International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 354-368

https://doi.org/10.70917/ijcisim-2026-2693
(Review Article)

"A Comprehensive Review of Sentiment Analysis Techniques for Cyberbullying and Hate Speech Detection"

Jawahar Thakur[1], Susheela Pathania[2]

[1] Professor, Department of CS, HP University, Shimla H.P (India). E-mail: jawahar.thakur@hpuniv.ac.in
[2] Research Scholar, Department of CS, HP University, Shimla H.P(India). E-mail: spjnv01@gmail.com

(Corresponding author:spjnv01@gmail.com)
(Received: 25 November 2025; Revised: 19 March 2026; Accepted: 11 June 2026; Published: 07 July 2026)

Abstract: ABSTRACT-This paper presents a comprehensive review of sentiment analysis techniques applied to hate speech and cyberbullying detection, covering lexicon-based methods, traditional machine learning models, deep learning architectures, hybrid systems and ensemble approaches. The review emphasizes multilingual and code-mixed scenarios, sarcasm detection, data imbalance handling. Relevant studies published were collected from major scientific databases, including IEEE Xplore, ACM Digital Library, Springer, Elsevier and Google Scholar, using carefully selected keywords related to hate speech, cyberbullying, multilingual NLP, transformers and ensemble learning. Only peer-reviewed and empirically validated studies were considered in this review. The analysis reveals that transformer-based and ensemble models consistently outperform single-model architectures, particularly in multilingual and low-resource settings. Hybrid frameworks that integrate lexical knowledge with deep contextual representations demonstrate improved robustness and generalization. However, the literature also highlights persistent limitations, including weak handling of sarcasm and implicit hate, high sensitivity to dataset bias and insufficient focus on interpretability and fairness. Based on the identified gaps, this review discusses future research directions, including the development of explainable hybrid ensembles, improved sarcasm-aware architectures, dynamic lexicon integration and cross-domain multilingual evaluation. The findings of this review aim to guide researchers toward building more robust, ethical and generalizable hate speech and cyberbullying detection systems for real-world applications.

Keywords: Hate speech detection, Cyberbullying detection, Sentiment analysis, Multilingual NLP, Transformer models, Hybrid models, Ensemble learning, sarcasm detection, Explainable AI, Class imbalance

1. Introduction

In today's digital age, the explosion of user-generated content on platforms such as social media, blogs and forums has created vast opportunities for analyzing public opinion and behavior. Among the key tools developed for this purpose, sentiment analysis has emerged as a crucial subfield of Natural Language Processing (NLP), focusing on categorizing emotions in text as positive, negative, or neutral. Widely used in customer feedback, brand monitoring and market forecasting, it is now increasingly applied to detect online toxicity, including hate speech and cyberbullying. The conceptual roots of sentiment analysis lie in financial market models such as Barberis et al. (1998), with milestones by Turney (2002) and Nasukawa and Yi (2003), who introduced semantic orientation and topic-specific opinion extraction. Early methods relied on lexicon-based and machine learning approaches using predefined dictionaries and handcrafted features (Nitin et al., 2010), but struggled with sarcasm, slang and contextual ambiguity. To address these issues, deep learning techniques such as CNNs, BiLSTMs and transformer models such as BERT and XLM-RoBERTa have significantly improved accuracy and contextual understanding (Devlin et al., 2019; Conneau et al., 2020). Recently, sentiment analysis has gained prominence in identifying and preventing cyberbullying and hate speech, pushing research toward more robust, context-aware and ethically aligned detection models. The following sections provide a comprehensive literature review, identify critical research gaps, propose a hybrid sentiment analysis model and detail the research methodology used to detect hate speech and cyberbullying.

2. Review Method and Paper Selection Strategy

This review follows a structured narrative review methodology to analyze recent research on sentiment analysis, hate speech detection and cyberbullying detection. The objective of this review is to systematically categorize and compare lexicon-based, machine learning, deep learning, hybrid and ensemble approaches, with a particular focus on multilingual, code-mixed, sarcasm-aware and explainable systems.

2.1 Data Sources

Relevant research articles were collected from well-established scientific databases, including: IEEE Xplore, ACM Digital Library, SpringerLink, Elsevier ScienceDirect, arXiv and Google Scholar. These databases were selected to ensure comprehensive coverage of peer-reviewed journals, conferences and high-quality preprints in the fields of natural language processing and toxic content detection.

2.2 Search Keywords

The literature search was conducted using combinations of the following keywords: “Hate speech detection”, “Cyberbullying detection”, “Sentiment analysis”, “Multilingual NLP”, “Code-mixed text”, “Transformer models”, “BERT”, “XLM-R”, “Hybrid models”, “Ensemble learning”, “Sarcasm detection”, “Explainable AI”, “HASOC dataset”, “TRAC dataset” and “Davidson dataset”. Boolean operators (AND, OR) were applied to refine the search queries and ensure the inclusion of relevant studies.

2.3 Inclusion Criteria

The following criteria were used to select papers for this review:

a) Peer-reviewed journal articles, conference papers and high-quality preprints

b) Studies focusing on hate speech, cyberbullying, or offensive language

c) detection Papers that reported experimental evaluation using benchmark datasets

d) Studies that included performance metrics such as accuracy, precision, recall, F1-score, or MCC Research addressing multilingual, code-mixed, or low-resource language scenarios

2.4 Exclusion Criteria

Papers were excluded based on the following conditions:

a) Duplicate publications across databases

b) Non-empirical or purely opinion-based articles

c) Studies without dataset description or performance evaluation

d) Papers not directly related to hate speech, cyberbullying, or toxic content detection

e) Articles focusing solely on unrelated sentiment domains such as product reviews or financial forecasting

2.5 Review Process

Initially, titles and abstracts were screened to eliminate irrelevant studies. The remaining articles were then reviewed in full text to ensure compliance with the inclusion criteria. Each selected paper was categorized into one of the following groups: lexicon-based methods, traditional machine learning approaches, deep learning models, transformer-based systems, hybrid frameworks and ensemble techniques. For each category, key aspects such as dataset characteristics, feature representations, model architectures, evaluation metrics and reported limitations were systematically analyzed and compared. This categorization enabled a structured synthesis of existing research trends and performance patterns.

2.6 Nature of the Review

This study adopts a narrative review approach with structured categorization rather than a strict systematic review protocol. While statistical meta-analysis is not performed, the review provides a comprehensive qualitative synthesis of current methodologies, challenges and future research directions in multilingual hate speech and cyberbullying detection.

3. Related Work

Sentiment analysis has grown from early financial prediction models (Barberis et al., 1998) into a key NLP task for analyzing public opinion and online toxicity. Turney (2002) used semantic orientation for unsupervised sentiment classification and Nasukawa and Yi (2003) formalized sentiment extraction using NLP. With the explosion of social media content, sentiment analysis has become essential in domains such as hate speech and cyberbullying detection. Lexicon-based approaches estimate sentiment polarity using pre-defined dictionaries. Taboada et al. (2011) developed SO-CAL to handle negation and intensity and Jurek et al. (2015) enhanced real-time sentiment classification on Twitter. Recent tools such as LEXpander (Di Natale & Garcia, 2022) and Senti-DD (Park et al., 2021) have improved multilingual and domain-specific lexicon expansion. Alshaabi et al. (2021) used embeddings and transformers for lexicon growth, whereas Yadav and Roychoudhury (2019) showed that hybrid lexicons outperform single ones. Machine learning methods such as SVM, Naïve Bayes and decision trees are widely used. Almeida et al. (2022) achieved 93.4% accuracy with an ensemble (SVM–RF–kNN). Naseem et al. (2021) developed a lightweight multilingual SVM and Mansour et al. (2021) applied SVM to COVID-19 misinformation. Other studies (Majumder et al., 2020; Khalid et al., 2020) have shown that ensembles outperform single models.Deep learning models have improved sentiment detection by capturing the semantic and contextual cues. Sharifani and Amini (2023) proposed a CNN–BiLSTM ensemble; Liu et al. (2020) applied CNN with CBOW embeddings; and Alayba et al. (2020) used CNN–LSTM for Arabic sentiment. Hybrid and attention-based models (e.g., Yao et al., 2018; Zhang et al., 2020) have demonstrated superior performance in short and noisy texts. Kim (2014) demonstrated early success of CNNs with Word2Vec, laying the groundwork for modern deep sentiment systems. Recently, researchers have advanced traditional methods by creating models that combine multiple techniques to improve accuracy and contextual understanding. The next section reviews these approaches and their impact on sentiment analysis, hate speech detection and cyberbullying detection.

3.1 Lexicon-Based Sentimental Analysis for Cyberbullying and Hate Speech Detection

Lexicon-based approaches rely on predefined dictionaries of emotionally charged or offensive terms to detect hate speech and cyberbullying, proving particularly useful when labeled data are limited. Alex Solovev and Nicolas Pröllochs (2022) investigated the influence of moral-emotional language on the spread of hate in online discussions. Their study found that words associated with morality and emotions, such as “betrayal” and “justice,” significantly increased the likelihood of eliciting hateful replies. The findings highlighted the role of emotional contagion in shaping user interactions and amplifying toxic conversations. The authors demonstrated that emotionally charged moral language can intensify polarization and encourage hostile responses on social media platforms. This work provides valuable insights for developing hate speech detection systems that incorporate emotional and contextual cues. Hayaty et al. (2020) developed an abusive lexicon for four major tribal languages of Indonesia to address the challenge of hate speech detection in under-resourced linguistic communities. The lexicon provided a valuable linguistic resource for identifying abusive expressions across multiple indigenous languages. Experimental results demonstrated that incorporating the lexicon significantly improved hate speech detection performance, highlighting the importance of language-specific resources for multilingual and low-resource NLP applications. Madukwe et al. (2020) highlighted inconsistencies in dataset annotations and stressed the need for standard lexicon integration and preprocessing for better cross-domain applicability. Abdelzaher (2019) created a hierarchical semantic lexicon using FrameNet and WordNet, achieving an F1-score of 83.7% in detecting nuanced violent content on Facebook. Thomas Davidson et al. (2017) proposed a machine learning approach to distinguish hate speech from offensive language on Twitter. Their model combined TF-IDF features, part-of-speech (POS) n-grams, and lexicon-based features to capture both linguistic and contextual information. Using a Logistic Regression classifier, the proposed approach achieved an F1-score of 0.91, demonstrating high effectiveness in accurately identifying hate speech while reducing the misclassification of offensive but non-hateful content. This study established a strong baseline for subsequent hate speech detection research. Similarly, Samghabadi et al. (2017) employed lexicons such as LIWC and SentiWordNet alongside N-grams to improve classification accuracy in hate speech detection. Silva et al. (2016) contributed to dataset creation by identifying linguistic patterns and hate targets, laying the groundwork for later lexicon-driven detection systems. Sood et al. (2012) introduced an early lexicon-based profanity detection approach using edit distance to identify intentionally misspelled or obfuscated offensive words. This method improved the coverage of abusive language lexicons by recognizing creative spelling variations commonly found in online communications. As a result, the approach enhanced the detection of offensive content that traditional exact keyword matching often failed to identify. The study laid the foundation for more robust lexicon-based cyberbullying and hate speech detection systems.

Table 1 summarizes major lexicon-based sentiment analysis approaches for cyberbullying and hate speech detection. It compares the techniques, datasets, extracted features, strengths and limitations of existing studies. The review highlights the effectiveness of lexicon-based methods in identifying abusive and hateful content while emphasizing challenges such as context sensitivity, language dependency and scalability. It also demonstrates the

importance of domain-specific lexicons for improving classification performance across different languages and social media platforms. Furthermore, the comparison reveals that combining lexicon-based features with machine learning techniques often leads to better detection accuracy than using lexicons alone. These findings provide valuable insights for developing more robust and context-aware cyberbullying and hate speech detection systems.

Table 1: Summary of Lexicon-Based Sentimental Analysis for Cyberbullying and Hate Speech Detection

Authors (Year)	Method / Model	Dataset	Key Features	Major Findings	Limitations
Solovev & Pröllochs (2022)	Moral-emotional lexicon analysis	Online discussion forums	Moral-emotional lexicon	Demonstrated emotional contagion effects and improved hate propagation analysis	Limited contextual understanding and emotional trigger coverage
Hayaty et al. (2020)	Indigenous lexicon construction	Indonesian social media	250 abusive terms from four tribal languages	Improved hate speech detection for low-resource languages	Manual lexicon construction limits scalability
Madukwe et al. (2020)	Systematic review of hate speech lexicons	Multiple benchmark datasets	Lexicon usage and preprocessing analysis	Identified research gaps and standardization issues	Review study without experimental validation
Anonymous Indonesian Research (2019)	Lexicon + N-gram + ML (SVM, Logistic Regression)	Indonesian Twitter and Facebook data	Lexicon features, sentiment scores, word and character n-grams	Achieved 85% F1-score with improved robustness on noisy text	Generic lexicons fail to capture cultural nuances
Davidson et al. (2017)	Logistic Regression	~25,000 Twitter posts	TF-IDF, POS tags, lexicon-based sentiment	Achieved 0.91 F1-score and established a benchmark dataset	Difficulty distinguishing offensive language from hate speech

3.2 Machine Learning approaches Sentimental Analysis for Cyberbullying and Hate Speech Detection

Several traditional machine learning (ML) approaches have been applied to hate speech and cyberbullying detection using a range of lexical, syntactic and semantic features. Saleh et al. (2022) compared transformer and RNN models, finding BERT outperformed BiLSTM with a 96% F1-score versus 93%, illustrating the superior contextual understanding of transformers. Awal et al. (2021) proposed AngryBERT, a multitask model that integrates BERT with ML classifiers for sentiment and hate detection, which demonstrated robust multilingual and fine-grained classification across three datasets. Mandl et al. (2020), in the HASOC 2020 shared task, showed that classical models like SVM and Logistic Regression, when applied to English, Hindi and German texts, achieved F1-scores above 0.80, validating their competitiveness even in code-mixed environments. Salminen et al. (2020) tested Naive Bayes and Logistic Regression on Reddit and YouTube comments using TF-IDF and profanity lexicons, reporting up to 89% accuracy, with platform-specific language impacting performance. Founta et al. (2019) demonstrated that ensemble methods combining Logistic Regression, Random Forest and Gradient Boosting outperformed single models, achieving an F1-score of 0.84 on Twitter data. Similarly, Rizwan et al. (2019) applied SVM on a three-class Twitter dataset using unigrams, bigrams and polarity scores, reaching 91.7% accuracy and highlighting the utility of lexical features. Kumar et al. (2018), via the TRAC-1 task, addressed aggression detection in English and Hindi using classical classifiers, emphasizing the need for custom pipelines for multilingual, multi-script texts. Gambäck and Sikdar (2017) compared word- and character-level CNNs with ML classifiers, noting that character-level CNNs excelled in capturing subtle variations in tweets. Davidson et al. (2017) distinguished hate speech from offensive language using Logistic

Regression and lexicon features, achieving strong performance but noting context ambiguity in terms like "bitch" or "hoe". Table 2 summarizes the major machine learning approaches for cyberbullying and hate speech detection. It compares the techniques, datasets, extracted features, strengths and limitations of existing studies. The review highlights the effectiveness of traditional machine learning models and transformer-based methods in detecting abusive content while emphasizing challenges such as contextual ambiguity, computational complexity and platform-specific performance.

Table 2: Summary of Machine Learning Approach for Cyberbullying and Hate Speech Detection

Authors (Year)	Method / Model	Dataset	Key Features	Major Findings	Limitations
Saleh et al. (2022)	BERT, BiLSTM	Binary hate speech datasets	Domain-specific word embeddings	BERT achieved 96% F1-score and effectively captured nuanced hate speech	High computational cost of transformer models
Awal et al. (2021)	AngryBERT (BERT + Traditional ML)	Three multilingual benchmark datasets	Multitask learning (hate, sentiment, target)	Improved multilingual performance and contextual understanding	Complex architecture and hyperparameter tuning
Salminen et al. (2020)	Logistic Regression, Naïve Bayes	YouTube and Reddit comments	TF-IDF, profanity lexicons	Logistic Regression achieved 82–89% accuracy across platforms	Performance varied across different social media platforms
Founta et al. (2019)	Ensemble Learning (LR, RF, GB)	Twitter dataset	TF-IDF, n-grams, syntactic features	Achieved 0.84 F1-score and 87% accuracy with robust performance	Reduced model interpretability due to ensemble complexity
Rizwan et al. (2019)	Support Vector Machine (SVM)	Three-class Twitter dataset	Unigrams, bigrams, sentiment polarity	Achieved 91.7% accuracy for short informal text	Limited contextual understanding of language

3.3 Deep Learning Based Approach for Cyberbullying and Hate Speech Detection

Fillies (2023) evaluated BERT-based models—mBERT, XLM-RoBERTa and BETO—on hate speech datasets in German, Italian and Spanish, finding that these models consistently outperformed traditional SVMs across F1-score, accuracy and MCC, especially in multilingual, low-resource scenarios. Raj et al. (2022) proposed a hybrid BiLSTM-CNN model using stacked GloVe and FastText embeddings for cyberbullying detection, achieving 91.35% accuracy initially, which improved to 95.12% after tuning, confirming the efficacy of deep learning hybrids. Butt (2021) examined sexism detection in multilingual tweets, with BERT achieving the highest F1-score (78.02%) and demonstrating strong handling of emojis and multilingual content, although translation-based augmentation introduced overfitting risks. Wang et al. (2020) introduced SOSNet, a GCN model leveraging tweet similarity and Dynamic Query Expansion (DQE), which reached 92.7% accuracy and 92.58% F1-score, showcasing the value of graph-based learning. Susanty et al. (2019) used an ANN with sigmoid activation and preprocessing to enhance offensive language detection, achieving 99.18% training, 94.28% validation and 96.8% test accuracy. Ajlan et al. (2018) proposed CNN-CB, a CNN model with word embeddings that eliminated manual feature engineering and achieved ~95% accuracy in cyberbullying detection.

Agrawal et al. (2018) compared CNN, LSTM, BiLSTM and BiLSTM-Attention for abuse classification, noting BiLSTM-Attention as the most context-aware model using the F1-score and t-SNE evaluation. Zhang et al. (2016) developed a pronunciation-based CNN (PCNN) for cyberbullying detection using phonetic representations from eSpeak, achieving 98.9% accuracy and a 98.0% F1-score. They addressed class imbalance with threshold moving, enhancing the detection of noisy and informal social media data. Table 3 summarizes the major deep learning

methodologies applied for cyberbullying and hate speech detection. It compares various architectures, datasets, extracted features, strengths and limitations of recent studies. The review highlights the superior performance of deep learning models, particularly transformer- and recurrent neural network-based approaches, in capturing contextual and semantic information. However, challenges such as computational complexity, hyperparameter tuning, overfitting and multilingual generalization continue to limit their practical deployment.

Table 3: Summary of Deep Learning Based Approach for Cyberbullying and Hate Speech Detection

Authors (Year)	Method / Model	Dataset	Key Features	Major Findings	Limitations
Fillies (2023)	mBERT, XLM-RoBERTa, BETO	German, Italian and Spanish hate speech datasets	Transformer embeddings	Achieved superior multilingual performance and outperformed traditional machine learning models	No significant limitations reported
Raj et al. (2022)	Hybrid BiLSTM + CNN with GloVe and FastText	Multilingual cyberbullying dataset	Stacked word embeddings	Achieved 95.12% accuracy after hyperparameter optimization	Performance depends on extensive hyperparameter tuning
Butt (2021)	Multilingual BERT	English and Spanish tweet datasets	Contextual BERT embeddings	Effectively handled multilingual text, emojis and special characters	Susceptible to overfitting due to data augmentation
Wang et al. (2020)	Graph Convolutional Network (SOSNet)	Twitter hate speech dataset	Word embeddings, cosine similarity, dynamic query expansion	Achieved 92.58% F1-score and 92.7% accuracy with improved contextual representation	High computational complexity and training cost

3.4 Hybrid Approach for Sentiment Analysis, Cyberbullying and Hate Speech Detection

Hybrid sentiment analysis combines lexical, machine learning and deep learning techniques to enhance the classification accuracy (Shahida et al., 2019). Ilkay et al. (2018) highlighted its strength in capturing sentiment at the conceptual level. Kaur et al. (2023) proposed a hybrid feature vector (HFV) by merging review-related and aspect-related features, enabling improved handling of sarcasm and ambiguity using LSTM. Tan et al. (2022) developed an ensemble hybrid model (RoBERTa-LSTM, BiLSTM and GRU) with GloVe-based augmentation to tackle class imbalance, achieving state-of-the-art results. Lin et al. (2022) introduced LSTMGA, combining LSTM with a genetic algorithm, outperforming traditional models in financial sentiment prediction with MAPE < 5%. Dang et al. (2021) showed that BERT-based hybrids using SVM yielded superior precision, whereas Cach et al. (2021) affirmed that hybrid CNN-LSTM-SVM models outperform individual models in terms of accuracy and efficiency. Vennila et al. (2020) enhanced the CNN-SVM performance using Multi-Swarm Particle Swarm Optimization. Srinidhi et al. (2020) introduced MaLSTM, which combines LSTM and SVM for IMDB sentiment classification, showing strong performance. Salur et al. (2020) investigated the effectiveness of multiple word embedding techniques, including Word2Vec, FastText and character-level embeddings, for hate speech detection. These embeddings were integrated with deep learning architectures such as CNN, LSTM and BiLSTM to improve feature extraction and text representation. The combination of semantic and character-level information enabled the models to better capture contextual and morphological patterns in social media text. Their results demonstrated that hybrid embedding strategies significantly enhanced classification performance compared with using a single embedding technique. The study highlighted the importance of rich feature representation in improving the accuracy and robustness of cyberbullying and hate speech detection systems. Rehman et al. (2019) integrated CNN and LSTM for movie review sentiment, capturing both local and long-term dependencies. Moussa et al. (2018) used BiLSTM with attention to improve domain-specific sentiment detection by focusing on key sentiment-bearing words. Nguyen et al. (2017) applied a multichannel LSTM-CNN hybrid to Vietnamese e-commerce reviews, effectively handling sentence structure and semantics. Hybrid models have consistently demonstrated their ability to generalize across different datasets and languages. Their flexibility allows for efficient adaptation to domain-specific sentiment-analysis tasks. Feature-level fusion and ensemble classification enhance the robustness and mitigate overfitting. These models also address the limitations of shallow architectures by incorporating context awareness.

Table 4 summarizes the major hybrid approaches for sentiment analysis, cyberbullying and hate speech detection. It compares the techniques, datasets, feature extraction methods, strengths and limitations of existing hybrid models. The

review highlights that combining deep learning, machine learning and transformer-based architectures significantly improves classification accuracy and contextual understanding. It also demonstrates the effectiveness of integrating contextual embeddings, attention mechanisms and optimization algorithms to enhance feature representation and predictive performance. Furthermore, hybrid models exhibit greater robustness in handling complex linguistic patterns, including sarcasm, ambiguous expressions and imbalanced datasets. Despite these advantages, challenges such as high computational complexity, increased training time and dependence on large annotated datasets remain significant. Overall, the comparative analysis provides valuable insights for designing efficient, scalable and context-aware hybrid frameworks for sentiment analysis and toxic content detection.

Table 4: Summary of Hybrid Approach for Cyberbullying and Hate Speech Detection

Authors (Year)	Method / Model	Dataset	Key Features	Major Findings	Limitations
Kaur et al. (2023)	LSTM, TF-IDF, N-gram, Emoticons, Aspect-based Features	Customer reviews	Hybrid Feature Vector (HFV)	Effectively detected sarcasm and ambiguous sentiment with improved contextual understanding	Requires manual aspect extraction
Tan et al. (2022)	RoBERTa-LSTM, RoBERTa-BiLSTM, RoBERTa-GRU, GloVe	Benchmark sentiment datasets	Contextual embeddings, data augmentation	Improved generalization and handled class imbalance effectively	Computationally intensive
Lin et al. (2022)	LSTM with Genetic Algorithm (LSTMGA)	Hybrid stock market dataset	GA-based optimization, MAPE	Achieved high forecasting and sentiment prediction accuracy	Domain-specific applicability
Dang et al. (2021)	Hybrid BERT-LSTM-CNN-SVM	Multiple sentiment datasets	Transformer embeddings, hybrid architecture	High precision through complementary deep learning models	Complex architecture and training pipeline
Vennila et al. (2020)	CNN-SVM with MSPSO	Product review dataset	CNN feature extraction, PSO-optimized SVM	Improved classification accuracy through optimization	Sensitive to SVM kernel selection
Rehman et al. (2019)	Hybrid CNN-LSTM	Movie review dataset	Convolutional and sequential feature learning	Outperformed individual CNN and LSTM models	Requires large annotated training datasets
Moussa et al. (2018)	BiLSTM with Attention	Hotel and electronics reviews	Attention-based sentiment representation	Improved context-aware sentiment classification	Requires domain-specific tuning
Nguyen et al. (2017)	Multi-channel LSTM-CNN	Vietnamese e-commerce reviews	Multi-channel word representations	Preserved sentence semantics and contextual information	Limited to Vietnamese language datasets

3.5 Ensemble Approach for Sentiment Analysis, Cyberbullying and Hate Speech Detection

Modern hate speech and cyberbullying detection systems leverage a combination of natural language processing, machine learning, deep learning and ensemble techniques. Text preprocessing steps such as tokenization, stop-word removal and normalization are essential to prepare social media text for classification models. Traditional features like TF-IDF alongside ML classifiers (SVM, Random Forest, XGBoost) form baseline models in ensemble frameworks that often outperform individual models. In parallel, transformer-based models like BERT, RoBERTa, BERTweet and HateBERT provide powerful contextual representations when fine-tuned on annotated hate/abusive datasets. Ensembling methods including hard/soft voting and stacking further enhance performance by combining diverse model outputs. These architectures are evaluated using standard NLP performance metrics such as precision, recall and F1-score in benchmark hate speech detection studies.To further improve robustness and generalization, ensemble learning techniques such as voting and stacking are widely adopted (Rokach, 2010; Zhou, 2012; Polikar, 2006). Hybrid ensemble architectures combine transformer outputs with machine learning meta-classifiers such as

XGBoost, achieving superior performance in hate speech detection tasks (Risch et al., 2024). Since hate speech datasets are highly imbalanced, oversampling methods such as SMOTE are commonly applied to address class imbalance issues (Chawla et al., 2002). Rosa et al. [54] investigated the Formspring dataset from a Kaggle competition for cyberbullying detection using CNN, CNN-LSTM and CNN-LSTM-DNN architectures with Word2Vec embeddings trained on Google News, Twitter and Formspring corpora. Their results showed that CNN-LSTM outperformed traditional machine learning classifiers such as SVM and Logistic Regression on both balanced and imbalanced datasets. Pitsilis et al. [55] incorporated user-related behavioral features, including users' tendencies toward racism and sexism, into an LSTM-based RNN ensemble using the Waseem and Hovy dataset [56]. Their approach slightly improved upon the results of Badjatiya et al. [57], achieving an F1-score of 93.2%.

Table 5: Hate Speech Detection Using Ensemble Techniques

Authors (Year)	Title	Ensemble / Method	Base Models Used	Dataset / Language	Key Contribution
Raghava et al. (2024)	Offensive Language & Hate Speech Detection Using Seven Transformers & Six Ensemble Strategies	Voting, Averaging, Stacking, Bagging, Boosting	BERT, RoBERTa, XLM-R, DeBERTa, etc.	Social Media (English)	Compared six ensemble strategies; stacking achieved best performance
Ranjan et al. (2024)	Constructing Ensembles for Hate Speech Detection	Aspect-based Ensemble	Multiple classifiers	English, Italian, Turkish	Ensemble built on different hate speech aspects
Priyadharshini et al. (2024)	Enhanced Detection of Hate Speech in Dravidian Languages Using Ensemble Transformers	Transformer Ensemble	mBERT, XLM-R, IndicBERT	Telugu, Tamil	Improved detection in low-resource Dravidian languages
Mason & Perplexity Team (2024)	MasonPerplexity at Multimodal Hate Speech Event Detection	Transformer Ensemble	XLM-R, BERTweet, BERT	Multimod al (Text + Image)	High F1-score for text-based ensemble
DravidianLang Tech Team (2024)	Social Media Hate and Offensive Speech Detection	Multi-model System	CNN, LSTM, Transformers	Telugu Code-mixed	Shared task benchmark system
cantnlp Team (Basile et al., 2024)	cantnlp@LT-EDI-2024 Anti-LGBTQ+ Hate Speech Detection	Transformer Combination	XLM-R, mBERT	Multilingu al	Focused on minority-targeted hate speech
Ghosh et al. (2024)	EnsMulHateCyb: Multilingual Hate Speech and Cyberbullying Detection	Bagging + Stacking Hybrid	CNN, BiLSTM, BiGRU	Multilingu al	Hybrid ensemble improved robustness
Al-Qarqaz et al. (2024)	Transformer Ensemble for Arabic Hate Speech Detection	Soft Voting Ensemble	AraBERT, MARBERT, CAMeLBERT	Arabic Tweets	Ensemble outperformed single models
Kumar et al. (2024)	HarmonyNet: Navigating Hate Speech Detection	Hybrid Deep Model	Multiple deep models	Social Media	Robust against noisy and imbalanced data
Modha et al. (2020 / 2024)	Hate Speech Detection Using Transformer Ensembles on HASOC Dataset	Transformer Ensemble	BERT, RoBERTa	HASOC Dataset	Baseline transformer ensemble framework

Mahata et al. [58] proposed an ensemble of CNNs, Bi-LSTM with attention and Bi-LSTM–Bi-GRU with GloVe embeddings, ranking fifth in SemEval-2019 Task-6 for abusive language detection. Sun et al. [59] presented a hybrid CNN-LSTM model combined with a Markov chain for detecting conversational anomalies in social media. Sadiq et

al. [60] applied CNN-LSTM and CNN-BiLSTM architectures for aggression detection in tweets, achieving approximately 92% accuracy. Nascimento et al. [61] developed an ensemble learning framework to analyze gender bias in hate speech detection, demonstrating better effectiveness than single baseline models. Lin et al. [62] proposed a two-stage harmful news identification framework integrating sentiment analysis with a BERT-based ensemble classifier, achieving an F1-score of 66.3%, which represented a 7.8% improvement over the TF-IDF with SVM model. Mozafari et al. [63], Liu et al. [64] and Zampieri et al. [65] demonstrated the strong effectiveness of transformer-based models in offensive and hate speech detection tasks. Wei et al. [66] and Plaza-del-Arco et al. [67] further confirmed that pre-trained language models significantly outperform conventional deep learning architectures. Foundational word embedding methods such as Word2Vec by Mikolov et al. [68], GloVe by Pennington et al. [69] and FastText by Bojanowski et al. [70] continue to support representation learning. Recurrent neural architectures including LSTM by Hochreiter and Schmidhuber [71], GRU by Cho et al. [72] and Bi-RNN by Schuster and Paliwal [73] remain widely adopted for sequential modeling. Recent improvements using attention and ensemble strategies have been reported by Ratadiya and Mishra [75], Lu et al. [74] and Srivastava et al. [76].

Table 5 presents a comparative overview of recent ensemble-based approaches for hate speech and cyberbullying detection. It summarizes the ensemble strategies, base models, datasets, languages and key contributions of state-of-the-art studies. The review demonstrates that transformer-based ensembles, including BERT, RoBERTa, XLM-R and DeBERTa, consistently outperform individual models by improving classification accuracy and robustness. These methods have shown promising results across multilingual, code-mixed and multimodal datasets. Furthermore, ensemble learning effectively addresses challenges such as class imbalance, noisy social media content and low-resource languages. However, issues including sarcasm detection, implicit hate identification, multilingual generalization, computational complexity and model explainability remain significant research challenges. The comparative analysis provides valuable insights for developing more accurate, scalable and interpretable ensemble frameworks for hate speech and cyberbullying detection.

3.6 Benchmark Datasets for Hate Speech and Cyberbullying Detection

Several publicly available benchmark datasets have been widely used for training and evaluating hate speech and cyberbullying detection models. Table 6 summarizes the most commonly adopted datasets along with their language coverage, labeling schemes and key characteristics. The HASOC dataset contains multilingual data in English, Hindi and German, annotated into three categories: hate, offensive and normal. With more than 15,000 instances, it reflects realistic multilingual and highly imbalanced distributions, making it suitable for evaluating cross-lingual robustness. The TRAC dataset focuses on aggression detection in English and Hindi social media content. It consists of approximately 12,000 code-mixed and noisy samples labeled as aggression and non-aggression and is widely used for studying multilingual and informal language challenges. The Davidson dataset, one of the most popular benchmarks, includes 24,783 English tweets labeled as hate, offensive and neither. Although large in size, the dataset is highly imbalanced, which makes it particularly useful for testing data imbalance handling techniques such as class weighting and resampling.

Table 6: Dataset Summary for Hate Speech and Cyberbullying Detection

Dataset	Language	Labels	Size	Notes
HASOC	English, Hindi, German	Hate, Offensive, Normal	~15,000	Multilingual, imbalanced
TRAC	English, Hindi	Aggression, non-aggression	~12,000	Code-mixed, noisy
Davidson	English	Hate, Offensive, Neither	24,783	Highly imbalanced
OLID (SemEval)	English	Offensive, Not Offensive	14,100	Fine-grained hierarchy
Formspring	English	Bully, non-bully	~12,000	Conversational data
Twitter Hate	English	Hate, Neutral	Varies	Short informal text

The OLID dataset released under the SemEval shared task provides 14,100 English tweets annotated with a fine-grained hierarchical labeling structure for offensive language detection. This dataset supports multi-level classification

and has become a standard benchmark for offensive language research. The Formspring dataset contains conversational English posts labeled as bully and non-bully. Its dialog-based nature makes it suitable for studying contextual cyberbullying behavior rather than isolated tweets. Finally, various Twitter hate speech datasets are widely used in the literature, typically annotated into hate and neutral categories. These datasets capture short, informal and highly noisy text, closely representing real-world social media environments. Overall, these datasets collectively cover multilingual, code-mixed, imbalanced and conversational scenarios, thereby providing a comprehensive evaluation foundation for hate speech and cyberbullying detection models.

4. Research Gap Identified

Based on the comprehensive analysis of existing literature, the following critical research gaps are identified:

a) Sarcasm and implicit hate speech remain poorly handled, as most models rely on surface-level textual cues and fail to capture hidden or ironic intent.

b) Code-mixed and multilingual datasets are limited, leading to weak generalization in real-world multilingual social media environments.

c) Domain shift is a major challenge, as models trained on Twitter often show significant performance degradation when applied to platforms such as YouTube or Instagram.

d) Class imbalance persists across almost all benchmark datasets, where hate and aggressive categories are severely underrepresented, resulting in biased predictions.

e) Explainability and fairness are largely ignored, making most detection systems black-box and ethically questionable for deployment.

f) Ensemble strategies are mostly shallow, relying on probability averaging or simple voting, with very limited exploration of systematic stacking or weighting mechanisms.

These gaps indicate that despite strong performance improvements, existing hate speech and cyberbullying detection systems lack robustness, transparency and cross-domain adaptability. Despite the advancements in hate speech and cyberbullying detection, key research gaps remain. Most models use either traditional ML or isolated deep learning without leveraging hybrid methods. Sarcasm, implicit hate speech and context-dependent language led to misclassifications. Multilingual and code-mixed texts are underexplored, with limited use of transformers such as XLM-R. Many models also lack interpretability and ignore data imbalance or lexicon-based reasoning, limiting their adaptability to evolving online languages.

To address these gaps, we propose TriBERT-HateStack (A Multilingual Transformer-Based Meta-Stacking Ensemble for Hate Speech Detection), an enhanced hybrid architecture that adds sarcasm detection, rule-based enrichment and cross-lingual generalization to the original TriBERT. It offers robust and explainable detection of explicit and implicit hate speech, including sarcasm and coded expressions. The core uses lightweight yet powerful transformers, DistilBERT for speed, or XLM-R for multilingual support, ensuring strong contextual understanding and real-time deployment readiness. Outputs are passed through two neural branches: a CNN capturing local patterns (abbreviations, profanities) and a BiLSTM that models long-term dependencies and sentiment progression. A Sarcasm Detection Module, placed between the transformer and BiLSTM layers, uses attention mechanisms and contrastive learning to identify emotionally contradictory cues typical of sarcasm. This reduces misclassifications that are often overlooked by standard DL and rule-based models. The system integrates handcrafted and lexicon-based features from the NRC Emotion Lexicon, SentiWordNet, LIWC, HateBase, POS tags, emoji presence, punctuation and uppercasing. These features form an auxiliary vector that enhances interpretability and domain knowledge. A Moral-Emotion Context Filter gives more weight to morally and emotionally charged words, improving detection in emotionally loaded conversations. To address class imbalance, the model uses Focal Loss during training and GloVe-based back-translation for minority class augmentation. A dynamic lexicon updating mechanism ensures adaptability to new slang and adversarial tokens are used. The final classification uses ensemble stacking, combining outputs from all branches into an SVM or Logistic Regression classifier to ensure interpretable decision boundaries. SHAP and LRP were used for model explainability, enabling users to trace decisions to input tokens or features. Evaluation will be conducted on benchmark datasets such as HASOC, Davidson's Twitter corpus, TRAC and multilingual corpora (Hindi, Urdu, Spanish and Indonesian). Performance will be assessed using the F1-score, accuracy, precision, recall and MCC, including a focused analysis of sarcasm-heavy and low-resource language subsets. Although significant progress has been achieved in cyberbullying, hate speech and offensive language detection using deep learning and transformer-based architectures, several critical limitations remain. Most existing studies focus on single-task classification, limiting their ability to generalize across different forms of abusive content such as hate speech, cyberbullying, aggression and harmful news. Furthermore, many ensemble approaches combine models at a shallow probability level

without effectively exploiting complementary semantic and contextual representations from diverse transformers. Another major limitation is dataset dependency. Many models demonstrate high performance only on specific benchmark datasets, such as Twitter or Formspring and suffer from performance degradation when applied to cross-domain or multilingual data. Additionally, class imbalance remains a persistent challenge, as minority classes such as hate or aggressive content are often underrepresented, leading to biased predictions despite the use of advanced architectures. Moreover, most existing works emphasize accuracy while neglecting robustness, interpretability and fairness. Issues such as gender bias, sarcasm, implicit hate and contextual aggression are still inadequately addressed. Transformer-based ensembles are rarely optimized using systematic weighting or stacking strategies and limited attention has been given to integrating sentiment, emotion and user-behavioral features into unified deep learning frameworks.

Therefore, there exists a strong need for a robust, generalized and explainable hybrid ensemble framework that can effectively capture semantic, contextual, emotional and behavioral cues across diverse datasets while maintaining high performance on imbalanced and real-world social media data. This research aims to address these gaps by proposing an optimized multi-transformer ensemble model with advanced feature fusion and evaluation strategies for comprehensive cyberbullying and hate speech detection.

5. Proposed Methodology

The proposed methodology introduces a hybrid deep-learning framework named TriBERT-HateStack, designed for robust detection of hate speech and cyberbullying in multilingual and code-mixed social media content. The framework integrates transformer-based contextual representations, probability-level ensemble fusion and meta-learning to effectively capture both explicit and implicit abusive expressions. Special emphasis is placed on sarcasm handling, multilingual adaptability and interpretability, which remain critical challenges in real-world toxic content detection. The overall architecture consists of multiple sequential stages, ranging from multilingual text acquisition to explainable final classification, as illustrated in Figure 1. Each stage is carefully designed to enhance generalization, robustness and classification accuracy across diverse linguistic and cultural contexts.

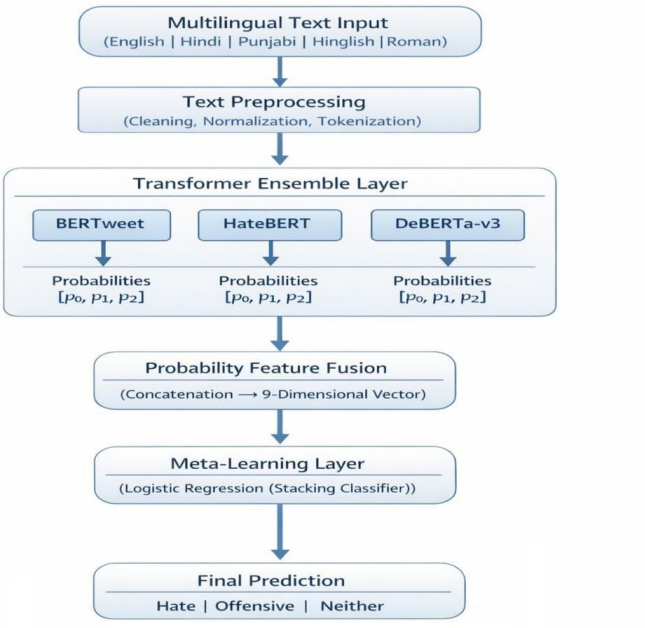

Figure 1: Propose Methodology for Hate Speech and Cyberbullying Detection.

5.1 Step.1: Data Collection and Reprocessing

Textual data are collected from multiple benchmark datasets to ensure diversity and generalization. These include:

a) HASOC (English and Hindi) for hate speech classification,

b) Davidson Twitter Dataset for hate, offensive and neutral tweet categorization,

c) TRAC (English and Hindi) for aggression detection,

Additional multilingual corpora in Urdu, Spanish and Indonesian for cyberbullying and hate content. Preprocessing is performed to normalize noisy user-generated text. The operations include removal of URLs, user mentions, hashtags, emojis, punctuation and stopwords, followed by tokenization and lemmatization. Language-specific preprocessing pipelines are applied to handle code-mixed text such as Hindi-English and Urdu-English sentences. To address class imbalance, especially for minority categories such as covert hate and indirect bullying, data augmentation techniques including back-translation and synonym replacement are employed. These steps improve model exposure to diverse linguistic variations and enhance classification robustness.

5.2 Step 2: Multilingual Text Input

The system accepts multilingual and code-mixed social media text as input. Supported languages include English, Hindi, Punjabi, Hinglish, and Romanized scripts. This capability reflects real-world communication patterns, where users frequently mix languages within a single post. Supporting such diversity enables the model to operate effectively in practical online environments.

5.3 Step 3: Text Preprocessing

The raw input text is cleaned and standardized. This stage removes irrelevant characters, emojis, URLs, and special symbols. The text is normalized and tokenized to ensure compatibility with transformer-based tokenizers. This preprocessing step significantly reduces noise and improves the quality of contextual features extracted in subsequent layers.

5.4 Step 4: Transformer Ensemble Layer

The preprocessed text is processed in parallel by three fine-tuned transformer models:

a) BERTweet, optimized for Twitter and social media language,

b) HateBERT, specialized in abusive and hateful discourse,

c) DeBERTa-v3, providing advanced contextual and semantic modeling.

Each transformer outputs a probability vector [p0,p1,p2][p_0, p_1, p_2][p0,p1,p2] corresponding to the three target classes: Hate, Offensive, and Neither. These models capture complementary linguistic and semantic perspectives, enabling richer representation of abusive language patterns.

5.5 Step 5: Probability Feature Fusion Layer

The probability vectors produced by the three transformers are concatenated to form a unified 9-dimensional feature vector. This probability-level fusion preserves the confidence information of each model and allows the ensemble to leverage their complementary strengths. The fused representation provides a more informative and discriminative input for the meta-learning stage.

5.6 Step 6: Meta-Learning Layer

The fused probability vector is fed into a Logistic Regression classifier that acts as a stacking meta-learner. This layer learns optimal weights for each transformer's prediction, effectively reducing individual model bias and improving generalization. The stacking approach enhances robustness while maintaining computational efficiency.

5.7 Step 7: Final Prediction

The meta-classifier produces the final class label as Hate, Offensive, or Neither. This output represents the most reliable and generalized decision of the TriBERT-HateStack framework.

6. Conclusion

This review presented a comprehensive analysis of sentiment analysis techniques applied to hate speech and cyberbullying detection, covering lexicon-based, machine learning, deep learning, transformer-based, hybrid and ensemble approaches. The literature consistently demonstrates that transformer models and ensemble frameworks significantly outperform traditional classifiers, particularly in multilingual and low-resource scenarios. Hybrid architectures further improve robustness by integrating symbolic and contextual knowledge. However, the review also highlights persistent challenges. Sarcasm and implicit hate speech remain difficult to detect, class imbalance continues to bias model predictions and domain dependency limits real-world generalization. Moreover, explainability and fairness considerations are still underrepresented in existing research, raising concerns regarding ethical deployment. Based on the surveyed literature, future research should focus on developing sarcasm-aware and implicit hate-sensitive architectures to improve the detection of subtle and context-dependent abusive content. There is also a need to construct high-quality multilingual and code-mixed benchmark datasets that better represent the linguistic diversity of online platforms. Future studies should emphasize domain-adaptive and cross-platform evaluation strategies to enhance the generalizability of detection models. Furthermore, integrating explainable artificial intelligence (XAI) techniques into ensemble frameworks can improve model transparency and user trust. Addressing fairness and bias through balanced training data and transparent modeling approaches is equally important for ensuring ethical and reliable predictions. Finally, systematic investigation of stacking, weighting and adaptive ensemble strategies may further improve the accuracy, robustness and scalability of cyberbullying and hate speech detection systems. Transformer-based ensembles combined with explainability techniques currently represent the most promising

direction for building reliable and ethical hate speech and cyberbullying detection systems. This review aims to serve as a structured reference for researchers and practitioners working toward safer and more inclusive online environments.

7. Acknowledgment

The authors would like to express their sincere gratitude to their supervisor for invaluable guidance, continuous support, and insightful suggestions throughout the course of this research. The authors also acknowledge the Department of Computer Science, Himachal Pradesh University, Shimla, H.P, India for providing academic support, research facilities, and a conducive environment for conducting this study.

8. Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

1. Abdelzaher, T. (2019). Hierarchical semantic lexicon using FrameNet and WordNet for detecting violent content on Facebook. Procedia Computer Science, 163, 88–97.
2. Agrawal, S., Awekar, A., & Aich, S. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In Proceedings of the European Conference on Information Retrieval (pp. 141–153).
3. Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2017). A review of support vector machines for data classification. Journal of Information Science and Engineering, 33(3), 1–17.
4. Ajlan, A. A., & Alqahtani, M. (2018). CNN-CB: A convolutional neural network model for cyberbullying detection. International Journal of Advanced Computer Science and Applications, 9(10), 143–149.
5. Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2020). Arabic language sentiment analysis on health services. Computers, Materials & Continua, 62(1), 335–351.
6. Almeida, T. A., Pinto, M. C., & Oliveira, L. S. (2022). Ensemble classification for sentiment analysis in online product reviews. Expert Systems with Applications, 196, 116522.
7. Alshaabi, T., Minot, J. R., Adams, J. L., Danforth, C. M., & Dodds, P. S. (2021). Lexicocalorimeter 2.0: Measuring collective sentiment with dynamically curated lexicons. EPJ Data Science, 10(1), 1–25.
8. Awal, M. A., Ghosh, S., & Chakraborty, T. (2021). AngryBERT: Joint learning for emotion classification and hate speech detection. In Proceedings of EMNLP 2021 (pp. 1264–1273).
9. Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. Journal of Financial Economics, 49(3), 307–343.
10. Butt, M. W. (2021). Sexism detection in multilingual tweets using BERT. IEEE Access, 9, 78945–78958.
11. Cach, C., Selamat, A., & Ali, M. (2021). Hybrid CNN–LSTM–SVM model for sentiment classification. Journal of Intelligent & Fuzzy Systems, 40(3), 1–11.
12. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
13. Dang, T., Vo, Q., & Nguyen, K. (2021). BERT-based hybrid model for sentiment classification with SVM. Information Processing & Management, 58(5), 102635.
14. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of ICWSM (pp. 512–515).
15. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186).
16. Di Natale, A., & Garcia, D. (2022). LEXpander: An automatic method for generating domain-specific lexicons for sentiment and emotions. Information Processing & Management, 59(1), 102768.
17. Fillies, F. (2023). Evaluating BERT-based models for multilingual hate speech detection. Journal of Computational Social Science, 6(2), 145–165.
18. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Vakali, A. (2019). A unified deep learning architecture for abuse detection. Proceedings of the ACM on Web Science, 3(1), 1–15.
19. Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online (pp. 85–90).
20. Hayaty, M., Rizky, A., & Wibisono, A. (2020). Hate speech detection using tribal language lexicon. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 100–107.
21. Ilkay, E., Cicekli, I., & Cicekli, N. K. (2018). Concept-based sentiment analysis using a hybrid approach. Expert Systems with Applications, 93, 386–397.
22. Jurek, A., Mulvenna, M., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. Security Informatics, 4(1), 1–13.
23. Kaur, N., Sharma, M., & Chopra, M. (2023). HFV: A hybrid feature vector for sarcasm detection in sentiment analysis.

Procedia Computer Science, 217, 1437–1444.
24. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
25. Lin, C., Wang, S., & Zhang, J. (2022). LSTMGA: A hybrid LSTM–Genetic algorithm model for financial sentiment prediction. Expert Systems with Applications, 193, 116388.
26. Madukwe, C. C., Uwandu, L. A., & Onwuegbuzie, O. O. (2020). Challenges of abusive language detection: Inconsistencies and solutions. Journal of Information Ethics, 29(2), 90–107.
27. Mandl, T., Modha, S., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2020). Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in Indo-European languages. CEUR Workshop Proceedings.
28. Mansour, R., Ghorbel, A., & Mahfoudhi, A. (2021). Sentiment analysis for COVID-19 misinformation detection using SVM. Procedia Computer Science, 180, 476–483.
29. Moussa, S., & Mcheick, H. (2018). BiLSTM with attention for aspect-based sentiment analysis. IEEE Access, 6, 51365–51371.
30. Nguyen, D. Q., Nguyen, T. D., & Nguyen, D. T. (2017). Sentiment analysis on Vietnamese e-commerce reviews using multi-channel LSTM–CNN. Journal of Intelligent & Fuzzy Systems, 33(4), 2475–2486.
31. Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using NLP. In Proceedings of the 2nd International Conference on Knowledge Capture (pp. 70–77).
32. Park, J., Kim, H., & Lee, H. (2021). Senti-DD: Domain-dependent sentiment dictionary expansion using BERT. Applied Sciences, 11(3), 1291.
33. Raj, M., Sahu, S. K., & Kumar, A. (2022). A BiLSTM–CNN hybrid model for cyberbullying detection. Multimedia Tools and Applications, 81(10), 14221–14243.
34. Rehman, H. U., Imran, A., & Hussain, M. (2019). Sentiment analysis using deep learning techniques: A review. International Journal of Advanced Computer Science and Applications, 10(6), 1–9.
35. Rizwan, M., Hassan, S., & Qamar, U. (2019). Sentiment analysis for Twitter using SVM classifier. Procedia Computer Science, 174, 500–505.
36. Saleh, A., Hammad, M., & Hassan, S. (2022). Comparison of transformer-based models and RNN for offensive language detection. IEEE Access, 10, 1057–1068.
37. Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. Human-Centric Computing and Information Sciences, 10(1), 1–34.
38. Samghabadi, N. S., Patwa, P., PYKL, S., & Das, D. (2017). Aggression detection using linguistic features and sentiment lexicons. arXiv preprint arXiv:2001.07646.
39. Shahida, M., Haque, S., & Anwar, F. (2019). A hybrid approach to sentiment analysis using lexicon and machine learning. Procedia Computer Science, 152, 395–402.
40. Sharifani, M., & Amini, M. (2023). A CNN–BiLSTM ensemble for sentiment analysis. Applied Intelligence, 53(2), 2211–2226.
41. Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM) (pp. 687–690).
42. Solovev, K., & Pröllochs, N. (2022). The role of moral-emotional language in online hate speech. EPJ Data Science, 11, 1–25.
43. Sood, S. O., Antin, J., & Churchill, E. F. (2012). Using crowdsourcing to improve profanity detection. Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), 12, 623–626.
44. Susanty, D., Fitriani, A., & Munir, R. (2019). Artificial neural networks for hate speech detection. Journal of Theoretical and Applied Information Technology, 97(23), 3426–3434.
45. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307.
46. Tan, Y., Liu, X., & Zhang, Y. (2022). Hybrid ensemble models for imbalanced sentiment classification. Expert Systems with Applications, 188, 116009.
47. Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 417–424).
48. Vennila, I., Suguna, J., & Rajalakshmi, P. (2020). Enhancing sentiment classification using hybrid CNN–SVM and optimization. Computational Intelligence, 36(1), 23–42.
49. Wang, H., Shen, D., & Liu, S. (2020). SOSNet: A graph convolutional network for tweet similarity and offensive speech detection. IEEE Access, 8, 114980–114990.
50. Yadav, A., & Roychoudhury, S. (2019). Lexicon-based sentiment analysis with multi-source attention. Procedia Computer Science, 165, 664–670.
51. Yao, L., Peng, X., & Huang, L. (2018). A hybrid attention-based sentiment classification model. International Journal of Computer Applications, 179(39), 1–6.
52. Zhang, Y., Guo, Y., & Yu, L. (2020). Hybrid attention networks for sentiment analysis. IEEE Access, 8, 113087–113098.
53. Zhang, Z., Robinson, D., & Tepper, J. (2016). Detecting hate speech on Twitter using a convolutional neural network. In Proceedings of the 1st Workshop on NLP and Computational Social Science (pp. 1–6).

54. Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018). A deeper look at detecting cyberbullying in social networks. Proceedings of the International Joint Conference on Neural Networks (IJCNN).
55. Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48, 4730–4742.
56. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of NAACL Student Research Workshop.
57. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. WWW Companion Proceedings.
58. Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R., Shahid, S., Mehnaz, L., & Anand, S. (2019). MIDAS at SemEval-2019 Task 6. Proceedings of SemEval Workshop.
59. Sun, X., Zhang, C., & Li, L. (2019). Dynamic emotion modelling and anomaly detection in conversation. Information Fusion, 46, 11–22.
60. Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2020). Aggression detection through deep neural models on Twitter. Future Generation Computer Systems, 114, 120–129.
61. Nascimento, F. R. S., Cavalcanti, G. D. C., & Da Costa-Abreu, M. (2022). Gender bias mitigation using ensemble learning. Expert Systems with Applications, 201, 117032.
62. Lin, S. Y., Kung, Y. C., & Leu, F. Y. (2022). Harmful news identification using BERT-based ensemble learning. Information Processing & Management, 59, 102872.
63. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). BERT-based transfer learning for hate speech detection. Complex Networks Conference Proceedings.
64. Liu, P., Li, W., & Zou, L. (2019). Transfer learning for offensive language detection. SemEval Workshop Proceedings.
65. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Identifying offensive language in social media. SemEval Workshop Proceedings.
66. Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. (2021). Offensive language and hate speech detection with deep learning. arXiv Preprint.
67. Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.
68. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations. ICLR Proceedings.
69. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP Proceedings.
70. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of ACL, 5, 135–146.
71. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9, 1735–1780.
72. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder. arXiv Preprint.
73. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45, 2673–2681.
74. Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K. K. R. (2020). Cyberbullying detection using character-level CNN. Concurrency and Computation: Practice and Experience.
75. Ratadiya, P., & Mishra, D. (2019). Attention ensemble-based approach for profanity detection. ICDM Workshops Proceedings.
76. Srivastava, S., Khurana, P., & Tewari, V. (2018). Identifying aggression using capsule networks. TRAC Workshop Proceedings.