	Wikipedia	English	GloVe and FastText embeddings	BERT+BiLSTM +CNN-LSTM	BiLSTM & GRU
Roy et al. (2022)	Twitter & YouTube	English-Tamil & English-Malayalam[1]	Word embeddings	BERT+DNN+ MuRIL	LR, RF, SVM, CNN, LSTM, BiLSTM, mBERT, XLM-R, MuRIL
Mahajan et al. (2024)	Twitter, Facebook, YouTube, Instagram and Forums	English, Bengali, Indonesian, Italian and Spanish	Word embeddings	Super Learner: BiGRU+BiLSTM +CNN-LSTM	BiGRU, BiLSTM, Stacked LSTM, XLM-R, AlBERT and BERT

[1] Code-mixed

Table 5 Studies that employed Deep Learning for HS detection.

References	Data source	Language	Features	Model	Outperformed
Asiri et al. (2022)	Twitter & Stormfront	English	GloVe embeddings	Attention BiLSTM	SVM, KNLPE-DNN, CG-DNN and CANL-NN
Karayiğit et al. (2021)	Instagram	Turkish	CBOW	CNN	SVM, NB, RF, LR, DT, AdaB, XGB
Akhter et al. (2022)	YouTube	Urdu	Word embeddings	CNN	LSTM, BiLSTM, LR, SVM and NB
Kamal et al. (2023)	Twitter & Fox News	English	GloVe embeddings, sentiment, hate lexicon, affective, syntatic and readability	Attention BiLSTM	DT, RF, DNN, CNN, LSTM, BiLSTM, GRU, BiGRU, BERT, Hate-BERT and ToxicBERT
Fazil et al. (2023)	Twitter	English	GloVe embeddings	Attention CNN-BiLSTM	BERT-LSTM, BiLSTM, CNN, LSTM, GRU and BERT
Priyadarshini et al. (2023)	Twitter	English	GloVe embeddings	LSTM	NB & DT
Keya et al. (2023)	Websites and Platforms	Bengali	BERT embeddings	GRU	KNN, XGB, SVM, RF, LR, LSTM-BERT and AdaB-BERT
Roy et al. (2020)	Twitter	English	GloVe embeddings	DCNN	LR, RF, NB, SVM, DT, GB, KNN, CNN and LSTM
Mozafari et al. (2020)	Twitter	English	BERT embeddings	BERT-CNN	SVM, BERT-BiLSTM and BERT
Aarthi and Chelliah (2023)	Twitter	English	Semantic, contextual and syntatic	CNN-BiGRU	SVM, Attention BiLSTM and MHA-BCNN
Ayo et al. (2020)	Twitter	English	TF-IDF word features and LSTM sentence features	NN with Cuckoo search	SVM,LR, GBDT, NN and CNN
Khan et al. (2022)	Twitter	English	BERT embeddings	Attention BiLSTM	DNN, CNN, LSTM, BiLSTM GRU, DCNN and BiGRU-Capsule Network
Anezi (2022)	Social Media	Arabic	Word2vec and GloVe embeddings	BiRNN	DT, MLP, NB and LR

Table 6 Studies that employed deep learning for HS detection.

References	Data source	Language	Features	Model	Outperformed
Dascălu and Hristea (2022)	Twitter, Gab, Reddit, Fox News and Stormfront	English	RoBERTa embeddings	RoBERTa-LSTM	KNN, SVM, DT, RF, LSTM, CNN, RNN, BiRNN and CNN-GRU
Sharmila et al. (2022)	Twitter	English	Word and position embeddings	Softmax	LSVC, MNB, KNN, AdaB, RF, DT, SGD, CNN, LSTM, GRU and BiLSTM
Khan et al. (2021)	Twitter	English	Word embeddings	Sequential CNN	LR, SVM, RNN and CNN-LSTM
A. T. Kabakus (2021)	Twitter	English	Word embeddings	CNN	LSTM, GRU, BiLSTM and CNN-BiLSTM
Khan et al. (2022)	Twitter	English	GloVe embeddings	CNN-BiGRU	LSTM, CNN, GRU, BiLSTM, BiGRU and DNN
Zhang and Luo (2019)	Twitter	English	Word2vec	CNN-Skipped CNN	SVM, GB and CNN+GRU
Pronoza et al. (2021)	Vkontakte	Russian	Linguistic	RuBERT-LSTM	NB, ML Ensemble, LSTM-GRU
Duwairi et al. (2021)	Twitter	Arabic	Skip-gram word embeddings	CNN	CNN and CNN-LSTM
Alshalan and Al-Khalifa (2020)	Twitter	Arabic	Word2vec	CNN	SVM, LR, GRU, CNN-GRU and BERT
Albadi et al. (2019)	Twitter	Arabic	CBOW	GRU	SVM
Pereira-Kohatsu et al. (2019)	Twitter	Spanish	TF-IDF and token embeddings	LSTM-MLP	SVM, RF, QDA and LDA
Kar and Debbarma (2023)	YouTube	English & German	Sentiment, semantic, unigram and pattern	Diagonal GRNN	RF, LR, NB, SVM, KNN and J48graft DT
Madhu et al. (2023)	Twitter	English-Hindi[1]	BERT embeddings	SentBERT-LSTM	NB, SVM, LR, KNN, CNN and LSTM
Mundra and Mittal (2023)	YouTube	English-Hindi[1]	Word2vec and FastText	BiLSTM	LR, XGB, CNN, LSTM and mBERT
Mundra and Mittal (2022)	YouTube	English-Hindi[1]	Word2vec and FastText	BiLSTM-CNN	LR, XGB, CNN, LSTM and mBERT

[1] Code-mixed

Table 7 Studies that employed transformer models for HS detection.

References	Data source	Language	Features	Model	Outperformed
Boulouard et al. (2022)	YouTube	Arabic	Transformer Embeddings	BERT	SVM, RF, NB, LR, LSVC, LSTM, AraBERT and mBERT
Bilal et al. (2023)	Twitter	Urdu	Transformer Embeddings	RuBERT	LR, SVM, LSVM, XGB, RF, DT, KNN, LSTM, BiLSTM, Attention BiLSTM, CNN and BERT-BiLSTM
Almaliki et al. (2023)	Twitter	Arabic	Transformer Embeddings	Arabic BERT-Mini Model	LSVC, MNB, BNB, KNN, SGD, DT, RF, SVC, CNN-LSTM and LSTM
Molero et al. (2023)	Twitter, Facebook, Instagram	Spanish	Transformer Embeddings	RoBERTuito	"ML: Linear SVM, SVM, RF, AdaBoost, GB, SGD, CNN, BiLSTM, XLM-RoBERTa and BETO
Casavantes et al. (2023)	Twitter	English	Transformer embeddings and tweet metadata	BERT	SVM & GRU
Arcila-Calderón et al. (2022)	Twitter	Spanish, Greek and Italian	Transformer Embeddings	BERT	NB, MNB, BNB, LR, SGD, SVC and RNN
Benítez-Andrades et al. (2022)	Twitter	Spanish	Transformer Embeddings	BETO	CNN, LSTM, CNN-LSTM and mBERT
Toliyat et al. (2022)	Twitter	English	Transformer Embeddings	BERT	NB, LR, SVM, KNN, DT, RF, XGB, LSTM, BiLSTM and CNN
Aurpa et al. (2021)	Facebook	Bangla	Transformer Embeddings	ELECTRA Base	BERT models

Table 8 Studies that employed transformer models for HS detection.

References	Data source	Language	Features	Model	Outperformed
Pronoza et al. (2021)	Vkontakte	Russian	Transformer Embeddings	Convers-RuBERT	NB, ML and LSTM-GRU
Arshad et al. (2023)	Twitter	Urdu	Transformer Embeddings	RoBERTa	KNN, RF, NB, LR, SVM, AdaB, NBSVM, CNN, LSTM, BiLSTM, Attention BiLSTM and BiGRU
Kaminska et al. (2023)	Twitter	English	Transformer Embeddings	RoBERTa	BERT, SBERT and USE
Subramanian et al. (2022)	YouTube	Tamil	Transformer Embeddings	XLM-RoBERTa Large	BNB, SVM, KNN, LR, mBERT, XLM-RobERTa base and large and Muril large
Valle-Cano et al. (2023)	Twitter	Spanish	Transformer Embeddings & tweet and user features	HaterBERT	mBERT and BETO
Plaza-del Arco et al. (2021)	Twitter	Spanish	Transformer Embeddings	BETO	LR, SVM, CNN, LSTM, BiLSTM, mBERT and RoBERTa
Ramponi et al. (2022)	Twitter	English & Italian	Transformer Embeddings	UmBERTo	DT, MNB, LSVC, LR, BERT, mBERT and XLM-RoBERTa
Perez et al. (2023)	Twitter	Spanish	Transformer Embeddings & title of article	BETO	–
Rodriguez-Sanchez et al. (2020)	Twitter	Spanish	Transformer Embeddings	mBERT	LR, SVM and RF
Vashistha and Zubiaga (2021)	Twitter	English & Hindi	Transformer Embeddings	BERT-LSTM	LR & BERT-CNN

Table 9 Studies that employed transformer models for HS detection.

References	Data source	Language	Features	Model	Outperformed
Awal et al. (2023)	Twitter, Reddit, Facebook, News	English, Spanish, German, Hindi, Italian, Arabic, Danish, Greek and Turkish	Transformer Embeddings	XLM-RoBERTa	mBERT & XLM-RoBERTa
Bigoulaeva et al. (2023)	Twitter & Stormfront	English & German	Transformer Embeddings	mBERT	CNN & BiLSTM
Dowlagar and Mamidi (2022)	Twitter & YouTube	English-Hindi, English-Bohra Hindi, English-Kannada and English-Tamil[1]	Transformer Embeddings	RoBERTa and mBERT	SVM, CNN, Bi-LSTM, mBERT and XLM-RoBERTa
Pamungkas et al. (2021)	Twitter & Facebook	English, Spanish, Portuguese, Italian, Indonesian, German, Hindi, French and Arabic	Transformer Embeddings	mBERT	LR
Liu et al. (2023)	Twitter	English, German and Chinese	Transformer Embeddings	XLM-RoBERTa	SVM, LR, BERT and mBERT
Fan et al. (2021)	Twitter & Wikipedia	English	Transformer Embeddings	BERT	mBERT, RoBERTa and DistilBERT
Kapil et al. (2023)	Twitter, Facebook, Reddit, Youtube and Stormfront	Hindi	Transformer Embeddings	mBERT and MuRIL	CNN, BiLSTM, XLM-RoBERTa and IndicBERT
Alrashidi et al. (2023)	Twitter	Arabic	Transformer Embeddings	MARBERT	SVM, NB, LSTM, CNN, CAMeLBERT, QARiB, ArabicBERT and AraBERT
Shanmugavadivel et al. (2022)	Twitter	English-Tamil[1]	Transformer Embeddings	Adapter-BERT	LR, CNN, BiLSTM, BERT and RoBERTa
Bhardwaj et al. (2023)	Social Media	Hindi	data	HindiBERT	IndicBERT, BERT and Ensemble five BERT models

[1] Code-mixed

Table 10 Studies that employed generative models for HS detection.

References	Data source	Language	Features	Model	Outperformed
Cohen et al. (2023)	GAB	English	BT and GPT-3 rephrasing	DeBERTa	Baseline without augmentation
Su et al. (2023)	Twitter, Wikipedia and GAB	English	RoBERTa embeddings	SSL-GAN	Several BERT models

Table 11 Studies that employed multi-task learning for HS detection

References	Data source	Language	Features	Model	Outperformed
Plaza-Del-Arco et al. (2021)	Twitter	Spanish	Transformer Embeddings	BETO MTL for HS, polarity and emotion	SVM, Ensemble model and BETO
Zampieri et al. (2023)	Twitter and GAB	English	Transformer Embeddings	RoBERTa MTL for post- and token-level offensiveness and other tasks	STL models
Min et al. (2023)	Twitter	English	BERT features	NN MTL for Hate and Emotion	SVM, LSTM, BiLSTM, GRU, CNN-GRU, BERT, GPT and RoBERTa

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported in part by the European Union under Grant CERV-2021-EQUAL (101049306). However the views and opinions expressed are those of the author(s) 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.

References

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 20(8): 2243–2268 (2021) https://doi.org/10.13052/jwe1540-9589.2082

Aarthi B, Chelliah BJ (2023) HATDO: hybrid archimedes tasmanian devil optimization CNN for classifying offensive comments and non-offensive comments. Neural Comput Appl 35(25):18395–18415. https://doi.org/10.1007/s00521-023-08657-z

Agarwal S, Chowdary CR (2021) Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.115632

Akhter MP, Jiangbin Z, Naqvi IR, AbdelMajeed M, Zia T (2022) Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimed Syst 28(6):1925–1940. https://doi.org/10.1007/s00530-021-00784-8

Al-Makhadmeh Z, Tolba A (2020) Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing 102(2):501–522. https://doi.org/10.1007/s00607-019-00745-0

Albadi N, Kurdi M, Mishra S (2019) Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. Social Netw Anal Min. https://doi.org/10.1007/s13278-019-0587-5

Aljero MKA, Dimililer N (2021) A novel stacked ensemble for hate speech recognition. Appl Sci (Switzerland). https://doi.org/10.3390/app112411684

Alkomah F, Ma X (2022) A literature review of textual hate speech detection methods and datasets. Information 13(6):273

Almaliki M, Almars AM, Gad I, Atlam E-S (2023) ABMM: Arabic BERT-mini model for hate-speech detection on social media. Electronics (Switzerland). https://doi.org/10.3390/electronics12041048

Alrashidi B, Jamal A, Alkhathlan A (2023) Abusive content detection in Arabic tweets using multi-task learning and transformer-based models. Appl Sci (Switzerland). https://doi.org/10.3390/app13105825

Alshalan R, Al-Khalifa H (2020) A deep learning approach for automatic hate speech detection in the Saudi twittersphere. Appl Sci (Switzerland) 10(23):1–16. https://doi.org/10.3390/app10238614

Anezi FYA (2022) Arabic hate speech detection using deep recurrent neural networks. Appl Sci (Switzerland). https://doi.org/10.3390/app12126010

Arcila-Calderón C, Amores JJ, SÃ¡nchez-Holgado P, Blanco-Herrero D (2021) Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish. Multimodal Technologies and Interaction **5**(10) https://doi.org/10.3390/mti5100063

Arcila-Calderón C, Amores JJ, SÃ¡nchez-Holgado P, Vrysis L, Vryzas N, Oller Alonso M (2022) 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) **14**(20)https://doi.org/10.3390/su142013094

Arshad MU, Ali R, Beg MO, Shahzad W (2023) UHated: hate speech detection in Urdu language using transfer learning. Language Resourc Eval 57(2):713–732. https://doi.org/10.1007/s10579-023-09642-7

Asiri Y, Halawani HT, Alghamdi HM, Abdalaha Hamza SH, Abdel-Khalek S, Mansour RF (2022) Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification. Applied Sciences (Switzerland) **12**(16) https://doi.org/10.3390/app12168000

Aurpa TT, Sadik R, Ahmed MS (2021) Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Soc Netw Anal Min 12(1):24. https://doi.org/10.1007/s13278-021-00852-x

Awal MR, Lee RK, Tanwar E, Garg T, Chakraborty T (2023) Model-agnostic meta-learning for multilingual hate speech detection. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2023.3252401

Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA (2020) Hate speech detection in twitter using hybrid embeddings and improved cuckoo search-based neural networks. Int J Intell Comput Cybernet 13(4):485–525. https://doi.org/10.1108/IJICC-06-2020-0061

Ayo FE, Folorunso O, Ibharalu FT, Osinuga IA, Abayomi-Alli A (2021) A probabilistic clustering model for hate speech classification in twitter. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.114762

Bansal S, Garimella V, Suhane A, Patro J, Mukherjee A (2020) 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: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1018–1023. Association for Computational Linguistics, Online . https://doi.org/10.18653/v1/2020.acl-main.96

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

Benítez-Andrades JA, González-Jiménez Á, López-Brea Á, Aveleira-Mata J, Alija-Pérez J-M, García-Ordás MT (2022) Detecting racism and xenophobia using deep learning models on Twitter data: CNN. LSTM BERT PeerJ Comput Sci. https://doi.org/10.7717/PEERJ-CS.906

Bhardwaj M, Sundriyal M, Bedi M, Akhtar MS, Chakraborty T (2023) HostileNet: multilabel hostile post detection in hindi. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2023.3244014

Bigoulaeva I, Hangya V, Gurevych I, Fraser A (2023) Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. Language Resour Eval. https://doi.org/10.1007/s10579-023-09637-4

Bilal M, Khan A, Jan S, Musa S, Ali S (2023) Roman Urdu hate speech detection using transformer-based model for cyber security applications. Sensors. https://doi.org/10.3390/s23083909

Boulouard Z, Ouaissa M, Ouaissa M, Krichen M, Almutiq M, Gasmi K (2022) Detecting hateful and offensive speech in Arabic social media using transfer learning. Appl Sci (Switzerland). https://doi.org/10.3390/app122412823

Carvalho P, Caled D, Silva C, Batista F, Ribeiro R (2023) The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments. Journal of Language Aggression and Conflict. https://doi.org/10.1075/jlac.00085.car

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

Carvalho P, Guerra R (2023) D3.2/D3.3 Annotation Guidelines OHS & OCS. Technical report, Iscte-Instituto UniversitÃ¡rio de Lisboa (May)

Casavantes M, Aragón ME, Gonzá¡lez LC, Montes-y-Gómez M Leveraging posts'and authors'metadata to spot several forms of abusive comments in Twitter. Journal of Intelligent Information Systems 61(2): 519–539 (2023) https://doi.org/10.1007/s10844-023-00779-z

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

Cruz RMO, Sousa WV, Cavalcanti GDC (2022) Selecting and combining complementary feature representations and classifiers for hate speech detection. Online Soc Netw Med. https://doi.org/10.1016/j.osnem.2021.100194

Dadu T, Pant K (2020) Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In: Herbelot A, Zhu X, Palmer A, Schneider N, May J, Shutova E (eds.) Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2183–2189. International Committee for Computational Linguistics, Barcelona (online). https://doi.org/10.18653/v1/2020.semeval-1.290

Dascălu Ş, Hristea F (2022) Towards a benchmarking system for comparing automatic hate speech detection with an intelligent baseline proposal. Mathematics. https://doi.org/10.3390/math10060945

Ding Y, Zhou X, Zhang X (2019) YNU_DYX at SemEval-2019 task 5: A stacked BiGRU model based on capsule network in detection of hate. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 535–539. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2096

Dowlagar S, Mamidi R (2022) Hate speech detection on code-mixed dataset using a fusion of custom and pre-trained models with profanity vector augmentation. SN Comput Sci. https://doi.org/10.1007/s42979-022-01189-8

Duwairi R, Hayajneh A, Quwaider M (2021) A deep learning framework for automatic detection of hate speech embedded in Arabic tweets. Arab J Sci Eng 46(4):4001–4014. https://doi.org/10.1007/s13369-021-05383-3

Fan H, Du W, Dahou A, Ewees AA, Yousri D, Elaziz MA, Elsheikh AH, Abualigah L, Al-Qaness MAA (2021) Social media toxicity classification using deep learning: Real-world application UK brexit. Electronics (Switzerland). https://doi.org/10.3390/electronics10111332

Fazil M, Khan S, Albahlal BM, Alotaibi RM, Siddiqui T, Shah MA (2023) Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. IEEE Access 11:16801–16811. https://doi.org/10.1109/ACCESS.2023.3246388

García-Díaz JA, Jiménez-Zafra SM, García-Cumbreras MA, Valencia-García R (2023) Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. Complex Intell Syst 9(3):2893–2914. https://doi.org/10.1007/s40747-022-00693-x

Google: Hate speech policy (2019). https://support.google.com/youtube/answer/2801939?hl=en Accessed 2024-19-01

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

Indurthi V, Syed B, Shrivastava M, Chakravartula N, Gupta M, Varma V (2019) FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 70–74. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2009

Kamal A, Anwar T, Sejwal VK, Fazil M (2023) BiCapsHate: attention to the linguistic context of hate via bidirectional capsules and hatebase. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2023.3236527

Kaminska O, Cornelis C, Hoste V (2023) Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony. Inf Sci 625:521–535. https://doi.org/10.1016/j.ins.2023.01.054

Kapil P, Kumari G, Ekbal A, Pal S, Chatterjee A, Vinutha BN (2023) HHSD: Hindi hate speech detection leveraging multi-task learning. IEEE Access 11:101460–101473. https://doi.org/10.1109/ACCESS.2023.3312993

Kar P, Debbarma S (2023) 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. Eng Appl Artif Intell. https://doi.org/10.1016/j.engappai.2023.107143

Karayiğit H, Aci Ç, Akdağlı A (2021) Detecting abusive instagram comments in Turkish using convolutional Neural network and machine learning methods. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2021.114802

Keya AJ, Kabir MM, Shammey NJ, Mridha MF, Islam MR, Watanobe Y (2023) G-BERT: an efficient method for identifying hate speech in Bengali texts on social media. IEEE Access 11:79697–79709. https://doi.org/10.1109/ACCESS.2023.3299021

Khairy M, Mahmoud TM, Omar A, Abd El-Hafeez T (2023) Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. Language Resour Eval. https://doi.org/10.1007/s10579-023-09683-y

Khan MUS, Abbas A, Rehman A, Nawaz R (2021) HateClassify: A Service Framework for Hate Speech Identification on Social Media. IEEE Internet Computing 25(1):40–49. https://doi.org/10.1109/MIC.2020.3037034

Khan S, Fazil M, Sejwal VK, Alshara MA, Alotaibi RM, Kamal A, Baig AR (2022) BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. Journal of King Saud

University - Computer and Information Sciences 34(7):4335–4344. https://doi.org/10.1016/j.jksuci.2022.05.006
Khan S, Kamal A, Fazil M, Alshara MA, Sejwal VK, Alotaibi RM, Baig AR, Alqahtani S (2022) HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network. IEEE Access 10, 7881–7894 https://doi.org/10.1109/ACCESS.2022.3143799
Lee E, Rustam F, Washington PB, Barakaz FE, Aljedaani W, Ashraf I (2022) racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble GCR-NN model. IEEE Access 10:9717–9728. https://doi.org/10.1109/ACCESS.2022.3144266
Li Q, Peng H, Li J, Xia C, Yang R, Lichao YuS, Philip S (2022) A survey on text classification: from traditional to deep learning | acm transactions on intelligent systems and technology. ACM Trans Intel Syst Technol 13(2):1–41
Liu L, Xu D, Zhao P, Zeng DD, Hu PJ-H, Zhang Q, Luo Y, Cao Z (2023) A cross-lingual transfer learning method for online COVID-19-related hate speech detection. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2023.121031
Madhu H, Satapara S, Modha S, Mandl T, Majumder P (2023) Detecting offensive speech in conversational code-mixed dialogue on social media: a contextual dataset and benchmark experiments. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.119342
Mahajan E, Mahajan H, Kumar S (2024) EnsMulHateCyb: multilingual hate speech and cyberbully detection in online social media. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2023.121228
Markov I, Gevers I, Daelemans W (2022) An ensemble approach forÂ dutch cross-domain hate speech detection. In: Rosso P, Basile V, Martínez R, Métais E, Meziane F (eds) Natural language processing and information systems. Springer, Cham, pp 3–15
Matos BC, Santos RB, Carvalho P, Ribeiro R, Batista F (2022) Comparing Different Approaches for Detecting Hate Speech in Online Portuguese Comments. In: Cordeiro, J.a., Pereira, M.J.a., Rodrigues, N.F., Pais, S.a. (eds.) 11th Symposium on Languages, Applications and Technologies (SLATE 2022). Open Access Series in Informatics (OASIcs), vol. 104, pp. 10–11012. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2022). https://doi.org/10.4230/OASIcs.SLATE.2022.10 . https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.10
Mazari AC, Boudoukhani N, Djeffal A (2023) BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput. https://doi.org/10.1007/s10586-022-03956-x
Media Smarts: Impact of Online Hate (2021). https://mediasmarts.ca/online-hate/impact-online-hate Accessed 2023-10-12
Meta: Hate Speech (2023). https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/ Accessed 2024-19-01
Min, C., Lin, H., Li, X., Zhao, H., Lu, J., Yang, L., Xu, B.: Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. Information Fusion 96, 214–223 (2023) https://doi.org/10.1016/j.inffus.2023.03.015
Mohapatra SK, Prasad S, Bebarta DK, Das TK, Srinivasan K, Hu Y-C (2021) Automatic hate speech detection in English-Odia code mixed social media data using machine learning techniques. Appl Sci (Switzerland). https://doi.org/10.3390/app11188575
Molero JM, Perez-Martin J, Rodrigo A, Penas A (2023) Offensive language detection in Spanish social media: testing from bag-of-words to transformers models. IEEE Access 11:95639–95652. https://doi.org/10.1109/ACCESS.2023.3310244
Montejo-Ráez A, Jiménez-Zafra SM, García-Cumbreras MA, Díaz-Galiano MC SINAI-DL at SemEval-2019 task 5: Recurrent networks and data augmentation by paraphrasing. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 480–483. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2085
Mozafari M, Farahbakhsh R, Crespi N (2020) Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS ONE. https://doi.org/10.1371/journal.pone.0237861
Mridha MF, Wadud MAH, Hamid MA, Monowar MM, Abdullah-Al-Wadud M, Alamri A (2021) L-Boost: identifying offensive texts from social media post in Bengali. IEEE Access 9:164681–164699. https://doi.org/10.1109/ACCESS.2021.3134154
Mullah NS, Zainon WMNW (2023) Improving detection accuracy of politically motivated cyber-hate using heterogeneous stacked ensemble (HSE) approach. J Ambient Intell Human Comput 14(9):12179–12190. https://doi.org/10.1007/s12652-022-03763-7
Mundra S, Mittal N (2022) FA-Net: fused attention-based network for Hindi English code-mixed offensive text classification. Social Netw Anal Min. https://doi.org/10.1007/s13278-022-00929-1
Mundra S, Mittal N (2023) CMHE-AN: Code mixed hybrid embedding based attention network for aggression identification in hindi english code-mixed text. Multimedia Tools and Applications 82(8):11337–11364. https://doi.org/10.1007/s11042-022-13668-4
Muneer A, Alwadain A, Ragab MG, Alqushaibi A (2023) Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information. https://doi.org/10.3390/info14080467
Oriola O, Kotze E (2020) Evaluating machine learning techniques for detecting offensive and hate speech in south African tweets. IEEE Access 8:21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:71
Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, McKenzie JE (2021) Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372
Pamungkas EW, Basile V, Patti V (2021) A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. Inf Process Manag. https://doi.org/10.1016/j.ipm.2021.102544
Parekh B (2012) Is there a case for banning hate speech? Cambridge University Press, Cambridge
Pereira-Kohatsu JC, Quijano-SÃ¡nchez L, Liberatore F, Camacho-Collados M (2019) Detecting and monitoring hate speech in twitter. Sensors (Switzerland) **19**(21) https://doi.org/10.3390/s19214654
Perez JM, Luque FM, Zayat D, Kondratzky M, Moro A, Serrati PS, Zajac J, Miguel P, Debandi N, Gravano A, Cotik V (2023) Assessing the impact of contextual information in hate speech detection. IEEE Access 11:30575–30590. https://doi.org/10.1109/ACCESS.2023.3258973
Pitropakis N, Kokot K, Gkatzia D, Ludwiniak R, Mylonas A, Kandias M (2020) Monitoring users'behavior: anti-immigration speech detection on twitter. Mach Learn Knowledge Extract 2(3):192–215. https://doi.org/10.3390/make2030011
Pitsilis GK, Ramampiaro H, Langseth H (2018) Effective hate-speech detection in twitter data using recurrent neural networks. Appl Intell 48(12):4730–4742. https://doi.org/10.1007/s10489-018-1242-y

Plaza-Del-Arco FM, Molina-Gonzalez MD, Urena-Lopez LA, Martin-Valdivia MT (2021) A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access 9, 112478–112489 https://doi.org/10.1109/ACCESS.2021.3103697

Plaza-Del-Arco F-M, Molina-GonzÃ¡lez MD, UreÃ±a-LÃ³pez LA, MartÃn-Valdivia MT (2020) Detecting misogyny and xenophobia in spanish tweets using language technologies. ACM Trans Internet Technol**20**(2)https://doi.org/10.1145/3369869

Plaza-del-Arco FM, Molina-González MD, Ureña-López LA, Martín-Valdivia MT (2021) Comparing pre-trained language models for Spanish hate speech detection. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.114120

Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V (2021) Resources and benchmark corpora for hate speech detection: a systematic review. Language Resour Eval 55(2):477–523

Priyadarshini I, Sahu S, Kumar R (2023) A transfer learning approach for detecting offensive and hate speech on social media platforms. Multimedi Tools Appl 82(18):27473–27499. https://doi.org/10.1007/s11042-023-14481-3

Pronoza E, Panicheva P, Koltsova O, Rosso P (2021) Detecting ethnicity-targeted hate speech in Russian social media texts. Inf Process Manage. https://doi.org/10.1016/j.ipm.2021.102674

Pérez JM, Luque FM (2019) Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In: May J, Shutova E, Herbelot A, Zhu X, Apidianaki M, Mohammad SM (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 64–69. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2008

Rajalakshmi R, Selvaraj S, Faerie Mattins R, Vasudevan P, Anand Kumar M (2023) HOTTEST: hate and offensive content identification in tamil using transformers and enhanced stemming. Comput Speech Language. https://doi.org/10.1016/j.csl.2022.101464

Ramponi A, Testa B, Tonelli S, Jezek E (2022) Addressing religious hate online: from taxonomy creation to automated detection. PeerJ Comput Sci. https://doi.org/10.7717/PEERJ-CS.1128

Raut R, Spezzano F (2023) Enhancing hate speech detection with user characteristics. Int J Data Sci Anal. https://doi.org/10.1007/s41060-023-00437-1

Rodriguez-Sanchez F, Carrillo-De-Albornoz J, Plaza L (2020) Automatic classification of sexism in social networks: an empirical study on twitter data. IEEE Access 8:219563–219576. https://doi.org/10.1109/ACCESS.2020.3042604

Roy PK, Bhawal S, Subalalitha CN (2022) Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Computer Speech Language. https://doi.org/10.1016/j.csl.2022.101386

Roy SS, Roy A, Samui P, Gandomi M, Gandomi AH (2023) Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2023.3260217

Roy PK, Tripathy AK, Das TK, Gao X-Z (2020) A framework for hate speech detection using deep convolutional neural network. IEEE Access 8:204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073

Saeed R, Afzal H, Rauf SA, Iltaf N (2023) Detection of offensive language and ITS severity for low resource language. ACM Trans Asian Low-Resour Language Inf Process. https://doi.org/10.1145/3580476

Santos RB, Matos BC, Carvalho P, Batista F, Ribeiro R (2022) Semi-supervised annotation of portuguese hate speech across social media domains. In: Cordeiro J, Pereira MJ, Rodrigues NF, Pais S (eds.) 11th symposium on languages, applications and technologies (SLATE 2022). Open Access Series in Informatics (OASIcs), vol. 104, pp. 11–11114. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany . https://doi.org/10.4230/OASIcs.SLATE.2022.11 . https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.11

Shanmugavadivel K, Sathishkumar VE, Raja S, Lingaiah TB, Neelakandan S, Subramanian M (2022) Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Sci Rep 12(1):21557. https://doi.org/10.1038/s41598-022-26092-3

Shannaq F, Hammo B, Faris H, Castillo-Valdivieso PA (2022) Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. IEEE Access 10:75018–75039. https://doi.org/10.1109/ACCESS.2022.3190960

Sharmila P, Anbananthen KSM, Chelliah D, Parthasarathy S, Kannan S (2022) PDHS: pattern-based deep hate speech detection with improved tweet representation. IEEE Access 10:105366–105376. https://doi.org/10.1109/ACCESS.2022.3210177

Siegel AA (2020). In: Persily N, Tucker JAE (eds) Online hate speech. Cambridge University, SSRC Anxieties of Democracy. Cambridge University Press

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

Statista: Number of social media users worldwide from 2017 to 2027 (2023). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Su X, Li Y, Branco P, Inkpen D (2023) SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media. Natural Language Engineering. https://doi.org/10.1017/S1351324923000396

Subramanian M, Ponnusamy R, Benhur S, Shanmugavadivel K, Ganesan A, Ravi D, Shanmugasundaram GK, Priyadharshini R, Chakravarthi BR (2022) Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. Comput Speech Language 76:101404. https://doi.org/10.1016/j.csl.2022.101404

Toliyat A, Levitan SI, Peng Z, Etemadpour R (2022) Asian hate speech detection on Twitter during COVID-19. Frontiers Artif Intell. https://doi.org/10.3389/frai.2022.932381

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

Turki T, Roy SS (2022) Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer. Appl Sci (Switzerland) **12**(13) https://doi.org/10.3390/app12136611

Twitter: Hateful Conduct (2023). https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy Accessed 2024-19-01

Tynes BM, Giang MT, Williams DR, Thompson GN (2008) Online racial discrimination and psychological adjustment among adolescents. J Adol Health Official Public Soc Adolesc Med 43(6):565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

United Nations: United Nations Strategy and Plan of Action on Hate Speech. Technical report, United Nations (2019). https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

Valle-Cano GD, Quijano-SÃ¡nchez L, Liberatore F, GÃ³mez J (2023) SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles. Expert Syst Appl **216**[SPACE]https://doi.org/10.1016/j.eswa.2022.119446
Vanetik N, Mimoun E (2022) Detection of racist language in French tweets. Information (Switzerland). https://doi.org/10.3390/info13070318
Vashistha N, Zubiaga A (2021) eOnline multilingual hate speech detection: experimenting with hindi and english social media. Information (Switzerland) 12(1):1–16. https://doi.org/10.3390/info12010005
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is All you Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., Long Beach, CA. https://doi.org/10.48550/arXiv.1706.03762
Wang S, Liu J, Ouyang X, Sun Y (2020) Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In: Herbelot A, Zhu X, Palmer A, Schneider N, May J, Shutova E (eds.) Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1448–1455. International Committee for Computational Linguistics, Barcelona (online). https://doi.org/10.18653/v1/2020.semeval-1.189
Watanabe H, Bouazizi M, Ohtsuki T (2018) Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access 6, 13825–13835 https://doi.org/10.1109/ACCESS.2018.2806394 . Conference Name: IEEE Access. Accessed 2023-10-11
Wiedemann G, Yimam SM, Biemann C (2020) UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In: Herbelot A, Zhu X, Palmer A, Schneider N, May J, Shutova E (eds.) Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1638–1644. International Committee for Computational Linguistics, Barcelona (online). https://doi.org/10.18653/v1/2020.semeval-1.213
Wiegand M, Siegel M, Ruppenhofer J (2018) 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), pp. 1–10. International Committee for Computational Linguistics, Viena (online) . https://www.lsv.uni-saarland.de/wpcontent/publications/2018/germeval2018_wiegand.pdf
Yin W, Zubiaga A (2021) Towards generalisable hate speech detection: a review on obstacles and solutions. Queen Mary University of London 7. https://doi.org/10.7717/peerj-cs.598
Zampieri M, Ranasinghe T, Sarkar D, Ororbia A (2023) Offensive language identification with multi-task learning. Journal of Intelligent Information Systems 60(3):613–630. https://doi.org/10.1007/s10844-023-00787-z
Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç.: SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., Shutova, E. (eds.) Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1425–1447. International Committee for Computational Linguistics, Barcelona (online) (2020). https://doi.org/10.18653/v1/2020.semeval-1.188
Zhang Z, Luo L (2019) Hate speech detection: A solved problem? The challenging case of long tail on Twitter. Semantic Web 10(5):925–945. https://doi.org/10.3233/SW-180338
Zhang M, He J, Ji T, Lu C-T (2024) Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection . https://arxiv.org/abs/2402.11406
Zhou Y, Yang Y, Liu H, Liu X, Savage N (2020) Deep learning based fusion approach for hate speech detection. IEEE Access 8:128923–128929. https://doi.org/10.1109/ACCESS.2020.3009244
Zhou X, Yong Y, Fan X, Ren G, Song Y, Diao Y, Yang L, Lin H (2021) Hate speech detection based on sentiment knowledge sharing. In: Annual Meeting of the Association for Computational Linguistics . https://api.semanticscholar.org/CorpusID:236459847
kNOwHATE: kNOwHATE (2023). https://knowhate.eu/pt-pt/ Accessed 2024-05-01

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.