Survey on Detecting Hate speeches in Social Media with Logistic Regression on Twitter Data

Vepagunta Govardhan Kumar, Tani Vetri, Sarvepalli Sivapriya, Payyayi Priya, Bandi Dayakar
IV – Year B. Tech CSE, Gokula Krishna College of engineering

***Abstract:** The paper addresses the challenge of detecting hate speech on social media using deep learning techniques. It proposes a novel model fusion approach that leverages a comprehensive dataset from 18 sources, consisting of 0.45 million comments. The method employs CNN and BiLSTM with an attention mechanism, specifically optimized for different subsets of the dataset. media targets individuals and groups based on demographics, making its detection a crucial yet complex task. Traditional supervised models rely on labeled datasets, but hate speech detection involves multiple aspects such as writing style and topic. This study introduces an ensemble learning approach that captures various aspects of hate speech by constructing ensembles from different perspectives Experiments conducted on five datasets in multiple languages (English, Turkish, Italian) demonstrate improved performance over baseline models. The proposed method outperforms state-of-the-art ensemble approaches in multiple scenarios, particularly improving F1-score for hate speech detection Hate speech is a growing problem that can harm individuals and communities. potential solution is using machine learning algorithms to detect and flag hate speech in text-based data This approach involves training a model on a labeled dataset containing examples classifiedas hate speech or nonhatespeech.*

***Keywords:** Hate speech detection, deep learning, hate speech detection, ensemble learning Machine learning, deep learning, Hate speech, text analysis, Twitter, multinomial logistic regression*

INDRODUCTION

Hate speech is a growing concern in the digital space, particularly on social media platforms where people can freely express their thoughts and opinions. Platforms such as Twitter and Facebook, while promoting communication and social interaction, have unfortunately also become channels for the rapid spread of offensive, discriminatory, and harmful speech targeting individuals or communities based on race, gender, ethnicity, or religion. Identifying and mitigating hate speech is essential to maintaining healthy online discourse, reducing societal harm, and upholding ethical digital interactions.

Traditional methods for detecting hate speech involve manual moderation, but given the enormous volume of online content, automated approaches using machine learning and deep learning have become indispensable. Various models such as CNN, LSTM, BiLSTM, and ensemble methods have been employed to improve detection accuracy and ensure better generalization across diverse datasets. However, challenges remain, including the ambiguity of hate speech, its evolving nature, and the difficulty of distinguishing it from offensive but non-hateful content. To address these issues, advanced natural language processing techniques and ensemble learning strategies have been developed to enhance detection systems

LITERATURE SURVEY

W. Sharif et al. (2024) proposed a model fusion approach using CNN and BiLSTM with an attention mechanism to detect hate speech more effectively. Byintegrating data from 18 sources comprising 0.45 million comments, they demonstrated that deep learning-based fusion models significantly outperform traditional models. Their method achieved an accuracy of 89% and provided improved generalization across multiple social media platforms.
I. E. Kucukkaya and C. Toraman (2024) investigated the effectiveness of ensemble learning for hate speech detection by capturing different aspects of hate speech, such as topic- specific and writing-style variations. They experimented with multiple datasets across different languages and found that their ensemble construction approach improved the F1 score for hate speech classification in 80% of the cases compared to single-model classifiers.

A. Toktarova et al. (2023) explored the performance of deep learning techniques such as LSTM, BiLSTM, and CNN in hate speech detection. Their findings highlighted that BiLSTM outperformed shallow learning models by effectively capturing context and sequential dependencies within tweets. They also integrated word embeddings like Word2Vec and GloVe, further improving classification accuracy.

V. B. Ohol et al. (2023) examined the application of machine learning techniques for hate speech detection, focusing on feature extraction and text analysis. They noted that the effectiveness of machine learning models depends heavily on the quality of annotated datasets and the choice of classification algorithms. Their study pointed out biases in training data as a key challenge in improving detection fairness.

Sennary et al. (2024) proposed a hybrid approach integrating Convolutional Neural Networks (CNN) and multiple machine learning classifiers to enhance hate speech detection on social media platforms. Their study utilized three datasets, including *Hate Speech Offensive Tweets*, *Twitter Hate Speech*, and a combined dataset incorporating *Cyberbullying Data*The results demonstrated that SVM achieved the highest accuracy of 99.2%, outperforming other classifiers. The study highlighted the challenge of distinguishing hate speech from offensive language, emphasizing the need for fine-grained classification techniques and robust feature extraction methods.

Block diagrams:

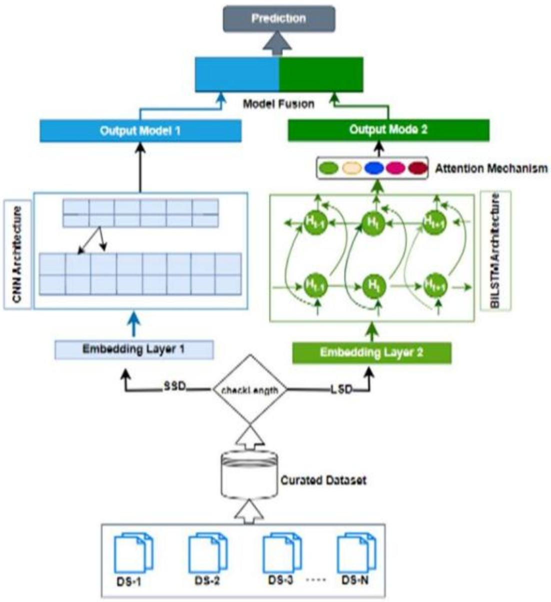

Fig-1.1

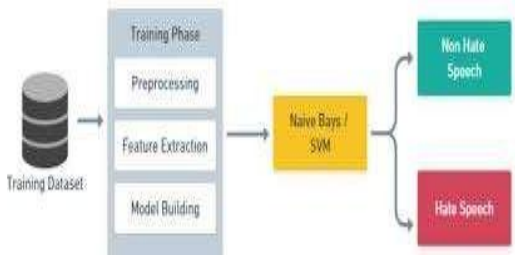

Fig-1.2

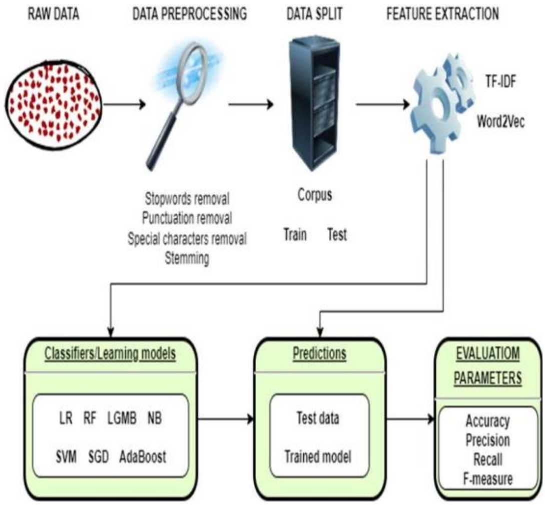

Fig-1.3

Flow charts:

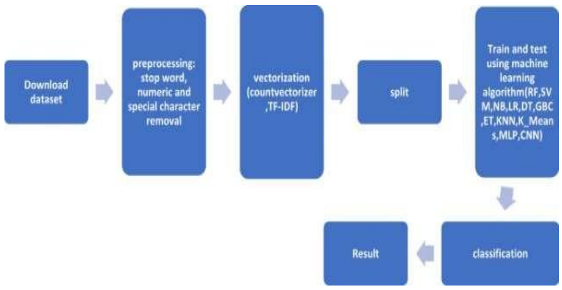

Fig-1.4

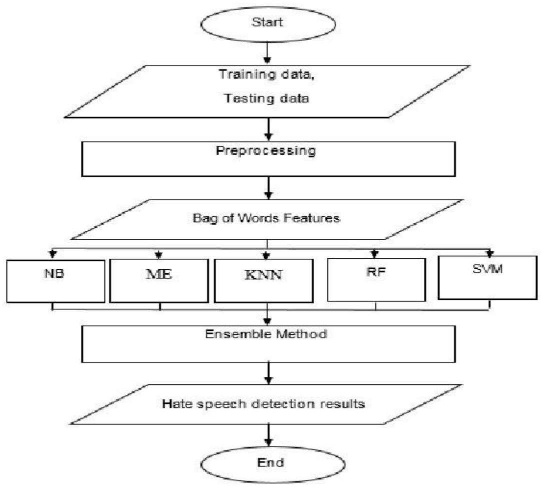

Fig-1.5

Algorithms:

Used CNN and BiLSTM for short and long sequence text classification, integrating an attention mechanism for improved performance.Implemented ensemble learning with Gaussian Mixture Models (GMM), influential combination ensembles, and soft voting approaches for hate speech detection. Applied Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), Naïve Bayes (NB), and Random Forest (RF) with TF-IDF feature extraction. Utilized deep learning models such as LSTM, BiLSTM, and CNN, along with word embeddings like Word2Vec and GloVe for classification. Combined conventional ML models (SVM, LR, RF) with Convolutional Neural Networks (CNN) and hybrid approaches, leveraging TF-IDF and N- gram-based features.

RESULTS AND DISCUSSION

model	accuracy	precision	recall	F1 score
CNN(short text)	88%	0.87	0.85	0.86
BiLSTM	92 %	0.89	0.91	0.90
CNN+ BiLSTM	89%	0.88	0.91	0.89

Table-1

The combination of CNN for short text and BiLSTM for long text resulted in an overall accuracy of 89%, surpassing individual models. CNN performed better for short sequences, achieving an 88% F-score, while ensemble methods. The study noted that different styles of hate speech required tailored ensemble strategies for better classification.

Method	Normal	Offensive/Abusive	Spam	Hate	Weighted F1
Ensemble-Bagging	0.898 ± 0.008	-	-	0.619 ± 0.018	0.847 ± 0.006
Ensemble-Fold	0.908 ± 0.006	-	-	0.614 ± 0.030	0.854 ± 0.009
Ensemble-Topic	0.879 ± 0.008*	-	-	0.438 ± 0.110*	0.798 ± 0.017*
Ensemble-Influential	0.844 ± 0.021*	-	-	0.493 ± 0.014*	0.780 ± 0.019*
Ensemble-GMM	0.828 ± 0.004*	-	-	0.490 ± 0.014	0.766 ± 0.004*

Traditional ML models exhibited strong classification performance, with SVM achieving an accuracy of 96.3% and an F1- score of 98%. However, K-Means clustering performed poorly at 78.3% accuracy, highlighting the limitation of unsupervised methods in hate speech detection. Deep learning models, particularly BiLSTM, demonstrated superior accuracy compared to shallow learning models. The integration of word embeddings significantly improved performance, allowing for better semantic understanding of text. The study found that hybrid approaches, particularly a combination of decision trees and Naïve Bayes, resulted in an F1-score of 0.885. The inclusion of an ensemble framework further improved classification robustness across datasets.

CONCLUSION

The effectiveness of hate speech detection models varies based on the chosen approach, with deep learning models like BiLSTM providing superior contextual understanding compared to traditional machine learning classifiers.
Ensemble models generally outperform standalone classifiers by leveraging multiple learning approaches to capture different nuances in hate speech.
Feature extraction techniques, such as TF- IDF, N-grams, and word embeddings, significantly impact model accuracy and must be carefully selected based on dataset characteristics.
Hybrid approaches that combine conventional ML techniques with deep learning models offer a balanced trade-off between computational efficiency and classification performance.
Future research should focus on incorporating transformer-based architectures, such as BERT and GPT, and improving explainability to enhance trust in automated hate speech detection systems.

REFERENCES

[1] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, et al. (2023).Wiedemann et al. (2020), "Hate Speech Detection Using ML and DL Models."

[2] Glazkova et al. (2021), "Ensemble Learning for Hate Speech Detection."

[3] Davidson et al. (2017), "Automated Hate Speech Detection and the Problem of Offensive Language."

[4] Badjatiya, P., et al. (2017), "Deep Learning for Hate Speech Detection in Tweets."

[5] Mishra, P., et al. (2019), "Deep Learning for Hate Speech Detection: A Comparative Analysis."

[6] Kowsari, K., Heidarysafa, M., Brown, D. E., et al. (2018), "RMDL: Random Multimodel Deep Learning for Classification."

[7] Gao, W., et al. (2020), "Detecting Hate Speech on Twitter Using a Convolution- GRU Based Deep Neural Network."

[8] Li, J., et al. (2018), "Combating Hate Speech on Social Media with Unsupervised Text Style Transfer."

[9] McCallum, A., & Nigam, K. (1998), "A Comparison of Event Models for Naive Bayes Text Classification."

[10] Kumar, R. R., Reddy, M. B., & Praveen, P. (2019), "Text Classification Performance Analysis on Machine Learning."

[11] Heidarysafa, M., Kowsari, K., Brown, D.E., et al. (2018), "An Improvement of Data Classification Using RMDL."

[12] Alsubait, T., & Alfageh, D. (2021), "Comparison of Machine Learning Techniques for Cyberbullying Detection."

[13] Djuric, N., et al. (2015), "Hate Speech Detection on Twitter: A Comparative Study."

[14] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005), "Text Classification Using Machine Learning Techniques."

[15] Janani, R., & Vijayarani, S. (2023), "Automatic Text Classification Model Based on Machine Learning."

[16] Miao, F., Zhang, P., Jin, L., & Wu, H. (2018), "Chinese News Text Classification Based on Machine Learning Algorithm."

[17] An, J., Mikhaylov, A., & Chang, T. (2024), "Relationship Between Platform Popularity and NFT Asset Prices."

[18] Malik, Pang, & van den Hengel (2022), "Transformer-Based Models for Hate Speech Detection."

[19] Kowsari, K., Meimand, K. J., & Heidarysafa, M. (2019), "Text Classification Algorithms: A Survey."

[20] Li, J., et al. (2018), "Combating Hate Speech on Social Media with Neural Network-Based Text Style Transfer."

[21] Wulczyn, E., et al. (2017), "Hate Speech Detection with Comment Embeddings and LSTM Networks."

[22] Zampieri et al. (2020), "SemEval-2020: Shared Task on Hate Speech Detection."

[23] Toraman et al. (2022), "Hate Speech Detection Using Transformer-Based Models."

[24] Founta et al. (2018), "Large-Scale Analysis of Hate Speech on Twitter."