Received: 29 March 2023 | Revised: 30 November 2023 | Accepted: 23 January 2024

DOI: 10.1111/exsy.13562

REVIEW

Expert Systems WILEY

Hate speech detection: A comprehensive review of recent works

Ankita Gandhi[1] | Param Ahir[1] | Kinjal Adhvaryu[2] | Pooja Shah[3] | Ritika Lohiya[4] | Erik Cambria[5] | Soujanya Poria[6] | Amir Hussain[7]

[1]School of Cyber Security and Digital Forensic, National Forensic Sciences University, Ganhinagar, Gujarat, India

[2]Computer Engineering, Shankarsinh Vaghela Bapu Institute of Technology, Ganhinagar, Gujarat, India

[3]School of Technology, Pandit Dindayal Energy University, Ganhinagar, Gujarat, India

[4]Faculty of Engineering, Sciences and Technology, Adani University, Ganhinagar, Gujarat, India

[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[6]Singapore University of Technology and Design, Singapore University of Technology and Design, Singapore, Singapore

[7]Edinburgh Napier University, Edinburgh, Scotland

Correspondence
Ankita Gandhi and Param Ahir, School of Cyber Security and Digital Forensic, National Forensic Sciences University, Gujarat, India.
Email: ankita.gandhi@nfsu.ac.in and param.ahir@nfsu.ac.in

Present address
Param Ahir, National Forensic Sciences University, Ganhinagar, Gujarat, India.

Abstract

There has been surge in the usage of Internet as well as social media platforms which has led to rise in online hate speech targeted on individual or group. In the recent years, hate speech has resulted in one of the challenging problems that can unfurl at a fast pace on digital platforms leading to various issues such as prejudice, violence and even genocide. Considering the acceptance of Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques in varied application domains, it would be intriguing to consider these techniques for automated hate speech detection. In literature, there have been efforts to recognize and categorize hate speech using varied Machine Learning (ML) and Deep Learning (DL) techniques. Hence, considering the need and provocations for hate speech detection we aim to present a comprehensive review that discusses fundamental taxonomy as well as recent advances in the field of online hate speech identification. There is a significant amount of literature related to the initial phases of hate speech detection. The background section provides a detailed explanation of the previous research. The subsequent section that follows is dedicated to examining the recent literature published from the year 2020 onwards. The paper presents some of the hate speech datasets considered for hate speech detection. Furthermore, the paper discusses different data modalities, namely, textual hate speech detection, multi-modal hate speech detection and multilingual hate speech detection. Apart from systematic review on hate speech detection, the paper also implement several multi-label models to compare the performance of hate speech detection by employing classic ML technique namely, Logistic Regression and DL technique namely, Long Short-Term Memory (LSTM) and a multiclass multi-label architecture. In the implemented architecture, we have derived two new elements to quantify the hatefulness and intensity of hatred to improve the results for hate speech detection using Indonesian tweet dataset. Empirical Analysis of the model reveals that the implemented approach outperforms and is able to achieve improved results for the underlying dataset.

KEYWORDS
CNN, deep learning, hate speech detection, LSTM, multi-label, multi-lingual

1 | INTRODUCTION

Automated hate speech detection is one of the important domains of Natural Language Processing (NLP) and Machine Learning (ML) that strives to design effective and efficient approaches to identify and remove hate speech from online platforms (Yin & Zubiaga, 2021). Hate speech refers to varied forms of communication that aims to demean, marginalize or harm a particular group of people based on their race, ethnicity, gender, sexual orientation, religion or other characteristics (Del Vigna et al., 2017). Automated hate speech detection is significant because hate speech can have serious negative consequences for the individuals and groups targeted by it, and can also contribute to the spread of misinformation and the erosion of civil discourse online (Del Vigna et al., 2017). Considering the amount of data generated on various communication platforms it is difficult to review the content manually, and hence, there is a need to escalate the research to develop techniques for automated hate speech detection.

There are several ways to detect hate speech which includes, manual review wherein trained human reviewers can manually review and identify hate speech by looking for certain keywords, phrases or patterns. ML algorithms can be trained to detect hate speech by analyzing text for patterns and features that are commonly associated with hate speech. These algorithms can be applied to automatically flag and remove hate speech from social media platforms, websites and forums. Further, these models are trained on large dataset of labelled text and can be fine-tuned to specific use cases. A combination of human review and machine-based detection can also be used to detect hate speech wherein, human reviewers can check flagged content and validate machine-based detection (Arango et al., 2019; Schmidt & Wiegand, 2017).

In literature, efforts has been made to identify and remove hate speech by applying ML techniques (Abro et al., 2020). However, automated hate speech detection can be regarded as a convoluted process as detection of hate speech is context specific and therefore, it would be a trivial task to discriminate between legitimate speech and hate speech (Abro et al., 2020). Comment toxicity detection is also done by (Kumar et al., 2021). The context of hate speech can be categorized as aggressive, abusive, sexist, homophobic and religious hate (Alkomah & Ma, 2022). Hate speech can be disseminated on social media through various methods, including posting, sharing, commenting, joining hate speech-promoting groups, using hashtags, live streaming and deploying bots. Users can post hateful content on their profiles, repost or send it, leave hateful comments (Boishakhi et al., 2021). Hence, pertaining to various mediums and categories, there still exists scope for improving and designing effective hate speech mechanism. It is important to note that detecting hate speech can be challenging as it is context-dependent, and even with advanced techniques, it's not always possible to detect hate speech with precise accuracy. This is the reason why one of the promising future research direction comes with explainable techniques (Mosca et al., 2021). Explainable hate speech detection refers to the ability to provide an explanation or justification for why a particular piece of text has been classified as hate speech (Meske & Bunde, 2023). Applying explainable techniques would be transfiguring hate speech detection task, because it is often difficult to understand why a particular piece of text has been classified as hate speech (Meske & Bunde, 2023). Explainable hate speech detection can be achieved through the use of advanced ML models, which are designed to provide insight into the reasoning behind the model's predictions. These models can be used to identify the specific words, phrases or features that contributes to the classification of a piece of text as hate speech (Abro et al., 2020). Additionally, explainable techniques can be designed using techniques such as feature extraction and attention-based mechanisms, which can highlight the most important words or phrases in the text that the model used to make its prediction. Feature extraction in explainable hate speech detection can also be used to improve the accuracy of hate speech detection by providing feedback to the model on what features and patterns it should pay more attention to. Thus, designing explainable models can help to reduce bias and improve the transparency of the decision making process.

Considering the significance and need of designing explainable approaches for hate speech detection, we aim to present a systematic review on hate speech detection considering various data modalities such as textual hate speech detection, multi-modal hate speech detection and multi-lingual hate speech detection. Moreover, we have also presented an explainable approach that incorporates an effective feature generation and extraction process by deriving two novel features, namely, hurtfulness and intensity of hurt. This is implemented on two classification models, namely, Long Short Term Memory (LSTM) and Logistic Regression (LR). Thus, the research contribution of our study can be summarized as follows.

- Our study gives a brief overview of hate speech detection of recent works.
- Summarizes various latest benchmark hate speech datasets used for hate speech detection.
- Presents a systematic literature review in different data modalities, namely, textual hate speech detection, multi-modal hate speech detection and multi-lingual hate speech detection.
- Implement an existing approach to compare the performance of hate speech detection.
- Discusses the current advancements in the field of hate speech detection.

To conduct a systematic and comprehensive literature survey on hate speech detection, a structured approach was employed. A diverse set of keywords was curated, encompassing terms such as ‘Hate speech detection’, ‘Hate speech classification’, ‘Online hate detection’, ‘Toxic comment classification’, ‘Abusive language detection’, ‘Discriminative speech detection’, ‘Machine learning for hate speech’, ‘Natural language processing for hate speech’ and ‘Deep learning for hate speech detection’. These keywords were diligently used to search major academic

databases, including Google Scholar, Scopus, IEEE Xplore and PubMed. The inclusion criteria were stringently defined: papers needed to primarily address hate speech detection, be published after 2020, demonstrate methodological rigor, be available in full text, and be written in English. The exclusion criteria ensured the omission of off-topic papers, those not disseminated in peer-reviewed journals or esteemed conferences, purely theoretical articles without practical application, and redundant studies. After identifying the relevant articles, each was meticulously reviewed to extract key findings, methodologies, datasets employed and performance metrics. This rigorous approach facilitated a comprehensive understanding of the prevailing research landscape in hate speech detection, highlighting current trends, gaps and avenues warranting future exploration. The rest of the sections of this paper are organized as follows: Section 2—Background elucidates the foundational principles and imperative for hate speech detection. Section 3—Summarizes the has speech datasets followed by Section 4 which is further divided into textual, multilingual and multimodal hate speech detection, comprehensively explores distinct approaches to hate speech detection. Section 5—Implemented Architecture with Benchmark Dataset, Pre-processing, Feature Generation and Extraction, Model Training, Experiment and Results offers a detailed account of the methodology employed in the study. Section 6—Current Advancements in the Field provides insights into recent trends and developments in hate speech detection. Section 7—Limitations candidly addresses potential constraints and biases in the research. Section 8—Future Directions explores potential avenues for future research, suggesting areas for enhancement and novel methodologies. Finally, Section 9—Conclusion succinctly encapsulates the core findings and their implications.

2 | BACKGROUND

The process of hate speech detection typically involves several steps, which can vary depending on the specific approach and technology used (Sood, Churchill, et al., 2012). However, generic hate speech detection process is as shown in Figure 1, which includes the following steps:

- Dataset collection: The data are gathered from varied sources which includes social media websites, news websites and online discussion boards. This is the initial stage in the process of detecting hate speech. The data gathered can be in any form such as text, pictures or videos.
- Pre-processing: After the data has been collected, it is pre-processed, which includes operations like lemmatization, stemming, stop-word removal and tokenization. The purpose of this stage is to standardise the data and prepare it for analysis.
- Dataset Labelling: Dataset labelling is the process of providing annotation to data for making it easier for the underlying model to understand and interpret data. Machine learning models are trained using this tagged data for further classifying the input data as hate or non-hate speech. Data annotation techniques include manual annotation, crowd-sourced annotation and semi-supervised learning-based annotations.
- Feature Engineering: Varied feature engineering techniques are applied on the pre-processed data to extract pertinent features such that unsequenced textual data are transformed into structured features. Apart from feature extraction, feature reduction techniques can also be applied to minimize the time and memory complexity. Various feature engineering techniques such as TF-IDF, BOW, PCA, LDA and many more can be applied to extract information such as word frequency, n-grams and grammatical patterns from text data.
- Model Learning: It is one of the crucial step in hate speech detection process wherein, a machine learning model or deep learning model is trained using the extracted features. Various machine learning classifiers can be applied for hate speech detection. Moreover, word embedding can be performed along with deep learning model as some of the deep learning models have embedding layer in neural network architecture. The evaluation of the trained model is performed using different evaluation metrics such as *f*-score, accuracy, precision and recall. The evaluation helps to determine effectiveness of the model in detecting hate speech.
- Classification and Detection: The trained model can be used to perform multi-class, multi-label or binary classification for hate speech detection.

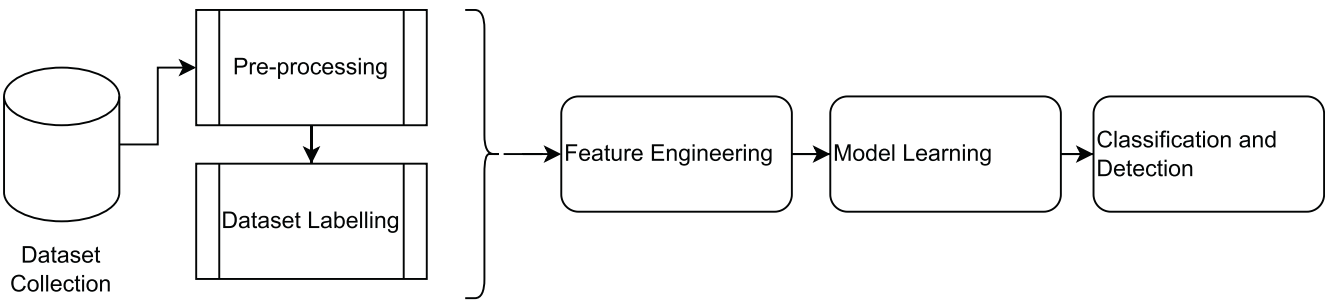

FIGURE 1 Generic process of hate speech detection.

TABLE 1 Major distinguishing features used in various approaches.

Approach base	Brief detail	References
Simple surface feature	Bags of words, unigrams, n-grams, and other surface features make up the feature set. Together with the frequency of URL mentions and punctuation, other data can be utilised to analyse comment and token lengths, capitalization, terms that are uncommon in English dictionaries, and the quantity of non-alpha numeric characters that are contained in tokens.	(Burnap & Williams, 2015; Burnap & Williams, 2016; Chen et al., 2012; Fosler-Lussier et al., 2012; Sood, Antin, et al., 2012; Warner & Hirschberg, 2012)
Word generalization	One may encounter a data sparsity issue because hate speech detection is typically used to brief chunks of text or even individual sentences. This is why several works use word clustering and word generalisation techniques to address this problem.	(Dinakar et al., 2012; Nobata et al., 2016; Warner & Hirschberg, 2012)
Sentiment analysis	Given the intimate connection between hate speech and sentiment analysis, it is plausible to presume that hate speech messages are typically associated with unfavourable sentiment. As a result, many systems incorporate sentiment analysis as an auxiliary categorization to recognise the relationship between hate speech and the latter.	(Burnap & Williams, 2015; Gitari et al., 2015; Sood, Churchill, et al., 2012)
Lexical resources	Several authors try to capitalise on the widespread belief that nasty texts contain certain bad phrases like slurs, insults, and so on by using their inclusion as a feature. Lexical resources with these predictive phrases are necessary to gather this kind of information.	(Nobata et al., 2016; Xiang et al., 2012)
Linguistic features	It relates to taking the deeper syntactic features into account	(Fosler-Lussier et al., 2012; Gitari et al., 2015)
Knowledge based features	It seems logical that the identification of a phenomenon as complicated as hate speech would benefit from integrating information on elements not directly related to language. Whether a message is hateful or innocuous can be greatly dependent on global knowledge. Prior to using the procedure, provide any domain-specific assertions.	(Dinakar et al., 2012)
Meta information	It may be quite predictive to know a little bit about the poster of a post. A user who has previously posted hate speech is permitted to do so once again. It is improbable that a user who has never written such messages will do so in the future.	(Xiang et al., 2012)
Multimodal information	Images, videos, and audio content are all included in contemporary social networking platforms in addition to text. Such non-textual material is frequently discussed as well, which makes it part of the discourse of a hate speech utterance. This external context can be used as a predictive function without requiring a written user comment.	(Hosseinmardi et al., 2015)

It is important to note that hate speech detection is a complex and ongoing process that requires continuous monitoring and improvement to keep up with the evolving nature of hate speech.

This generic methodology extends to multiple approaches spanning from different types of features used for detection. Thus hate speech detection can be classified broadly in to the different categories like textual, multimodal and multilingual hate speech detection. Table 1 describes various features leading to various approaches to deal with hate speech detection problem.

2.1 | Types of hate speech detection

There are different types of hate speech detection basis on different context like multi-label, multilingual and multimodal which are described in brief as follows.

- Binary Hate Speech Detection: It is hate speech classification problem wherein the trained model classifies the underlying data as hate speech or non-hate speech. Varied binary classifier can be applied for binary classification, where various attributes and features are considered for learning and classification. However, automatic binary classification of hate speech often tends to neglect distinct attributes of other sentimental types such as aggressive and racist.
- Multi-Label Hate Speech Detection: It is hate speech classification problem wherein the trained model classifies the underlying data as various types of hate speech or non-hate speech. The dataset can consist of various labels such as cyber-bullying, racism, radicalization, sexism, racial, to name a few. Hence, the trained model learns based on given attributes and features and classifies the underlying input data in most appropriate label defined in the underlying dataset.
- Textual Hate Speech Detection: Textual hate speech detection identifies and categorizes hate speech in written content. These models, which include machine learning and deep learning algorithms, are trained on labelled data to differentiate between hate speech, offensive language and non-offensive information.
- Multi-Lingual Hate Speech Detection: In literature, there has been focus on developing generalized hate speech detection models that may remove cultural nuances by utilize language agnostic embedding to classify data in multiple languages. Thus, multi-lingual hate speech detection aims that comparing the performance of models trained on mono-lingual data compared to multi-lingual data for hate speech detection.
- Multimodal Hate Speech Detection: Multimodal hate speech detection extends the scope of detection beyond textual content by including various forms of multimedia such as images, videos and audio. These models recognises that hate speech can present itself through various channels and that proper identification necessitates a thorough, multi-dimensional strategy. Multimodal models strive to provide a more comprehensive understanding of hate speech by using visual and auditory clues alongside textual information. This allows to capture occurrences that text-only techniques may overlook.

3 | HATE SPEECH DETECTION DATASETS

The increased usage of social media platform has resulted in direct influence of hate speech and offensive language. Furthermore, for detection of hate speech various datasets, in varied languages are used. In this section, we aim to describe some of the hate speech datasets and provide insights for the same. The summary of hate speech datasets is presented in Table 2. An annotated hate speech dataset in Albanian language is created (Nurce et al., 2021). Here, the dataset consist of abusive language and hate speech in hierarchical schema, which is formed using user-generated posts on various social media platforms such as Instagram and YouTube. The difficulty and ambiguity in Arabic language and content has affected the provision of required resources for abusive language detection in Arabic. Hence, first publicly available dataset named as Levantine Hate Speech and Abusive dataset is developed (Mulki et al., 2019) using Twitter as social media platform. Here, the objective was to create a benchmark datasets for identifying toxic language on online platforms. Furthermore, the designed dataset is annotated using various annotation metrics such as Cohen's Kappa (k) and Krippendorff's alpha α for demonstrating the consistency of the annotations.

Apart from hate speech and abusive language, online sexism is also increasing with surge in the usage of social media platforms. One of the first sexism dataset in Chinese language is developed (Jiang et al., 2022), named as Sina Weibo Sexism Review dataset. The dataset consist of large number of sex and hate related lexicons in Chinese language along with abusive and gender related terms. The designed dataset is annotated and consist of varied class labels at different granularity levels. A English text-based hate speech dataset is developed by considering three conversational AI systems, namely, open domain-based social bot, rule-based chatbot and task-based conversational AI system in (Curry et al., 2021). The developed dataset is finely annotated with hierarchical schema that consist of different types of abuse and severity of abuse. A synthetic dataset is designed by considering trained annotators that generate dynamic content for better hate speech detection. The data creation is performed in four rounds wherein the first round consist of data created by humans that mimic the real world social media posts and other rounds is split into original content and content created by perturbations. The dataset consist of more than 40,000 entries with more than 54% of hate and abusive comments, which substantially higher compared to other existing datasets.

A refined version of OffensEval dataset is designed (Caselli et al., 2020) to address the challenges of varied abusive language phenomena such as hate speech, cyber bullying and offensive language. The dataset is developed with a focus to differentiate various abusive language phenomena and hence, the dataset is annotated with explicit and implicit abusive language in English. A multimodal meme dataset is designed (Suryawanshi et al., 2020), that considers two modality, namely, image and text, for hate speech detection. Meme being a new form of communication on various social media platforms, is one of the recent mediums for posting hate and abusive content on online platforms. Hence, in this dataset, two modalities are combined to detect hate speech, offensive and aggressive comments. For developing the multi-modal meme dataset, meme related to 2016 US presidential elections were considered.

A textual hate speech dataset named as ETHOS is developed with two variants, namely, binary and multi-labelled (Mollas et al., 2020) for hate speech detection. The dataset is created by considering various posts and comments on social media platforms, namely, YouTube and Reddit. The dataset is also validated through Figure-Eight crowd sourcing platform. The dataset is annotated using designed annotation protocol. A multi-labelled dataset based on the posts of Indonesian Twitter platform is designed (Ibrohim & Budi, 2019). Here, in the designed dataset varied

TABLE 2 Hate speech datasets.

References	Name	Year	Description	Size	Level of annotations	Platform	Source link
(Nurce et al., 2021)	Detecting abusive Albanian	2022	Hierarchical (offensive/not; untargeted/targeted; person/group/other)	11,874	Posts	Instagram, Youtube	https://doi.org/10.6084/m9.figshare.19333298.v1
(Jiang et al., 2022)	SWSR	2022	Sexism detection on social media in Chinese	8969 comments from 1527 weibos	Posts	Sina Weibos	https://doi.org/10.5281/zenodo.4773875
(Curry et al., 2021)	ConvAbuse	2021	Abuse detection in conversational AI	4185	Messages	Carbonbot on Facebook Messenger and E.L.I.Z.A. chatbots	https://github.com/amandacurry/convabuse
(Vidgen et al., 2020)	Learning From the Worst (Dynamically generated hate speech dataset)	2020	Multi-category hate speech detection	41,255	Posts	Synthetic Dataset	https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset
(Caselli et al., 2020)	AbuseEval v1.0	2020	Explicitness annotation of offensive and abusive content	14,100	Tweets	Twitter	https://github.com/tommasoc80/AbuseEval
(Suryawanshi et al., 2020)	MultiOFF	2020	Hate per se (related to 2016 U.S. presidential election)	743	Posts	Kaggle, Reddit, Facebook, Twitter and Instagram	https://github.com/bharathichezhiyan/Multimodal-Meme-Classification-Identifying-Offensive-Content-in-Image-and-Text
(Mollas et al., 2020)	ETHOS	2020	Binary dataset consisting of gender, race, national origin, disability, religion, sexual orientation	998	Posts	YouTube, Reddit	https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset
(Mulki et al., 2019)	L-HSAB	2019	Ternary (hate, abusive, normal)	5846	Posts	Twitter	https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
(Ibrohim & Budi, 2019a)	Multi-labelled Dataset	2019	Multi-labelled hate speech and abusive language detection in Indonesian Twitter	13,169	Posts	Twitter	https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection
(Ibrohim & Budi, 2018)	Abusive Language Dataset	2018	Abusive language detection in Indonesian social media	2016	Posts	Twitter	https://github.com/okkyibrohim/id-abusive-language-detection

categories are considered such as religion, race, physical disability, sexism, slander, to name a few. Moreover, the designed dataset is evaluated using different Machine learning (ML) and data transformation methods. A preliminary study describing creation and evaluation of abusive language dataset based on Indonesian social media platform is presented (Ibrohim & Budi, 2018). Here, posts from Twitter are considered for creating the dataset and are annotated in hierarchical schema for detecting abusive language.

4 | HATE SPEECH DETECTION IN DIFFERENT DATA MODALITIES

While textual analysis remains foundational, the rise of multilingual and multimodal content on digital platforms has propelled the need for specialized techniques. In this section, we explore approaches and advancements in addressing hate speech across these diverse data modalities.

4.1 | Textual hate speech detection

A majority of hateful content that is shared on social media platforms today takes the form of written content. Despite the fact that social media platforms encourage users to express themselves freely, these platforms have sadly become vehicles for the dissemination of intolerance, hate speech and the start of threats and bribery against other users. A great number of social media platforms are having trouble figuring out how to address the conflicting goals of suppressing hate speech and maintaining users' rights to share their views. Hate speech that takes place online, such as in blog posts and comment threads, has the potential to have detrimental effects on society. Attempts are currently being made to identify instances of hate speech using a variety of complicated methods. One way is a keyword-based method (Saleem et al., 2017) which compares a piece of writing to a database of words that are considered objectionable. One of the most significant limitations of the strategy is that it requires regular database updates to be performed. Mining user metadata (Waseem & Hovy, 2016) is another technique for detecting hate speech. This technique is used to determine who is responsible for publishing offensive information. This method is helpful for identifying repeat criminals; nevertheless, it can generate bias based on the past activities of the individuals being investigated. Currently, there is a significant amount of trust placed on methods that are based on machine learning. Throughout the course of its history, this technique has made use of a wide variety of classifiers, such as Naive Bayes (Kiilu et al., 2018) and Support Vector Machines (Hana et al., 2020). Deep learning, reinforcement learning and transfer learning are three types of learning-based methods that have become increasingly common as a result of the availability of such large datasets.

Figure 2 depicts the process for recognizing text based hate speech using machine learning. The initial stage in detecting hate speech in a text is pre-processing, which entails removing non-speech elements such as stop words and punctuation. This procedure is known as data cleaning. The extraction of textual features is the next phase, which can be achieved using various techniques of machine learning. One Hot Encoding (Wang, 2021) presents a simple technique in which each word is encoded as either 0 or 1. The TF-IDF (Christian et al., 2016) text representation can be used to characterize the frequency with which individual words appear in a given document. Word2Vec (Mori et al., 1999) is a word embedding method that converts a given word into a vector of integers. Document relevance importance measure (DRIM) (Radlinski et al., 2009) is a technique for measuring the importance of a word in a set of documents. After obtaining the features, they are transferred to a classifier, which can be a binary classifier or a multi-class classifier. Model is updated based on results of evaluation metrics during the final step.

Various recent papers with keywords such as hate-speech detection, hate-speech detection using machine learning, textual hate-speech detection and hate-speech detection using deep learning are considered for this study. Roy et al. (2020) classified data into hate speech and non-hate speech classes using a deep CNN with LSTM and a 10 cross fold validation technique. This approach works well with imbalance datasets and requires raw text data. Rodrguez-Sánchez et al. (2020) created a Spanish dataset to detect sexist tweets and applied Bi-LSTM and multilingual BERT to classify text into three categories: Sexist, Non-Sexist and Doubtful. Oriola and Kotzé created a multi-tier meta-learning model with

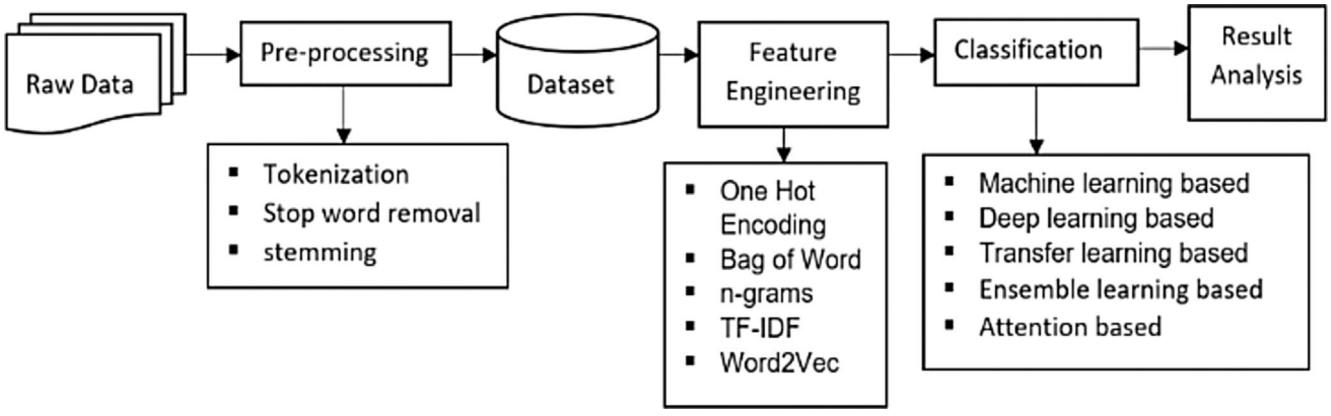

FIGURE 2 Process flow for text based hate speech detection using machine learning.

ensemble classifiers (Oriola & Kotzé, 2020) to identify South African text as hate speech, offensive speech or free speech. In this method basic machine learning methods are used in the ensemble classifier. Plaza-Del-Arco et al. (2021) detect hate speech from a Spanish corpus using the one language based Transformer-based model BETO. Alatawi et al. (2021) detect hate speech from white supremacists and classify text as explicit white supremacy, implicit white supremacy, other hate speech, and neutral using bi-LSTM and domain specific word embedding. Baydogan and Alatas (2021) created an autonomous hate speech detection system by combining the Ant Lion Optimization algorithm and the Moth Flame Optimization algorithm. Mridha et al. (2021) created the L-Boost technique, a modified version of AdaBoost techniques with bidirectional long short-term memory and BERT, to classify Bengali text into two categories: normal and offensive. Alzamzami and El Saddik (2021) created a BERT-based feed forward network to classify text into hate and non-hate. Qureshi and Sabih (2021) examined outcomes from different classifiers, including Random Forest, SVM, logistic regression, decision tree, gradient boosting, cat boost and MLP. Bilal et al. (2022) constructed a context-aware bi-LSTM and attention-based model to detect hate from an Urdu corpus.

Lee et al. (2022) employed stacked ensemble model-based Gated Convolutional Recurrent-Neural Networks (GCR-NNs) for racism detection in an English corpus. Khan, Kamal, et al. (2022) developed a capsule network-based Convolutional and Bi-Directional Gated Recurrent Unit classifier. Shannaq et al. (2022) use a genetic algorithm and XGBoost to detect hate speech in Arabic. Sharmila et al. (2022) devised the Dual-level Cross Attention approach to classify material into three categories: hateful, offensive and neither. In Table 3, a detailed summary of various recent hate speech detection research is given. From all the methods (Lee et al., 2022) stacked ensemble model-based Gated Convolutional Recurrent-Neural Networks (GCR-NNs) provides a good accuracy score. This can be further explored by combining multilingual and multimodal models.

4.2 | Multilingual hate speech detection

The task of locating instances of hate speech in literature published in numerous languages is known as multilingual hate speech detection. This is essential for maintaining a polite and secure online community since hate speech may have negative effects on people and communities. It employs natural language processing methods, such as machine learning and deep learning, to analyze text for patterns and linguistic cues that suggest the presence of hate speech. The goal is to accurately and impartially identify hate speech, regardless of the language in which it is written. The detailed steps for hate detection in multilingual context is shown in Figure 3. For multilingual hate detection, a number of recent advancements in the field have been reviewed. In the study by Oriola and Kotzé (2020), various machine learning based approaches like support vector machine, logistic regression, random forest, hyper-parameter optimization, ensemble approach and gradient boosting algorithms are analyzed. Character n-grams, word n-grams, negative sentiment, syntactic-based features and their hybrid methods, are extracted and analyzed. The findings indicated that while optimal gradient boosting with word n-gram worked best in terms of detecting hate speech, optimized support vector machines with character n-gram performed best in terms of detecting hate speech, with a true positive rate of 0.894. Another work by Aluru et al. (2020) employed deep learning models to create classifiers for multilingual hate speech categorization using 16 datasets from 9 languages. Many tests are conducted for a range of languages in low and high resource, monolingual and multilingual contexts. They found that BERT models are more successful for high resource situations while LASER + LR are for low resource situations. In a research by Vashistha and Zubiaga (2020), experimentation is conducted on a combined dataset made up of six datasets in English, Hindi and Code-mixed Hindi. These models obtain performance that is equivalent to or better than a wide variety of baseline monolingual models. The model operates in a near-real-time online context and produces competitive performance on pooled data. A set of practical tests for multilingual hate speech detection models are created in the work by Röttger et al. (2022) under the name MULTILINGUAL HATCHECK (MHC). Ten new languages—Arabic, Dutch, French, German, Hindi, Italian, Mandarin, Polish, Portuguese and Spanish—are added to the English-language HATCHECK by MHC. MHC has the broadest language coverage of any hate speech dataset. Native language specialists developed 36,582 test cases that contrast hateful and non-hateful content for all the languages. As a result, MHC presents a challenge to hate speech detection algorithms and makes it possible to assess model quality more effectively. Zero-shot cross-lingual transfer learning challenges have been highlighted in Montariol et al. (2022) utilizing rigorous experimental conditions. To enhance zero-shot translation of hate speech detection models across languages, a novel model is suggested to train on multilingual auxiliary tasks, such as sentiment analysis, named entity identification, and tasks dependent on syntactic information. Auxiliary task's favourable contribution to bridging the linguistic and cultural divide across languages in hate speech is also shown to assist hate speech detection models by serving as a cross-lingual knowledge proxy. A comparison of the monolingual and multilingual BERT models is done (Velankar et al., 2022). With an emphasis on the Marathi language, models on datasets for sentiment analysis, basic text categorization and hate speech detection are assessed. Standard multilingual models like mBERT, indicBERT and xlm-RoBERTa are compared to the Marathi-only versions namely MahaBERT, MahaALBERT and MahaRoBERTa. In five distinct downstream fine-tuning trials, Marathi monolingual models beat the multilingual BERT versions. By freezing the BERT encoder layers, sentence embedding from these models are also possible. Sentence embedding from multilingual equivalents do not offer the same level of rich representation as MahaBERT-based models for monolingual data. A Marathi sentiment classification datasets, L3Cube-MahaSent, as well as Marathi headline and article classification datasets have all been taken into consideration. In Table 4, a detailed summary of various recent hate speech detection research is given.

TABLE 3 Summary of textual hate speech detection.

Sr. No.	Year	Dataset	Classifier	Key objectives	Accuracy	Advantages	Limitations
(Roy et al., 2020)	2020	Twitter hate speech dataset	Deep Convolutional Neural Network and LSTM with 10-fold cross-validation	Built a text-only deep convolutional neural network that can accurately forecast unbalanced datasets.	0.93	On unbalanced datasets, the system out-predicted state-of-the-art approaches.	English-language model uses only twitter dataset. Non-hate-speech model findings are poor.
(Rodrguez-Sánchez et al., 2020)	2020	MeTwo	Bi-LSTM and Multilingual BERT	Used various machine learning and deep learning models to recognise sexist statements and attitudes in Spanish.	0.74	Modern Spanish language model. Created simple keyword databases.	'Most "SEXIST" tweets are misclassified. Keyword-based datasets can prejudice words'.
(Zhou et al., 2020)	2020	SemEval 2019 Task 5	ELMo, BERT and CNN	ELMo, BERT, and CNN classifiers were used to generate fusion models to improve model performance at low cost.	0.75	Due to fusion, this ensamble process is simpler and cheaper than others.	Model uses one language and one source data. This model produces low results compared to others.
(Oriola & Kotzé, 2020)	2021	InterTASS, EmoEvent, HatEval, MEX-A3T	Monolingual transformer-based model BETO	Created a multi-task learning model MTLsent + emo for Hate Speech detection using the monolingual transformer-based model BETO.	-	This is the first work to combine multitask learning with transformer-based model.	Multi-task learning's low performance with smaller data renders it inappropriate for regional languages without a huge dataset.
(Plaza-Del-Arco et al., 2021)	2021	Twitter White Supremacy Dataset, Stromfront dataset	BiLSTM, BERT	A bidirectional long short-term memory (BiLSTM) model was used to understand white supremacist coded language.	0.74	Text is classified beyond hate and non-hate.	BERT cannot detect misspellings or slang. The dataset is imbalanced and unfixed.
(Alatawi et al., 2021)	2021	Twitter, online web forums	Ant Lion Optimization (ALO) algorithm and Moth Flame Optimization (MFO) algorithm	Novel fitness function and data structure for Ant lion and moth flame optimization meta-heuristic.	0.92	More textual document jobs can use this representation style and fitness function. Optimizes model simultaneously.	Dice and cosine are not examined. English only model.

(Continues)

TABLE 3 (Continued)

Sr. No.	Year	Dataset	Classifier	Key objectives	Accuracy	Advantages	Limitations
(Baydogan & Alatas, 2021)	2021	Own dataset	L-Boost, LSTM Model	Combining modified LSTM and AdaBoost-BERT models for Bengali corpus, created L-Boost ensemble classifier.	0.95	This model outperforms all the other previous ensemble techniques.	Photos, pdfs, videos, and audio are not considered because Bengali is spoken differently in different regions.
(Mridha et al., 2021)	2021	HatEval 2019, OffensEval 2019, Antigoni Dataset 2018, Waseem and Hovy 2016	Feed forward Neural Network	Created a real-time pandemic cyber-social framework using unsupervised and supervised learning approaches.	0.86	For domain-specific corpus, LDA-style topic interpretation and labelling improves the real-time model.	Labelling other than pandemic text requires domain customization.
(Alzamzami & El Saddik, 2021)	2021	-	Support Vector Machines (SVM) and Multinomial NaïveBayes (MNB)	Dynamic stop words filtering, VGFSS, SMOTE, SVM, and MNB solve sparsity, dimensionality, and class imbalance.	-	Availability of Urdu corpus. The key Twitter sentiment analysis issues—highly skewed classes, high-dimensional feature vectors, and low data density—were addressed.	This model has class imbalance and class skew problem.
(Qureshi & Sabih, 2021)	2022	Roman Urdu hate speech dataset	Bi-LSTM with Attention Layer	Context-aware Roman and Urdu Hate Speech detection model using Bi-LSTM with an attention layer and updated word2vec word embeddings.	0.72	This model employs an attention technique to include context-aware features.	This approach cannot recognise sarcasm or implicit hate speech. Roman Urdu's lexical diversity makes certain harsh remarks sound funny.
(Bilal et al., 2022)	2022	-	Gated Convolutional Recurrent-Neural Networks (GCR-NN)	Gated convolutional recurrent neural networks were created by integrating GRU, CNN and RNN (GCR-NN).	0.97	This ensemble model is state-of-art-model specifically for racism detection	This model can only identify racism.
(Lee et al., 2022)	2022	DS1, DS2	HCovBi-Caps model	Created a convolutional BiGRU and Capsule network-based deep learning model.	0.87	HCovBi-Caps works well with class imbalance dataset	HCovBi-Caps also ignores sentiment and user profile data.

TABLE 3 (Continued)

Sr. No.	Year	Dataset	Classifier	Key objectives	Accuracy	Advantages	Limitations
(Shannaq et al., 2022)	2022	Arabic cyber bullying corpus dataset	XGBoost and SVM, and a genetic algorithm (GA)	Created a genetic algorithm-based cyber bullying detection model from Arabic dataset.	0.88	This model performs well with low-resource datasets.	Neural network optimization uses only a transfer learning-based model.
(Sharmila et al., 2022)	2022	-	Dual-level CrossAttention	Pattern-based Deep Hate Speech (PDHS) detection methodology uses cross-attention encoder and dual-level attention mechanism.	0.81	Dual-level CrossAttention increases feature embedding pattern representation.	The model cannot handle unstructured social media data or multimodal elements.
(Wang et al., 2022)	2022	Chinese hate speech dataset	BERT	Created a political hate speech lexicon and trained AI classifiers to detect it.	0.73	This system detects Chinese hate speech using deep learning and lexicons.	This dataset covers political news comments only. The hate speech lexicon in this study is limited to 153 items.

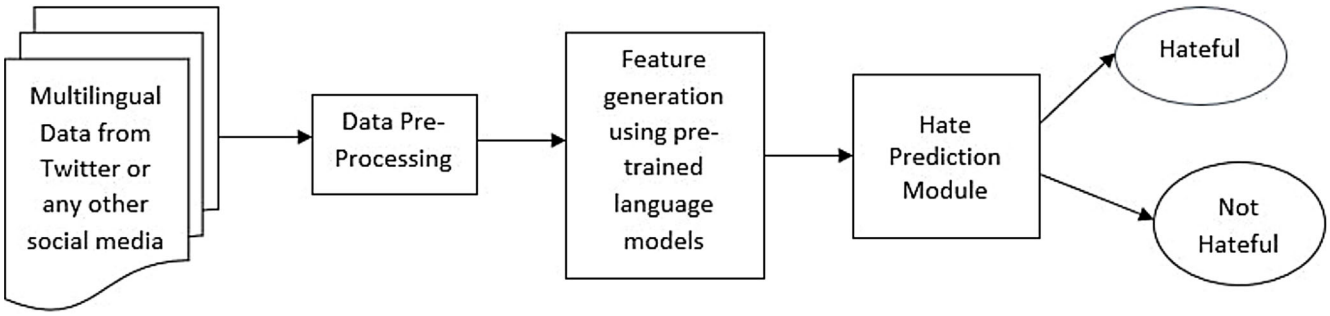

FIGURE 3 Process flow of multilingual hate speech detection using machine learning.

4.3 | Multimodal hate speech detection

In recent time, multimodal that is using different modalities like audio, visual and textual data processing is gaining much more traction. In study by Gandhi et al. (2021), number of multimodal data processing applications are discussed. Classifying hateful memes is a difficult process. It is difficult to manually identify every meme from the vast Internet data. The research of combining two modalities to find objectionable information is still in its early stages. The most common kind of memes are text-based visual memes. Ordinarily, a statement or image would not have any unique emotional significance, but when they are put together, they take on new meaning. As a result, hateful memes start to appear and pose a severe threat to contemporary civilization. The material appears to be normal and innocuous when we evaluate the text and image as separate characteristics. Memes are components of cultural or behavioural systems that spread vastly across the internet. Memes are increasingly powerful on social media platforms and appear in a variety of styles and formats, including in forms of photos, videos and posts. An eye-catching issue is the abundance of memes on the Internet. Not only memes can represent people's inherent emotions, but they also have the potential to hurt someone's feelings. The detection of hate speech, offensive material and aggressive material has been thoroughly studied in a single medium like text or

TABLE 4 Summary of multilingual hate speech detection.

References	Model Name	Year	Dataset	Key objective	Accuracy	Advantages	Limitation
(Oriola & Kotzé, 2020)	-	2020	Twitter	English tweets from South Africa, machine learning methods for hate speech detection	0.92	SMOTE for imbalance, feature extraction techniques	Contextual awareness, key word attention
(Aluru et al., 2020)	LASER, mBERT	2020	16 datasets, 9 languages	Multilingual hate speech classification	0.83	LASER + LR for low resources, efficient for low-resource languages	Challenges in annotation, hidden context
(Vashistha & Zubiaga, 2020)	-	2021	Combined dataset	Multilingual Hindi–English hate speech detection	0.9	Real-time evaluation tool, co-occurrence and contextual embedding	Incorporate more languages, try different models
(Mozafari et al., 2022)	-	2022	Various datasets	Few-shot hate speech detection in low-resource languages	-	MAML and Proto-MAML for effective performance	Experiment with different target languages
(Röttger et al., 2022)	MHC	2022	34 languages	Practical tests for multilingual hate speech detection	-	Diagnostic tool for model deficiencies	Limited support for protected groups
(Montariol et al., 2022)	XLM-R, XLM-T	2022	HatEval + AMI dataset	Cross-lingual transfer of hate speech models, impact of language model and task training	-	NER and sentiment analysis training, limited improvement in hate speech against women	Fine-tuning effects on data
(Velankar et al., 2022)	MahaBERT, and so on	2022	Various Marathi datasets	Comparison of multilingual and monolingual models for Marathi	-	Monolingual models outperform in downstream tasks	Need for better language embedding models

picture. Nevertheless, memes make it even more difficult since they implicitly reflect sarcasm and humour. Memes are not necessarily offensive if we merely look at the words or the image. Consequently, combining the two modalities is required to determine if a certain meme is offensive or not. The underlying meaning of the meme is quite rude and offensive, though, when taken as a whole. Identifying hateful memes from text and image-based multimodal data are a binary classification challenge. The samples of several hateful and non-hateful memes images and texts are shown in Figure 4. The detailed steps for hate detection in multi-modal memes utilising texts and images is shown in Figure 5. This method, known as multi-modal fusion, combines textual and visual data from social network memes. The different Meme types are first used as an input, and then pre-processing is done on the data. Data are cleansed and chosen using dimensionality reduction during the pre-processing stage in accordance with the needs of the task. Subsequently, several visual and textual feature extraction methods are used to extract features. Following the extraction of the features, a multi-modal feature vector is created, and the data are then divided into two categories—hateful and non hateful—using several classification algorithms. Several research studies have looked upon multi-modal hate detecting memes. A research by Kiela et al. (2020) uses subliminal cues and binary classification to identify hate speech in memes as its main objective. An early fusion deep learning approach to address the classification problem (Suryawanshi et al. 2020) is employed by combining the text and image modalities of a meme. They also presented the 743 memes that make up the new MultiOFF dataset, which was categorised as offensive or not offensive. In a different study by Vlad et al. (2020), the multi-modal multi-task learning architecture for hate detection in Italian memes for two distinct tasks is developed. Meme Detection, Task 1, focuses on identifying memes from a set of photos. The second task, 'Hate Speech Identification' aims to categorise images according to their intended use by looking at their content and determining if they are hateful. The Thai textual meme detection as a latest research challenge is employed (Mookdarsanit & Mookdarsanit, 2021). This article classifies mis-spelt Thai terms as synonyms to train pre-trained language models since they are commonly used in social media. The dataset for training the irregular Thai texts is expanded by multiplying them by various power law distributions and rotating them at various angles. Single Shot Detector (SSD) can identify both regular and irregular Thai text in memes. DisMultiHate framework, established (Lee et al. 2021), untangles representations of entities connected to hate speech, including race

FIGURE 4 Samples of hateful memes (Kiela et al., 2020).

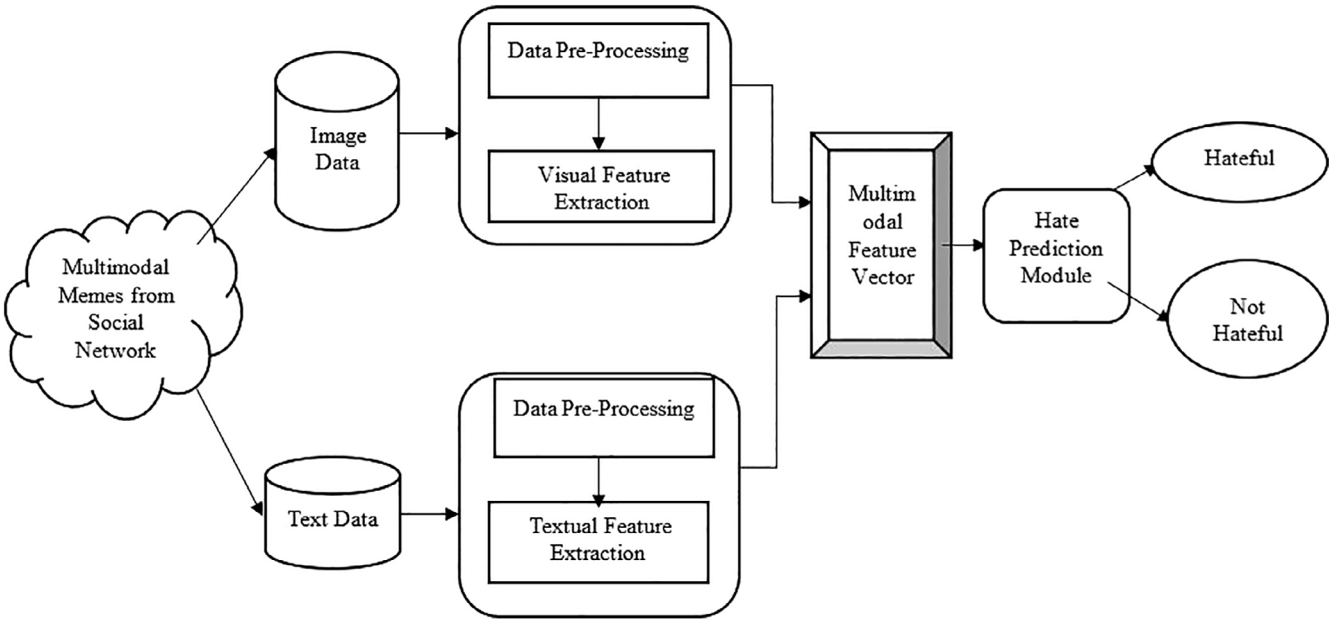

FIGURE 5 Process flow of multimodal hate speech detection using machine learning.

and gender, in memes to enhance the hateful content classification. Backward guiding approach and an adaptive weight adjustment strategy were employed in a different study by Ma et al. (2022) to record consistency and variability between various modalities. Without using extra data or labels created by humans, the self-supervised unimodal auxiliary label generation module improved the feature learning capabilities. For text and picture categorization, BERT and RESNET serve as the foundation. Hateful memes (Chen & Pan, 2022) cannot be correctly recognised by examining only embedded text or graphics. To identify a hostile meme, the system must have excellent visual and language fusion capabilities. Using a public dataset, the Transformer-based Vision-Language pre-training model OSCAR+ attained a 0.768 AUROC score. A significant alignment between textual and visual information was achieved using the optical character recognition (OCR) technology and the detection model VinVL. Several techniques, including Convolutional Neural Network (CNN), Bidirectional Encoder (BERTF) and stacked LSTM, have been used (Khedkar et al., 2022) for identification. In order to help, optical character recognition (OCR) technology was applied. The study found that the speaker's emotional state and how it affected the spoken words were the most crucial characteristics in classifying hate speech. The results show that including emotional attributes significantly improves the ability to identify hateful multimedia content over text-based models. Also, this work developed a new Hate Speech Detection Video Dataset (HSDVD), which was compiled for multimodal learning because no such dataset is currently available. It focus primarily on language which records semantic data and vocal signal which encodes paraverbal data through voice tone, pitch and tempo. Detailed Summary of advanced multimodal fusion methods is prepared by Gandhi et al. (2023). In Table 5 A detailed summary of various recent hate speech detection research is given.

TABLE 5 Summary of multimodal hate speech detection.

References	Model name	Year	Dataset	Fusion technique	Key objective	Accuracy	Advantages	Limitation
(Kiela et al., 2020)	ViLBERT CC, Visual BERT COCO	2020	-	Early fusion	Multimodal (Text and Image) hate speech detection	65.90, 69.47	Works exclusively with multimodal data	Advanced fusion methods can be explored
(Suryawanshi et al., 2020)	CNNText + VGG16	2020	MultiOff	Early fusion	Offensive meme detection	0.48	New method for offensive meme classification	Limited to image and text, not suitable for multi-domain data
(Vlad et al., 2020)	ResNet152 + VGCNItalianBERT	2020	-	Late fusion	Meme and hate speech detection	0.87	Utilizes meme meta-information for improved performance	Specific to Italian language, More advanced visual architectures could be employed
(Mookdarsanit & Mookdarsanit, 2021)	Multi-tasking GPT-2	2021	Wisesight's Dataset	-	Thai hateful text Understanding	0.78	Improved localization and recognition techniques	Specific to Thai language, Consider cultural elements for classification.
(Lee et al., 2021)	DisMultiHate	2021	FHM, MultiOFF	-	Enhanced categorization of hate speech in memes	75.8	Extracts target information effectively	Consider adding sophisticated reasoning modules for better categorization
(Ma et al., 2022)	-	2022	FAI Multimodal Hateful Memes Dataset	Late fusion	Offensive meme identification	66.30	Improved generalization and feature learning capabilities	Explore advanced fusion techniques and backbone models
(Chen & Pan, 2022)	OSCAR + RF	2022	FHM Dataset	-	Crossmodal learning for memes	68.73	Enhanced efficacy with better alignment of textual and visual information	Model re-training needed for evolving hateful messages
(Khedkar et al., 2022)	Stacked LSTM + VGG16	2022	FHM Dataset	-	OCR-based text processing for meme classification	52.9	Utilizes pre-trained word embeddings for contextual meaning	Different Image Extraction and Fusion methods can be explored.
(Rana & Jha, 2022)	MML	2022	HSDVD	Joint representation	Hate speech detection in videos	92.62	Multimodal framework analyzing speaker's tone and words	Consider visual features for improved detection.
(Zhuang & Zhang, 2022)	-	2022	Memotion 2.0 Dataset	Feature fusion/ early fusion	Text-image model for meme classification	0.4891	Utilizes VGG-16 and GloVe for text-image models	Improved multi-modal feature fusion techniques can be applied.

TABLE 6 Label classification in Indonesian tweet dataset.

Label	Occurrence	Percentage
General	5561	42
Abusive	5043	38
Individual	3575	27
Group	1986	15
Religion	793	6
Race	566	4
Physical	323	2
Gender	306	2
Other	3740	28

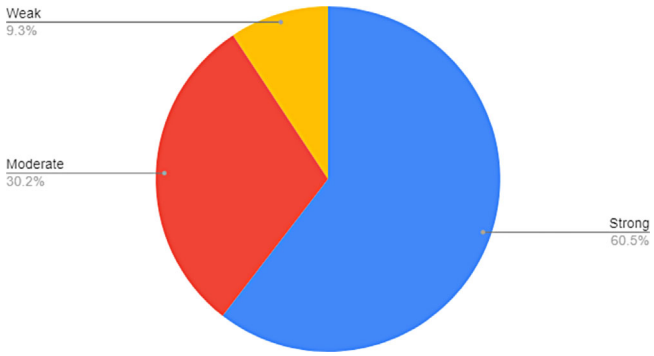

FIGURE 6 Distribution of hurt intensity.

5 | IMPLEMENTED ARCHITECTURE

The tendency of previous models to classify information into only two categories, 'hate' and 'non-hate', was one of the most notable concerns about those models. The difficulty in detecting hate speech might be seen as a problem of multi-class categorization. Another issue is that the intensity of hatred indicated in statistics is not always acknowledged. This makes it harder for platforms to defend why they block particular types of content. The implemented model is centred on these two challenges.

5.1 | Benchmark dataset

An Indonesian tweet dataset containing abusive language and hate speech is used in this model (Ibrohim & Budi, 2019). The data were obtained from Twitter. This is a dataset with multiple labels, with categories such as 'Hatred Speech', 'Abusive', 'Individual hate', 'Group hate', 'Religious hate', 'Race-based hate', 'Physical attribute-based hate', 'Gender-based hate' and 'other' types of hatred. The distribution of labels for various tweets is given in the Table 6. There are three further classes in the dataset that describe the level of hatred as weak, moderate and strong, respectively. Their general distribution in the dataset is available in Figure 6 and wordcloud from the dataset is shown in Figure 7.

5.2 | Pre-processing

The first step in the pre-processing stage is to eliminate any special characters and unnecessary spaces. Stopwords are removed during the following phase with the help of the Indonesian Stopwords list, which can be accessed at (https://github.com/stopwords-iso/stopwords-id). Sastraw Python is used for stemming. Sastrawi Python is a simple module that allows users to abbreviate derivative words in Bahasa Indonesia (the Indonesian language) to their simplest form (stem). A range of data exploration approaches are used to check for imbalances in the data and to verify the data.

FIGURE 7 Wordcloud derived from the dataset containing hurtful words.

5.3 | Feature generation and extraction

For this model two new derived features are generated, hurtfulness and intensity of hurt. Hurtfulness is derived from the various hate classes and intensity is derived from weak, moderate and strong hate labels in the dataset, following Equation (1) is for hurtfulness and Equation (2) is for the intensity of hurt

$$\sum_{i=0}^{n} l_1 + l_2 + \ldots l_m \text{ where } i, m > 0 \tag{1}$$

$$\sum_{i=0}^{n} wk_i + (md_i{}^{*}5) + (st_i{}^{*}10) \text{ where } i > 0 \tag{2}$$

Here 'n' is the number of tweets in Equations (1) and (2). 'l' is the label in Equation (1) where there are a total of m labels. In Equation (2), wk represents the weak class value, md represents the moderate class value, and st represents the strong class value. In order to extract features, multiple methods are implemented across all of the models. Bag of Words and distributed Bert are utilised in deep learning model building. In our feature extraction process, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) representation in the logistic regression base model. TF-IDF is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents or a corpus. The rationale behind using TF-IDF stems from its ability to downweight common terms (which are frequent across many documents) while giving more importance to terms that are significant in a specific document. In the context of hate speech detection, common terms that may not carry specific sentiments or intent get lower weights, while terms that might be indicative of hate speech get appropriately emphasized. By employing TF-IDF, we aimed to capture the nuanced differences between standard language usage and potential hate speech patterns, making it a valuable tool in our feature generation process.

5.4 | Model training

To discern the efficacy of different techniques in hate speech detection, we extracted features from a dataset containing tweets. The initial approach employed logistic regression, with two distinct variants: one that permits duplicate terms and another that filters them out. Our exploration then pivoted towards deep learning techniques. The first of these models relied on a Bag of Words representation for feature extraction, followed by training using an LSTM. Another sophisticated approach integrated the power of distributed BERT (Sohn & Lee, 2019) in tandem with a BiLSTM attention-based CNN (Khan, Fazil, et al., 2022).

5.5 | Experimental setup

All experiments ran in an environment powered by an NVIDIA GTX 1080 Ti GPU and an Intel i7 processor. The setup used Ubuntu 18.04 with Python 3.7, and the models were implemented and evaluated using TensorFlow v2.4 and Scikit-learn v0.23. For our experiments, we employed a random split, allocating 80% of the data for training and 20% for testing, ensuring that the distribution of labels in both splits mirrored the original

dataset. Our baseline model utilized a basic Bag of Words representation trained on a logistic regression classifier. In contrast, one variant incorporated TF-IDF weighted features, motivated by its potential to emphasize discriminative terms. Another variant explored the capabilities of LSTM, drawing from recent successes in sequence modelling for text data.

5.6 | Results and discussion

Model evaluation leaned on three pivotal metrics: accuracy (Equation 3), recall (Equation 4) and the F1-Score (Equation 5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

While accuracy offers a broad perspective on performance, it may falter in imbalanced datasets, potentially yielding inflated results by primarily predicting the majority class. To circumvent this, recall, which zeroes in on the accurate prediction of actual positive cases, and F1-Score, which harmonizes recall, was also employed. For a more granular evaluation, our dataset, with its driving features, was used for training several deep learning models. The initial model, an LSTM, underwent four variations, each with distinct features and configurations. The second sequential deep learning model, leveraged an attention-based mask coupled with distributed BERT to enhance its performance. Detailed accuracies, F1-Score and Recall for each model are presented in Table 8. Notably, the LSTM-based model outperformed its counterparts. The amalgamation of intensity score and hurtfulness with other features amplified accuracy. However, their isolated use curtailed the model's accuracy.

By employing these metrics, we aim to provide a comprehensive assessment of our models, ensuring that they not only predict accurately but also capture the nuances of the minority class effectively. Results are given in Table 7. This dataset with driving features is also used in the training of several deep-learning models. The first model was an LSTM model. Four variations are implemented in this model. The baseline features from the dataset are used in the first variation. The categorical cross-entropy loss function is employed in this model, and the Adam optimizer is used for optimization. Two more models are utilised for comparison, with only hurtfulness and hurt intensity being employed. The binary cross entropy loss function is employed in these two models. All features, including derived features, are incorporated in the final model, the activation function is softmax, and categorical cross entropy is employed for multi-class classification. The second model is a sequential model which uses an attention-based mask and distributed BERT to improve model performance. Table 8 contains the accuracy received for all of these

TABLE 7 Performance of logistic regression-based model.

Evaluation	Model_Type	HS	Abusive	HS_ Individual	HS _Group	HS_ Religion	HS _Race	HS_ Physical	HS_ Gender	HS_ Other
Accuracy	Repeated_ World	0.82155	0.89511	0.80066	0.87186	0.9402	0.96678	0.97437	0.9739	0.8187
	Remove_Repeated	0.56194	0.61035	0.71381	0.82962	0.91837	0.944	0.95491	0.95776	0.701
Recall	Repeated_World	0.72698	0.7949	0.40635	0.22917	0.1295	0.19753	0.10345	0.03704	0.47209
	Remove_Repeated	0.02677	0.02549	0.02843	0.02381	0.02158	0.01235	0.01724	0.03704	0.02871
F1-Score	Repeated_World	0.72698	0.7949	0.40635	0.22917	0.1295	0.19753	0.10345	0.03704	0.47209
	Remove_Repeated	0.02677	0.02549	0.02843	0.02381	0.02158	0.01235	0.01724	0.03704	0.02871

TABLE 8 Result comparison of deep learning based models.

Model	Accuracy	Recall	F1-Score
LSTM + baseline	0.7063	0.65	0.68
LSTM + hurtfulness	0.135	0.80	0.25
LSTM + hurt_intesity	0.255	0.75	0.35
LSTM + hurtfulness + hurt_intesity	0.8789	0.85	0.86
Sequential + attention	0.8347	0.80	0.82

models. The LSTM-based model has the highest accuracy. Using intensity score and hurtfulness in combination with other features improves accuracy, but when used alone, the model's accuracy decreases. All sequential models produce good results, and we used distributed BERT in that model. From the following results it can be observed that adding hurtfullness and intensity of hurt can significantly improve the results of previous work. In the future, we intend to create a model that combines distributed BERT with LSTM and these derived features. These deep learning models can provide better results if trained on a machine with higher computer performance.

There was a presence of class imbalance in our dataset, where three class labels significantly outnumber the others. Such imbalances can lead to classifiers that are biased towards the majority class, potentially compromising the detection of the minority class, which is often of greater interest, especially in tasks like hate speech detection. To address this, we explored both oversampling and downsampling techniques. Oversampling, specifically the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), was used to create synthetic samples in the minority class, balancing out the class distribution. Conversely, downsampling involved randomly reducing the majority class instances to match the minority class count. After employing these resampling techniques, we retrained our models and observed a 3% increase in recall for the minority class with a slight decrease in overall accuracy. While these techniques can help mitigate the effects of class imbalance, they also introduce other considerations, such as the potential overfitting introduced by oversampling. Given these findings, future work will delve deeper into optimizing the balance between classes and investigating more advanced techniques for imbalance mitigation.

6 | CURRENT ADVANCEMENTS IN THE FIELD

The landscape of research in this domain has expanded significantly, reaching a magnitude that makes summarizing every model an impractical endeavour. In 2021 alone, approximately 135 papers delved into this field, followed by 132 papers in 2022, and an additional 01 paper until January 2023. Within this vast expanse, this section highlights selected automatic hate speech detection models that have undergone benchmarking.

In the study by Zhou et al. (2021), they explored the use of multitask learning for detecting hate speech. They combined different hate speech features using a special attention method, and various parts of the model shared information to improve the understanding of sentiments. This allowed them to make better use of both the data about the sentiments of the target and external sentiment resources. Their findings showed that sharing knowledge about sentiments improved the performance of hate speech detection compared to the basic approach.

For reliable hate speech detection estimates the study by Miok et al. (2022) uses a Bayesian approach with Monte Carlo dropout in transformer model attention layers is recommended. This addresses the challenge of multilingual hate speech detection. Monte Carlo dropout, specifically Mutiscale Contextual Dual (MCD) within transformer attention layers, is employed during both training and prediction phases, resulting in two novel architectures: BAN (fully trained from scratch with dropout) and MCD BERT (dropout with a pre-trained BERT model during tuning and prediction).

In study by Hartvigsen et al. (2022), they introduced TOXIGEN, a novel dataset containing 274k statements about 13 minority groups, spanning harmful and helpful content. They used a large language model to generate such text. TOXIGEN covers implicit harmful content across various demographic groups more comprehensively than previous human-written resources. Human evaluators had difficulty distinguishing between machine-generated and human-written text in TOXIGEN, with 94.5% of toxic cases identified as hate speech. The study also showed significant improvements in fine-tuning a toxicity classifier on public datasets and highlighted TOXIGEN's potential in combating artificial toxicity, particularly for MoH (Map only Hindi)-2022. The issue of code-switching text in the area of hate speech detection is studied by Sharma et al. (2022). Analysis of hate speech in Hindi-English code-switched language is the main goal of the research. The approach analyses text representation strategies to capture accurate text. 'MoH' or (Map Only Hindi), which stands for 'Love' in Hindi, to contain the structure of data while still allowing it to be used with current algorithms. The 'MoH' pipeline uses Multilingual Bert and MuRIL language models that have been adjusted after language identification and Roman to Devanagari Hindi translation utilising a knowledge base of Roman Hindi terms. There were several experiments performed. The performance of 'MoH' mapped text is studied in the first experiment using traditional machine learning models, which results in an average improvement of 13% in F1 scores. The second indicates a 6% improvement in performance when comparing the proposed work's scores to those of the baseline models. The third contrasts the suggested 'MoH' approach with different data simulations made with the use of the current transliteration library. Here, 'MoH' performs 15% better than the competition. DSR (Design Science Research)-2022, the design of user interfaces for AI-based design assistance systems and how users perceive it in the context of hate speech detection are key real-world challenges that have been addressed in the work by Meske and Bunde (2023), for which a scientifically supported and evaluated design is provided. Design knowledge that had been generated and assessed was codified as prescriptive knowledge and condensed into an explanatory design theory. A collection of reusable design principles has been created by them. Furthermore, 86% of the 641 participants who had personally experienced hate speech underscored the necessity for greater research in the battle against hateful material in the digital arena, which is highlighted by this study. In study by García-Díaz et al. (2023), the authors explores the effectiveness of hate speech detection in Spanish by analyzing key characteristics and their potential combinations for more precise algorithms. It examines the linguistic traits associated with different categories of hate speech. It suggests future research directions which includes incorporating cross-validation techniques into the experimentation pipeline, addressing sub-

tasks related to hate speech such as target identification and media aspects, focusing on longer content, and emphasizing mistake analysis and explainability tools. The study also aims to employ tools like SHAP (SHapley Additive explanations) to assess the contribution of each feature within the neural network. While model-independent metrics have been used to validate the use of linguistic characteristics for hate speech categorization, these features are evaluated separately from the neural network.

In study by Chiril et al. (2022), they advocate a multifaceted approach to hate speech detection. It explores models that identify not only broad topics like racism, xenophobia, sexism, and misogyny but also specific hate speech targets. This approach moves beyond binary classification to achieve a finer level of granularity. The study also assesses the impact of affective information from resources like HurtLex, SenticNet and EmoSenticNet on identifying hate speech instances. The findings indicate that a multi-task approach outperforms single-task models in detecting both the hatefulness and topic of a tweet within a multi-label classification framework. Models incorporating EmoSenticNet emotions, SenticNet's first-level emotions, or a blend of SenticNet and other topic-specific datasets perform better than single-task models in detecting both the topic and hatefulness of a tweet.

In study by Mehta and Passi (2022), two datasets were used in this research study's demonstration of XAI-based hate speech identification. Data pre-treatment was done to remove any discrepancies from the data, clean the tweet's; content, tokenize and lemmatize the text, and more. Moreover, categorical variables were condensed to produce a clean dataset for training. The datasets were subjected to exploratory data analysis in order to find distinct trends and insights. The Google Jigsaw dataset was subjected to the application of a number of previously developed models, including decision trees, k-nearest neighbours, multinomial naive Bayes, random forests, logistic regression, and long short-term memory (LSTM), with LSTM achieving the highest accuracy of 97.6%. The HateXplain dataset was subjected to explainable methodologies like LIME (local interpretable model—agnostic explanations). Using the ERASER (evaluating rationales and simple English reasoning) benchmark, BERT + ANN (Artificial Neural Network) and BERT + MLP (Multi-Layer Perceptron) variants of the BERT (bidirectional encoder representations from transformers) model both achieved accuracy of 93.55% and 93.67%, respectively.

The study presented in Karayiğit et al. (2022), provides a Multilingual Bidirectional Encoder Representations from Transformers (M-BERT) model that is trained to recognise homophobic or related hate speech in Turkish comments on social media (i.e., sexist, severe humiliation, and defecation expressions). Instagram comments were gathered to create the Homophobic-Abusive Turkish Comments (HATC) dataset, which was used to train the detection algorithms. The Abusive Turkish Comments (ATC) dataset that was created in our previous work was integrated with the HATC dataset after being manually classified at the sentence level. The resampling approach was employed to balance the HATC dataset, and two versions of the dataset—resHATC and original HATC—were utilised in the tests. The M-BERT model was then evaluated for performance against DL-based models, including Long-Short Term Memory, Bidirectional Long-Short Term Memory (BiLSTM), Gated Recurrent Unit, Traditional Machine Learning (TML) classifiers, including Support Vector Machine, Naive Bayes, Random Forest, and Ensemble Classifiers, including Adaptive Boosting, eXtreme Gradient Boosting, and Gradient Boosting. Performance criteria such as F1-score, accuracy, and recall were used to assess the detection model's performance. The M-BERT model on the HATC dataset produced the best results.

The study presented in Khan, Fazil, et al. (2022), a unique BiLSTM with CNN and Hierarchical ATtention-based deep learning model called BiChat is introduced. It is used to learn Twitter representations in order to identify hate speech. The tweets are entered into the proposed model, which then runs them through a BERT layer and an attention-aware deep convolutional layer. An attention-aware Bidirectional LSTM network is used to further process the convolutional encoded representation. Via a softmax layer, the model assigns the tweet either a hostile or neutral label.

In study by Hoang et al. (2023), the first human-annotated corpus with 26k spans on 11k comments is the Vi-HOS (Vietnamese Hatred and Offensive Spans) dataset. There are precise annotation standards and definitions of offensive and hostile spans in Vietnamese. Several cutting-edge models are used in experiments. Particularly, XLM-RLarge and PhoBERTLarge both had the greatest F1-scores for Single span identification and All spans detection, respectively. Lastly, for future study, our error analysis highlights the challenges in identifying particular types of spans in our data.

In study Nagar et al. (2023), they proposed a system inclusive of textual, social context, and authorial language aspects. This strengthens the capturing of intricacies of hate speech and increase detection accuracy. This paradigm blends text analysis with social network analysis to formalise the notion that someone's offensive content is influenced by their social circle. To jointly learn the unified features of writers using a social network, language features, and profile information, this method employs a variational graph auto-encoder. Also, the framework is versatile and may include any text encoder as a plug-in to acquire the textual properties of the content in order to support current and forthcoming language models. Model generates good results for two different Twitter datasets.

In this study Awal et al. (2023), a new model HateMAML is developed which is a model-independent, meta-learning-based system that efficiently detects hate speech in languages with limited resources. HateMAML employs a self-supervision technique to get beyond the problem of data scarcity and delivers improved LM initialization for quick cross-lingual transfer or domain generalisation to other datasets of hate speech. In five datasets spanning eight different low-resource languages, extensive experiments are run. According to the results, HateMAML outperforms the cutting-edge baselines in the cross-domain multilingual transfer situation by more than 3. In this study Paula et al. (2023), they applied six transformer models to the challenge of detecting hate speech in Arabic: AraBERT, AraELECTRA, Albert-Arabic, AraGPT2, mBERT and XLMRoBERTa. In order to aggregate the output of the transformer and enhancE THE outcomes, Majority Vote and Highest Sum ensembles were

utilized. Majority Vote ensemble outperformed all other models when utilising the five-fold cross-validation method on the training data. By applying the test data as the foundation, Majority Vote to carry out the official projection was employed.

7 | LIMITATIONS

One of the major limitations in automatic hate speech detection is the lack of standardized datasets. The development of these systems heavily relies on labelled datasets, but the absence of a standardized dataset for hate speech detection makes it difficult to compare the performance of different models. Another challenge is identifying the context in which hate speech is used. Since hate speech is often context-dependent, understanding the context in which it is used is crucial for accurate detection. However, automatic hate speech detection systems can struggle with identifying the context, leading to false positives or false negatives. Bias is another limitation in hate speech detection, as the training data may be imbalanced or biased towards certain groups, leading to biased results. Additionally, the algorithms used in these systems can be biased, leading to unfair or inaccurate results. Hate speech can take many forms, including subtle and indirect expressions, which automatic hate speech detection systems may struggle to identify accurately. Multilingualism can be a significant challenge for these models, as the nuances and complexities of each language can vary greatly. Hate speech can also vary based on social and cultural differences. Detecting hate speech requires an understanding of these differences, which can be challenging for models to grasp. The evolution of language, including the emergence of new words and phrases, can pose a challenge for hate speech detection models, as they may not be able to keep up with the changes in language. Hate speech can often be communicated through ambiguous language that makes it difficult for machines to detect. For example, sarcasm, irony or humour can be used to mask hate speech, making it hard for algorithms to identify it.

8 | FUTURE DIRECTIONS

Future work on automated hate speech detection will prioritise multilingualism. Because hate speech appears in a variety of languages, it is vital to construct models that can recognise hate speech in a variety of languages. Another priority should be to improve contextual knowledge of hate speech, which includes detecting sarcasm, irony, and other forms of indirect hate speech. Explainability is important for future development, as it is crucial to create user confidence by providing explicit explanations of the detection system's decisions. The building of robust models that are resilient to adversarial attacks and can detect hate speech even when attackers aim to avoid detection also needs to be focused on. Another focus should be on real-time detection, as detecting hate speech in real-time is critical for preventing harm and promoting online safety. It is also crucial to develop multimodal algorithms capable of recognising hate speech across several modalities, such as text, images, and videos. Furthermore, addressing bias in both the data and the algorithms used in automatic hate speech detection systems is an important topic for future research in order to ensure fair and accurate results. With these kinds of advancements, there are several potential applications for automatic hate speech detection that can be built in the future. A social media monitoring tool that detects hate speech in real time and flags it for examination by human moderators could be one such use. A chatbot that can intervene in online chats and give users tools to assist them in comprehending the harm caused by hate speech could be another application. Furthermore, automatic hate speech detection can be included in messaging apps and online forums to prevent hate speech from spreading and toxic environments from developing. There is no doubt that as technology advances, more novel applications for automatic hate speech detection will be developed in order to promote safer and more inclusive online communities. Not only limited to hate speech only, speech can be classified using different labels for identification of Online sexual violence, suicidal ideation also. Some research studies including (Khatua et al., 2018), has focused on identification of gender-based violence on Twitter using #Metoo movement. Deep learning-based lexical approaches were used to categorise sexual assaults in terms of their locations and perpetrators. This method can distinguish between many groups with a fair amount of accuracy. However, a thorough investigation demonstrates that several instances in a tweet can make classification difficult. Future research should go deeper into it. In another study Khatua et al. (2019), to better understand the public acceptability of sexual minorities in the Indian setting, LGBT-related Twitter discussion were analyzed using DL-based algorithms. Proposed aspect extraction method enables to comprehend the root causes of why a particular segment of society is wary of the LGBT population. From a broad perspective, Ji et al. (2020) examines the various techniques for detecting suicidal ideation, including clinical approaches like patient-physician interaction and medical signal sensing, textual content analysis techniques like lexicon-based filtering and wordcloud visualisation, feature engineering techniques like tabular, textual, and affective features, and deep learning-based representation learning techniques like CNN- and LSTM-based text encoders. They introduce four key domain-specific applications: questionnaires, electronic health records, suicide notes, and online user content. Further directions are also suggested, including the use of cutting-edge learning strategies, interpretable intention understanding, temporal detection, and proactive conversational intervention. In another study (Kansara & Adhvaryu, 2022), one other important problem fake news detection is also discussed in detail. In another study Jaafar and Lachiri (2023), for another serious concern of detecting aggression in surveillance multimodal fusion is used. It combines four multimodal methods, including audio, video, text, and extra-information with

Deep Learning techniques. They analysed how acoustic, visual, and textual features are combine, as well as a group of five meta-features that have an impact on the fusion process and the level of aggression.

9 | CONCLUSIONS

Throughout the last decade, there has been a lot of interest in hate speech detection research. Studies have shown that it is particularly useful for automatic identification and categorization of offensive content. A number of noteworthy contributions, such as the application of text and picture fusion approaches to improve efficiency and performance, have been found. Additional categories in which advancements in hate speech detection have been recorded include: variation in the number of unimodal or bimodal modalities, variation in the number of languages employed, and hate speech recognition in text as well as multimodal data. This timely assessment highlights recent advancements in hate speech detection architectures. Latest advances in the discipline were reviewed. Ultimately, the implemented architecture categorises hate content into many categories and classes using classic machine learning and deep learning methods. These models dramatically improve results.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Ankita Gandhi https://orcid.org/0009-0007-0464-1728
Pooja Shah https://orcid.org/0000-0002-9478-7021
Amir Hussain https://orcid.org/0000-0002-8080-082X

REFERENCES

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491.

Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *IEEE Access*, *9*, 106363–106374.

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

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Alzamzami, F., & El Saddik, A. (2021). Monitoring cyber SentiHate social behavior during COVID-19 pandemic in North America. *IEEE Access*, *9*, 91184–91208.

Arango, A., Pérez, J., & Poblete, B. (2019). *Hate speech detection is not as easy as you may think: A closer look at model validation* (pp. 45–54).

Awal, M. R., Lee, R. K. W., Tanwar, E., Garg, T., & Chakraborty, T. (2023). Model-agnostic meta-learning for multilingual hate speech detection. *arXiv preprint arXiv:2303.02513*.

Baydogan, C., & Alatas, B. (2021). Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. *IEEE Access*, *9*, 110047–110062.

Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*, *10*, 121133–121151.

Boishakhi, F. T., Shill, P. C., & Alam, M. G. R. (2021). *Multi-modal hate speech detection using machine learning*. IEEE (pp. 4496–4499).

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Burnap, P., & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*, 1–15.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). *I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language* (pp. 6193–6202).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y., & Pan, F. (2022). Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLoS One*, *17*(9), e0274300.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety*. IEEE (pp. 71–80).

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, *14*, 322–352.

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285–294.

Curry, A. C., Abercrombie, G., & Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. *arXiv preprint arXiv:2109.09483*.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). *Hate me, hate me not: Hate speech detection on Facebook* (pp. 86–95).

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1–30.

Fosler-Lussier, E., Riloff, E., & Bangalore, S. (2012). Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies.

Gandhi, A., Adhvaryu, K., & Khanduja, V. (2021). *Multimodal sentiment analysis: review, application domains and future directions*. IEEE (pp. 1–5).

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444.

García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., & Valencia-Garca, R. (2023). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. *Complex & Intelligent Systems*, *9*(3), 2893–2914.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230.

Hana, K. M., Al Faraby, S., Bramantoro, A., et al. (2020). *Multi-label classification of Indonesian hate speech on twitter using support vector machines*. IEEE (pp. 1–7).

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Hoang, P. G., Luu, C. D., Tran, K. Q., Van Nguyen, K., & Nguyen, N. L. T. (2023). ViHOS: Hate speech spans detection for Vietnamese. *arXiv preprint arXiv:2301.10186*.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the Instagram social network. *arXiv preprint arXiv:1503.03909*.

Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science*, *135*, 222–229.

Ibrohim, M. O., & Budi, I. (2019). *Multi-label hate speech and abusive language detection in Indonesian Twitter* (pp. 46–57).

Jaafar, N., & Lachiri, Z. (2023). Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, *211*, 118523.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, *8*(1), 214–226.

Jiang, A., Yang, X., Liu, Y., & Zubiaga, A. (2022). SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, *27*, 100182.

Kansara, P. M., & Adhvaryu, K. U. (2022). *Identify fake data or misinformation in near real-time using big data and sentiment analytics*. IEEE (pp. 489–496).

Karayiğit, H., Akdagli, A., & Aci, Ç. İ. (2022). Homophobic and hate speech detection using multilingual-BERT model on Turkish social media. *Information Technology and Control*, *51*(2), 356–375.

Khan, S., Fazil, M., Sejwal, V. K., et al. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4335–4344.

Khan, S., Kamal, A., Fazil, M., et al. (2022). HCovBi-caps: Hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network. *IEEE Access*, *10*, 7881–7894.

Khatua, A., Cambria, E., Ghosh, K., Chaki, N., & Khatua, A. (2019). *Tweeting in support of LGBT? A deep learning approach* (pp. 342–345).

Khatua, A., Cambria, E., & Khatua, A. (2018). *Sounds of silence breakers: Exploring sexual violence on Twitter*. IEEE (pp. 397–400).

Khedkar, S., Karsi, P., Ahuja, D., & Bahrani, A. (2022). *Hateful memes, offensive or non-offensive!* (pp. 609–621). Springer.

Kiela, D., Firooz, H., Mohan, A., et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, *33*, 2611–2624.

Kiilu, K. K., Okeyo, G., Rimiru, R., & Ogada, K. (2018). Using Nave Bayes algorithm in detection of hate tweets. *International Journal of Scientific and Research Publications*, *8*(3), 99–107.

Kumar, A. J., Abirami, S., Trueman, T. E., Cambria, E. (2021). Comment Toxicity Detection via a Multichannel Convolutional Bidirectional Gated Recurrent Unit. *Neurocomputing*, *441*, 272–278.

Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Aljedaani, W., & Ashraf, I. (2022). Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble GCR-NN model. *IEEE Access*, *10*, 9717–9728.

Lee, R. K. W., Cao, R., Fan, Z., Jiang, J., & Chong, W. H. (2021). *Disentangling hate in online memes* (pp. 5138–5147).

Ma, Z., Yao, S., Wu, L., Gao, S., & Zhang, Y. (2022). Hateful memes detection based on multi-task learning. *Mathematics*, *10*(23), 4525.

Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). *Algorithms*, *15*(8), 291.

Meske, C., & Bunde, E. (2023). Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers*, *25*(2), 743–773.

Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *Cognitive Computation*, *14*, 353–371.

Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

Montariol, S., Riabi, A., & Seddah, D. (2022). Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. *arXiv preprint arXiv:2210.13029*.

Mookdarsanit, L., & Mookdarsanit, P. (2021). Combating the hate speech in Thai textual memes. *Indonesian Journal of Electrical Engineering and Computer Science*, *21*(3), 1493–1502.

Mori, Y., Takahashi, H., & Oka, R. (1999). *Image-to-word transformation based on dividing and vector quantizing images with words* (pp. 1–9). Citeseer.

Mosca, E., Wich, M., & Groh, G. (2021). *Understanding and interpreting the impact of user context in hate speech detection* (pp. 91–102).

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, *10*, 14880–14896.

Mridha, M. F., Wadud, M. A. H., Hamid, M. A., Monowar, M. M., Abdullah-Al-Wadud, M., & Alamri, A. (2021). L-Boost: Identifying offensive texts from social media post in Bengali. *IEEE Access*, *9*, 164681–164699.

Mulki, H., Haddad, H., Ali, C. B., & Alshabani, H. (2019). *L-hsab: A levantine twitter dataset for hate speech and abusive language* (pp. 111–118).

Nagar, S., Barbhuiya, F. A., & Dey, K. (2023). Towards more robust hate speech detection: using social context and user data. *Social Network Analysis and Mining*, *13*(1), 47.
Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). *Abusive language detection in online user content* (pp. 145–153).
Nurce, E., Keci, J., & Derczynski, L. (2021). *Detecting abusive Albanian.*
Oriola, O., & Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, *8*, 21496–21509.
Paula, d A F M., Bensalem, I., Rosso, P., & Zaghouani, W. (2023). Transformers and Ensemble methods: A solution for hate speech detection in Arabic languages. *arXiv preprint arXiv:2303.09823*.
Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martn-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, *9*, 112478–112489.
Qureshi, K. A., & Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access*, *9*, 109465–109477.
Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). *Redundancy, diversity and interdependent document relevance* (pp. 46–52). ACM.
Rana, A., & Jha, S. (2022). Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.
Rodrguez-Sánchez, F., Albornoz, C.-d J., & Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, *8*, 219563–219576.
Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). Multilingual Hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*.
Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, *8*, 204951–204962.
Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.
Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing* (pp. 1–10).
Shannaq, F., Hammo, B., Faris, H., & Castillo-Valdivieso, P. A. (2022). Offensive language detection in arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. *IEEE Access*, *10*, 75018–75039.
Sharma, A., Kabra, A., & Jain, M. (2022). Ceasing hate with moh: Hate speech detection in hindi–english code-switched language. *Information Processing & Management*, *59*(1), 102760.
Sharmila, P., Anbananthen, K. S. M., Chelliah, D., Parthasarathy, S., & Kannan, S. (2022). PDHS: Pattern-based deep hate speech detection with improved tweet representation. *IEEE Access*, *10*, 105366–105376.
Sohn, H., & Lee, H. (2019). *Mc-bert4hate: Hate speech detection using multi-channel BERT for different languages and translations* (pp. 551–559). IEEE.
Sood, S., Antin, J., & Churchill, E. (2012). *Profanity use in online communities* (pp. 1481–1490).
Sood, S. O., Churchill, E. F., & Antin, J. (2012). Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, *63*(2), 270–285.
Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). *Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text* (pp. 32–41).
Vashistha, N., & Zubiaga, A. (2020). Online multilingual hate speech detection: experimenting with Hindi and English social media. *Information*, *12*(1), 5.
Velankar, A., Patil, H., & Joshi, R. (2022). *Mono vs multilingual BERT for hate speech detection and text classification: A case study in Marathi* (pp. 121–128). Springer.
Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2020). Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.
Vlad, G. A., Zaharia, G. E., Cercel, D. C., & Dascalu, M. (2020). UPB@ DANKMEMES: Italian memes analysis-employing visual models and graph convolutional networks for meme identification and hate speech detection. *EVALITA Evaluation of NLP and Speech Tools for Italian*. p. 288.
Wang, C. C., Day, M. Y., & Wu, C. L. (2022). Political hate speech detection and lexicon building: A study in Taiwan. *IEEE Access*, *10*, 44337–44346.
Wang, D. (2021). Semantic representation and inference for NLP. *arXiv preprint arXiv:2106.08117*.
Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web (pp. 19–26).
Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter (pp. 88–93).
Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus* (pp. 1980–1984).
Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598.
Zhou, X., Yong, Y., Fan, X., et al. (2021). *Hate speech detection based on sentiment knowledge sharing* (pp. 7158–7166).
Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. *IEEE Access*, *8*, 128923–128929.
Zhuang, Y., & Zhang, Y. (2022). *Yet at memotion 2.0 2022: Hate speech detection combining BiLSTM and fully connected layers*.

AUTHOR BIOGRAPHIES

Ankita Gandhi is a researcher and currently pursuing her PhD in affective computing, with a focus on artificial intelligence and natural language processing. She completed her MTech in Computer Science and Technology, from L.D. College of Engineering, Gujarat Technological University in 2013. Her research areas include Deep Learning, Computer Vision and Natural Language Processing. She has published several research papers in international journals and conferences. Her work is highly appreciated by research community.

Param Ahir is a researcher currently pursuing a PhD in medical image analysis, with a focus on artificial intelligence and big data analytics. She holds a Master's degree in Information Technology from L. D. College of Engineering and has published several papers in the field of computer

vision and medical image analysis. Her work has been presented at national and international conferences and well-received by the academic community.

Kinjal Adhvaryu has received her PhD degree from School of Engineering, RK University, Rajkot, India. She is currently a principal professor in computer engineering department at Shankersinh Vaghela Bapu Institute of Technology, Gandhinagar, India. Her research interest includes Knowledge in Wireless Communication, Soft Computing, Software Engineering and Artificial Intelligence.

Pooja Shah is Assistant Professor at the Department of Computer Science and Engineering under School of Technology, Pandit Deendayal Energy University. With an experience of more than 17 years as academician she is member of ACM, life member of CSI and ISTE and senior member IEEE. Her research interests are in the domain of cyber security, image processing, machine learning and deep learning.

Ritika Lohiya has pursued her PhD from Nirma University, Ahmedabad, India. She completed her MTech in Information Network Security from the Department of Computer Science and Engineering, Nirma University in 2015. Her research areas include Machine Learning, Intrusion Detection System and Evolutionary Computation. She has published several research papers in international journals and conferences. She has been a reviewer for many international journals.

Erik Cambria is the Founder of SenticNet, a Singapore-based company offering B2B sentiment analysis services, and aProfessor at Nanyang Technological University, where he also holds the appointment of Provost Chair in Computer Science and Engineering. Prior to joining NTU, he worked at Microsoft Research Asia (Beijing) and HP Labs India (Bangalore) and earned his PhD through a joint programme between the University of Stirling and MIT Media Lab. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems and financial forecasting. He is recipient of several awards, for example, IEEE Outstanding Career Award, was listed among the AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future. He is an IEEE Fellow, Associate Editor of many top-tier AI journals, for example, Information Fusion and IEEE Transactions on Affective Computing, and is involved in various international conferences as keynote speaker, program chair and senior program committee member.

Soujanya Poria holds a PhD in Computer Science. His main areas of research interest are NLP and sentiment analysis. At present, He is working as an assistant professor at the Singapore University of Technology and Design (SUTD), Singapore. He is also part of the Institute of High-Performance Computing (IHPC), ASTAR as a senior scientist. Before joining SUTD, He has worked at NTU where I was awarded the prestigious NTU presidential postdoctoral fellowship.

Amir Hussain received his BE (highest 1st Class Honours with distinction) and PhD degrees, from the University of Strathclyde, Glasgow, UK, in 1992 and 1997, respectively. Following postdoctoral and academic positions at the Universities of West of Scotland (EPSRC postdoctoral fellow: 1996–1998), Dundee (Research Lecturer: 1998–2000) and Stirling (Lecturer: 2000–2004; Senior Lecturer: 2004–2008; Reader: 2008–2012; Professor: 2012–2018), respectively, he joined Edinburgh Napier University (ENU) in 2018 as a Professor in the School of Computing. He is founding Director of the Centre for AI and Robotics (CAIR) and Head of the Data Science and Cyber Analytics (DSCA) Research Group (managing over 20 academics and research staff). He is also founding Head of the Cognitive Big Data Analytics (CogBiD) Research Lab, and co-Lead of the Centre for Cardio-Vascular Health (with the School of Health and Social Care).

How to cite this article: Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., Poria, S., & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, e13562. https://doi.org/10.1111/exsy.13562