

Insights of learning approach towards determination of potentially objectional communication in social networking

Praneetha Garagadakuppe Nanjundappa^{1,2}, Kamalakshi Naganna²

¹Department of Computer Science and Engineering, Saphthagiri College of Engineering, Visvesvaraya Technological University, Belagavi, India

²Department of Computer Science and Engineering, M.S. Ramaiah University of Applied Sciences, Bengaluru, India

Article Info

Article history:

Received Nov 28, 2023

Revised Feb 9, 2024

Accepted Mar 2, 2024

Keywords:

Hateful speech

Machine learning

Offensive speech

Sentiment analysis

Social network

ABSTRACT

Over the last decade, sentiment analysis has evolved significantly towards extracting the contextual knowledge associated with the communication exchanged in social networks. Irrespective of various approaches to natural language processing and constantly evolving machine learning, sentiment analysis has inherent shortcomings, which further act as an obstacle to determining hateful and offensive speech exchanged in social networks. Therefore, this paper offers a compact yet granular insight into the effectiveness of existing sentiment analysis approaches used distinctly for determining hateful and offensive speech with particular emphasis on machine learning-based methodologies. The paper further contributes towards research trend analysis followed by distinct highlights of the research gap. The paper offers a learning outcome that significantly benefits future researchers investigating the same field.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Praneetha Garagadakuppe Nanjundappa

Department of Computer Science and Engineering, M.S. Ramaiah University of Applied Science

Bengaluru, India

Email: praneethaguddi@gmail.com

1. INTRODUCTION

The communication taking place within a social network is from the perspective of the user's response, which could be text, audio, video, images [1]. With rising concerns about cybersecurity, it is essential to understand the significance of meaningful and logically communicated information in terms of personal opinion. However, certain forms of communicated information in social networks harm everyone. This study, therefore, discusses one such harmful effect in social networks in the form of hateful and offensive speech that induces discrimination and misinformation [2]. From the perspective of social networks, hateful speech is a type of communication that is meant to intimidate, dehumanize, and demean the individual or group based on personal characteristics, sexual orientation, gender, religion, ethnicity, and race [3]. Hateful speech often takes the shape of harassment, threats, slurs, and insults. The negative impact of hateful speech is that it propagates an environment of insecurity and unsolicitation for online users, making the environment more hostile [4]. On the other hand, offensive speech is usually considered inappropriate communication with disrespect and rudeness during communication in social networks [5]. The majority of offensive speech on social networks deploys derogatory opinions about an individual or group, often using sexually explicit comments, vulgarity and profanity. One way to solve the associated problem is to develop a system to identify and remove such hateful and offensive statements in social networks. One effective way is to use sentiment analysis, which harnesses the potential of natural language processing (NLP) to process such objectional text [6]. As a social network is a vast network with massive information

streams, it is humanly impossible to develop a scheme to identify it from one interface only. Hence, machine learning is the only practical solution where an intelligent algorithm can be developed to identify and remove such objectional text after confirming its degree of severity [7]. Various sentiment analysis schemes can be adopted to determine such text indexes concerning hateful and offensive text [8]. The adoption of sentiment analysis can also be used to monitor the patterns and trends of such objectional terms over time to address the systemic issues of discrimination and objectional speech. Irrespective of various available sentiment analysis models, there are challenges, too.

The notable problems associated with the existing form of sentiment analysis are as follows, viz. i) The first problem is associated with fluctuating quality and availability of data in social networks. ii) The existing sentiment analysis models suffer from bias and subjectivity as the opinion shared by a user has a higher impact on the data. iii) The third problem in sentiment analysis is associated with the incapability of supporting multilingualism owing to a discrete set of syntax, vocabulary, and grammar. iv) The fourth issue is related to language variability because a user's speech pattern is essentially governed by their cultural background, location, gender, and age. v) The final issues of sentiment analysis are associated with the context and ambiguity of the text, leading to wrong interpretation by the machine.

To realize the above-mentioned clear statement of the problem, it is necessary to exhibit and discuss some relevant literature. The discussion presented by Alkomah and Ma [9] discussed various detection schemes of textual hate speech to signify that there is increasing attention to recent research on this topic. According to the study outcome, hybrid models are used more, followed by lexicon-based models, while the lowest trend of other conventional machine learning schemes is noted. Another significant contributory finding is presented by Kovacs *et al.* [10]. The authors have presented an NLP-based deep learning methodology that jointly uses recurrent and convolution layers for autonomous detection of hateful speech over social media. Jahan and Oussalah [11] have presented a similar study investigating NLP methods. The study towards the detection of offensive text is carried out by Pradhan *et al.* [12], where analysis of the Dataset is carried out along with a briefing of some current approaches. The study outcome infers that long short-term memory (LSTM) offers better outcomes than others. A similar direction of research towards hateful and offensive speech is also carried out by Sokolova *et al.* [13], illustrating the distinction of varied learning models. This study's outcome showcases that open end problems are still associated with this topic.

The proposed study contributes towards offering a solution by providing a compact yet distinct visualization of the effectiveness of existing methodologies used to detect hateful and offensive statements in social networks. The new value of the research is presented in the proposed study in the form of the following contribution viz. i) the proposed schemes present a compact review of methodologies used for identifying and classifying hateful speech and offensive speech, as well as combined detection of both of them, which is not reported in any existing review work, ii) the proposed scheme elaborates the methodologies exclusively concerning problems being addressed and notable benefits and limiting factors to judge the effectiveness of existing schemes of machine learning, iii) the study also offers a direct insight to the research gap which is significantly essential information for any researchers attempting to investigate on same topic. The following section discusses the method used in the proposed scheme.

2. METHOD

The proposed study uses a desk research methodology to check various reputed periodicals, peer-reviewed journals, and databases carefully. Figure 1 highlights the methods deployed in a proposed study where the identification of information is in the form of explicit implementation models where machine learning has been used for hateful and offensive speech analysis. Further, preliminary screening is conducted by reviewing the title and fast paper scanning. The elimination of duplicates is carried out for similar papers published in short and long papers by the same author or two different authors but with the same methodology. Finally, screening of the abstract is carried out based on exclusion criteria and inclusion criteria. The exclusion criteria are i) papers published before a decade and ii) theoretical papers. The inclusion criteria are i) the paper must have implementation models and ii) clear result highlights to show the strength of stated models. Finally, the complete articles are assessed where the emphasis is given to the adopted methodology, parameters used for analysis, the dataset used for the experiment, and results achieved. In this review process, it is studied whether the implemented machine learning models have been used as it is or if any form of indicative novel features are being implemented to obtain a better accuracy performance. Finally, this step also contributes towards the next step of extracting learning outcomes where further multiple criteria are sought. The first criterion is to assess the strength of the Dataset to find out the sustainability of the models. If the models are assessed on a small Dataset, they cannot be eventually inferred as a robust model. The second criterion is to assess the numerical and graphical outcomes of the study. The analysis assesses if a potential comparative analysis or benchmarking is being carried out. A lack of potential comparison could also infer the low applicability of the outcome, as well a lack of benchmarking will also

render the applicability of the presented methodology to be further low. Finally, all this information is collected in one place to understand the open-ended issues. The prime factor assessed in this process is whether that model has offered any novel solution or solves the existing problems of sentiment analysis used for detecting and classifying hateful and offensive speech in social networks.

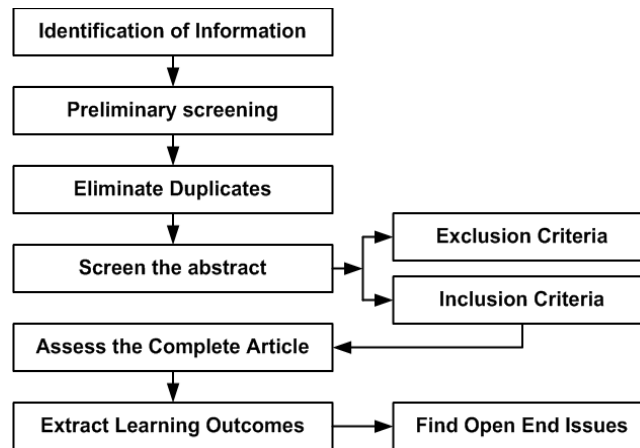


Figure 1. Method adopted in the proposed study

3. RESULTS

This section discusses the results obtained from the proposed study. As the proposed study reviews existing methodologies associated with hate and offensive speech detection, the outcomes obtained from this study are inclined towards reviewing the strength and effectiveness of existing schemes. The section discusses existing studies towards hateful speech detection, offensive speech detection, and combined studies, followed by research trends and gaps.

3.1. Existing studies towards hateful speech detection

At present, certain distinct studies are investigating the effectiveness of detection approaches for hateful speech. The work carried out by Abro *et al.* [14] has performed a comparative study where machine learning schemes have been used for the involuntary detection of hate speech. The study investigated feature extraction methods for Doc2vec, word2vec, and bigram for multiple machine learning schemes to find that better detection performance is exhibited by random forest and support vector machines (SVM). In contrast, inferior performance is exhibited by the k-nearest neighbor (KNN) scheme. Agarwal and Chowdary [15] have used the ensemble learning method to mitigate the hateful speech associated with the recent pandemic. The scheme has used recurrent neural network (RNN), long short-term memory (LSTM), and Softmax activation, while the Adam optimizer is used in the last stage of feature extraction. Aljero and Dimililer also carry out a similar approach [16], where multiple classification schemes are used, viz. XGBoost, logistic regression (LR), and SVM, while universal encoding features and word2vec are used for training. The work presented by Arco *et al.* [17] has integrated emotional knowledge and polarity-based detection to recognise hate speech in Spanish text from Twitter. The scheme further uses a multi-task learning scheme for the shared transformed encoder. The work presented by Biradar *et al.* [18] has used bidirectional encoder representations from transformers (BERT) for evaluating bilingual code-mixed data for mitigating hate speech. A similar direction of the problem of bilanguage code-mixed data was also investigated by Sreelakshmi *et al.* [19], where the deployment of SVM is used along with radial basis function (RBF). The study also uses deep learning to interpret and extract features. Fauzi and Yuniarti [20] have used ensemble and multiple classification approaches to identify hateful speech in Indonesian tweets. Mukherjee and Das [21] have adopted a pre-training mechanism for transformer-based methods for extracting the context of text sequences. At the same time, the model was proven to be efficient in determining hate speech compared to conventional LSTM and convolution neural network (CNN). Salminen *et al.* [22] have used multiple machine learning-based classifiers to detect hate speech on social media. The study uses BERT to incorporate sophisticated linguistic features. Singh *et al.* [23] have presented a mechanism for extracting knowledge from protected attributes from unstructured forms of social media data. Further, the adoption of BERT is also witnessed in the work of Wahl and Skjastad [24]. The limitations associated with above-mentioned studies

are computational complexity is not examined, not fine-grained dataset, suffers from data imbalance issue, limited to spanish corpora, detection doesn't involve sarcasm, the higher computational cost, highly iterative scheme, more case studies and a broader dataset are required, language specific, sub-optimal feature analysis, challenges in inferring predictive value, lacks comparative analysis, and no benchmarking.

3.2. Existing studies towards offensive speech detection

It has been noticed that nearly similar forms of methodologies have been used for the detection of offensive speech propagating within a social network. The adoption of deep learning has been witnessed in the work of Bansal [25] towards identifying offensive comments in social media. The study suggests the efficiency of adopting CNN with the LSTM model for better performance in detection. Gemes *et al.* [26] have used hybrid and rule-based methods to detect offensive text where graph patterns and semantic parsing have been used. Mehmood *et al.* [27] have used Naïve Bayes, LR, and extra tree to detect offensive language in Urdu over social media. Further, the model has used multiple machine learning approaches, e.g., LSTM, gated recurrent unit, fully connected network, CNN, and SVM. Features were developed using term frequency (TF) and inverse document frequency (IDF) along with bag-of-words (BoW). A similar form of the problem is also discussed by Mridha *et al.* [28], considering Bengali offensive text over social media using revised AdaBoost with LSTM and BERT. Ranasinghe and Zampieri [29] have investigated detecting offensive language from multilingual families of text, where a word embedding of cross-lingual context is used. A similar direction of adopting the research problem is also investigated by Shanmugavadivel *et al.* [30], where deep learning has been used as a BERT-based approach. Shannaq *et al.* [31] have developed an intelligent predictive scheme for classifying offensive text in Arabic. This scheme uses SVM, XGBoost, and a genetic algorithm (GA). Souza and Abreu [32] have used naïve Bayes and SVM to classify offensive text from social data. Suryawanshi *et al.* [33] have developed a mechanism to perform labelling of offensive data where the training is carried out using CNN, bidirectional LSTM (BiLSTM), ensemble embedding, and BERT. Similar adoption of BiLSTM and CNN is also studied by Wiedemann *et al.* [34] using transfer learning. Wu *et al.* [35] have developed an automated decoding technique where a classifier can detect offensive language on multiple datasets using a unique experimental setup. The limitations associated with above-mentioned studies are highly iterative process, specific to english and germany, the narrowed dataset, no analysis towards complexity, lack of specific case study, lack of trained data, lower accuracy sentiment analysis model, lower accuracy rate, no benchmarking, yet to be benchmarked, less focus on the pre-training task, decoding specific to encoding scheme and hence not generalized.

3.3. Existing studies towards hateful and offensive speech detection

There are also certain studies where hateful and offensive speech detection is carried out jointly. The work carried out by Boulouard *et al.* [36] has used transfer learning to identify both hateful and offensive Arabic speech. Further, the study also contributes towards comparing different variants of frequently used BERT models for training followed by classification. Mozafari *et al.* [37] used a meta-learning approach based on metric and optimization, while the model differentiated tasks for detecting hate and offensive speech on cross-lingual factors. Regarding geographic-specific social network data, Oriola and Kotze [38] have used multiple machine learning approaches, e.g., gradient boosting, random forest, LR, and SVM. The outcome exhibited better performance of SVM in detecting hateful and offensive speech. Watanabe *et al.* [39] have used patterns and unigrams obtained from training data to perform feature training in machine learning. The table's information showcases that most schemes are characterized by beneficial and limiting attributes. This fact foretells that existing approaches still have more extensive scope towards future amendments. Further, it can be noticed that machine learning is one of the dominant schemes in this part of predictive analysis, where similar learning strategies can be used for both hateful and offensive text detection and classification. The limitations associated with above-mentioned studies are need extensive benchmarking, no contextual analysis, model applicability restricted, and yet to be benchmarked.

3.4. Research trend

The discussion towards the existing research methodologies for detecting hateful speech and offensive speech carried out in prior sub-sections are some critical publications with unique attempts towards performance improvement. However, more studies have been published in the last ten years in various research-based journals. Table 1 highlights the same statistics from various reputed publications. From the score obtained in Table 1, it can be noticed that there are approximately 3722 research publications towards hateful speech. At the same time, there are a massive number of 23,797 publications in research towards offensive speech detection. Further, there are 2807 research articles published for combined hateful and offensive speech detection. A closer look into this tabulated score shows a smaller number of research works for combined criterion and towards hateful speech in contrast to offensive speech detection in the existing

system. It is evident from these scores that complications towards determining hateful speech are higher than offensive speech.

Further, the trend of frequent adoption of machine learning towards detection of hateful and offensive speech is found to be XGBoost, LR, SVM, BERT, convolution neural network (CNN), long short-term memory (LSTM), transfer learning, Naïve Bayes, and random forest. Apart from the information in this table, not much work has been done on each learning scheme, most recently. It is also noted that the total number of unique publications is analyzed to arrive at a specific number of manuscripts where particular machine learning has a dominant role in analysis as shown in Figure 2.

According to this outcome in Figure 2, the adoption of SVM and BERT is constantly on the rise. At the same time, CNN and XG-Boost are witnessed to be less adopted. After SVM and BERT, the following frequently adopted machine learning methods are transfer learning and LSTM, while consecutive usage patterns are also found for Naïve Bayes and random forest. However, they are less dominantly used. Table 2 showcases the adopted Dataset for determining hateful and offensive speech, where HatEval and Kaggle are the most frequently adopted in the existing scheme.

However, other datasets are also increasingly adopted in research work. The Dataset for HatEval is mainly associated with the promotion of violence and hatred, especially targeting immigrants and women. The Dataset for Founta and Davidson consists of text associated with insult or humiliation towards a group or an individual considering gender, disability, sexual orientation, ethnic origin, religion, and race. The Dataset of Waseem is associated with hate speech associated with criticism towards minorities. OLDI dataset is exclusively used for offensiveness and abusiveness, while the Golbeck dataset represents harassment-based text. AbuseEval Dataset is used for both implicit and explicit abuses. It is to be noted that almost all these datasets in Table 2 have their source from the social network Twitter.

Table 1. Research trend in publication

Publication	Hateful Speech	Offensive Speech	Combined
IEEE	5	9	3
MDPI	77	11	11
TF	1512	21562	591
Springer	9	16	8
Elsevier	16	25	14
ACM	2103	2174	2180

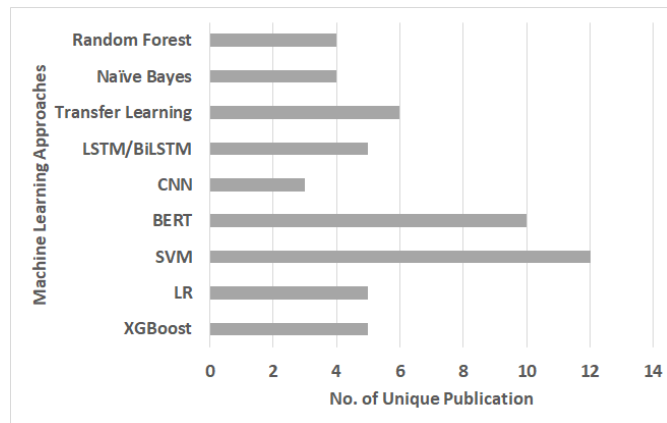


Figure 2. Research trend of machine learning adoption

Table 2. Dataset in adoption

Dataset	Adopted in Studies
HatEval	Fortuna <i>et al.</i> [40], Poletto <i>et al.</i> [41]
Founta	Wich <i>et al.</i> [42]
Davidson	Davidson <i>et al.</i> [43]
Waseem	Waseem and Hovy [44]
Kaggle	Kaggle [45]
Golbeck	Golbeck <i>et al.</i> [46]
AbuseEval	Caselli <i>et al.</i> [47]
OLID	Zampiere <i>et al.</i> [48]

3.5. Research gap

The identified research gap from the insights to existing schemes of detection of hateful and offensive speech is as follows: i) It has been noted that machine learning approaches have been dominantly used to detect hateful speech; however, most adopt multi-tasking mechanisms, leading to positive and negative correlations. The presence of a negative correlation results in artefacts that degrade the classification performance. This fact is found unaddressed in existing schemes. ii) The adoption of machine learning (especially deep learning approaches) is incapable of aggregating and refining varied contextual information. This causes increasing steps of iteration, causing computational complexity. Hence, accuracy is not reported to be highly optimal in such cases. iii) None of the existing schemes are reported to extract potential information from the corpus of textual data, which is critically important. Identification of different variants of hateful speeches is quite a computationally challenging task. It cannot differentiate itself from offensive terms or specific identity-related terms. iv) Adopting a deep learning-based scheme for offensive speech detection is carried out mainly using LSTM, and CNN. However, conventional deep learning methods adopt embedding words in a static form that cannot address polysemy problems. Such an approach further degrades classifier performance. v) Finally, a closer look into existing schemes shows increasing trends in using BERT schemes and their different variants. Such schemes are well known for addressing the problems associated with word ambiguity using its contextual word embedding approach; however, features associated with characters are often ignored in this process, further degrading the accuracy.

4. CONCLUSION

Sentiment analysis has proven effective in various applications using NLP and machine learning approaches. However, it encounters a significant challenge concerning various aspects, yet an open-ended issue. Hence, adopting sentiment analysis in its current state towards determining hateful and offensive speech in social networks is still a more extensive set of challenges. This paper discusses the effectiveness of machine learning schemes towards the determination of such objectional speech. The significant contribution of this paper are introduced here in the form of learning outcomes: i) machine learning is one potential solution towards determining and distinguishing objectional text in social network, but majority of existing implementation models has not been witnessed with any novel scheme of machine learning addressing to various constraints associated with it, ii) there is no balance between the accuracy being accomplished and computational complexity associated with the scheme, which is mainly due to selection of imbalanced data as well as inappropriate learning operation, which requires immediate attention, iii) there is a need of an alternative solution of learning as well as incorporation of intelligent features in training operation which can offer better consistency over exponential dynamicity of contextual problems in such objectional text, iv) there is a need to evolve up with a good labelling of Dataset as well as need to construct a novel dataset which overcomes the problems of existing imbalance data, and v) there is a need of an integrated modelling which can actually distinguish between determination of hateful speech from offensive speech over various extensive test cases. Our future work will address the research gap presented in this study by evolving a novel computational framework of sentiment analysis, where a novel machine-learning framework will be constructed to solve the problem of determining hateful and offensive speech in social networks.

FUNDING DECLARATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

REFERENCES





- [1] K. Rrmoku, B. Selimi, and L. Ahmedi, "Provenance and social network analysis for recommender systems: a literature review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 5383–5392, 2022, doi: 10.11591/ijece.v12i5.pp5383-5392.
- [2] T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, "Abusive comment identification on Indonesian social media data using hybrid deep learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 3, p. 895, 2022, doi: 10.11591/ijai.v11.i3.pp895-904.
- [3] P. Jougleux, "Redefining freedom of speech in the digital environment from an EU law perspective," *International Journal of Electronic Governance*, vol. 11, no. 3/4, p. 1, 2019, doi: 10.1504/IJEG.2019.10023323.
- [4] H. EL-Zayady, M. S. Mohamed, K. Badran, and G. Salama, "A hybrid approach based on personality traits for hate speech detection in Arabic social media," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1979–1988, 2023, doi: 10.11591/ijece.v13i2.pp1979-1988.
- [5] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, "Synonym based feature expansion for Indonesian hate speech detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1105–1112, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1105-1112.

- [6] P. Tanwar and P. Rai, "A proposed system for opinion mining using machine learning, NLP and classifiers," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 726–733, 2020, doi: 10.11591/ijai.v9.i4.pp726-733.
- [7] P. Patil and D. N. Naik, "Hate speech detection and analysis using machine learning," *Artificial Intelligence in Information and Communication Technologies, Healthcare and Education: A Roadmap Ahead*. Chapman and Hall/CRC, pp. 149–156, 2022. doi: 10.1201/9781003342755-16.
- [8] A. Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa, and J. Baldwin, "Sentiment analysis of comments in social media," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 5917–5922, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5917-5922.
- [9] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, May 2022, doi: 10.3390/info13060273.
- [10] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources," *SN Computer Science*, vol. 2, no. 2, p. 95, Apr. 2021, doi: 10.1007/s42979-021-00457-3.
- [11] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, 2023, doi: 10.1016/j.neucom.2023.126232.
- [12] R. Pradhan, A. Chaturvedi, A. Tripathi, and D. K. Sharma, "A Review on Offensive Language Detection," in *Lecture Notes in Networks and Systems*, vol. 94, Springer Singapore, 2020, pp. 433–439. doi: 10.1007/978-981-15-0694-9_41.
- [13] Z. Sokolová, J. Staš, and J. Juhár, "Review of recent trends in the detection of hate speech and offensive language on social media," *Acta Electrotechnica et Informatica*, vol. 22, no. 4, pp. 18–24, Dec. 2022, doi: 10.2478/aei-2022-0018.
- [14] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic hate speech detection using machine learning: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 484–491, 2020, doi: 10.14569/IJACSA.2020.0110861.
- [15] S. Agarwal and C. R. Chowdhary, "Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19," *Expert Systems with Applications*, vol. 185, p. 115632, Dec. 2021, doi: 10.1016/j.eswa.2021.115632.
- [16] M. K. A. Aljero and N. Dimililer, "A novel stacked ensemble for hate speech recognition," *Applied Sciences*, vol. 11, no. 24, p. 11684, Dec. 2021, doi: 10.3390/app112411684.
- [17] F. M. Plaza-Del-Arco, M. D. Molina-Gonzalez, L. A. Urena-Lopez, and M. T. Martin-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," *IEEE Access*, vol. 9, pp. 112478–112489, 2021, doi: 10.1109/ACCESS.2021.3103697.
- [18] S. Biradar, S. Saumya, and A. Chauhan, "Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 87, 2022, doi: 10.1007/s13278-022-00920-w.
- [19] K. Sreelakshmi, B. Premjith, and K. P. Soman, "Detection of hate speech text in Hindi-English code-mixed data," *Procedia Computer Science*, vol. 171, pp. 737–744, 2020, doi: 10.1016/j.procs.2020.04.080.
- [20] M. A. Fauzi and A. Yuniarti, "Ensemble method for Indonesian Twitter hate speech detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 1, pp. 294–299, Jul. 2018, doi: 10.11591/ijeecs.v11.i1.pp294-299.
- [21] S. Mukherjee and S. Das, "Application of Transformer-based language models to detect hate speech in social media," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 278–286, Dec. 2021, doi: 10.47852/bonviewJCCE2022010102.
- [22] J. Salminen, M. Hopf, S. A. Chowdhury, S. Gyo Jung, H. Almerexhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, 2020, doi: 10.1186/s13673-019-0205-6.
- [23] A. Singh, J. Chen, L. Zhang, A. Rasekh, I. Golbin, and A. Rao, "Independent ethical assessment of text classification models: A hate speech detection case study," *arxiv.org (Preprint)*, 2021, [Online]. Available: <http://arxiv.org/abs/2108.07627>
- [24] M. Wahl and S. G. Skjåstad, "Detecting hate speech in Norwegian texts using BERT semi-supervised anomaly detection," Norwegian University of Science & Technology, 2021. [Online]. Available: <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2833641>
- [25] P. Bansal, "Detection of offensive YouTube comments, a performance comparison of deep learning approaches," Technological University Dublin, 2019.
- [26] K. Gémes, Á. Kovács, and G. Recski, "Offensive text detection across languages and datasets using rule-based and hybrid methods," *CEUR Workshop Proceedings*, vol. 3318, 2022.
- [27] A. Mehmood *et al.*, "Threatening URDU language detection from tweets using machine learning," *Applied Sciences*, vol. 12, no. 20, p. 10342, Oct. 2022, doi: 10.3390/app122010342.
- [28] M. F. Mridha, M. A. H. Wadud, M. A. Hamid, M. M. Monowar, M. Abdullah-Al-Wadud, and A. Alamri, "L-Boost: Identifying offensive texts from social media post in Bengali," *IEEE Access*, vol. 9, pp. 164681–164699, 2021, doi: 10.1109/ACCESS.2021.3134154.
- [29] T. Ranasinghe and M. Zampieri, "An evaluation of multilingual offensive language identification methods for the languages of India," *Information*, vol. 12, no. 8, p. 306, Jul. 2021, doi: 10.3390/info12080306.
- [30] K. Shanmugavadivel, V. E. Sathishkumar, S. Raja, T. B. Lingaiah, S. Neelakandan, and M. Subramanian, "Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data," *Scientific Reports*, vol. 12, no. 1, p. 21557, Dec. 2022, doi: 10.1038/s41598-022-26092-3.
- [31] F. Shannaq, B. Hammo, H. Faris, and P. A. Castillo-Valdivieso, "Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings," *IEEE Access*, vol. 10, pp. 75018–75039, 2022, doi: 10.1109/ACCESS.2022.3190960.
- [32] G. A. De Souza and M. Da Costa-Abreu, "Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020, pp. 1–6. doi: 10.1109/IJCNN48605.2020.9207652.
- [33] S. Suryawanshi, M. Arcan, and P. Buitelaar, "NUIG at SemEval-2020 task 12: Pseudo labelling for offensive content classification," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: International Committee for Computational Linguistics, 2020, pp. 1598–1604. doi: 10.18653/v1/2020.semeval-1.208.
- [34] G. Wiedemann, E. Ruppert, R. Jindal, and C. Biemann, "Transfer learning from LDA to BiLSTM-CNN for offensive language detection in Twitter," *arxiv.org (Preprint)*, 2018, [Online]. Available: <http://arxiv.org/abs/1811.02906>
- [35] Z. Wu, N. Kambhatla, and A. Sarkar, "Decipherment for adversarial offensive language detection," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 149–159. doi: 10.18653/v1/W18-5119.
- [36] Z. Boulouard, M. Ouaisa, M. Ouaisa, M. Krichen, M. Almutiq, and K. Gasmi, "Detecting hateful and offensive speech in Arabic social media using transfer learning," *Applied Sciences*, vol. 12, no. 24, p. 12823, 2022, doi: 10.3390/app122412823.





- [37] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-lingual few-shot hate speech and offensive language detection using meta learning," *IEEE Access*, vol. 10, pp. 14880–14896, 2022, doi: 10.1109/ACCESS.2022.3147588.
- [38] O. Oriola and E. Kotze, "Evaluating machine learning techniques for detecting offensive and hate speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.
- [39] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.
- [40] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?," *Information Processing and Management*, vol. 58, no. 3, p. 102524, 2021, doi: 10.1016/j.ipm.2021.102524.
- [41] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477–523, 2021, doi: 10.1007/s10579-020-09502-8.
- [42] M. Wich, T. Eder, H. Al Kuwaty, and G. Groh, "Bias and comparison framework for abusive language datasets," *AI and Ethics*, vol. 2, no. 1, pp. 79–101, 2022, doi: 10.1007/s43681-021-00081-0.
- [43] T. Davidson, D. Warmley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/icwsm.v11i1.14955.
- [44] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 88–93. doi: 10.18653/v1/N16-2013.
- [45] Kaggle.com, "Dynamically Generated Hate Speech Dataset," kaggle.com. [Online]. Available: <https://www.kaggle.com/datasets/usharengaraju/dynamically-generated-hate-speech-dataset>
- [46] J. Golbeck *et al.*, "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on Web Science Conference*, New York, NY, USA: ACM, Jun. 2017, pp. 229–233. doi: 10.1145/3091478.3091509.
- [47] T. Caselli, V. Basile, J. Mitrovic, I. Kartoziya, and M. Granitzer, "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 6193–6202, 2020.
- [48] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1415–1420. doi: 10.18653/v1/N19-1144.

BIOGRAPHIES OF AUTHORS



Praneetha Garagadakuppe Nanjundappa     received the B.E. Engineering degree in Computer Science and Engineering from GSKJTI affiliated to VTU in 2013 and M.Tech. degree in Computer Science and Engineering from in 2015 from Saphagiri College of Engineering affiliated to VTU and currently pursuing Ph.D. under VTU University in machine learning domain. Currently, she is an Assistant Professor at the Department of Computer Science and Engineering, Ramaiah University of Applied Science. She is a member of The International Association of Engineers. Her research interests include machine learning and natural language processing. She can be contacted at email: praneethaguddi@gmail.com.



Dr. Kamalakshi Naganna     is a Professor and Head of the Department of Computer Science & Engineering at Saphagiri College of Engineering Bengaluru. She received her M.Tech. from M.S. Ramaiah Institute of Technology which was affiliated to VTU, Ph.D. in the area of Image Processing from University of Mysore. She has more than 26 years of experience. She has published quite a number of research papers in International Conferences and Journals. She is a member of Indian Society of Technical Education (ISTE), IAENG and Computer Society of India. She has guided many M.Tech. students. She has been actively involved in Research work in the area of image processing, block chain technology and machine learning. She has guided 5 KSCST funded projects. She was in Editorial Board of MASAUM Journal of Image Processing (MJIP). She is in the Organizing and Advisory committees of various conferences and also reviewer of IEEE papers. She has contributed 2 chapters of springer book. She can be contacted at email: kamalnags@gmail.com.