International Conference on Current Research Trends
ICCRT 2025 - Hybrid (Hafizabad, Pakistan & Online)
20-21 December 2025

Hate Speech Detection Using Ensemble Approach and Embedding Technique

Bira Alam [1*], Fatima Abbas [1], Nafees Ayub [2]

[1]Riphah International University Faisalabad, Pakistan

[2]Government College University Faisalabad, Pakistan

* Corresponding Email: bira.alam@riphahfsd.edu.pk

Published Online: 24 January 2026 by JCPP *DOI: https://doi.org/10.64060/ICPP.02*

ABSTRACT

The rising trend of hate speech on the internet is being major concern to the internet security and societal coexistence, which will requires efficient automated detection tools. As much as different machine learning strategies have been suggested, the issue of the high accuracy using limited data still remains a challenge. This paper introduces a collective detecting frame of hate speech that was tested on the Ethos Binary dataset. Various machine learning classifiers such as K-Nearest Neighbor, Naive Bayes, Logistic Regression, and Decision Tree are used with pre-trained GloVe word embeddings in order to extract semantic representations of textual data. The models are also trained and tested in various hyperparameter configurations, to be robust. The experimental findings indicate that Decision Tree classifier is much better than other models, with a precision of 87, recall of 93, F1-score of 90 and an overall accuracy of 91. The results have shown that an ensemble learning approach with embedding techniques has the potential to greatly increase the performance of hate speech detection. This work helps to enhance viable and scalable solutions to the content of moderating dangerous content online.

Keywords*:* Hate Speech Detection; Ensemble Learning; Word Embeddings; GloV; Ethos Dataset; Machine Learning

1. Introduction

The high presence of social media sites had been greatly contributed to the spread of user-created information, including hate speech would discriminates against the user or any group of people based on factors like race, religion, gender, and ethnicity. The definition of hate speech is consistently an aggressive, abusive, or offensive language that is directed to the covered groups, and is usually aimed at talking down and causing hostility [1]. This kind of content does not only undermine online conversation, but also adds to the psychological damage, social rejection, and, in the worst scenario, real-life aggression.

Hate speech has been increased in the digital environment because of the speed with which it is shared on social networks like Facebook, Twitter, Instagram, and YouTube. The hate speech can be very tricky to pick up because it is usually propagated using misconceptions and inflammatory stories as noted in [2]. In addition, hate speech may be expressed not only in explicit ways, direct threats and slurs, but also in implicit ways, including sarcasm, coded language, and contextual abuse, which are more difficult to reveal by the conventional rule-based methods. [3] further pointed out that hate speech differs according to the categories of the target such as race, religion, gender and sexual orientation, which prevents the necessity of strong automated detection models.

Machine learning has become a viable option to overcome the flaws of the manual moderation and rule-based filtering. The classical machine learning algorithms are especially applicable to smaller datasets, on which the deep learning algorithms may be affected by overfitting and expensive computations. The above models have been shown to work in the past, namely Naive Bayes, Logistic Regression, Decision Trees, and ensemble methods in text classification tasks [4]. However, feature representation to a significant extent defines performance and the ability of the models to encode semantic background.

This paper is devoted to the Ethos Binary dataset, which is a under-researched hate speech corpus of short user-created remarks that are classified as hateful or non-hateful[1]. The dataset is relatively small and thus poses a challenge in terms of good classification and thus is a good testbed in terms of testing ensemble-based methods. This study is motivated by the earlier studies that have focused on the development of strong machine learning systems[2]. The ensemble approach used in this study also involves several classical classifiers that have been trained using GloVe-based embedding. The voting mechanism is used to aggregate the ensemble predictions to enhance generalization and elimination of the bias of each model.

A number of recent research papers have investigated sophisticated ways of detecting hate speech. To combine learning of related abusive language tasks, [3] suggested a multi-task learning framework, which performed better. Transformer-based models like DistilBERT were proved to be rather effective, whereas ensemble classifiers like AdaBoost could be more effective than single models proved to be, as demonstrated at[4]. It has been proven that ensemble classifiers such as AdaBoost could perform better in single models [5] .A profound ensemble-based NLP framework was proposed by [6] and it was superior to the traditional methods. These methods are good in that they can get high results although they usually use large scales of data or complex structures. Conversely, the paper focuses on efficiency and effectiveness on a scarcely studied dataset based on classical models with improvements of ensemble learning and word embeddings.

The main goal of the study will be to enhance the accuracy and reliability of hate speech classification on the Ethos dataset with the aid of thorough text preprocessing, ready-to-use GloVe embeddings, and an ensemble-based classification system. This work can advance scalable and useful hate speech detection systems that can be used to provide safer and more inclusive online communities because it focuses on a relatively unexploited dataset.

2. Literature Review

The section will summarize the existing studies in hate speech detection in different languages and social media platforms, with a focus on the methodologies, data sets and performance metrics.

The past few years have seen research on various methods of detecting hate speech. A modular neural classifier in cross-lingual detection on English, Italian, and German, the proposal by [7] comprised LSTM, GRU, and Bi-LSTM, word vectors, n-grams, sentiment lexicon, emojis, and social network features. [8] trained CNN, RNN, and BERT-based models on Arabic tweets, and compared the hybrid and single architectures to find the best detection.

In low-resource languages[4] addressed Urdu tweets by combining sentiment analysis with SMOTE, dynamic stop words filtering, and feature selection techniques, evaluating SVM and Multinomial Naïve Bayes. In a similar way[5] used the LSTM and Bi-LSTM networks to detect hate speech on social media, utilizing preprocessing algorithms, including stemming, tokenization and one-hot encoding, and word representations.

The Roman Urdu Hate Speech corpus was created by [9]and the study used NB, LR, RF, SVM, and CNN among the supervised learning methods. Transformer based methods have become increasingly popular too. [10][11] have also identified the difficulties in identifying subtle hate speech, with BERT embeddings and neural networks or ensemble deep learning models. Some of the researches have been directed toward regional and low-resource languages.

Ensemble learning and hybrid architectures have been brought out to the forefront [12]. The article [13] exhibited stacked English tweet recordings utilizing SVM, LR, and XGBoost with word2vec and universal encoding.

The other studies pointed out difficulties in the accuracy of classification, interpretability, and cross-lingual applicability. Some promising results were obtained with methods that incorporate both traditional machine learning and deep learning, transformers, and attention mechanisms used on

languages, including Arabic, Urdu, Roman Urdu, Indonesian, and Dravidian code-mixed datasets. The following Table 1 shows a comparison of the literature.

Table 1: Suggested Comparison Table for Literature Review

Author(s)	Year	Language/Dataset	Methodology	Features	Key Findings
Corazza et al[6]	2020	English, Italian, German	LSTM, GRU, Bi-LSTM	Word vectors, n-grams, emojis, sentiment lexicon	Cross-lingual detection with modular classifier
Alshalan& Al-Khalifa[11]	2020	Arabic	CNN, RNN, BERT	Word embeddings	Hybrid models improved detection accuracy
Paul& Bora[12]	2021	English	LSTM, Bi-LSTM	One-hot encoding, word embeddings	Sequence modeling effective for text sequences
Tița& Zubiaga[13]	2021	Multilingual	mBERT, XLM-RoBERTa	Transformer layers, dropout	Fine-tuned models improved class imbalance handling
Mahajan et al.[14]	2024	Multilingual	CNN-LSTM, BiLSTM, BiGRU	GloVe embeddings	Hybrid ensemble model outperformed single models

3.Methodology

This paper has used Ethos Binary Dataset, which is a highly curated dataset that is used to do binary classification in the detection of hate speech. Two major columns make up the dataset: comments and isHate and are separated by semicolons to be easily processed as structured data. It comprises 998 comments out of which 433 ones have been classified as hate speech (1) and 565 as non-hate (0). Although the dataset is rather small, it is of good quality and structured to be referred to as a good resource to develop, test, and evaluate machine learning models. Its size can be easily preprocessed, extracted features, and tested on various model settings, which is adequate to the methodology used, although its weaknesses must be admitted.

3.1 Data Preprocessing

The raw textual data requires preprocessing so that this data can be converted into a uniform format that can be read by machine learning algorithms. In the case of Ethos data, the steps taken were:

Tokenization: The keras tokenizer was used to segment text data into individual tokens and each word was associated to a unique integer in terms of frequency. The process gave 3,885 different tokens.

Stopword Elimination**:** Instances of the most frequent word with minimal semantic value, including is, the, and, etc., were filtered out by list of NLTK stopwords with a customized filtering function and only informative words were extracted.

Stemming: The Porter Stemmer of NLTK was used to convert words to their root form, standardizing variations and enhancing the efficiency and consistency of the downstream feature representation.

The Figure 1 showcases the most common words in hateful messages of Ethos-based dataset, which underlines the use of offensive and discriminative words that trigger the necessity of automated hate speech recognition.

Figure.1 Word Cloud of Ethos[15]

3.2 Data Splitting

The dataset was divided into two sections (training and testing) to facilitate the quality training and testing of the model. This will guarantee a proper trade-off between model learning and unbiased evaluation of performance, despite small dataset.

3.3 Embedding Technique

Pre-trained GloVe (Global Vectors for Word Representation) embeddings were used to get semantic word representations. The tokens were mapped to a semantical relationship and contextual meaning of a 100-dimensional dense vector space. By using the GloVe embeddings to supplement a small dataset, it is possible to overcome certain weaknesses of these small data sets to a certain extent and offer rich, context-sensitive feature representations.

3.4 Ensemble Learning

In order to improve predictive performance, ensemble learning technique was used, which involves the integration of various machine learning algorithms. The learning algorithms that are part of the ensemble have a different purpose of hate speech detection.

K-Nearest Neighbor classifier is an effective model to detect hateful content as it uses the similarity of text samples in the feature space to determine hateful words, thus an effective model to identify patterns in which hateful words resemble those previously encountered.

Naive bayes takes advantage of the fact that there are probabilistic associations between words and class labels, which enables it to identify those words that occur frequently as offensive or discriminatory.

The reasons for the use of Logistic Regression Logistic Regression is used to model a textual characteristic relationship to a label of a classification that has a linear decision boundary and provides stability and interpretability in classification decisions.

This is because the Decision Tree classifier will capture non-linear interactions of the features, which will allow it to identify the complex linguistic design and contextual patterns common with hate speech.

The ensemble arrangement is complementary and strengthens the weaknesses of the individual classifiers by combining these strengths.

Every model was trained on Ethos and several cases with varied random seeds were pooled together using a soft Voting Classifier. This measure enhances prediction stability and accuracy which is especially desirable in light of a relatively small dataset.

3.5 Evaluation Metrics

Precision, recall, F1-score and accuracy were used as quantitative measures to evaluate model performance.

Precision: The ratio of correctly predicted hate comments of all the hate comments made.

Recall: Percentile of correct identification of actual hate comments.

F1-Score: Precision and recall in a given sample, weighted towards their harmonic mean, or balance between false negatives and positives.

Precision: The total percentage of correct comments.

All these metrics give a holistic analysis of the model performance to ensure high and interpretable results of the performance even with a small dataset. To give a summary of the hat speech detention framework to be proposed, Figure 2 illustrates the entire methodology workflow embraced in this study. The diagram gives an overview of the key steps of the strategy, such as preprocessing of data, feature extraction based on GlovE embeddings, training of an ensemble model, and evaluation.

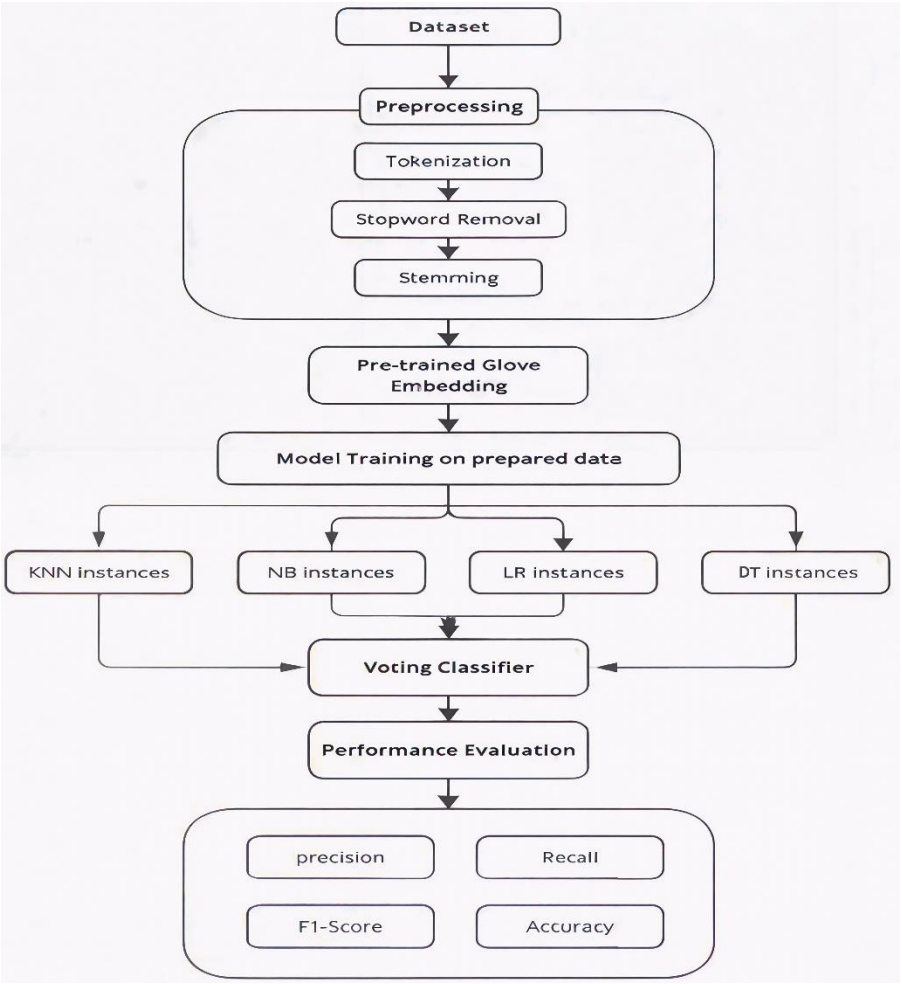

Figure.2 Methodology Diagram

3.6 Methodology Workflow

In Figure 2. It is initiated by Ethos data collection, then there is data preprocessing involving tokenizing, removing stopwords, and stemming. Pre-trained GloVe embeddings are subsequently used to transform usable texts into numerical feature vectors. Several machine learning models KNN, Naive Bayes, Logistic Regression, and Decision Tree are independently trained and assembled into an ensemble with the help of a soft Voting Classifier. Lastly, the precision, recall, F1-score and accuracy are used to assess the performance of hate speech detection with the ensemble model.

3.7 Limitations

The Ethos dataset (998 comments) is quite small in its size, which can limit its ability to be used to generalize to the larger or more heterogeneous online context. However, its structured design, accurate labelling, and representative response of hate and non-hate speech makes it sufficient to develop and test the offered machine learning model. GloVe embeddings and ensemble learning are also used to reduce the weaknesses of the dataset, increase robustness of the model, and predictive confidence.

4. Results

The current section states the results of the experiment conducted on the Ethos Binary Dataset. The effectiveness of four ensemble-based machine learning classifiers, K-Nearest Neighbor (KNN), Naive Bayes (NB), Logistic Regression (LR), and Decision Tree (DT), were determined with the help of the common metrics: precision, recall, F1-score, accuracy, and ROC-AUC. The conclusions are presented in a table form to be more understandable and easier to compare.

4.1 Performance Comparison

Table 2 is a report of the quantitative result of both models on the Ethos dataset. The Decision Tree ensemble performs better in all the evaluation metrics than the other classifiers.

Table 2: Performance Comparison of Machine Learning Models on the Ethos Dataset

Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	ROC–AUC
K-Nearest Neighbor	70	74	72	75	0.87
Naive Bayes	67	67	67	73	0.76
Logistic Regression	66	64	65	71	0.76
Decision Tree	**87**	**93**	**90**	**91**	**0.93**

4.2 Discussion of Results

These findings prove that the Decision Tree ensemble model is the most accurate (91) and has the highest F1-score (90), which means that it is better able to detect instances of hate speech at the expense of false classification. Its high recall (93%) entails the fact that it has a high sensitivity in terms of identifying the hateful content, which is extremely crucial in hate speech detection work.

KNN has matched performance in terms of accuracy regarding precision and recall, but its accuracy is lower than that of the Decision Tree. Naive Bayes and Logistic Regression have relatively high lower performance, and perhaps this can be explained by the fact that they have stronger assumptions concerning the independence of features and the linear separability, respectively.

The large ROC–AUC (0.93) of the Decision Tree also supports this fact and proves that the decision tree is strong and classifies well on the Ethos data. These results imply that ensemble-based Decision Tree models that have been trained on semantic Glov embeddings are highly appropriate in detecting hate speech despite being trained on relatively small data sets.

5 Conclusion

This paper examined the usefulness of ensemble-based machine learning models in hate speech recognition with the Ethos Binary Dataset and pre-trained GloVe word embedding. There were four classifiers, namely Naive Bayes, K-Nearest Neighbor, Logistic Regression and Decision Tree, whose performance was tested on standard performance metrics.

The experimental outcomes show that Decision Tree ensemble model has the best results, having the highest accuracy at 91, precision at 87, recall at 93, and F1-score at 90. These results suggest that the Decision Trees, in combination with the semantic word representations, can be especially useful in doing the hate speech classification. KNN also demonstrated good performance whereas Naive Bayes and Logistic Regression performed with lower effectiveness.

Though the Ethos dataset is not that large, its systematic labeling and certain classification of hate and non-hate content show that it can be evaluated using specific methods. Pre-trained GloVe embeddings and ensemble learning were used, which contributed to the overcoming of data size constraints and the enhancement of the model robustness.

The research in the future can consider the investigation of bigger and more varied data to contribute to the generalization improvement further. Also, more complex ensemble strategies, deep neural networks, and imbalance-correcting methods might also enhance the accuracy of detection. On the whole, this paper has shown that well-developed ensemble techniques can give credible and understandable decisions to automated hate speech detection to help to make the internet a safer place.

References

[1] S. Mukherjee and S. Das, "Application of Transformer-Based Language Models to Detect Hate Speech in Social Media," vol. 2, no. December 2021, pp. 278–286, 2023, doi: 10.47852/bonviewJCCE2022010102.

[2] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," Inf., vol. 13, no. 6, pp. 1–22, 2022, doi: 10.3390/info13060273.

[3] A. Haque, "Hate Speech Detection in Social Media Using the Ensemble Learning Technique Hate Speech Detection in Social Media Using the Ensemble Learning Technique," 2023.

[4] K. U. Wijaya and E. B. Setiawan, "Hate Speech Detection Using Convolutional Neural Network and Gated Recurrent Unit with FastText Feature Expansion on Twitter," vol. 9, no. 3, pp. 619–631, 2023, doi: 10.26555/jiteki.v9i3.26532.

[5] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," J. Intell. Syst., vol. 30, no. 1, pp. 578–591, 2021, doi: 10.1515/jisys-2020-0060.

[6] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, and A. Abayomi-Alli, "A probabilistic clustering model for hate speech classification in twitter," Expert Syst. Appl., vol. 173, no. February, p. 114762, 2021, doi: 10.1016/j.eswa.2021.114762.

[7] D. Mody, Y. D. Huang, and T. E. Alves de Oliveira, "A curated dataset for hate speech detection on social media text," Data Br., vol. 46, p. 108832, 2023, doi: 10.1016/j.dib.2022.108832.

[8] M. Almaliki, A. M. Almars, I. Gad, and E. S. Atlam, "ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media," Electron., vol. 12, no. 4, pp. 1–16, 2023, doi: 10.3390/electronics12041048.

[9] R. T. Mutanga, N. Naicker, and O. O. Olugbara, "Detecting Hate Speech on Twitter Network using Ensemble Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 331–339, 2022, doi: 10.14569/IJACSA.2022.0130341.

[10] Z. Al-Makhadmeh and A. Tolba, "Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach," Computing, vol. 102, no. 2, pp. 501–522, 2020, doi: 10.1007/s00607-019-00745-0.

[11] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," Appl. Sci., vol. 10, no. 23, pp. 1–16, 2020, doi: 10.3390/app10238614.

[12] C. Paul and P. Bora, "Detecting Hate Speech using Deep Learning Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 2, pp. 619–623, 2021, doi: 10.14569/IJACSA.2021.0120278.

[13] T. Tiţa and A. Zubiaga, "Cross-lingual Hate Speech Detection using Transformer Models," arXiv Prepr. arXiv2111.00981, 2021, [Online]. Available: https://arxiv.org/abs/2111.00981%0Ahttps://arxiv.org/pdf/2111.00981

[14] E. Mahajan, H. Mahajan, and S. Kumar, "EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media," Expert Syst. Appl., vol. 236, no. May 2023, p. 121228, 2024, doi: 10.1016/j.eswa.2023.121228.

[15] A. Chhabra and D. K. Vishwakarma, "A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset," 2023 Int. Conf. Innov. Trends Inf. Technol. ICITIIT 2023, pp. 1–4, 2023, doi: 10.1109/ICITIIT57246.2023.10068574.

Declaration

Conflict of Study: 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.

Author Contribution Statement: All authors contributed equally.

Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.

Acknowledgement: Not Applicable

Declaration of AI Use: Not Applicable

Author(s) Bio / Authors' Note

Bira Alam *from Riphah International University, Faisalabad, Pakistan. Email:* bira.alam@riphahfsd.edu.pk

Fatima Abbas *from Riphah International University, Faisalabad, Pakistan. Email:* fatima.abbas@riphahfsd.edu.pk

Nafees Ayub *from Government College University Faisalabad, Pakistan. Email:* nafees.ayub@gcuf.edu.pk