FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 2, 2026, 533-542

Systematic Literature Review on Hate Speech Detection Technology with English and Mixed Language on Social Media Space

Peter C. ANYAORA[1*], Emily T. KORMENE[2], Andrew A. UDUIMOH[3], Temple C. OKEAHIALAM[4], Callistus T. IKWUAZOM[5], Lasotte B.-M. YAKUBU[6], Franklyn O. OFOH[7], Rukayat B. AHMED[8]

[1*,2,3]Department Cyber Security Science, Federal University of Technology, Minna Niger State, Nigeria
[4,5]Department of Information Technology, Federal University of Technology, Minna, Nigeria
[6]Department of Computer Science, Federal University of Technology, Minna, Nigeria
[7]Department of Cyber Security Science, Federal University of Technology, Owerri, Nigeria
[8]Department of Cyber Security Science and Data Protection, Federal Polytechnic Offa, Nigeria

[1*]p.anyaora@futminna.edu.ng, [2]tswemy@gmail.com, [3]a.uduimoh@futminna.edu.ng, [4]t.okeahialam@futminna.edu.ng, [5]calistus.tochukwu@futminna.edu.ng, [6]y.lasotte@futminna.edu.ng, [7]onyinyechi.ofoh@futo.edu.ng, [8]rukayat.ahmed@fedpoffaonline.edu.ng

Abstract

This comprehensive overview of the literature review assesses the developments in hate speech identification with an emphasis on English-language content, the dynamics of hate and offensive speech detection online has evolved. There is need to analyse and indicate these trends to enable researchers to be able to proffer better solutions, however several researchers also consider datasets with mixed languages but no mixed datasets of Nigerian pidgin English or languages. The PRISMA framework is used in the review, which covers papers from 2018 to 2024, to guarantee methodological rigor. Important discoveries emphasize how difficult it is becoming to identify hate speech because of changing linguistic conventions like code-switching and emoji usage. Machine learning techniques, especially deep learning models such as BERT and LSTM, have demonstrated notable achievements in identifying hate speech. However, there are still difficulties in differentiating hate speech from content that is offensive but not hateful because it is more of semantic and contextual rather lexical. In order to increase model generalizability, the paper also emphasizes the necessity of diverse datasets that include various linguistic subtleties and cultural situations. In the end, even though the accuracy of existing models is excellent, there is always a need for better ways to deal with biases and imbalances in the data. In the future, studies focusing on multilingual and mixed-language datasets should investigate interpretability as a means of improving the efficacy and impartiality of hate speech detection algorithms.

Keywords: *Hate speech, social media, English language, Multilingual, mixed-language, Offensive speech.*

1.0 Introduction

With the emergence of social media platforms, which have allowed people to voice their thoughts globally, hate and offensive speech has grown to be a serious issue. These platforms encourage connection and communication, but they have also turned into havens for hate speech and other dangerous information. Hate speech has the potential to inspire violence, encourage prejudice and injure people psychologically, thus it is imperative to identify and deal with it [20].It is difficult to automatically identify hate speech since internet language usage is dynamic and ever-changing. Hate speech can be nuanced, using slang, sarcasm or coded language, which makes it challenging for conventional detection methods to reliably identify damaging information [4]. Additionally, hate speech often incorporates contextual elements such as emojis, emoticons and abbreviations, further complicating its detection.

To address this issue, there have been variety of approaches from sophisticated models to conventional machine learning strategies. The calibre and variety of the datasets used for training have a significant impact on how well these models perform. In order to enable more precise detection methods, recent initiatives have concentrated on compiling extensive datasets that reflect contemporary hate speech trends, such as emoticons and contractions [3]. Detecting hate speech in text and images has also been investigated using techniques like Latent Semantic Analysis (LSA) [1]. LSA assists in determining whether a certain text contains hate speech by examining the context and word relationships. Despite advancements, current models frequently perform well on particular datasets but struggle to generalize to new or multilingual data. Research has demonstrated that biases in the datasets, such as an excessive dependence on a limited number of users or sources for hate speech content, cause many models to overfit [3]. Lack of extensive and varied datasets covering hate speech in its various manifestations is one of the main challenges, particularly in languages with limited resources. Because of the disparity between hateful and non-hateful content, it is challenging to gather enough training data, which makes it more difficult to train efficient models [27]. It can be challenging to identify harmful language on social media since many users try to get around detection systems by changing or hiding nasty phrases for instance by using

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

symbols like "f**k" or "fcuk"). Hate speech identification becomes much more challenging when regional dialects, foreign languages and code-mixed text are used, as models must take linguistic variances into consideration[27].

2.0 Related works

Methods such as EasyMixup, which improve training data by combining various text samples, strengthen hate speech recognition models in more than one language. This approach exhibits notable improvements in identifying hate speech, both overt and covert, particularly in contexts with limited resources [27]. Instead of depending only on keywords, models that frame hate speech detection as an entailment problem that is, figuring out whether a statement implies a hateful intent are better able to capture the underlying meaning of text. [13] method focuses on teaching hate speech detection models to forecast the hidden reasoning behind specific decisions, thus increasing the explainability and fairness of these models. By taking into account the surrounding context instead of depending just on specific phrases, this strategy trains models to detect hate speech in a more human-like manner. It is imperative that hate speech detection models exhibit not just high performance but also fairness and interpretability [13]. Inadvertently reinforcing discrimination can occur when a model misclassifies harmless communication as hateful due to certain phrases for instance racial slurs used in a non-hateful context). Thus, adding human justifications as demonstrated by the material requirement planning (MRP) approach helps lessen prejudice and offers more lucid justifications for why a given comment is labelled as hate speech.

[20] offered a carefully selected dataset designed to find hate speech on social media. There are 451,709 sentences in the sample; 371,452 of them are categorized as hateful, and 80,250 as not hateful. It comes from a number of places, such as GitHub and Kaggle. In order to produce a custom vocabulary and a balanced dataset, the dataset also includes augmented examples. The goal of the effort is to use natural language processing techniques to help train machine learning models for the identification of hate speech, [9] addresses the difficulty of identifying hate speech on internet platforms that has the potential to incite violence or social disturbance. This hate speech is focused in various circumstances at different environments and populations, one of the repercussions of hate speech is its further reflect in the physical violence that occurs offline. In order to categorize hate speech in photos, [1] conducted text analysis on the images using machine learning methods, namely Latent Semantic Analysis (LSA). They investigate how this technique might be used to differentiate between content that is hateful and that is not, emphasizing on enhancing detection via contextual word analysis.

Additionally, they employed the Latent Semantic Analysis (LSA) method to analyse the text embedded in the photos in order to detect hate speech. They suggested a technique for identifying hate speech in photographs with the goal of addressing the understudied field of image-based hate speech identification. The study describes how well the LSA approach works and ends with findings that indicate a 57.9% accuracy rate in identifying hate speech. [4] examined the difficulties in detecting hate speech on social media, emphasizing that cutting-edge models frequently perform well on particular datasets but have difficulty generalizing. In their re-examination of validation techniques, [4] identify problems like as sample biases and data overfitting that cause performance measures to be exaggerated. They make the case for improved cross-dataset testing and suggest baseline procedures for evaluating hate speech detection models across languages.

[19] research noted the difficulties in detecting offensive language that is specifically targeted, with a particular emphasis on the problem of data imbalance. Class imbalances are a problem for both traditional machine learning and deep learning methods, which results in poorly generalizable models. To solve this, the study investigates several techniques such as synthetic sampling, oversampling and under sampling. [19] investigates the effects of various methods on the generalizability and performance of the model. 43 models with different rates of under-sampling and oversampling were examined in the study. The model's performance on the test dataset was shown to be enhanced by oversampling; the top-performing model obtained an F1 score of 0.9565 on the test set and demonstrated strong performance in cross-dataset generalization.

[8] research offers a refined version of BERT designed to identify hateful and hurtful speech. Several datasets of foul language are combined to build a strong model that can generalize across various offensive speech forms (racism, misogyny, homophobia). ProperBERT is appropriate for real-world deployment since it is more compact and effective. The study emphasizes how well it can identify hate speech while striking a compromise between mobility and performance, they got a general accuracy of 88% on their test set. Also [2] proposes a solution that focuses on increasing the availability of labelled data, particularly for underrepresented languages like Hindi and Bangla. It also suggests a new LSTM-Autoencoder network to detect cyberbullying on social media using synthetic data. The models that the [2] experiment which include LSTM, BiLSTM, BERT, GPT-2, and Word2Vec. They also use machine-translated synthetic data. The model compares its performance to previous models by analysing three different types of data: noisy, semi-noisy, and noise-free, across several datasets in Hindi, Bangla, and English. With a 95% accuracy on the English dataset and high accuracy of 91% on the Bangla and 91% on the Hindi datasets, the suggested LSTM-Autoencoder performs better than other models. It continued to operate well even with loud, totally machine-translated data.

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

Furthermore [5] research looks at whether a multilingual strategy can enhance the model's capacity to identify sexism in Bengali language using BERT-based models. To gain insight into the decision-making process of the models, [5] employ BERT, BanglaBERT, and mBERT models along with multilingual training in Bengali, Hindi and English. Additionally, they integrate LIME for interpretability. Annotated sexist remarks are included in the dataset. With 91.5% accuracy, the BERT model was the most accurate, followed by mBERT (90.8% after multilingual training). According to a study, learning multiple languages marginally enhances one's ability to recognize misogyny in Bengali. The Graph Neural Network (GNN) model MTBullyGNN, which integrates many social media data formats into a graph structure, is presented in this research [15]. It is intended to be used for multitask learning-based cyberbullying detection on social media platforms. A graph with nodes representing users and edges representing their interactions is the input format used by the GNN model. [15] suggests using multitask learning, the model integrates tasks like sentiment analysis and aggression detection. MTBullyGNN performs more accurately than more conventional models like BiLSTM and CNN, proving the efficacy of graph-based techniques in the identification of cyberbullying.

[23] presented ensemble learning categorization using Term Frequency-Inverse Document Frequency (TF-IDF) with a majority voting in the process of identifying hate speech, models yielded results with 95% accuracy and 0.95 F-Measure. Also [24] provides a model for the more precise identification of hate speech on an internet platform by enhancing the CNN feature selection layer with Bidirectional Encoder Representations from Transformers (BERT) and detecting hate speech using a hybrid of CNN and Long-Short-Term Memory (LSTM). The accuracy of the proposed BERT-CNN-LSTM model was 96.1%. This provides a new and more accurate tool that, when incorporated into organizations and governments security policies, will help protect user's rights and privacy and shield them from the psychological and physical harm that comes with hate speech spread online.

3.0 Methodology

The research considered the search period from January 1st, 2018 to October 2024[29]. The study's objective was met through the application of selection criteria, research questions and a research methodology. Over time, a great deal of work has been made in developing methods for identifying hate speech.

3.1 Research Questions

1. What linguistic features for instance specific keywords, phrases, or grammatical structures) are most indicative of hate speech in social media?
2. How effective are existing machine learning models in distinguishing hate speech from other types of offensive or controversial language on social media?
3. To what extent do contextual factors and social factors enhance the accuracy of hate speech detection on social media generally?

3.2 Phases of the Study and the Protocol

The present review was carried out in compliance with the Preferred Reporting Items for the systematic literature reviews (SLRs) and meta-analyses (PRISMA) guideline [29].

3.3 Study's eligibility requirements and exclusions

To decide which research papers should be selected for review, the study employed a five-point criterion. These was described in Table 1.

3.4 Sources of knowledge and the Search Method

The IEEE Xplore Digital Library, Elsevier, Research Gate, SagePub, Springer, ACL, MDPL, PlosOne, and Google Scholar databases were used for the manual search. The search terms were manually searched across the selected online databases and sources in order to uncover fresh and important materials for the study. "Hate and offensive speech detection," "social media," "Machine learning," "English and mixed language," are the key terms used to search for article related to the field of study. Using the Prisma framework as our methodology, Figure one (1) to provide an explanation for the inclusion and removal of various research publications on the detection of hate speech that were retrieved from several research publishing databases. After removing duplicates, the systematic literature review identified 180 articles at the identification stage. The study covered the following dates: 2018 to 2024; IEEE Explore (69), Research Gate (21), Google Scholar (70), Elsevier (2), Springer (3), MDPL (4), ACL (10), and Springer (3). At the screening stage, 98 articles were removed from the original 180 articles, leaving 82 articles.

Articles that met the eligibility requirements for the SLR within a seven-year timeframe were accepted for consideration. The SLR research project resulted in fifty-five (55) complete research publications, with the exception of twenty-seven (27) publications that were rejected because some of them contained hate speech that was solely in other languages entirely than English language.

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

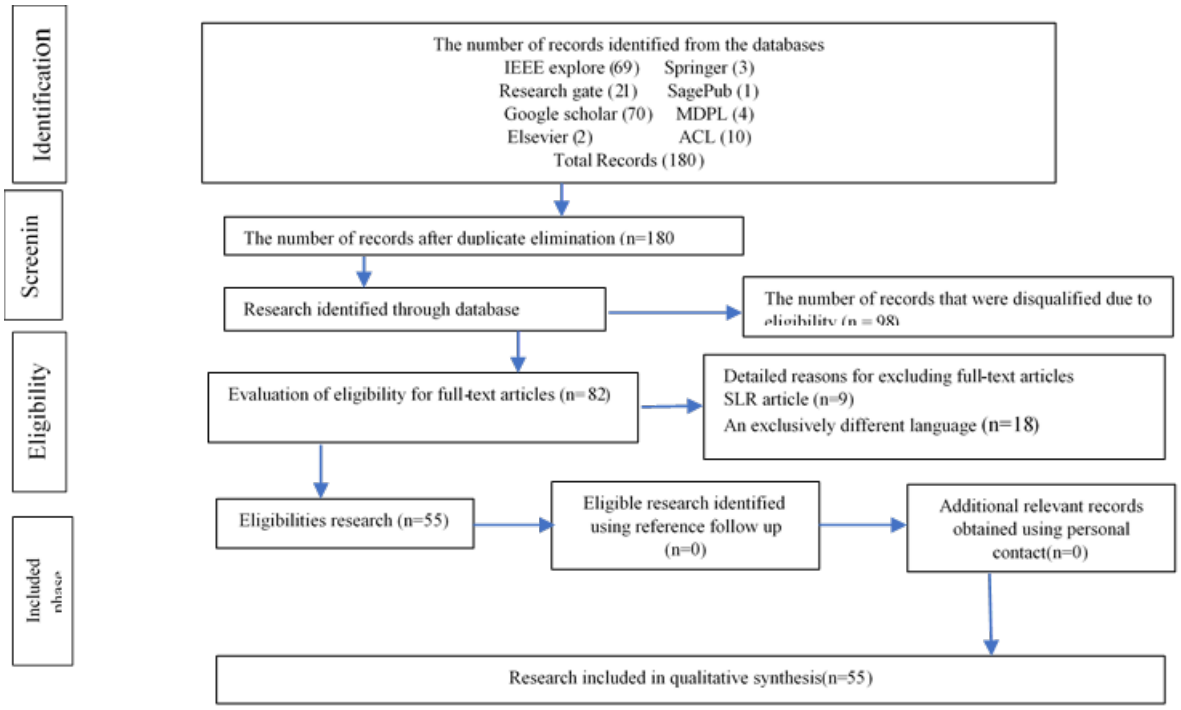

Figure 1: The Hate speech Prisma Framework

Table 1: Selecting which research papers to publish in this systematic literature review involved a number of considerations, which are summarized and justified in Table 1: Requirements for research publications' Article integration and removal.

Table 1: Requirements for research publications' Article integration and removal.

SN	Criteria	Explanation/Justification
1.	a study that is published that is original and not a review or survey	Hate speech techniques and methods, including their application and operation, must be covered in the research articles.
2	The proposed remedies ought to be able to identify or stop hate speech.	This study aims to provide guidance to policy makers and security software developers for the creation of safer and non-toxic social media space policies and systems.
3	The publication needs to be a whole paper.	Brief papers don't provide enough information regarding the recommended repair.
4	The recognised language is use of English	The use of English is only accepted for published article.
5	The publication of must occur between 2018 and 2024.	The SLR is valid from 2018 to 2024.

4.0 Results and Discussion

This section relays the outcomes of the systematic literature reviews carried out in this research. The systematic literature review has been able derive answers to the research questions, as regards the linguistic features most indicative of hate speech on social media are quite numerous for instance specific hateful or derogatory keywords and phrases, slang and offensive expression, coded or disguised language, regional dialects, code-mixed/multilingual expression, emojis, emoticons and abbreviations. But then underlying hateful intent or hate speech indicative of linguistic features are not merely individual hate-related keywords, but combinations of semantic, contextual and lexical features to truly know if it is a hate or non-hate speech. Next, examining how well current machine-learning algorithms differentiate hate speech from other offensive or contentious words, from Appendix A the research show that existing model can be highly effective under the conditions in which they were trained and evaluated, models may perform well with a particular dataset while struggling with another. some of

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

the factors which affect them are multilingual data, code-mixed language, imbalanced dataset, unseen linguistic patterns and biased datasets.

The effectiveness in differentiating hate speech from merely offensive or controversial language remains challenging. Performance depends on dataset quality, class balance, language, context for training of this machine learning algorithms. Also, more importantly evaluation methodology used for instance precision, F1-score, recall, cross-dataset validation, multilingual performance and performance on offensive-but-not-hateful samples as it is very pertinent not to rely only on accuracy. These brings us to the third question if contextual and social factors enhance the accuracy of hate speech detection on social media, this will be possible when considering surrounding textual context like conversation within a context, emojis, code-switching, user profile information where available. Hate speech detection is moving from simple keyword-based identification toward context-aware, semantic and deep-learning approaches. The major remaining challenges are distinguishing hate from offensive language, dataset imbalance and bias, multilingual and code-mixed language and generalization to unseen data. The appendix A relate some of the literature that was reviewed and their success.

4.1 Descriptive meta-analysis performance of some reported model performance on hate-speech.

Looking at the Appendix A and Figure 2, thirteen literatures to explain more the issue of model's performance. With a mean reported accuracy of almost 92.91% and a median of 95.00% among the 13 accuracy findings shown in the graph, the descriptive meta-analysis shows that hate speech detection research has made significant strides overall. The broad performance range of 74.00% to 99.12%, however, shows significant heterogeneity between investigations. The results imply that reported performance is influenced by model architecture, dataset features, language diversity, class distribution, and evaluation methods. More significantly, the analysis shows that practical reliability of hate speech detection systems cannot be established by high accuracy alone. Future research should focus on the ongoing issues raised by the systematic review, such as differentiating hate speech from offensive language, managing multilingual and code-mixed content, resolving class imbalance, enhancing cross-dataset generalization, and creating interpretable models.

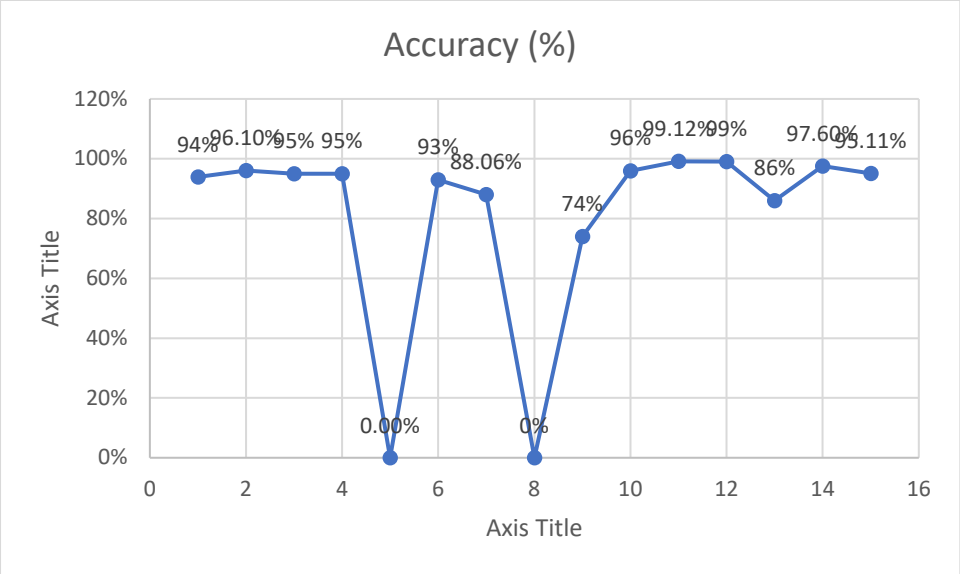

Figure 2: Review of literature model performance

4.2 Limitations, Open Issues and Future research Directions

4.2.1 Limited diversity of datasets

The main issues affecting datasets available for hates and offensive speech detection spans from the facts a dataset may contain thousands of comments but still fail to represents different cultures, regions, demographic group, dialects and forms of hate speech [7]. An instance is a model trained primarily on English twitter/X may perform well during testing but perform poorly on Nigerian social-media conversation involving pidgin English, Hausa English, Yoruba-english, igbo-english, slang and code-switching.

4.2.2 Class imbalance

There are lots of researches that had use an imbalance dataset for training their model, there is need to use oversampling, under-sampling and synthetic samplings as approaches to address imbalance [17]. Future researchers should not simply report accuracy but explore other minority metric for hate-speech evaluation.

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

4.2.3 Difficulty distinguishing hate speech from offensive speech

The review has quite indicated how various machine learning models has done significantly done in detecting hate speech but differentiating hateful from offensive content remains problematic. Semantic and contextual meaning across various demographic area particularly important here [21].

4.2.4 Open Issues on Hate-speech Review

Researchers should look towards context- aware hate-speech detection, it's pertinent to builds models that understand the context surrounding a statement, rather than classifying isolated sentences. Also, multilingual and low-resource languages should be investigated and build dataset that will contain instances beyond English, for instance English + Nigerian pidgin + hausa + igbo + code-mixed forms. There is also a need to have an updated dataset as trends of communication keep growing. Addressing the dataset bias will help models perform better.

4.2.5 Future Research Directions (2026 and beyond)

Looking at the combination of current literatures, the following research directions are proposed.

i. Multimodal detection.
ii. Cross-dataset generalization.
iii. Multilingual/code-mixed hate speech.
iv. Dataset diversity and representations.
v. Class imbalance.

5 Conclusion and Recommendation

The systematic literature review on hate speech detection and issues provides recent discussion and awareness to readers, cyberspace users, social media users and academic researchers on the importance of hate speech detection, also various areas hate speech are perpetuated for example race, religion, ethnicity, athlete, women, tribe, sexual preference. This hate speech tends to materialize physically at different communities by resulting in violence. Curbing such kind of character (Hate speech) on the social media space cannot be over emphasized, building more solution and extensive dataset will help curb or reduce this hate speech minimally. The review suggests researchers should look at building a real-life behaviour multilingual context aware dataset for the Nigerian social media space.

APPENDIX A

S/N	Author	Title	Methodology used	Result	Language Dataset
1	[22]	Detection of Offensive Language in Code-Mixed Text on Social Media	Machine learning (transformer based-model)	They had f1 score of 65%	Mixed of English and Tamil
2	[25]	Detecting Derogatory Comments on Women using Transformer-Based Models	Machine learning (BanglaBERT, XLM-RoBERTa, m-BERT, and DistilBERT)	F1 score of 94% and 86.1	English and Bengali
3	[10]	Smart Language Checker: A Machine Learning Solution for Offensive Language detection in social media	Machine-learning (bag-of-words (BoW), Vectorization-Term frequency-inverse document frequency (TF-IDF), and Tokenization methods, random forest	Accuracy of 93.90%	English
5	[24]	Enhanced Convolutional Neural Network Model Using Bidirectional Encoder Representations from Transformers and Long Short-Term Memory for Hate Speech Detection	Machine-learning-BBidirectional-Encoder Representations Transformers,CNN,LSTM	Accuracy 96.1%	English
6	[23]	Hate and Offensive Speech Detection Using Term Frequency- Inverse	Machine-learning-Term Frequency-Inverse Document Frequency (TF-	Accuracy 95%	English

S/N	Author	Title	Methodology used	Result	Language Dataset
		Document Frequency (TF-IDF) and Majority Voting Ensemble Machine Learning Algorithms	IDF) with a majority voting ensemble learning		
7	[2]	A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data Mst	Machine-learning- Long Short-Term Memory (LSTM), Bidirectional Long Short- Term Memory (BiLSTM), LSTM-Autoencoder, Word2vec-Bidirectional Encoder Representations from Transformers (BERT), and Generative Pre-trained Transformer 2 (GPT-2) models	Accuracy 95%	Mixed-Hindi, Bangla, and English datasets
8	[5]	Bengali Misogyny Identification with Deep Learning and LIME Shafakat	Machine-learning- mBERT	Accuracy 90.3%	Mixed-Hindi, Bangla, and English datasets
9	[18]	Offense Detection Using BERT and CNN	Machine learning -BERT and CNN	Accuracy 85% & 95%	English
10	[16]	Detection of Cyberbullying on Social Media Code Mixed Data Kavisha	Machine learning- Naive Bayes, Logistic Regression, SVM, Decision Tree, KNN and Random Forest	accuracy of 88.06%	Bengali-English, Tamil-English, and Kannada-English languages
11	[12]	HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network	Machine learning-Deep learning	Accuracy of 93%	English
12	[18]	deep learning techniques for spamming and cyberbullying detection	Machine learning	Accuracy of 61.06% to 97.4%	English and Tanglish
13	[3]	Detecting Hate Speech Against Athletes in Social Media	Machine learning	Accuracy of 74%	English
14	[11]	Automatic Detection of Multilingual Misogynistic Content in Social Media Data Based on Machine Learning Approach	Machine learning	Accuracy of 96%	English and Hindi
15	[28]	Interpretable Multi-Modal Hate Speech Detection Prashanth	Machine learning- deep neural multi-modal model	F1 score of 0.90%	English
16	[6]	Enhanced Seagull Optimization with Natural Language Processing Based Hate	Machine learning- Enhanced Seagull Optimization with Natural Language Processing	Accuracy 99.12% & 99%	English

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

S/N	Author	Title	Methodology used	Result	Language Dataset
		Speech Detection and Classificatio			
17	[9]	An approach to detect abusive content incorporating Word2Vec and Multilayer Perceptron 2022	Machine learning-word2vec model and compositional vector model to analyze text, deep learning	Accuracy 86%	English
18	[14]	An Evaluation of Automation on Misogyny Identification (AMI) and Deep-Learning Approaches for Hate Speech Highlight	Machine learning-graph convolutional networks,natural language processing	Accuracy 85.175%	Spanish, English, and Hindi languages
19	[17]	Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)	Machine learning	Accuracy 97.6%	English
20	[26]	A Framework for Hate Speech Detection Using Deep Convolutional Neural Network	Deep Convolutional Neural Network (DCNN)	precision, recall and F1-score value as 0.97, 0.88, 0.92	English

Reference

[1] Ahmad Niam, I. M., Irawan, B., Setianingsih, C., & Putra, B. P. (2018). Hate Speech Detection Using Latent Semantic Analysis (LSA) Method Based on Image. *Proceedings - 2018 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018*, 166–171. https://doi.org/10.1109/ICCEREC.2018.8712111

[2] Akter, M. S., Shahriar, H., Cuzzocrea, A., Wu, F., & Rodriguez-Cardenas, J. (2023). A Trustable LSTM-Autoencoder Network for Cyberbullying Detection on Social Media Using Synthetic Data. *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023*, 5418–5427. https://doi.org/10.1109/BigData59044.2023.10386719

[3] Alsagheer, D., Mansourifar, H., Dehshibi, M. M., & Shi, W. (2022). Detecting Hate Speech Against Athletes in Social Media. *2022 International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2022*, 75–81. https://doi.org/10.1109/IDSTA55301.2022.9923132

[4] Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, *105*, 101584. https://doi.org/10.1016/j.is.2020.101584

[5] Arnob, S. S., Ahad Shikder, M. A., Ovey, T. A., Rhythm, E. R., & Rasel, A. A. (2023). Bengali Misogyny Identification with Deep Learning and LIME. *Proceeding - COMNETSAT 2023: IEEE International Conference on Communication, Networks and Satellite*, 285–292. https://doi.org/10.1109/COMNETSAT59769.2023.10420536

[6] Asiri, Y., Halawani, H. T., Alghamdi, H. M., Abdalaha Hamza, S. H., Abdel-Khalek, S., & Mansour, R. F. (2022). Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification. *Applied Sciences (Switzerland)*, *12*(16). https://doi.org/10.3390/app12168000

[7] Boulouard, Z., Ouaissa, M., Ouaissa, M., Krichen, M., Almutiq, M., & Gasmi, K. (2022). Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning. *Applied Sciences (Switzerland)*, *12*(24). https://doi.org/10.3390/app122412823

[8] Diesenreiter, C., Krauss, O., Sandler, S., & Stöckl, A. (2022). ProperBERT - Proactive Recognition of Offensive Phrasing for Effective Regulation. *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2022*, (November), 16–18. https://doi.org/10.1109/ICECCME55909.2022.9987933

[9] Ghosal, S., Jain, A., & Tayal, D. K. (2022). An approach to detect abusive content incorporating Word2Vec and Multilayer Perceptron. *IBSSC 2022 - IEEE Bombay Section Signature Conference*, 1–5. https://doi.org/10.1109/IBSSC56953.2022.10037274

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026

[10] Kavitha, S., Anchitaalagammai, J. V., Murali, S., Deepalakshmi, R., Himal, L. R., & Suryakanth, M. S. (2023). Smart Language Checker: A Machine Learning Solution for Offensive Language detection in Social Media. *2023 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2023*, 1–6. https://doi.org/10.1109/ICDSAAI59313.2023.10452454

[11] Kaware, P. D., & Raut, A. B. (2024). Automatic Detection of Multilingual Misogynistic Content in Social Media Data Based on Machine Learning Approach. *2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024*, 1–7. https://doi.org/10.1109/ICICACS60521.2024.10499136

[12] Khan, S., Kamal, A., Fazil, M., Alshara, M. A., Sejwal, V. K., Alotaibi, R. M., Baig, A. R., & Alqahtani, S. (2022). HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network. *IEEE Access*, *10*, 7881–7894. https://doi.org/10.1109/ACCESS.2022.3143799

[13] Kim, J., Lee, B., & Sohn, K. A. (2022). Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection. *Proceedings - International Conference on Computational Linguistics, COLING*, *29*(1), 6644–6655.

[14] Li, K. (2022). An Evaluation of Automation on Misogyny Identification(AMI) and Deep-Learning Approaches for Hate Speech -Highlight on Graph Convolutional Networks and Neural Networks. *Proceedings - 2022 International Conference on Computers, Information Processing and Advanced Education, CIPAE 2022*, 239–244. https://doi.org/10.1109/CIPAE55637.2022.00058

[15] Maity, K., Sen, T., Saha, S., & Bhattacharyya, P. (2024). MTBullyGNN: A Graph Neural Network-Based Multitask Framework for Cyberbullying Detection. *IEEE Transactions on Computational Social Systems*, *11*(1), 849–858. https://doi.org/10.1109/TCSS.2022.3230974

[16] Mathur, K., Mehta, K. N., Shivakumar, K., & Uma, D. (2022). Detection of Cyberbullying on Social Media Code Mixed Data. *Proceedings - 2022 IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2022*, 16–23. https://doi.org/10.1109/CCEM57073.2022.00011

[17] Mazari, A. C., & Kheddar, H. (2023). Deep Learning-based Analysis of Algerian Dialect Dataset Targeted Hate Speech, Offensive Language and Cyberbullying. *International Journal of Computing and Digital Systems*, *13*(1), 965–972. https://doi.org/10.12785/ijcds/130177

[18] Meenakshi, M., Shyam Babu, P., & Hemamalini, V. (2023). Deep Learning Techniques for Spamming and Cyberbullying Detection. *Proceedings of the 1st IEEE International Conference on Networking and Communications 2023, ICNWC 2023*, 1–10. https://doi.org/10.1109/ICNWC57852.2023.10127460

[17] Mehta, H. (2022). *Social media hate speech detection using explainable AI*. https://zone.biblio.laurentian.ca/handle/10219/4022

[18] Mehta, M., Gada, D., Sharma, R., Chavan, K., & Kanani, P. (2022). Offense Detection Using BERT and CNN. *2022 IEEE 3rd Global Conference for Advancement in Technology, GCAT 2022*, (October), 2–7. https://doi.org/10.1109/GCAT55367.2022.9971862

[19] Mifsud, R., Deka, L., & Lahiri, I. (2023). An Optimised BERT Pretraining Approach for Identification of Targeted Offensive Language: Data Imbalance and Potential Solutions. *2023 4th International Conference on Computing and Communication Systems, I3CS 2023*, 1–8. https://doi.org/10.1109/I3CS58314.2023.10127515

[20] Mody, D., Huang, Y. D., & Alves de Oliveira, T. E. (2023). A curated dataset for hate speech detection on social media text. *Data in Brief*, *46*, 108832. https://doi.org/10.1016/j.dib.2022.108832

[21] 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. https://doi.org/10.1109/ACCESS.2021.3134154

[22] Nithya, M., Packiam, R. S. L., & Murugappan, S. A. (2024). Detection of Offensive Language in Code-Mixed Text on Social Media. *2024 3rd International Conference on Artificial Intelligence for Internet of Things, AIIoT 2024*, (AIIoT), 1–5. https://doi.org/10.1109/AIIoT58432.2024.10574567

[23] Okechukwu, C., Idris, I., Ojeniyi, J. A., Adebayo, O. S., & Olarere, M. (2023). Hate and Offensive Speech Detection Using Term Frequency - Inverse Document Frequency (TF - IDF) and Majority Voting Ensemble Machine Learning Algorithms. *International Engineering Conference, Science , Cyber Security*, (Iec). http://repository.futminna.edu.ng:8080/jspui/bitstream/123456789/18703/1/Hate and Offensive Speech Detection.pdf

[24] Okechukwu, C., Nwodo, C., Anyaora, P. C., Hosea, G., & Subairu, S. O. (2024). *Communication Conference Enhanced Convolutional Neural Network Model Using Bidirectional Encoder Representations from Transformers and Long Short-Term Memory for Hate Speech Detection*. (I3c), 1–13.

[25] Prithila, S. J., Tonima, F. H., Oishi, T. T., Islam, M. N., Rhythm, E. R., Amit, A. M., & Rasel, A. A. (2023). Detecting Derogatory Comments on Women using Transformer-Based Models. *Proceeding - COMNETSAT 2023: IEEE International Conference on Communication, Networks and Satellite*, 278–284. https://doi.org/10.1109/COMNETSAT59769.2023.10420692

[26] 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. https://doi.org/10.1109/ACCESS.2020.3037073

[27] Roychowdhury, S., & Gupta, V. (2023). Data-Efficient Methods For Improving Hate Speech Detection. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, 125–132. https://doi.org/10.18653/v1/2023.findings-eacl.9

[28] Vijayaraghavan, P., Larochelle, H., & Roy, D. (2021). *Interpretable Multi-Modal Hate Speech Detection*. http://arxiv.org/abs/2103.01616

[29] Page, M. J., McKenzie, J. E., Bossuyt, P., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Medicina Fluminensis*, *57*(4), 444–465. https://doi.org/10.21860/medflum2021_264903

Received: 02/06/2026; Revised: 25/08/2026; Accepted: 26/08/2026; Published: 10/09/2026