Journal of Big Data

SURVEY Open Access

Detecting hate in diversity: a survey of multilingual code-mixed image and video analysis

Hafiz Muhammad Raza Ur Rehman[1†], Mahpara Saleem[2†], Muhammad Zeeshan Jhandir[2*], Eduardo Silva Alvarado[3,4,5], Helena Garay[3,6,7] and Imran Ashraf[1*]

†Hafiz Muhammad Raza Ur Rehman and Mahpara Saleem contributed equally to this work.

*Correspondence: zeeshan.jhandir@iub.edu.pk; imranashraf92@gmail.com

[1] Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea
[2] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[3] Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain
[4] Universidad Internacional Iberoamericana, 24560 Campeche, México
[5] Fundación Universitaria Internacional de Colombia Bogotá, Bogotá, Colombia
[6] Universidade Internacional do Cuanza, Cuito, Bié, Angola
[7] Universidad de La Romana, La Romana, República Dominicana

Abstract

The proliferation of damaging content on social media in today's digital environment has increased the need for efficient hate speech identification systems. A thorough examination of hate speech detection methods in a variety of settings, such as code-mixed, multilingual, visual, audio, and textual scenarios, is presented in this paper. Unlike previous research focusing on single modalities, our study thoroughly examines hate speech identification across multiple forms. We classify the numerous types of hate speech, showing how it appears on different platforms and emphasizing the unique difficulties in multi-modal and multilingual settings. We fill research gaps by assessing a variety of methods, including deep learning, machine learning, and natural language processing, especially for complicated data like code-mixed and cross-lingual text. Additionally, we offer key technique comparisons, suggesting future research avenues that prioritize multi-modal analysis and ethical data handling, while acknowledging its benefits and drawbacks. This study attempts to promote scholarly research and real-world applications on social media platforms by acting as an essential resource for improving hate speech identification across various data sources.

Keywords: Hate speech detection, Social media, Feature engineering, Deep learning, Multimodal data analysis

Introduction

In the current digital era, social media (SM) allows us to instantly interact with individuals around the world and share our ideas with them. People like to spend their time on SM platforms since they cater to each person's demands [1, 2]. The rapid growth of SM users has facilitated the easy sharing of opinions across various platforms. The popularity of these platforms is attributed, in part, to the diverse array of information formats they offer, including audio, video, and images [3]. However, hate speech and false information are greatly disseminated through these platforms [4]. Nowadays, individuals seek connections through SM platforms like Facebook, Twitter, and community forums, which provide them the ability to freely express

themselves and share information [5–8]. Several approaches have been taken to counter hate speech, which is a category of offensive material [9]. The hatred expressed on the internet as hate speech [6, 10–15] aggression [16], abusive language [17, 18], cyberbullying [19–21], sexism [22–24], racism [25, 26], radicalization [27], discrimination [28] and flaming [29]. Hate speech is outlined as abusive speech that targets particular group traits, such as 'religion', 'ethnicity', or 'gender' [10]. Any offense that is motivated, in part or whole, by bias toward an attribute of a group of people is referred to as hate speech [30]. Hate speech is defined as words used to disparage, degrade, or offend members of a certain group or to convey hatred for that group [31]. Vulnerability, inherent in both physical and social aspects of humanity [32], signifies the innate susceptibility to harm. Further, specific attributes or affiliations are linked to potential harm, such as physical and social vulnerability, lack of capacity, and association with marginalized groups. Table 1 depicts a thorough comparison of different hate speech concepts with related ideas.

Regarding these explanations, we define hate speech as follows, 'Hate speech is a type of language used to express feelings of enmity or aggression toward a group of people or a specific individual based on attributes like race, nationality, gender, religion, or ethnicity'. Researchers have focused on major platforms like Facebook, Twitter, etc. [43]. Table 2 contains the definitions of Hate Speech on various SM platforms. People in multilingual nations do not always convey their thoughts in just one language. Owing to limited English proficiency, numerous SM users incorporate words from their local/native languages in Roman script, alongside English, to express their messages. This linguistic blend is commonly referred to as code-mixed text [44]. When words, phrases, or morphemes from one language are incorporated into a spoken or written expression in a different language, this is known as code-mixing [45]. While significant efforts have been directed toward identifying hate speech within textual data, the exploration of hate-speech detection in videos and images has been comparatively limited [46–48].

The primary aim of this work is to conduct a comprehensive analysis of hate speech identification in textual, audio, and visual formats is carried out in this paper, taking into account the difficulties associated with code-mixing and multilingual environments. We discuss the difficulties these elements present and how they affect the ability to identify hate speech. Our work fills gaps in the literature and offers a useful resource for aspiring researchers by covering all facets of hate speech, including detection in videos and social media, paving the way for further significant research in this vital area.

Motivation

The spread of hate speech in the modern SM environment is a primary social problem that calls for effective detection and mitigation systems. The field is still fragmented and lacks a thorough synthesis of methodology and conclusions, even with the combined efforts of researchers. This gap will be filled by a survey report that addresses hate-speech detection in SM combines existing research, evaluates approaches, and clarifies ethical implications. The text's objective is to improve the effectiveness and moral integrity of hate speech mitigation initiatives in online spaces by promoting comprehensive knowledge and directing future research paths.

Table 1 Comparing the definition of hate speech with related ideas

"Concept"	Concept's definition	Distinction/similarity from hate-speech
Abusive language	The phrase 'abusive language' refers to harsh communication, which includes profanity, hate speech, and other offensive phrases [17]	Hate speech is a type of Abusive language
Cyberbullying	Cyberbullying, which is frequently referred to as the online equivalent of traditional bullying, comprises aggressive and harassing behaviors directed against someone who may be unable to effectively defend themselves [33]	Hate speech is an offensive language that is specifically targeted at a distinctive, uncontrollable characteristic of a group of individuals
Discrimination	Discrimination against specific groups of individuals is a result of hate speech, which also threatens equality. Typically, immigrants and women are the major targets [28]	Hate speech embodies an aggressive type of discrimination
Flaming	Flaming refers to comments that are hostile, offensive, and threatening that can annoy and offend other forum users, often known as trolls [29]	Hate speech, as opposed to flame, can occur in any situation
Profanity	Offensive or derogatory phrase or term [34]	Hate speech may involve the use of profanity, but it is not an absolute requirement
Extremism	The ideology associated with extremists or hate organizations often advocate violence and seek to divide populations meanwhile attempting to restore their perceived status. This typically involves portraying outgroups as either perpetrators or inferior groups [35]	Hate speech is a common feature in extremist discourses
Radicalization	Radicalization and hate-speech are terms that are frequently used interchangeably. Some authors connect religiously motivated hate speech and radicalization. Radical organizations were referred to as 'cyber extremists by Wadhwa and Bhatia [36]	Radical discourses, similar to extremism, can incorporate hate speech. However, in radical discourses, subjects such as war, religion, and negative emotions are frequently addressed, while hate speech may adopt a more discreet approach, often rooted in stereotypes [37]
Racism	Tribalism, regionalism, xenophobia (particularly toward migrant labor), nativism (hostility toward immigrants and refugees), and any prejudice towards a certain tribe or region are all forms of racial offense [38]	Racism and hate speech are related concepts, but they differ in focus and scope
Sexism	Sexism is the term for prejudice or discrimination that is directed mostly towards women [39]	While sexism can impact anyone, its primary impact is disproportionately felt by women and girls
Religious hate-speech	In nations with the greatest rates of social crime, religious hate-speech is regarded as a reason for crime [40]	Religious hate-speech is a subset of hate-speech that specifically targets individuals or groups based on their religious beliefs or affiliations
Toxic language	Toxic language is an impolite, disrespectful, or irrational comment that has the potential to make a person leave a discussion [41]	While not all offensive remarks necessarily include hate speech, certain instances of hate speech can provoke more extensive discussions among people
Hybrid hate speech	This kind of hate speech does not focus on just one type of target. Instead, it includes expressions of hatred that can affect many different groups at the same time. For example, it could involve harassment related to religion that harms both Hindus and Muslims, without singling out one group specifically [42]	It is a fusion of multiple types of prejudice or bias within a single expression or communication

Table 2 Definitions of hate speech on social media platforms' documentation

Source	Definition
Twitter (April 2023) (https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy)	"You are prohibited to attack individuals based on their race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability"
Facebook (May 26, 2023) (https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/)	"We see hate-speech as when someone directly hurts others not ideas or groups because of things like race, where they're from, disabilities, religion, caste, gender, or serious illness"
YouTube (June 5, 2019) (https://support.google.com/youtube/answer/2801939?hl=en)	"On YouTube, we do not permit hate speech. We take down content that encourages violence or spreads hate towards people or groups due to various factors. If you come across content that goes against this rule, please report it."
Instagram (Feb 11, 2021) (https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram)	"Hate-speech, according to our standards, involves directly targeting individuals, not ideas or organizations due to specific qualities we consider protected: race, national origin, disability, religion, caste, sexual orientation, gender, gender identity, and serious illness"

Survey methodology

The approach of gathering pertinent contributions from the computer science literature is described in this section, with emphasis on the identification and evaluation of Hate Speech.

Search process

The search process, following Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines, as illustrated in Fig. 1, began with the identification of 1500 papers from well-known databases and publishers, including Web of Science, Semantic Scholar, and the ACL. Using keywords like "hate speech detection," "classification," "survey, review," "multimodal," "audio," "video," "code-mixed," and "multilingual," we conducted an exhaustive search to capture a comprehensive range of studies.

After removing duplicates, 800 papers were excluded, leaving 900 unique papers. Screening based on abstracts and titles led to the removal of 400 papers deemed irrelevant. This resulted in 300 papers being considered for further review based on inclusion criteria. Following a detailed assessment, 175 papers were excluded based on our exclusion criteria, resulting in a final selection of 125 papers suitable for inclusion in our survey on hate-speech detection in social media.

Inclusion criteria

Inclusion criteria encompassed papers in various languages, with a primary emphasis on hate speech and its categories. This covered reviews, survey articles, proposed methods, methodologies, as well as comparative and evaluation studies. Included datasets were in English and other languages such as Arabic, and Dutch, covering diverse forms of hate content like images, text, videos, multilingual, and code-mixed data.

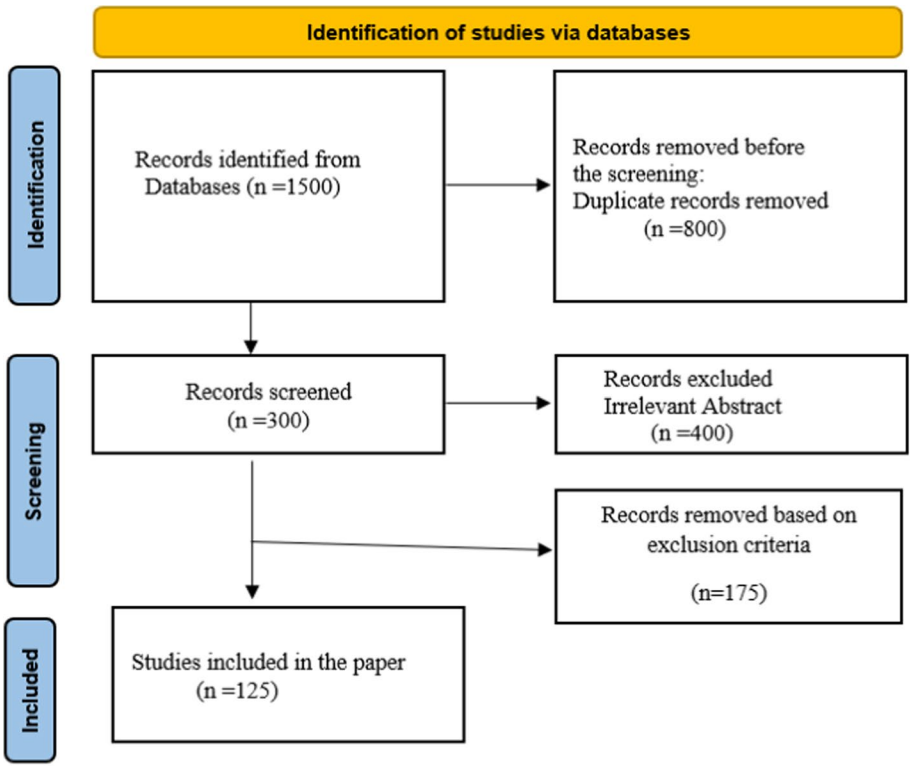

Fig. 1 PRISMA method to streamline article screening

Exclusion criteria

In our literature survey on hate-speech detection in SM, we have established exclusion criteria to refine the review's scope. One vital criterion is excluding studies solely focusing on hate-speech detection outside SM platforms, such as forums or private messaging systems. This decision aims to maintain the relevance of included studies to the specific context of S.M. Additionally, we are excluding studies with narrow domain focus unrelated to hate-speech detection in SM, such as those centered on specific online communities. This ensures the broad applicability of surveyed studies. Furthermore, we will exclude duplicate publications or studies substantially overlapping with others already included, enhancing the survey's comprehensiveness. Additionally, studies not reporting approaches to address hate speech-related issues, including those not focusing on Machine Learning techniques, NLP (Natural Language Processing), or performance evaluations, are also excluded, along with those not focusing on hate speech at all.

Structure of the paper

The Introduction of the paper outlines the importance of hate speech detection and explains the survey methodology used to select relevant studies. It describes how studies were included or excluded based on certain criteria, such as their focus on text, multilingual, and multimodal detection methods, and provides a roadmap for the rest of the paper.

In the section titled Hate Speech Detection Procedure and Prior Surveys, the paper reviews existing methods, including detection in audio and video content. It also

discusses the data collection and preprocessing steps for creating datasets used in hate speech detection. A review of prior surveys highlights trends and gaps in the field, especially regarding the challenges of detecting hate speech in code-mixed data, and the complexities of working with multilingual and multimodal data. A comparison of detection methods across these contexts is also provided.

The Proposed Methodology introduces a new framework for improving hate speech detection by integrating cross-lingual models and multimodal data fusion. This approach aims to improve detection accuracy, especially for social media content that includes both text and visual elements.

The Practical Applications of Hate Speech Detection section discusses how these detection methods are applied in real-world scenarios, such as social media monitoring and content moderation. Case studies are used to illustrate both the successes and challenges of current systems.

In the section on Challenges, Solutions, and Future Directions, the paper explores ongoing challenges, such as ethical issues and the limitations of current methods. It also suggests solutions, including new AI techniques, and looks at future research directions to improve hate speech detection in multilingual and multimodal settings.

The paper concludes by summarizing key findings, stressing the need for continued research, and offering a call for improved methods, datasets, and evaluation metrics in the field of hate speech detection.

Hate speech detection procedure and prior surveys on hate speech detection

Hate speech identification in audio and video

The workflow for detecting hate speech in this study is multifaceted, incorporating audio feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC) and Chroma Vectors. These techniques are instrumental in capturing the subtle tones and rhythms that may indicate the presence of hate speech [47, 48]. Research has demonstrated that combining text, audio, and visual elements through multimodal learning enhances detection accuracy. This approach allows the system to leverage context from multiple sources, making it better equipped to handle diverse forms of hate speech, including the nuanced cues found in tone and facial expressions [48].

Emotion recognition techniques play a crucial role in this process by identifying sentiments like anger or disgust that may arise in audio and video content. Such emotional cues are often overlooked in traditional text-only analyses, yet they provide valuable insights into the intent behind the words spoken or depicted [6, 48]. Furthermore, improving transcription accuracy is vital for effectively detecting hate speech in spoken content. Misclassifications can occur due to transcription errors, so implementing noise reduction techniques and customizing Automatic Speech Recognition (ASR) models for colloquial or accented speech is essential [48, 49].

To foster transparency, explainable AI methods are utilized in hate speech detection models, clarifying the rationale behind classifying certain instances as hate speech. This is particularly beneficial for ambiguous cases, helping to differentiate between hate speech and expressions of humor [48]. Moreover, utilizing diverse datasets that encompass various languages, accents, and dialects contributes to the model's generalization abilities, ultimately making it more robust and less biased [48, 50].

A dedicated hate-speech detection system is developed with an emphasis on spoken content within videos. The system follows these steps [47]:

i. Extract audio from the video.
ii. Convert the audio into text format.
iii. Train machine learning models using text-based features to
iv. classify videos as either normal or containing hate speech.

Data collection

Researchers employ web scraping techniques, ensuring they adhere to platform terms of service, to access social media data through APIs. The creation of datasets, which involves gathering diverse data from social media posts, is crucial. Data is annotated manually or automatically to indicate the presence of hate speech, all while ethical considerations guide adherence to privacy regulations and standards [6, 51, 52].

Data preprocessing

Data preprocessing poses significant challenges due to the inherent noise within the data. To prepare the data for analysis without losing critical information, several preprocessing techniques are applied. The initial phase involves collecting information primarily from social media platforms like Facebook, Twitter, YouTube, and Instagram. Once the data is collected, it undergoes various preprocessing steps, including removing stop words, stemming, and lemmatization, to ensure it is in a usable format [49].

Dataset for detecting hate speech

Before delving into the strategies and methods used for detecting hate speech in social media comments, we conducted a comprehensive survey of the datasets commonly utilized by research communities to develop and evaluate their models. The relevant datasets pertaining to hate speech are detailed in Table 3.

Review of prior surveys

Significant progress has been accomplished in the recognition of hate speech through various methodologies, including NLP, machine learning models, deep learning architectures, language models, and other associated techniques [39]. Current survey and review articles are presented in this section, placing a strong emphasis on the noteworthy contributions made by these studies [4, 38, 68–78].

In accordance with the current understanding, the initial survey on hate speech detection conducted by [77] provides a short and comprehensive overview of the primary areas that have been investigated to automatically recognize hate speech in online content, mainly focusing on NLP. Our survey, however, extends beyond English-only content to include multilingual and code-mixed data, enhancing the scope across languages. The second study [71] emphasizes the complexity of hate speech by discussing nuances such as the use of humor and subtle forms of discrimination. While insightful, it lacks our comprehensive multimedia approach that includes both text and video data. The study [69] carried out a thorough review of the literature

Table 3 Datasets used in hate speech detection

References	Year	Source	Language
[25]	2016	Twitter	English
[6]	2016	Twitter	English
[31]	2017	Twitter	English
[53]	2017	Twitter	English
[54]	2018	Twitter	English
[55]	2018	Instagram	Indonesian
[56]	2018	Twitter	Italian
[57]	2019	Twitter	English, Spanish
[58]	2019	Twitter	Portuguese
[59]	2019	Twitter	Indonesian
[60]	2019	Social media comments	English
[61]	2020	Twitter	Arabic
[62]	2020	Twitter	English, German, Spanish, French, Greek
[63]	2021	Amazon	English
[64]	2021	Hatebase Twitter	English
[65]	2021	Twitter	Spanish-English
[66]	2022	Facebook	English
[67]	2023	Twitter	English

on methods and strategies for detecting hate speech, concentrating on eight methods that are extensively employed in automatic hate detection: dictionary searches, BoW (bag of words), N-grams, TF-IDF (term frequency-inverse document frequency), sentiment analysis, part of speech, rule-based approaches, and template-based approaches. Our survey extends this approach by incorporating machine learning and deep learning techniques across multiple data types.

To the best of our knowledge, the initial survey on the topic of multilingual hate speech detection was conducted by [38]. This study investigates the detection of hate speech in social networks using a multilingual corpus, primarily focusing on the Arabic language. Our survey builds on this by including additional languages, code-mixed data, and multimedia content such as video. The study by [79] primarily focuses on the use of machine learning models for hate speech detection, although it notes inefficiencies in real-time prediction accuracy. Our survey addresses real-time detection challenges and integrates a multilingual, multimedia perspective to provide a more robust approach. Further, the study by [80] focuses on language-specific studies in Asian languages for automated hate-speech detection. Our survey expands this by including a wider array of language families and non-text data types such as images and video. The study by [73] points out limitations related to small and unreliable hate-speech datasets. Our survey emphasizes dataset credibility and suggests standardized, multilingual datasets for improved reliability. The work by [81] is focused solely on models that address explicit hate speech detection. This survey includes both explicit and implicit forms, alongside multimedia formats, for a more comprehensive view. Lastly, [82] overlooks cultural nuances and language-specific

communication styles necessary for effective detection. Our survey, in contrast, integrates these cultural nuances and includes code-mixed data to enhance detection accuracy.

In contrast to previous surveys, our goal is to conduct a comprehensive and inclusive investigation that encompasses all aspects of hate speech. This involves examining text, video, multilingual content, and code-mixed expressions on social media platforms. While earlier surveys have focused primarily on either textual or visual elements, they often overlook the code-mixed and multilingual dimensions. This survey aims to provide a comprehensive analysis that considers all these dimensions within a single study. A thorough comparative analysis of existing surveys is given in Table 4.

Detecting hate speech in code-mixed data

Mixed-language text is commonly utilized on the SM platform [84]. Numerous studies have focused on detecting hate, offensive, and aggressive content in English tweets. However, a tiny amount of research has been accomplished in recognizing hate speech in Hindi-English code-mixed tweets, primarily due to language

Table 4 Comparative analysis of existing surveys

Title	Year	Type of data	Technique	Research focus
[77]	2017	Text	NLP	Focuses on English-only content, while our survey includes multilingual and code-mixed data for broader analysis across languages
[83]	2018	Video	DL	Analyzes 3D human data for video-based hate speech detection, while our survey integrates text, images, and code-mixed data as well
[38]	2019	Text (Multilingua)	ML, DL	Focus on Arabic-language data and multilingual aspects in text, whereas our study includes a wider variety of languages, code-mixed data, and multimedia
[69]	2019	Text	NLP	Emphasis on NLP techniques only, while our survey extends this by including ML and DL techniques across multiple data types
[79]	2020	Text	ML	Primarily focuses on ML models with noted inefficiencies in real-time predictions, whereas we address real-time detection alongside a multilingual, multimedia perspective
[80]	2021	Text	NLP	Focuses on language-specific studies in Asian languages, while our examination expands to more language families and includes non-text data types such as images and video
[73]	2022	Text	ML	Points out limitations in small and unreliable datasets, whereas our study emphasizes dataset credibility and recommends standardized, multilingual datasets
[81]	2023	Text	ML, DL	Focused on explicit hate speech models, whereas our survey also includes implicit forms and multimedia formats for a more comprehensive view
[82]	2023	Text	Traditional learning, ML, DL	Overlooks cultural nuances and language-specific communication styles, while our survey integrates these cultural nuances and includes code-mixed data
Current survey	2024	Text, Images, Video, audio Multilingual, Code-mixed	NLP, ML, DL	A comprehensive survey covering text, video, multilingual, and code-mixed aspects of hate speech on social media

challenges and a scarcity of suitable datasets [85]. Code-mixing, or the blending of words and phrases from multiple languages within a single text, is prevalent in multilingual communities and is particularly common in social media interactions. This phenomenon complicates hate speech detection, as the sentiment or intensity of certain expressions can change dramatically when a different language is introduced. For instance, embedding a Hindi word in an English sentence can amplify or alter the sentiment of the phrase in ways that monolingual models might fail to detect. There is a scarcity of research conducted in the domain of code-mixed language, specifically with references to studies [45, 84–91].

The study [90] conducted the inaugural investigation on hate-speech detection in Hindi-English tweets. The authors gathered 4575 tweets, annotated by two linguists and validated through Cohen's Kappa coefficient for inter-annotation agreement. The data underwent preprocessing, extracting features like punctuation and emoticon counts, characters and words N-grams, and word2vec of lexicon words. To handle the feature-rich matrix, they employed dimensionality reduction using chi-square. Classification was performed using SVM and RF, achieving accuracies of 71.7% and 66.7%, respectively.

The authors [85] collected and classified 10,000 code-mixed data samples into non-hate and hate categories. They utilized the FastText word embedding from Facebook to represent the data and employed an SVM with RBF (radial basis function) for classification. The proposed method was compared with other representation methodologies, such as word2vec and doc2vec, highlighting fastText's superior performance in the classification task. The study also found that character-level features are particularly effective for code-mixed data, suggesting that linguistic nuances at the sub-word level are essential for capturing the specific meaning of hate speech.

The authors in [45] carried out the second study on this dataset, conducting two sets of experiments using DL models. These experiments involved a sub-word level LSTM model and an attention-based hierarchical LSTM model, focusing on phonemic sub-words. The sub-word level LSTM model achieved an accuracy of 69.8%, while the attention-based hierarchical-LSTM model on phonemic sub-words achieved an accuracy of 66.6%. The study also included an experimental comparison of their model's performance against existing works.

[86] focused on the categorization of objectionable tweets in code-mixed Hindi-English. A novel dataset comprising 3000 data points underwent manual annotation, classifying tweets into categories like 'Abusive,' 'Non-offensive,' and 'Hate-inducing.' The authors adopted a transfer learning approach, utilizing CNN and LSTM models as the architecture. Initially, the models were trained on offensive English tweets, and the weights were extracted. Subsequently, the model underwent further training on Hinglish tweets. Using techniques like transfer learning enables better performance on code-mixed data, as the models become capable of understanding embedded sentiment or hatefulness specific to Hindi-English interactions. This capability is crucial for effective hate speech detection in multicultural settings where multiple languages are often blended. Various word embedding techniques, including Glove, fastText, and Twitter word2vec, were employed, along with additional inputs like LIWC and sentiment score as features (Table 5).

Table 5 Examples of code-mixing affecting sentiment and hate speech intensity [92]

Original sentence	Code-mixed sentence	Effect of code-mixing	Hate intensity scale (0–2)	Intensity level
"I received a nice gift today"	"I received a nice gift today, *meri pyaari maa*"	The phrase "*meri pyaari maa*" (meaning "my dear mother" in Hindi) adds affection, suggesting a deeper emotional connection	0 (No hate)	0 (No hate)
"This politician is a complete failure"	"This politician is a complete *bekaar*"	The Hindi word "*bekaar*" (meaning "useless") intensifies the insult, adding a stronger negative tone specific to Hindi speakers	1 (Mild hate)	1 (Mild hate)
"You are such a terrible person"	"you are such a *ghyta* person"	The word "*ghyta*" (meaning "useless" or "worthless" in Urdu) heightens the insult, conveying a strong negative sentiment	2 (High hate)	2 (High hate)

Real-life use cases of hate speech detection

Hate speech detection has a large and diverse range of applications and can potentially be used in several industries as well. Here are just a few use-case examples.

News industry

People comment on various articles published by writers on the internet. Hate speech detection has become essential in the news industry for regulating online interactions to safeguard users and promote healthy dialogue. AI-based tools have been used by news platforms like "*The Guardian*" and "*The New York Times*" to screen user comments, detect offensive language, and uphold ethical standards [93]. These techniques are also used by media outlets to control politically sensitive conversations and shield reporters from offensive content, particularly when covering elections [94]. These developments in hate speech detection highlight how important it is for the news industry.

E-commerce organizations

In order to keep users secure and welcome, hate speech identification is essential for e-commerce companies. Customer reviews, product feedback, and discussion forums on e-commerce platforms like "Amazon" and "eBay" might occasionally contain offensive language or discriminatory sentiments [95]. These platforms ensure that user-generated content complies with community standards by using AI-driven hate speech identification to foster trust between buyers and sellers. Moreover, e-commerce platforms can enhance user experiences and protect brand reputation by employing real-time automatic moderation techniques to identify and remove inappropriate words [96].

Education

In the field of education, where creating a welcoming and courteous learning environment is crucial, hate speech detection has several applications. Toxic or hateful language is frequently encountered in online learning environments, discussion boards, and virtual classrooms, which can impede the learning process and drive students away. Teachers can ensure polite and productive discourse through the utilization of hate speech identification tools to control conversations. Moreover, research uses these technologies to examine student behavior and create treatments that foster empathy and digital citizenship in students [97]. In the digital age, these initiatives help create learning environments that are safer and more encouraging.

Social media platforms

Social media platforms have been increasingly relying on hate speech detection technologies to foster safer online environments. Social media sites such as Facebook, Twitter, and YouTube employ automated algorithms to identify and eliminate hate speech, guaranteeing adherence to legal and community standards [98]. In particular, Twitter employs machine learning algorithms to detect abusive tweets and punish violators by suspending or banning their accounts [94]. Similarly, Facebook uses human moderation and AI-powered technologies to combat hate speech in several languages, particularly during political events and crises [99]. These methods have been essential for preventing the spread of extremist beliefs and minimizing the psychological damage brought on by exposure to harmful content, even beyond moderation [31].

Online gaming platforms

It is essential to detect hate speech on online gaming platforms like Twitch, Xbox Live, and PlayStation Network. Hate speech and other toxic conduct are common in gaming communities, which can negatively impact the user experience [100]. In order to automatically identify instances of racism, sexism, and hate speech in chat, Twitch employs machine learning algorithms. Players who participate in such behavior are either warned or banned. "Practical applications of hate speech detection methods" section provides a detailed overview of the cutting-edge algorithms utilized for hate speech detection in real-world applications. Table 6 shows the text concerning real-life text examples of multilingual hate speech.

Multi-lingual and multi-modal Hate-speech

Exploration of the multi-lingual dimension of hate speech is a recently emerging research area. The study [28] addresses challenges in identifying hate speech across multiple languages. It proposes an efficient framework for identifying hate speech in low-resource languages, analyzing data from 16 sources in 9 languages. The authors recommend models effective in high-resource settings, emphasizing the use of translation and multilingual BERT for enhanced detection. The authors in [101] employed a Twitter hate-speech corpus encompassing five languages, annotated with demographic information. Their study utilizing this dataset explores the demographic bias

Table 6 Real-life examples of multilingual hate speech [57]

ID	Tweet text (original)	Tweet text (English translation)	Hate speech	Target type	Aggressiveness
32411	Callate @vikidonda y la gran puta madre que te repario. Que le diste a la politica... nada. Basura	Shut up @vikidonda you motherfucker. What did you do for politics... nothing. Trash	1	Individual	Aggressive
5823	Women are equal and deserve respect. Just kidding, they should suck my dick	Same	1	Generic	Aggressive
1890	Sick barstewards! This is what happens when we put up the refugees welcome signs! They not only rape our wives but our mothers too!	Same	1	Generic	Aggressive
33033	@RyanAbe This is inhumane Karma is a bitch she'll get around these brainless heartless assholes!	Same	0	Generic	Non-aggressive
33119	Soy un sudaca haciendo sudokus	I am a "sudaca" (slur for South American) doing sudokus	1	Generic	Non-aggressive
945	@EmmanuelMacron Hello?? Stop groping my nation. Migrant crisis is a long-prepared plan to alter Europe's identity	Same	0	Generic	Non-aggressive

present in hate-speech classification. The author [4, 102] conducted an exhaustive review encompassing the domains of multi-modal and multi-lingual automated hate-speech detection. Table 7 shows a comprehensive analysis of research papers focusing on the classification of hate speech with respect to multiple languages.

Comparison of hate speech detection methods across multilingual and code-mixed contexts

Evaluation metrics

Evaluation metrics serve as tools to evaluate the effectiveness of a model, system, or process in diverse domains such as ML data and DL. These metrics offer both quantitative and qualitative perspectives on the performance of a system, aiding in model comparison, informed decision-making, and overall performance enhancement. Various performance metrics are utilized for evaluating the usefulness of the created classifier. Below is a brief discussion on some typical performance measures used in text categorization [79].

Precision

Precision, alternatively termed the positive predicted value, represents the ratio of true positive predictions to the total predicted positives.

Table 7 Comparison of hate speech detection methods across multilingual and code-mixed contexts

Method	Strengths	Weaknesses	Effectiveness in multilingual contexts	Effectiveness in code-mixed contexts
Dictionary-based	Easy to implement; good for detecting explicit hate speech using predefined terms and slurs	Cannot adapt to new slang or contextually nuanced language; high rate of false negatives for subtle hate speech	Moderate—Requires dictionaries in multiple languages	Low—Code-mixing complicates matching due to mixed-language tokens
Bag of words (BoW)	Simplistic model, good for short text analysis	Ignores word order and context, leading to potential misclassification; computationally inefficient at large scales	Moderate—Limited accuracy in languages with different syntactic structures	Low—Loses meaning when processing code-mixed language without context
N-grams	Better contextual representation than BoW by capturing adjacent word sequences	Computationally heavy at higher N values; struggles with non-contiguous dependencies and context beyond N-grams	Moderate—Accuracy improves in certain language structures	Low—Code-mixing disrupts the N-gram sequence, leading to inaccuracies
Word embeddings	Captures semantic relationships, allowing for synonym recognition and nuanced language understanding	Requires large training datasets in each language; embeddings may vary significantly across languages	High—Multilingual embeddings (e.g., FastText, BERT) improve detection	Moderate—Struggles with hybrid expressions unless embeddings are language-agnostic
TF-IDF	Effective for identifying significant terms within specific hate speech contexts	Lacks context sensitivity, and can struggle with language-specific frequency patterns	Moderate—Requires training on large multilingual datasets	Low—Difficulty distinguishing relevant terms in blended code-mixed language
Machine learning (ML)	Adaptable to various features, allowing for detection in multilingual and code-mixed settings with feature engineering	Requires extensive labeled data; high computational costs; risks bias without language diversity in training data	High—Multilingual ML models like SVM and RF are widely used	Moderate—Code-mixed adaptation depends heavily on engineered features
Deep learning (DL)	High accuracy with large datasets; models like CNNs and LSTMs excel in capturing context and complex relationships	Computationally expensive; often considered black-box models with limited interpretability	High—Multilingual DL models (e.g., multilingual BERT, mBERT) perform well	Moderate—Code-mixed performance varies; LSTM with attention can improve results

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{1}$$

Recall

Recall is the fraction of true positive predictions among the predicted positive instances. For NLP-based model performance evaluation, it has many advantages like [103];

- Identification of Relevant Information: It ensures that the model correctly identifies positive instances, such as relevant documents or sentiment-bearing sentences.
- Performance Benchmarking: It provides a baseline for comparing different models or algorithms in terms of their ability to detect relevant information in text data.
- Usefulness in Binary Classification: It helps in understanding how well a model can distinguish between positive and negative cases.

It is calculated as follows.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{2}$$

However, in the context of NLP, it has some limitations.

- Imbalance: In cases of imbalanced classes, the recall may not provide a complete picture of model performance.
- Context Dependency: NLP tasks often involve understanding language in context, which can be complex and ambiguous. Recall may not capture the nuances of contextual understanding, leading to misinterpretations.

F1-Score

The F1-Score, which is the harmonic mean of precision and recall, assigns equal significance to both precision and recall.

$$F1\text{-}Score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \tag{3}$$

Accuracy

Accuracy represents the count of accurately classified instances, including both true positives and true negatives.

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total Population}} \tag{4}$$

For a more exhaustive description of performance metrics, the readers can delve into a detailed examination of the various measurements used to evaluate the efficiency of a system, process, or activity [118].

Table 8 shows the various hate speech models with different feature selection and evaluation matrices. It illustrates that hate-speech detection has been conducted with

Table 8 A comparison of methods with feature selection and evaluation metrics for detecting hate speech

References	Model	Evaluation metrics	Feature extraction
Oriola and Kotze [1]	ML	Precision, Recall, F1	N-Gram
Djuric et al. [10]	NLP		BOW
Ombui [104]	ML, DL	Precision, Recall, F1, Acc	W.E, TF-IDF
Nugroho et al. [105]	ML, DL	Precision, Recall, F1, Prediction	
Zhang and Luo [106]	DL	Precision, Recall, F1	W.E
Al-Ibrahim et al. [107]	ML, DL	F1, Acc	
Corazza et al. [102]	DL		N-Gram, W.E
Warner and Hirschberg [12]	NLP, ML	Precision, Recall, Acc	N-Gram
Liu and Forss [108]	NLP		N-Gram, TF-IDF
Pawar et al. [109]	ML		TF-IDF
Aljero and Dimililer [110]	ML	Precision, Recall, F1, Acc	TF-IDF
Kwok and Wang [14]	ML	Acc	BOW, N-Gram
Li et al. [111]	NLP, DL	Precision, Recall, F1	
Sharma et al. [15]	NLP, ML	Acc, KPP	
Kumar et al. [112]	DL	Acc	
Nobata et al. [113]	ML	Acc	TF-IDF
Dinakar et al. [33]	NLP, ML	F1, Acc	TF-IDF
Wachs et al. [20]	NLP		
Watanabe et al. [2]	NLP, ML	Precision, Recall, F1, Acc	N-Gram
Hosseinmardi et al. [21]	NLP	KPP, Prediction	BOW
Chatzakou et al. [22]	NLP	Prediction	
Waseem and Hovy [6]	NLP	Precision, Recall, F1, KPP	N-Gram
Aluru et al. [28]	NLP, DL	Acc, Prediction	BOW
Guermazi et al. [29]	NLP, ML	Acc	
Kamal et al. [11]	NLP, DL	Precision, Recall, F1	W.E
Roy et al. [3]	DL	Precision, Recall, F1	N-Gram, TF-IDF
Saumya et al. [114]	NLP, DL	F1, Acc	W.E
DelVigna et al. [34]	ML, DL	Precision, Recall, F1, Acc	N-Gram
Albadi et al. [40]	NLP, ML, DL	Precision, Recall, F1, Acc, AUC	N-Gram
Wu et al. [115]	NLP, ML	Precision, Recall, F1, Acc, AUC	N-Gram, TF-IDF
Boishakhi [48]	NLP, ML	Precision, Recall, F1, Acc	TF-IDF
Badjatiya et al. [51]	NLP, DL	Precision, Recall, F1, Acc	BOW, N-Gram, TF-IDF
Unsvag and Gambäck [49]	NLP	Precision, Recall, F1	N-Gram
Masadeh et al. [50]	ML, DL	Precision, Recall, F1, Acc	BOW
Sanguinetti et al. [56]	NLP	Acc, Prediction	
Waseem and Hovy [6]	NLP	Precision, Recall, F1	N-Gram
Fortuna and Nunes [71]	NLP, DL	F1, Prediction	BOW
Ibrohim and Budi [59]	NLP, ML	Acc	N-Gram
Tang and Dalzel [60]	NLP, ML	Acc, Prediction	TF-IDF
Qureshi and Sabih [64]	NLP, ML, DL	F1, Acc, KPP, AUC	N-Gram
Rodriguez et al. [66]	NLP, ML	Precision, Recall, F1	TF-IDF
Salminen et al. [7]	ML, DL	Precision, Recall, F1	BOW, TF-IDF
Abro et al. [79]	NLP, ML	Precision, Recall, F1	N-Gram, TF-IDF
DeAlcantara et al. [9]	DL	Precision, Recall, F1, KPP, AUC	N-Gram, W.E
Faris et al. [116]	DL	Precision, Recall, F1	W.E
Mundra and Mittal[117]	NLP, DL	Acc	N-Gram

different learning models while using different evaluation metrics and feature extraction techniques.

Proposed framework for hate speech detection

We propose a novel approach that combines multilingual text analysis with multimodal data fusion to solve the difficulties of hate speech identification in multilingual and multimodal situations. This framework integrates vision-language models like the Contrastive Language-Image Pretraining model, which is intended to analyze both text and images, with cross-lingual representations from sophisticated language models (such as multilingual BERT or XLM-R). Transfer learning between languages is made possible by the framework's initial processing of text in different languages using shared semantic regions. In order to align visual signals with textual information, a crucial step in comprehending context and intent in social media content, it then applies multimodal fusion algorithms to image and video data. Important issues like code-switching, linguistic variety, and the ambiguity of visual material are addressed by this integrated approach. By addressing important issues like code-switching, ambiguity in visual content, and language variety, this integrated approach improves detection accuracy and generalizability across many media forms and languages. This approach provides a more thorough way to identify hate speech in the varied and ever-changing landscapes of contemporary digital platforms by combining text and visual analysis. The process and interactions between the components are made clearer with the help of a graphic representation of this framework.

Practical applications of hate speech detection methods

Detecting hate speech is essential to fostering a polite and safe atmosphere in online forums, where a lot of conversation takes place every day. Harmful language can increase online aggression, strengthen social differences, and cause people and groups to become estranged from one another. Several cutting-edge techniques have been created and implemented to identify and stop hate speech in real-time, ensuring that platforms follow their community norms and policies, in an effort to lessen these effects. RoBERTa [119], cross-platform hate speech detection [120], and the hate speech automated recognition and evaluation (HARE) [121] framework are some of the best methods used in the existing literature for hate speech detection. These algorithms help to identify hate speech in a more efficient manner. Another important aspect of these models is scalability which is an essential trait needed for practical applications of hate speech detection methods. For completeness, RoBERTA, cross-platform hate speech detection, and HARE algorithms are described here.

Algorithm 1 RoBERTa model for hate speech detection [119].

Input: Image and text data collected from social media platforms.
Parameters:

- Data Preprocessing: Cleaning and normalizing the data to ensure consistency.
- Model Selection: Choosing the Swin Transformer for image data and RoBERTa for text data.

i. Collect image and text data from social media platforms.
ii. Preprocess data:
    (a) Clean and normalize text data (remove special characters, lowercasing).
    (b) Resize and normalize images to a consistent format.
iii. Use the Swin Transformer model to extract visual embeddings from images.
iv. Use the RoBERTa model to extract text embeddings from tweets.
v. Combine visual embeddings and text embeddings into a unified feature set for analysis.
vi. Classify the unified feature set as hate speech or non-hate speech using a classification model.
vii. Evaluate model performance using metrics such as accuracy and F1-score.

Algorithm 2 Cross-platform hate speech detection [120].

Input: Text data from multiple social media platforms (YouTube, Twitter, Gab) in English and German.
Parameters:

- Platform Diversity: Ensure data collection spans different platforms for a comprehensive dataset.
- Data Volume: Aim for a substantial amount of data to train the model effectively.

i. Collect data from multiple platforms (YouTube, Twitter, Gab) in English and German.
ii. Preprocess the text data from the platforms to ensure consistency.
iii. Combine datasets from different platforms to increase data volume and diversity.
iv. Train a classification model on the combined dataset to detect hate speech.
v. Evaluate model performance with F1-scores for both English and German comments.
vi. Analyze results to understand the effect of cross-platform data on hate speech detection.

Cross-platform hate speech detection helps detect hateful speech from different languages including English and German. Text data from multiple social media platforms like YouTube, Twitter, Gab, etc. has been used with this algorithm.

Algorithm 3 HARE framework for hate speech detection [121].

Input: Diverse datasets with human annotations for training the hate speech detection model.
Parameters:

- Annotation Quality: Ensure the datasets are annotated accurately by humans for effective learning.
- Contextual Information: Utilize additional contextual data to improve classification accuracy.

i. Gather diverse datasets with human annotations for hate speech detection.
ii. Use big language models to enhance reasoning capabilities in hate speech classification tasks.
iii. Incorporate contextual information through free-text comments to provide additional context for the model.
iv. Train the model on benchmarks like SBIC and Implicit Hate to evaluate its performance.
v. Evaluate the explainability and flexibility of the model across different benchmarks.
vi. Refine the model based on evaluation feedback and results from benchmarking.

Table 9 illustrates that hate-speech detection has been conducted in various languages, with English being the most commonly studied language. Additionally, Twitter datasets have been frequently employed as primary data sources for hate-speech research. Furthermore, SVM emerges as the predominant algorithm of choice for hate-speech detection in these studies.

Challenges, solutions, and future directions in hate speech detection

Hate speech detection has faced many challenges throughout its development. Based on previous work here are some key challenges encountered in hate speech detection.

Complexity of hate-speech detection/recognition

Hate-speech detection involves a more complex process than mere keyword spotting [113]. There is no standard definition available to define hate that is universally accepted due to different cultures [109, 128]. Machines might have difficulty correctly recognizing such content given the lack of human agreement on how to label hate speech [14]. Successful completion of the task necessitates a deep understanding of culture and societal frameworks [129]. Furthermore, implicit hate speech, which can be conveyed through insinuation or subtlety, poses additional challenges for detection systems. This type of hate speech often relies heavily on context, making it difficult for algorithms to identify without comprehensive contextual analysis.

Solution: Implement models that incorporate contextual understanding, such as Transformer-based models like BERT, to capture subtle nuances in implicit hate speech. Additionally, incorporating cultural context through labeled datasets specific to different regions can improve model accuracy in detecting culturally sensitive hate speech.

Linguistic complexity and contextual considerations

Hate speech frequently has a high level of verbal fluency and grammatical accuracy despite its hateful nature. It can flow effortlessly across sentence boundaries, and sarcasm is frequently used in it [113]. In light of recent events, studies have also focused on identifying hate speech related to the COVID-19 pandemic [110, 111]. Implicit hate speech relies significantly on the surrounding context. Without a thorough grasp of the context, distinguishing between harmful and harmless statements can be a difficult task [130].

Table 9 A comprehensive analysis of research papers focusing on the classification of hate-speech in literature (2012–2023)

References	Year	Citations	Feature	Classifier	Dataset	Language	Evaluation (accuracy, precision, recall, F1)
[12]	2012	776	Template-based strategy	SVM	Yahoo and American Jewish Congress		0.68, 0.6, 0.63, –
[122]	2012	749	Lexical and syntactic	Rule-based	YouTube	English	0.98, 0.94, –, –
[14]	2013	498	N-gram	Naïve Bayes	Twitter	English	–, –, –, –
[108]	2014	41	TF-IDF, sentiment Analysis, N-grams, topic Similarity	Naïve Bayes	Web pages	English	0.97, 0.82, – , –
[13]	2015	126	N-gram, type Dependencies	DT, RF, SVM	Twitter	English	0.89, 0.69, 0.95, –
[123]	2015	485	Rule-based approach, sentiment analysis typed dependencies	Non-supervised	Web pages	English	0.65, 0.64, 0.65, –
[37]	2015	164	Linguistic, Term Frequency	KNN, SVM	Twitter	English	–, –, 0.83, –
[124]	2015	76	Profile and tweet-based features, bag of words, N-gram, TF-IDF	Naïve Bayes	Twitter	Arabic	0.85, 0.85, 0.85, –
[125]	2016	109	Dictionaries	SVM	Facebook	Dutch	0.49, 0.43, 0.46, –
[6]	2016	1635	User-features	Logistic Regression	Twitter	English	0.72, 0.77, 0.73, –
[51]	2017	1258	Char n-grams, TF-IDF, BOW	SVM, GBDT, DNN, CNN, LR, RF	Twitter	English	0.93, 0.93, 0.93, –
[34]	2017	432	Pos, sentiment polarity	LSTM'SVM	Facebook	English	0.833, 0.872, 0.851, –
[2]	2018	348	Pattern-based, unigrams, sentiment feature	SVM	Twitter	English	0.88, 0.87, 0.87, –
[126]	2019	123	Word n-grams, semantic sequence	LSTM, GRU	Facebook	Amharic	–, –, 0.97, 0.92
[127]	2019	705	Hierarchical annotation schema	SVM, CNN	(OLID)Tweet	English	–, –, –, –

Moreover, the challenge of code-switching, where speakers alternate between languages within a single conversation, can further complicate the identification of hate speech. Traditional models may struggle to recognize hate speech patterns when mixed languages are involved, leading to higher rates of false negatives.

Solution: To handle code-switching, consider using multilingual embeddings and specialized models trained on code-switched datasets. Additionally, leveraging

sarcasm and sentiment detection algorithms can improve the recognition of context-dependent hate speech.

Dataset challenges

The utilization of self-generated datasets by most authors poses a challenge in terms of assessing the credibility of these datasets and the reliability of the results achieved through them [82]. Problems with datasets arise when most of the data comes from a single user, which can make the latest technology seem better than it actually is. Moreover, detecting hate speech is tricky because it's hard for both people and computers to identify [131]. Difficulties arise in achieving domain generalization due to the inability to secure a dataset that thoroughly encompasses all aspects of hate speech across numerous SM platforms [132].

Additionally, the lack of representative datasets for minority languages and dialects limits the ability of models to detect hate speech effectively in those contexts. The underrepresentation of certain communities in training data can lead to biased outcomes and reinforce existing stereotypes.

Solution: Expand dataset diversity by including data from multiple social media platforms and minority languages. Efforts should be made to collect balanced datasets that represent a wide array of languages, dialects, and user demographics to improve the model's fairness and generalization.

Multilingualism and political discrimination

Achieving domain generalization and detecting hate content across various social media platforms pose significant challenges. This part explores the complexities of detecting hate speech in code-mixed social media posts, particularly in languages like Hindi-English [85]. In order to detect hate speech, it can be difficult to define political discrimination. Researching political discrimination has a number of challenges, such as understanding what constitutes discrimination, potential risks, and issues with freedom of speech [133].

Furthermore, language-specific idioms and culturally specific references can create hurdles in understanding and categorizing hate speech accurately across different languages. This necessitates a nuanced approach that considers local contexts and expressions of hate.

Solution: Develop localized hate speech detection models that understand culturally specific idioms and expressions. Employ techniques such as translation-based approaches and use pre-trained multilingual models to aid in detecting hate speech in multiple languages.

Code-mixed data and ML model challenges

We also highlighted challenges uncovered during our survey. Given the predominant use of supervised learning algorithms, there is a clear necessity to incorporate semi-supervised or unsupervised methods to enhance the robustness of the analysis. Code-mixed data can present unique challenges for NLP and understanding due to the blending of different linguistic elements. Moreover, for the accurate detection of hate speech in code-mixed data, the creation of appropriate datasets, including code-mixed languages

like Urdu-English, is imperative. Therefore, as more people use multiple languages, we need datasets that can handle the challenges of this linguistic diversity. Continuous monitoring and updates to the model are crucial to address evolving hate-speech patterns effectively.

The presence of code-mixing in online content presents unique challenges for hate speech detection systems. As illustrated in Table 5, code-mixed language can obscure or amplify sentiments depending on the cultural and linguistic context, which often leads to misinterpretation by standard hate speech detection models. Traditional machine learning models, trained primarily on monolingual data, may miss or misclassify the nuances introduced by code-mixed phrases, making it necessary to develop specialized models for these linguistic contexts.

Solution: Develop and use code-mixed datasets that reflect linguistic diversity, including examples of blended languages. Experiment with unsupervised and semi-supervised models to address cases where labeled data is scarce, and periodically update models to adapt to new patterns in hate speech.

Global hate speech

The global problem of hate speech necessitates a quick response. Different regions target specific communities; for instance, Asians use hurtful words like "mullah" against Muslims, and Europe directs derogatory terms at Africans. This challenge needs careful attention because people are singled out for who they are. This is unjust. Muslims and Africans are not the only targets—indigenous groups in South America and certain religious communities in the Middle East also face this. We must act decisively. Starting with understanding the targeted groups and their reasons can make a difference. Promoting respect and comprehension is key. Governments, communities, and individuals must unite against hate speech. Together, we can build a secure and inclusive world.

Enhancing hate speech detection

To boost the precision of hate-speech detection, a future direction involves implementing Multi-modal Analysis, which integrates text, audio, and visual data, considering factors like voice tone, gestures, and facial expressions. Creating comprehensive multilingual hate-speech datasets that encompass diverse languages and cultures is essential for in-depth research. Researchers should account for cultural differences in the linguistic expressions of anger and hate when constructing online hate speech models. For enhancing the identification of hate-speech instances in code-mixed data, annotation methods should be refined, taking into consideration the complexities of mixed languages. Addressing the scarcity of balanced datasets for online hate speech is a significant challenge.

Addressing data challenges

Expediting hate-speech detection and analysis can be achieved by harnessing unlabeled data for unsupervised Machine Learning models, as the data labeling process is time-consuming. Furthermore, there is a need for further exploration and study of Deep-learning models to advance hate-speech research. In addition to the above, developing effective strategies for handling Disproportionated datasets, for example utilizing

advanced sampling approaches, emphasizing Cost-Sensitive learning, and embracing ensemble techniques, can lead to enhanced model performance and more precise hate-speech classification.

Enhancing model interpretability

It's critical to find suitable techniques for integrating non-textual elements, like images or emojis, into hate-speech detection models to adapt to the evolving nature of hate speech and enhance a whole classification system's effectiveness. Additionally, investigating techniques to enhance the interpretability of Deep-learning models, for instance, incorporating attention-based mechanisms or employing explainable AI approaches, is pivotal for providing valuable insights into these models' decision-making processes, which can enhance trust and promote their adoption in real-world applications.

Conclusions

In this survey, we extensively examined recent developments in text-based hate speech detection systems, addressing related topics such as cyberbullying, abusive language, discrimination, sexism, extremism, and radicalization on social media platforms. While several surveys have addressed hate speech detection, previous studies often failed to provide a comprehensive overview of the latest advancements. To fill a significant gap in the literature, this survey specifically focuses on publications that deal with code-mixed data, where multiple languages are prevalent in online comments. We investigated both deep learning and machine learning models for hate speech detection, including feature extraction methods, experimental datasets, and performance evaluation metrics. Our findings reveal that the Support Vector Machine (SVM) algorithm remains one of the most widely used models, with TF-IDF features being prevalent in previous research.

Notably, there is a shift occurring in the research community towards the increased use of deep learning models for hate speech detection. Techniques that combine various word embedding methods with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have gained popularity. Our survey indicates that deep learning models consistently outperform traditional machine learning models, such as SVM, Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) when utilizing word2Vec, GloVe, and FastText. Moreover, ensemble deep learning models often yield even better results, marking a significant trend in the field.

Despite these advancements, several challenges persist that require coordinated efforts from the research community. Detecting hate speech is complex, especially in code-mixed data, necessitating a deep understanding of social and cultural contexts. The linguistic complexity and contextual nuances further complicate this task. Although large datasets are available, preprocessing and labeling them remain labor-intensive endeavors. Furthermore, concerns regarding the credibility of existing datasets limit the applicability of proposed frameworks for hate speech detection. Multilingualism and political bias present additional challenges in detecting hate speech on social media. Additionally, the opaque nature of machine learning and deep learning models, often referred to as "black boxes", restricts transparency in decision-making processes. There is an urgent need for explainable artificial intelligence approaches to enhance the reliability of these models.

Overall, this survey synthesizes recent developments in hate speech detection, highlighting the significance of addressing various forms, including code-mixed data, multilingual settings, and social context. "We also emphasize the necessity for transparency in machine learning processes to improve the reliability of these models. Our findings pave the way for more effective and equitable hate speech detection methods in critical areas.

Author contributions
HMRUR conceptualization, data curation, and writing—the original manuscript. MS formal analysis, data curation, and methodology. ZJ methodology, software, project administration. ESA funding acquisition, investigation, and visualization. HG validation, investigation, and project administration. IA supervision, validation, and writing—review and editing. All authors reviewed and approved the manuscript.

Funding
This research was funded by the European University of Atlantic.

Availability of data and materials
The dataset can be requested from the authors.

Declarations

Competing interests
The authors declare no competing interests.

Received: 2 June 2024 Accepted: 18 April 2025
Published online: 03 May 2025

References

1. Oriola O, Kotze E. Evaluating machine learning techniques for detecting offensive and hate speech in South African Tweets. IEEE Access. 2020;8:21496–509. https://doi.org/10.1109/ACCESS.2020.2968173.
2. Watanabe H, Bouazizi M, Ohtsuki T. Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access. 2018;6:13825–35. https://doi.org/10.1109/ACCESS.2018.2806394.
3. Roy PK, Tripathy AK, Das TK, Gao X-Z. A framework for hate speech detection using deep convolutional neural network. IEEE Access. 2020;8:204951–62. https://doi.org/10.1109/ACCESS.2020.3037073.
4. Chhabra A, Vishwakarma DK. A literature survey on multimodal and multilingual automatic hate speech identification. Multimed Syst. 2023;29:1203–30. https://doi.org/10.1007/s00530-023-01051-8.
5. Kumar A, Tyagi V, Das S. Deep, learning for hate speech detection in social media. In: IEEE 4th international conference on computing power and communication technologies GUCON. 2021;2021(1–4):2021. https://doi.org/10.1109/GUCON50781.2021.9573687.
6. Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: HLT-NAACL 2016–2016 conference of the North American chapter association computational linguistics: human language technologies proceedings of the student Research work. 2016. p. 88–93. https://doi.org/10.18653/v1/n16-2013.
7. Salminen J, et al. Developing an online hate classifier for multiple social media platforms. Human centric Comput Inf Sci. 2020;10:1–34. https://doi.org/10.1186/s13673-019-0205-6.
8. Tyagi V, Kumar A, Das S. Sentiment Analysis on Twitter Data Using Deep Learning approach. In: Proceedings of the—IEEE 2020 2nd international conference on advances in computing, communication. Control networking, ICACCCN 2020;2020. p. 187–90. https://doi.org/10.1109/ICACCCN51052.2020.9362853
9. de Alcântara CS, Feijó D, Moreira VP. Offensive video detection: dataset and baseline results. In: Lr. 2020—12th international conference on language resources and evaluation conference of proceedings; 2020. p. 4309–19.
10. Djuric N, et al. Hate speech detection with comment embeddings; 2015. p. 29–30. https://doi.org/10.1145/2740908.2742760arXiv:1405.4053.
11. Kamal A, Anwar T, Sejwal VK, Fazil M. Bicapshate: attention to the linguistic context of hate via bidirectional capsules and hatebase. IEEE Trans Comput Soc Syst. 2024;11:1781–92. https://doi.org/10.1109/TCSS.2023.3236527.
12. Warner W, Hirschberg J. Detecting hate speech on the world wide web. In: Proceeding LSM '12 Proc Second Work Lang Soc Media; 2012. p. 19–26.
13. Burnap P, Williams ML. Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. Policy Internet. 2015;7:223–42. https://doi.org/10.1002/poi3.85.
14. Kwok I, Wang Y. Locate the hate: Detecting tweets against blacks. In: Proceedings of the 27th AAAI conference on artificial intelligence AAAI 2013; 2013. p. 1621–2. https://doi.org/10.1609/aaai.v27i1.8539.
15. Sharma S, Agrawal S, Shrivastava M. Degree based classification of harmful speech using Twitter data. arXiv:1806.04197v1.

16. Kumar S, Hamilton WL, Leskovec J, Jurafsky D. Community interaction and conflict on the web. In: Web conference 2018—proceedings of the world wide web conference WWW 2018; 2018. p. 933–43. https://doi.org/10.1145/3178876.3186141. arXiv:1803.03697.
17. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. In: 25th international world wide web conference WWW 2016; 2016. p. 145–53. https://doi.org/10.1145/2872427.2883062.
18. Kaur S, Singh S, Kaushal S. Abusive content detection in online user-generated data: a survey. Procedia CIRP. 2021;189:274–81. https://doi.org/10.1016/j.procs.2021.05.098.
19. Ptaszynski M, et al. Expert-annotated dataset to study cyberbullying in polish language. Data. 2024. https://doi.org/10.3390/data9010001.
20. Wachs S, Wright MF, Vazsonyi AT. Understanding the overlap between cyberbullying and cyberhate perpetration: moderating effects of toxic online disinhibition. Crim Behav Ment Heal. 2019;29:179–88. https://doi.org/10.1002/cbm.2116.
21. Hosseinmardi H, et al. Analyzing labeled cyberbullying incidents on the Instagram social network. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015;9471:49–66. https://doi.org/10.1007/978-3-319-27433-1_4.
22. Chatzakou D, et al. Measuring #Gamergate: a tale of hate, sexism, and bullying. In: 26th international world wide web conference 2017, WWW 2017 Companion; 2017. p. 1285–90. https://doi.org/10.1145/3041021.3053890arXiv:1702.07784.
23. Singh K, Vajrobol V, Aggarwal N. Iic_team@multimodal hate speech event detection 2023: detection of hate speech and targets using xlm-roberta-base. In: CASE; 2023.
24. Jha A, Mamidi R. W17-2902; 2017. p. 7–16.
25. Waseem Z. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In: NLP + CSS 2016—EMNLP 2016 workshop on natural language processing and computational social science proceedings of work; 2016. p. 138–42. https://doi.org/10.18653/v1/w16-5618.
26. Maruf A, et al. Hate speech detection in the Bengali language: a comprehensive survey. J Big Data. 2024;11:97.
27. Vogel I, Meghana M. Profiling hate speech spreaders on twitter: Svm vs. bi-lstm; 2021.
28. Aluru SS, Mathew B, Saha P, Mukherjee A. Deep learning models for multilingual hate speech detection; 2020. p. 1–16. arXiv:2004.06465v3.
29. Guermazi R, Hammami M, Hamadou AB. Using a semi-automatic keyword dictionary for improving violent web site filtering. In: Proceedings of the international conference on signal image technologies internet based system SITIS 2007; 2007. p. 337–44. https://doi.org/10.1109/SITIS.2007.137.
30. Silva L, Mondal M, Correa D, Benevenuto F, Weber I. Analyzing the targets of hate in online social media. In: Proceedings of the 10th international conference web social media, ICWSM 2016; 2016. p. 687–90. https://doi.org/10.1609/icwsm.v10i1.14811. arXiv:1603.07709.
31. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th international conference on web social media, ICWSM 2017; 2017. p. 512–5. https://doi.org/10.1609/icwsm.v11i1.14955. arXiv:1703.04009.
32. Rogers W, Mackenzie C, Dodds S. Why bioethics needs a concept of vulnerability. JFAB Int J Fem I Approaches Bioeth. 2012;5:11–38. https://doi.org/10.3138/ijfab.5.2.11.
33. Dinakar K, Jones B, Havasi C, Lieberman H, Picard R. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst. 2012;2:1–30. https://doi.org/10.1145/2362394.2362400.
34. Del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. Hate me, hate me not: hate speech detection on Facebook. CEUR Workshop Proc. 2017;1816:86–95.
35. McNamee LG, Peterson BL, Peña J. A call to educate, participate, invoke and indict: understanding the communication of online hate groups. Commun Monogr. 2010;77:257–80. https://doi.org/10.1080/03637751003758227.
36. Wadhwa P, Bhatia MPS. Tracking on-line radicalization using investigative data mining. In: 2013 National confernce on communication; 2013. p. 1–5. https://doi.org/10.1109/NCC.2013.6488046.
37. Agarwal S, Sureka A. Using KNN and SVM based one-class classifier for detecting online radicalization on twitter. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015;8956:431–42. https://doi.org/10.1007/978-3-319-14977-6_47.
38. Al-Hassan A, Al-Dossari H. Detection of hate speech in social networks: a survey on multilingual corpus; 2019. p. 83–100. https://doi.org/10.5121/csit.2019.90208.
39. Parihar AS, Thapa S, Mishra S. Hate speech detection using natural language processing: applications and challenges. In: Proceedings of the 5th international conference on trends electronics informatics, ICOEI 2021; 2021. p. 1302–8. https://doi.org/10.1109/ICOEI51242.2021.9452882 (Institute of Electrical and Electronics Engineers Inc., 2021).
40. Albadi N, Kurdi M, Mishra S. Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. In: Proceedings of the 2018 IEEE/ACM international conference advance social networks analysis mining, ASONAM 2018; 2018. p. 69–76. https://doi.org/10.1109/ASONAM.2018.8508247.
41. JIGSAW. Perspective API; 2017. https://www.perspectiveapi.com. Accessed 22 Feb 2020.
42. Chetty N, Alathur S. Hate speech review in the context of online social networks. Aggress Violent Behav. 2018;40:108–18. https://doi.org/10.1016/j.avb.2018.05.003.
43. Reis VD. A survey of machine learning based techniques for hate speech detection on Twitter Uma pesquisa sobre técnicas de aprendizado de máquina para detecção de discurso de ódio no Twitter; 2023. p. 3605–24. https://doi.org/10.54033/cadpedv20n8-030.
44. Madhu H, Satapara S, Modha S, Mandl T, Majumder P. Detecting offensive speech in conversational code-mixed dialogue on social media: a contextual dataset and benchmark experiments. Expert Syst Appl. 2023;215: 119342. https://doi.org/10.1016/j.eswa.2022.119342.
45. Santosh TY, Aravind KV. Hate speech detection in Hindi-English code-mixed social media text. In: ACM international conference proceeding series; 2019. p. 310–3. https://doi.org/10.1145/3297001.3297048.
46. Lippe P, et al. A multimodal framework for the detection of hateful memes; 2020. arXiv:2012.12871.

47. Wu CS, Bhandary U. Detection of hate speech in videos using machine learning. In: Proceedings—2020 international conference on computing science computing intelligence CSCI 2020; 2020. p. 585–90. https://doi.org/10.1109/CSCI51800.2020.00104.
48. Boishakhi FT. Detection Multi-modal Hate Speech, using Machine Learning. In: IEEE international conference Big Data (Big Data); 2021. p. 4496–9. https://doi.org/10.1109/BigData52589.2021.9671955.
49. Unsvåg EF, Gambäck B. The effects of user features on Twitter hate speech detection. In: 2nd Workshop on Abusive Language Online. Proceedings of the Workshop, co-located with EMNLP 2018; 2018. p. 75–85. https://doi.org/10.18653/v1/w18-5110.
50. Masadeh M, Davanager HJ, Muaad AY. A novel machine learning-based framework for detecting religious Arabic hatred speech in social networks. Int J Adv Comput Sci Appl. 2022;13:767–76. https://doi.org/10.14569/IJACSA.2022.0130991.
51. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In: 26th international world wide web conference 2017, WWW 2017 companion; 2017. p. 759–60. https://doi.org/10.1145/3041021.3054223. arXiv:1706.00188.
52. Burnap P, Williams ML. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci. 2016. https://doi.org/10.1140/epjds/s13688-016-0072-6.
53. Golbeck JA. large human-labeled corpus for online harassment research. In: WebSci 2017—proceedings et al. ACM web science conference 2017; 2017. p. 229–33. https://doi.org/10.1145/3091478.3091509.
54. Founta AM. et al. Large scale crowdsourcing and characterization of twitter abusive behavior. In: 12th international AAAI conference web social media, ICWSM 2018; 2018. p. 491–500. https://doi.org/10.1609/icwsm.v12i1.14991. arXiv:1802.00393.
55. Pratiwi NI, Budi I, Alfina I. Hate speech detection on Indonesian instagram comments using FastText approach. In: 2018 international conference advance computing science information system ICACSIS 2018; 2019. p. 447–50. https://doi.org/10.1109/ICACSIS.2018.8618182.
56. Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M. An Italian twitter corpus of hate speech against immigrants. In: Lr 2018—11th international confernce language resource evaluating; 2019. p. 2798–805.
57. Basile V, et al. SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter. In: NAACL HLT 2019—international workshop Semant. evaluating SemEval 2019, proceegind of the 13th workshop; 2019. p. 54–63. https://doi.org/10.18653/v1/s19-2007.
58. Fortuna P, Rocha da Silva J, Soler-Company J, Wanner L, Nunes S. A hierarchically-labeled portuguese hate speech dataset; 2019. p. 94–104. https://doi.org/10.18653/v1/w19-3510.
59. Ibrohim MO, Budi I. Multi-label hate speech and abusive language detection in Indonesian Twitter; 2019. p. 46–57. https://doi.org/10.18653/v1/w19-3506.
60. Tang Y, Dalzell N. Classifying hate speech using a two-layer model. Stat Public Policy. 2019;6:80–6. https://doi.org/10.1080/2330443X.2019.1660285.
61. Alsafari S, Sadaoui S, Mouhoub M. Hate and offensive speech detection on Arabic social media. Online Soc Netw Media. 2020;19: 100096. https://doi.org/10.1016/j.osnem.2020.100096.
62. Charitidis P, Doropoulos S, Vologiannidis S, Papastergiou I, Karakeva S. Towards countering hate speech against journalists on social media. Online Soc Netw Media. 2020;17: 100071. https://doi.org/10.1016/j.osnem.2020.100071. arXiv:1912.04106..
63. Chen J, et al. A classified feature representation three-way decision model for sentiment analysis. Appl Intell. 2022;52:7995–8007. https://doi.org/10.1007/s10489-021-02809-1.
64. Qureshi KA, Sabih M. Un-compromised credibility: social media based multi-class hate speech classification for text. IEEE Access. 2021;9:109465–77. https://doi.org/10.1109/ACCESS.2021.3101977.
65. Plaza-Del-Arco FM, Molina-Gonzalez MD, Urena-Lopez LA, Martin-Valdivia MT. A multi-task learning approach to hate speech detection leveraging sentiment analysis. IEEE Access. 2021;9:112478–89. https://doi.org/10.1109/ACCESS.2021.3103697.
66. Rodriguez A, Chen YL, Argueta C. FADOHS: framework for detection and integration of unstructured data of hate speech on Facebook using sentiment and emotion analysis. IEEE Access. 2022;10:22400–19. https://doi.org/10.1109/ACCESS.2022.3151098.
67. Yuan L, Wang T, Ferraro G, Suominen H, Rizoiu MA. Transfer learning for hate speech detection in social media. J Comput Soc Sci. 2023. https://doi.org/10.1007/s42001-023-00224-9.
68. Mansur Z, Omar N, Tiun S. Twitter hate speech detection: a systematic review of methods, taxonomy analysis, challenges, and opportunities. IEEE Access. 2023. https://doi.org/10.1109/ACCESS.2023.3239375.
69. Alrehili A. Automatic hate speech detection on social media: A brief survey. In: Proceedings of the IEEE/ACS international conference computing system application AICCSA. 2019-Novem; 2019. p. 1–6. https://doi.org/10.1109/AICCSA47632.2019.9035228.
70. Mohiyaddeen SS. Automatic hate speech detection a literature review. Int J Eng Manag Res. 2021. https://doi.org/10.31033/ijemr.11.2.17.
71. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Comput Surv. 2018. https://doi.org/10.1145/3232676.
72. Shushkevich E, Cardiff J. Automatic misogyny detection in social media: a survey. Comput y Sist. 2019;23:1159–64. https://doi.org/10.13053/CyS-23-4-3299.
73. Alkomah F, Ma X. A literature review of textual hate speech detection methods and datasets. Information. 2022. https://doi.org/10.3390/info13060273.
74. Simon H, Yusuf Baha B, Garba EJ. Trends in machine learning on automatic detection of hate speech on social media platforms: a systematic review. FUW Trends Sci Technol J. 2022;7:1–016.
75. Istaiteh O, Al-Omoush R, Tedmori S. Racist and sexist hate speech detection: literature review. In: 2020 international conference intelligence data science technologies application IDSTA 2020; 2020. p. 95–9. https://doi.org/10.1109/IDSTA50958.2020.9264052.

76. Mullah NS, Zainon WMNW. Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access. 2021. https://doi.org/10.1109/ACCESS.2021.3089515.
77. Schmidt A, Wiegand M. A Survey on Hate Speech Detection using Natural Language Processing. In: Soc. 2017—5th international workshop Nat. language processing soc. media, Proceedings work. AFNLP SIG Soc. 2017. p. 1–10. https://doi.org/10.18653/v1/w17-1101.
78. Hee MS, et al. Recent advances in hate speech moderation: multimodality and the role of large models; 2024. arXiv:2401.16727.
79. Abro S, et al. Automatic hate speech detection using machine learning: a comparative study. Int J Adv Comput Sci Appl. 2020;11:484–91. https://doi.org/10.14569/IJACSA.2020.0110861.
80. Dhanya LK, Balakrishnan K. Hate speech detection in Asian languages: a survey. In: ICCISc 2021–2021 international confernece communication control information science proceedings. 2021;1:1–5. https://doi.org/10.1109/ICCISc52257.2021.9484922.
81. Subramanian M, Easwaramoorthy Sathiskumar V, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Eng J. 2023;80:110–21. https://doi.org/10.1016/j.aej.2023.08.038.
82. Anjum, Katarya R. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. Int J Inf Secur. 2023. https://doi.org/10.1007/s10207-023-00755-2.
83. Berretti S, Daoudi M, Turaga P, Basu A. Representation, analysis, and recognition of 3D humans: a survey. ACM Trans Multimed Comput Commun Appl. 2018;14:1–36. https://doi.org/10.1145/3182179.
84. Nayak R, Joshi R. Contextual hate speech detection in code mixed text using transformer based approaches. CEUR workshop proceedings; 2021. 3159:217–25. arXiv:2110.09338.
85. Sreelakshmi K, Premjith B, Soman KP. Detection of hate speech text in Hindi-English code-mixed data. Procedia Comput Sci. 2020;171:737–44. https://doi.org/10.1016/j.procs.2020.04.080.
86. Mathur P, Sawhney R, Ayyar M, Shah RR. Did you offend me? Classification of Offensive Tweets in Hinglish Language. In: 2nd workshop on abusive language online—proceedings workshop co-located with EMNLP 2018; 2018. p. 138–48. https://doi.org/10.18653/v1/w18-5118.
87. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: 15th conference Europe chapter association computing linguistics EACL 2017—proceedings conference 2017;2:427–31. https://doi.org/10.18653/v1/e17-2068. arXiv:1607.01759.
88. Kamble S, Joshi A. Hate speech detection from code-mixed Hindi-English Tweets Using Deep Learning Models; 2018. arXiv:1811.05145.
89. Chopra A, Sharma DK, Jha A, Ghosh U. A framework for online hate speech detection on code-mixed Hindi-English text and Hindi text in Devanagari. In: ACM transactions on Asian low-resource language information processing; 2023. p. 22. https://doi.org/10.1145/3568673.
90. Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M. A dataset of Hindi-English code-mixed social media text for hate speech detection. In: Proceedings 2nd workshop computing model. PFople's opinion personality emotional social media, PEOPLES 2018 2018 conference North America chapter association computing linguistics human language T; 2018. p. 36–41. https://doi.org/10.18653/v1/w18-1105.
91. Vashistha N, Zubiaga A. Online multilingual hate speech detection: experimenting with Hindi and English social media. Information. 2021;12:1–16. https://doi.org/10.3390/info12010005.
92. Yadav A, Garg T, Klemen M, Ulcar M. Code-mixed sentiment and hate-speech prediction. p. 1–12. arXiv:2405.12929v1.
93. Chen Z, Zhou L, Zhu W, et al. Hate speech detection in online platforms. Proc ACL. 2019;2:149–63.
94. Vidgen B, Derczynski L. Directions for hate speech research: challenges and applications. In: Online harassment symposium; 2020. p. 1–15.
95. Alshalan R, Al-Mohanna H. Hate speech detection in customer reviews: challenges and opportunities. J E-Commerce Res. 2022;13:205–20.
96. Zhang Z, Luo Y. Detecting hate speech in online platforms with neural networks. Neural Inf Process Syst. 2018;31:2232–42.
97. Chandrasekaran H, Kumar V. Detecting and mitigating hate speech in educational forums: a machine learning approach. J Educ Technol. 2021;18:25–40.
98. Gorwa R. The platform governance triangle: conceptualizing the challenges of regulating social media platforms. Internet Policy Rev. 2020;9:1–22.
99. Zuckerberg M. Social media's role in addressing hate speech. Facebook community standards report; 2018. p. 1–8.
100. Haapoja J, Laaksonen S-M, Lampinen A. Gaming algorithmic hate-speech detection: stakes, parties, and moves. Soc Media Soc. 2020. https://doi.org/10.1177/2056305120924778.
101. Huang X, Xing L, Dernoncourt F, Paul MJ. Hate speech recognition; 2019. arXiv:2002.10361v2.
102. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S. A multilingual evaluation for online hate speech detection. ACM Trans Internet Technol. 2020. https://doi.org/10.1145/3377323.
103. Diaz F, Mitra B. Recall, robustness, and lexicographic evaluation; 2024. arXiv:2302.11370.
104. Ombui E. Hate speech detection in code-switched text messages; 2019. https://doi.org/10.1109/ISMSIT.2019.8932845.
105. Nugroho K, et al Improving random forest method to detect hatespeech and offensive word. In: 2019 international conference information communication technology ICOIACT 2019; 2019. p. 514–8. https://doi.org/10.1109/ICOIACT46704.2019.8938451.
106. Zhang Z, Luo L. Hate speech detection: a solved problem? The challenging case of long tail on Twitter. Semant Web. 2019;10:925–45. https://doi.org/10.3233/SW-180338. arXiv:1803.03662.
107. Al-Ibrahim RM, Ali MZ, Najadat HM. Detection of hateful social media content for Arabic language. ACM Trans Asian Low Resour Lang Inf Process. 2023. https://doi.org/10.1145/3592792.

108. Liu S, Forss T. Combining, N-gram based similarity analysis with sentiment analysis in web content classification. In: KDIR proceedings of the international conference knowledge discovery and information retrieval. 2014;530–537:2014. https://doi.org/10.5220/0005170305300537.
109. Pawar AB, Gawali P, Gite M, Jawale MA, William P. Challenges for hate speech recognition system: approach based on solution. In: International conference on sustainable computing data communication system ICSCDS 2022—Proceedings; 2022. p. 699–704. https://doi.org/10.1109/ICSCDS53736.2022.9760739 (Institute of Electrical and Electronics Engineers Inc., 2022).
110. Aljero MKA, Dimililer N. Genetic programming approach to detect hate speech in social media. IEEE Access. 2021;9:115115–25. https://doi.org/10.1109/ACCESS.2021.3104535.
111. Li M, et al. COVID-HateBERT: A Pre-trained Language Model for COVID-19 related Hate Speech Detection. In: Proceedings of the 20th IEEE international conference machine learning application ICMLA 2021; 2021. p. 233–8. https://doi.org/10.1109/ICMLA52953.2021.00043.
112. Kumar C, Yadav RK, Namdeo V. A review on hate speech recognition on social media. Int J Innov Res Technol Manag. 2020;3404:50–7.
113. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Hate speech detection using natural language processing: applications and challenges. In: 25th international world wide web conference WWW 2016; 2016. p. 145–53. https://doi.org/10.1145/2872427.2883062
114. Saumya S, Kumar A, Singh JP. Filtering offensive language from multilingual social media contents: a deep learning approach. Eng Appl Artif Intell. 2024;133: 108159. https://doi.org/10.1016/j.engappai.2024.108159.
115. Wu H, Zhang Z, Shi S, Wu Q, Song H. Phrase dependency relational graph attention network for aspect-based sentiment analysis. Knowl Based Syst. 2022;236: 107736. https://doi.org/10.1016/J.KNOSYS.2021.107736.
116. Faris H, Aljarah I, Habib M, Castillo PA. Hate speech detection using word embedding and deep learning in the Arabic language context. Int Conf Pattern Recognit Appl Methods. 2020;1:453–60. https://doi.org/10.5220/0008954004530460.
117. Mundra S, Mittal N. Evaluation of text representation method to detect cyber aggression in Hindi English code mixed social media text; 2021. p. 402–9. https://doi.org/10.1145/3474124.3474185.
118. Seliya N, Khoshgoftaar TM, Van Hulse J. A study on the relationships of classifier performance metrics. In: Proceedings of the international conference on tools with Artificial Intelligence ICTAI; 2009. p. 59–66. https://doi.org/10.1109/ICTAI.2009.25.
119. Wang X, et al. A multimodal approach to hate speech detection. In: 2024 multimodal hate speech event identification challenge; 2024.
120. Shahi A, et al. Cross-platform hate speech identification: a study on youtube, twitter, and gab. In: 2023 conference on social media analysis; 2023.
121. Yang Y, et al. Hare: harnessing language models for explainable hate speech detection. 2023 J Artif Intell Res; 2023.
122. Chen Y, Zhou Y, Zhu S, Xu H. Detecting, offensive language in social media to protect adolescent online safety. In: Proceedings of the 2012 ASE, IEEE international conference on privacy, security risk trust. ASE/IEEE international conference social computing social. 2012;2012(71–80):2012. https://doi.org/10.1109/SocialCom-PASSAT.2012.55.
123. Gitari ND, Zuping Z, Damien H, Long J. A lexicon-based approach for hate speech detection. Int J Multimed Ubiquitous Eng. 2015;10:215–30. https://doi.org/10.14257/ijmue.2015.10.4.21.
124. Abozinadah EA, Mbaziira AV, Jones JHJ. Detection of abusive accounts with Arabic Tweets. Int J Knowl Eng. 2015;1:113–9. https://doi.org/10.7763/ijke.2015.v1.19.
125. Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W. A dictionary-based approach to racism detection in dutch social media; 2016. arXiv:1608.08738.
126. Mossie Z, Wang JH. Vulnerable community identification using hate speech detection on social media. Inf Process Manag. 2020;57: 102087. https://doi.org/10.1016/j.ipm.2019.102087.
127. Zampieri M, et al. Predicting the type and target of offensive social media posts in Marathi. Soc Netw Anal Min. 2022;12:1415–20. https://doi.org/10.1007/s13278-022-00906-8.
128. Saleem HM, Dillon KP, Benesch S, Ruths D. A web of hate: tackling hateful speech in online social spaces; 2014. arXiv:1709.10159v1.
129. Raisi E, Huang B. Cyberbullying identification using participant-vocabulary consistency; 2016. p. 46–50. arXiv:1606.08084.
130. Subramanian M, Easwaramoorthy V, Deepalakshmi G, Cho J, Manikandan G. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Eng J. 2023;80:110–21. https://doi.org/10.1016/j.aej.2023.08.038.
131. Nascimento FR, Cavalcanti GD, Da Costa-Abreu M. Exploring automatic hate speech detection on social media: a focus on content-based analysis. SAGE Open. 2023;13:19. https://doi.org/10.1177/21582440231181311.
132. Malik JS, Pang G, van den Hengel A. Deep learning for hate speech detection: a comparative study; 2022. arXiv:2202.09517.
133. Kagne P. Political hate speech detection using machine learning. Int J Sci Res Eng Manag. 2023;07:1–11. https://doi.org/10.55041/ijsrem26514.

Publisher's Note