[1]N. Solomon Praveen Kumar

[2]Dr. M. S. Mythili

EAF-HSD: Ensemble Adaptive Fuzzy Logic-Based Hate Speech Detection on Social Media

Journal of Electrical Systems

Abstract: - With rise in incidence of hate speeches on social media platforms, this has become an important issue where sociability and human beings are among the greatest elements that are affected. Addressing this issue, an advanced algorithm for sentiment analysis and hate speech detection is proposed in this paper. This technique involves a pre-processing of social media data, which encompasses adaptive fuzzy logic-enhanced DBSCAN clustering for topic detection as well as semantic pattern recognition indicative of hate speech. An ensemble learning using Naive Bayes and Random Forest is used to detect the hate speech.The results confirm that the proposed EAF-HSD approach aids in reaching higher accuracy than the existing approaches. The accuracy of 94% shown in the model that has been proposed as far as classifying offensive language is concerned as the best of all and it has an overall accuracy of 92%, which shown a massive breakthrough over the baseline models. he proposed work introduces novel advancements in hate speech detection on social media platforms. Integrating adaptive fuzzy logic with adaptive DBSCAN clustering method to capture the intricate hate speech patterns. An ensemble learning framework combines diverse classifiers is used to classify the hate speech and non-hate speech accurately.

Keywords: Hate Speech, Sentiment analysis, Ensemble learning, DBSCAN, Adaptive fuzzy.

1. Introduction

On the digital age, social media has revolutionized the modes of communication, relationship building, and information dissemination [1]. Social media platforms such as Facebook, Twitter and Instagram are the ones that have paved way for people to connect with global networks and bring their diverse thoughts to the mainstream [2]. The rise of such social media has offered people the chance to have a strong voice and participate in the public debate [3]. However, the newly acquired right of speech has also enabled the diffusion of internet-based hate speech [4]. The concept of "hate speech" includes diverse forms of communication that express dislike or contempt for a certain group of people because of their race, religion, gender, ethnicity, sexual orientation, etc. [5]. In the recent times, there has a increase in the number of hate speech on the internet. This has been evident through the use of social media platforms by individuals to share defamatory words, precipitate violent acts, and spread negative ideologies [6].

1.1 Motivation

The widespread existence of cyber-bullying on social media platforms has become a significant societal issue [7]. Hate speech has harmful consequences that go beyond the digital domain, leading to tangible outcomes such as prejudice, aggression, and societal fragmentation [8]. To tackle this problem, it is necessary to develop strategies that utilize state-of-the-art technologies to accurately detect the hate speech in social-media data

1.2 Problem Statement

The problem statement centres on low performance of the existing hate speech detection methods due to their inaccuracy while identifying and categorising the hateful content in the social platforms. The pre-existing methods are facing this problem because the hate speech is dynamic and complex, the numbers of false positives and negatives are high. Manual moderation is not an option because of the fact that user-generated content is too large. Consequently, instead of human intervention, auto-detectors of hate speech should be drawn up using sophisticated tools like natural language processing and machine learning algorithms. These systems aim to effectively analyze and classify the hate speech found in social-media.

1.3 Objectives

The primary objective of this proposed model is to develop a novel approach for hate speech detection on social media platforms. This approach will leverage advanced techniques from Natural Language Processing (NLP), Machine Learning (ML), and ensemble learning to accurately identify and classify hate speech content.

[1]Research Scholar, Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, India. solomon@bhc.edu.in

[2]Associate Professor, Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, India. mythili.ca@bhc.edu.in

1.4 Significance of the Study

This research uses natural language processing and machine learning to differentiate hate speech. The current project focuses on online security and inclusivity, with the aim of educating the diverse community and making the digital world a safe space for everyone. These systems perform a crucial function of analyzing and categorizing web content on different issues which serves to build a safer digital platform. Also, the outcomes of the study can be the starting point for social media platforms to immediately develop the strategy that will be aiming at counteracting the hate speech.

1.5 Contributions

The paper presents multiple facets of the contribution. To begin, this paper presents a revolutionary strategy for the detection of hate speech that combines adaptive fuzzy logic-based clustering and ensemble learning techniques. This approach therefore constitutes a shift from traditional methods and a more holistic and complex insight into the content of hate speech. Besides, it carries out a comprehensive empirical analysis of the suggested approach using the social media datasets collected from the real world, showing that it is applicable and successful in hate speech detection. Finally, it gives the perspectives on the societal effects of online hate speech and propose multidisciplinary cooperation as a way to tackle this complex problem.

1.6 Organization of the Article

The remainder of this article is structured as follows: In Section 2, it reviews the literature on hate speech detection by discussing the state-of-the-art methods as well as their weaknesses. Section 3 presents the proposed methodology, which include adaptive fuzzy logic-based clustering and ensemble learning framework. Section 4 contains the empirical evaluation of the proposed methodology with real-time social media datasets. Finally, section 5 brings the article to its end by summarizing the main points, implications for further research and closing remarks.

2. Related Works

Rafael et al. [9] justified the prevalence of hate speech in social media and indicated the purpose of ML against this phenomenon. The researchers highlighted that the existing articles concentrated on ML methods to identify between hate speech, sarcasm, and offensive language. Even though many ML models have been suggested, they use a single type of feature extraction or classification algorithm. Authors brought forward the argument that composite feature extraction techniques and classification algorithms should be used to improve hate speech detection. They presents a framework to find how these techniques interact and they show that it is effective in the selection of combining techniques that develop a Multiple Classifiers System (MCS). The experimental study of the team which used 4 hate speech classification datasets [10-13] showed that the proposed framework overtook other heuristics in terms of accuracy, proving its importance in creating HSMAs for hate speech detection.

Iorliam et al. [14] presented the comparative analysis of Hate Speech classification while using deep learning approach with LSTM and CNN technique. Data collected during the trial show that LSTM classifier achieved the classification accuracy of 92.47% and CNN classifier got the classification accuracy 92.74%. The outcomes of this research reveal that deep learning strategies are capable of perfect identification of the hate speech from normal conversation.

The purpose of the research work [15] was to assess shallow machine learning and deep learning methods for the cyberbullying detection issue. For this use, they deployed three deep and six shallow learning algorithms. The results showed LSTM to be the most suitable for cyber-bullying detection, mainly in accuracy and recall indicators.

The contribution of using domain-specific word embeddings as features and a bidirectional LSTM-based deep model as classifier for automatic hate speech detection was evaluated by Hind Saleh et al. [16]. This technique implied a process of association of the negative meanings of words to facilitate the detection of coded language. In addition, the investigators delved into the application of transfer learning model, BERT, in the context of hate speech detection, as a task of binary classification. The contextual findings showed that the incorporation of domain-specific embedding of words with Bidirectional LSTM-based deep model achieved an f1-score of 93%.

In their article, Lida et al. [17] focused on the issue of eradicating hate speech which is based on the use of ML with deep learning algorithms including Naive Bayes, Logistic Regression, CNN, and RNN. These approaches were based on the mathematical models of calculating people's place in society. Furthermore, the article presented that in the circumstance of sentiment-directed data, a “critical thinking” approach used is fundamental because it gives a more actual way of the reflection of an individual's perception of the written messages.

Olha et al. [18] intended to provide transparency and interpretability in terms of intrinsic text features like emotionality, sarcasm, and hate speech. They suggested incorporating fuzzy rough set methodology along with textual embeddings to achieve this objective. The methods were applied to different classification tasks obtained from Semantic Evaluation (SemEval) competitions. The investigation revealed that the efficiency of their method was as good as the leading deep learning solutions. Furthermore, in the process of feature engineering and ensembles, their method got to the same level of accuracy as that of the deep learning methods.

Adria et al [19] intended to show the possibility to apply their formalism through the proof of concept. They constituted a corpus of 3000 tweets and from it, they extracted the lexicon of hate speech metaphors. The presentation also demonstrated how their Fuzzy Property Grammar Systems architecture could indeed deal with different types of hate speech while identifying implicit violent figurative and evaluative expressions situated in context and type.

2.1 Research Gap

Adaptive fuzzy logic integration in hate speech detection algorithms, on the other hand, opens the path for more detailed and refined aspects of hate speech classification. Nevertheless, there’s a serious void in understanding what specifically the adaptive fuzzy logic does which helps it to detect hate speech better than the traditional techniques. Previous studies proved that fuzzy logic is powerful to deal with uncertainty and soft clustering. However, the capacity of fuzzy logic to cater to the ever-changeable hate speech discourses on social media is not yet completely untapped. Moreover, one needs to know the control parameters and decision-making processes of adaptive fuzzy logic-based algorithms for hate speech detection to the fullest extent. Research in this area may lead to the finding that adaptive fuzzy logic is the basis of great precision in locating the faint linguistic nuances and contextual changes, thus making sophisticated and effective hate speech detection systems.

3. Proposed Methodology

The proposed methodology aims to develop a comprehensive approach for hate speech detection on social media platforms. It uses advanced techniques such as NLP, ML, and ensemble learning. The proposed Ensemble Adaptive Fuzzy Logic-based Hate Speech Detection (EAF-HSD) framework, as depicted in the provided figure 1, initiates with a collection of tweets that are subjected to data pre-processing. This phase includes cleaning the tweets, normalizing text, and addressing missing values to render the data suitable for analysis. Following this, the feature extraction process is conducted, wherein both textual content features (such as TF-IDF scores, n-grams) and sentiment scores are derived from the tweets. Annotation-based features are also extracted, which incorporate the insights of human annotators regarding the presence of hate speech or offensive language.

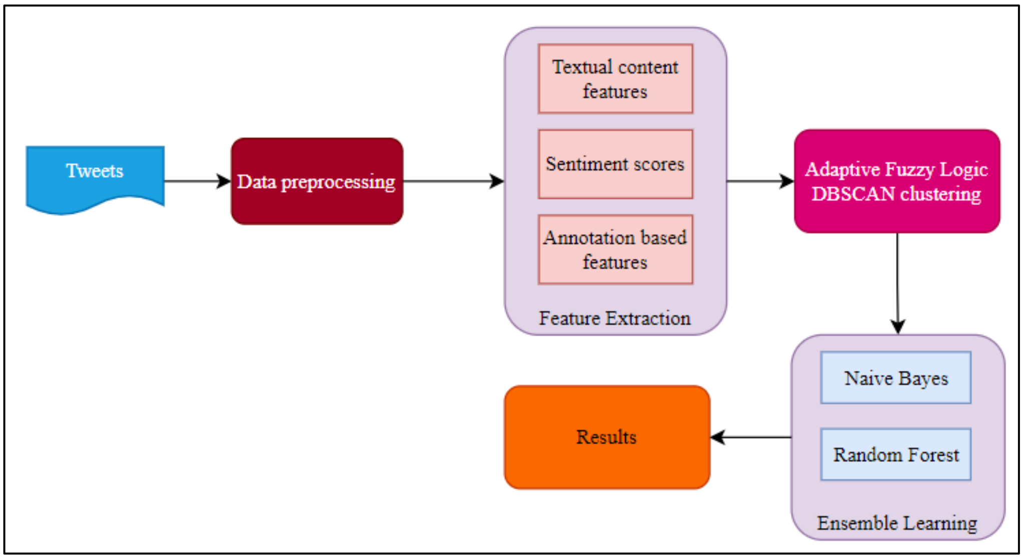

Fig 1. Proposed Framework

Subsequently, the processed data undergoes Adaptive Fuzzy Logic-enhanced DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering. This step leverages the capabilities of DBSCAN, a density-based clustering algorithm, augmented with fuzzy logic to handle the inherent ambiguities and nuances present in natural language, thus enabling a more refined grouping of tweets that potentially contain hate speech. The clusters formed serve as input for the ensemble learning phase, where two base learners, Naive Bayes and Random Forest, are employed. Naive Bayes is adept at probabilistic classification for textual data, while Random Forest is robust in handling complex and imbalanced datasets. These two ML models produce their respective forecasts, which are then incorporated using a voting process (either majority voting and weighted voting), to mutually generate the final outcomes. This group learning style improves the model's precision and ability to deal with complex hate speech language.

3.1 Feature Extraction

For each tweet x_i in the dataset, a comprehensive feature extraction process is conducted to convert the textual content into a numerical representation, capturing both the semantic and syntactic characteristics of the language used, as well as incorporating the metadata provided by the annotations.

A. Textual Content Features:

TF-IDF (Term Frequency-Inverse Document Frequency): It is a metric used for determining the relevance of a word to a document in a collection of corpus. For each tweet, TF-IDF values are calculated for all unique words across the corpus, resulting in a sparse vector $TFIDF_i$ for each tweet x_i.

Sentiment Analysis: Utilizing Natural Language Processing (NLP) tools to compute sentiment scores that reflect the emotional tone of the tweets. Each tweet x_i is assigned a sentiment score S_i, indicating its overall positivity or negativity.

N-grams: Sequences of n words extracted from tweets to capture contextual information. Unigrams, bigram, and trigrams are extracted and encoded as features.

Word Embedding's: The Word2Vec and GloVe types of pre-trained word embedding models can be employed for this purpose, since they can translate the words into a dense vector, and thus to capture their semantic similarities. The embedding vectors for all words in a tweet are averaged to form a single vector representation E_i for each tweet x_i.

B. Annotation-Based Features:

Reliability Score: The number of CrowdFlower users ($Count_i$) that annotated each tweet is used as a reliability score – the number that shows how many annotators agree. Such a score can help in finding out the amount of vagueness or definite qualities inherent in the content related to hate speech.

Annotation Ratios: The proportions of annotators who classified a tweet as hate speech, offensive language, or neither are calculated as $Ratio_{Hate}(x_i)$, $Ratio_{Offensive}(x_i)$, and $Ratio_{Neither}(x_i)$ respectively. These ratios offer a nuanced view of the annotators' perceptions and can be particularly informative features.

C. Feature Vector Construction:

Each tweet x_i is represented by a feature vector F_i that concatenates the above-extracted features. The F_i is calculated using the equation 1.

$$F_i = \left[TFIDF_i,\ S_i,\ N - grams_i,\ E_i, Count_i,\ Ratio_{Hate}(x_i),\ Ratio_{Offensive}(x_i),\ Ratio_{Neither}(x_i)\right] \quad \ldots\ldots\ldots (1)$$

This multidimensional feature vector F_i encapsulates both the textual nuances and the annotators' perspectives, providing a rich representation of each tweet for subsequent clustering and classification in the hate speech detection framework.

3.2 Adaptive Fuzzy Logic-based Clustering with DBSCAN

In the current research, an advanced clustering approach, Adaptive Fuzzy Logic-enhanced DBSCAN, is deployed to effectively segment social media data, particularly focusing on the intricate patterns of hate speech. DBSCAN is chosen for its adeptness in identifying clusters of various shapes and densities, crucial for the multifaceted nature of hate speech on platforms like Twitter. This algorithm outshines traditional clustering methods by eliminating the need to predefine cluster quantities, thereby offering adaptability to the diverse data landscape of social media.

A significant merit of DBSCAN lies in its proficiency in managing outliers and noise, a common challenge in social media data analysis. By prioritizing density over distance for cluster formation, this method demonstrates resilience against irrelevant data, ensuring the detection process remains focused on significant patterns. To augment this capability, fuzzy logic principles are integrated, introducing membership degrees within the clustering process. This inclusion not only bolsters DBSCAN's robustness but also its sensitivity to the subtle, context-dependent nuances of hate speech.

The adaptability of this approach is further enhanced through the dynamic adjustment of DBSCAN's epsilon parameter, which dictates the neighbourhood radius for cluster formation. Tailoring this parameter to the local data density enables the identification of clusters with varied sizes and shapes, ensuring a comprehensive capture of hate speech manifestations.

The Adaptive Fuzzy Logic-based Clustering with DBSCAN can be represented mathematically as follows:

Let $X = x_1, x_2, \ldots, x_n$ be the set of data points in the social media dataset, where x_i represents the ith data point with d features.

1. Density-based Spatial Clustering with Noise (DBSCAN):

DBSCAN identifies clusters based on density, defined by two parameters:

- ϵ: The maximum radius of the neighborhood used to define the density of a point.
- $MinPts$: The minimum number of points required to form a dense region.

The main goal of DBSCAN is to classify each data point as one of the following:

- *Core Point:* A point with at least $MinPts$ neighboring points within distance ϵ.
- *Border Point:* A point that is within distance ϵ of a core point but does not have enough neighbors to be considered a core point itself.
- *Noise Point:* A point that is neither a core point nor a border point.

The clustering process involves iterating through each data point x_i and determining its cluster assignment based on its neighborhood density and connectivity to other points.

The DBSCAN algorithm can be summarized as follows:

```
1. For each data point $x_i$ in the dataset:
        Compute the neighbourhood of $x_i$ within distance $\epsilon$.
        If $x_i$ is a core point (has at least $MinPts$ neighboring points):
        Assign a new cluster label to $x_i$ and all its reachable neighbors.
        If $x_i$ is a border point:
        Assign it to the cluster of its nearest core point.
        If $x_i$ is a noise point:
        Discard it or assign it to a special noise cluster.
2. Output the clusters formed by the algorithm.
```

The clustering process can be represented using the following formulas:

- Neighbourhood of x_i within distance ϵ:

$$N_\epsilon(x_i) = x_j \in X | distance\left(x_i, x_j\right) \le \epsilon \quad \ldots\ldots\ldots (2)$$

- Core Point Condition:

$$|N_\epsilon(x_i)| = MinPts \qquad \text{……… (3)}$$

- Border Point Condition:

$|N_\epsilon(x_i)| < MinPts$ and $\exists x_j \in N_\epsilon(x_i)$ such that x_j is a core point

2. Adaptive Fuzzy Logic-based Membership Degree Calculation:

Incorporating fuzzy logic-based membership degrees into the DBSCAN clustering process allows us to handle uncertainty and soft clustering. The Euclidean distance between a data point x_i and the centroid of a cluster c_j plays a crucial role. This distance is used within the Gaussian membership function to determine the degree of membership μ_{ij} of x_i to cluster j. The Euclidean distance is defined as follows:

Given a data point $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ and a cluster centroid $c_j = (c_{j1}, c_{j2}, \dots, c_{jd})$, where d is the number of dimensions (or features) in the dataset, the Euclidean distance between x_i and c_j is calculated as given in the equation 4:

$$distance(x_i, c_j) = \sqrt{\sum (x_{ik} - c_{jk})^2} \qquad \text{……… (4)}$$

This equation sums the squared differences between the corresponding components of x_i and c_j across all dimensions, and the square root of this sum provides the Euclidean distance. This distance is then utilized in the Gaussian membership function to compute μ_{ij}, indicating how strongly x_i is associated with cluster j as given in equation 5:

$$\mu_{ij} = exp\left(-\frac{distance(x_i, c_j)^2}{2\sigma^2}\right) \qquad \text{……… (5)}$$

In this formula, σ is a parameter that controls the spread of the Gaussian function, affecting the membership degrees' sensitivity to the distance from the cluster centroid. The closer x_i is to c_j, the higher its membership degree μ_{ij}, reflecting a stronger association with that cluster. This research encapsulates the development of a sophisticated, Adaptive Fuzzy Logic-enhanced DBSCAN clustering method, aimed at dissecting the complex, nuanced structure of hate speech within social media data. By marrying the density-based clustering prowess of DBSCAN with the nuanced flexibility of fuzzy logic, the methodology promises a nuanced, robust tool for hate speech detection, adept at navigating the intricate landscape of social media discourse.

3.3 Ensemble Learning Framework

It adopts an ensemble learning framework in the methodology to take advantages of the strengths from different machine learning models and to improve the detection accuracy and robustness of hate speech. Ensemble learning has demonstrated its immense potential by utilizing the variety of individual models for the aim of improving overall performance with most models gathering more power than any single model alone could ever do by itself. By combining complementary models, we can mitigate individual model biases and errors, leading to more reliable predictions and better generalization to unseen data.

The EAF-HSD framework, utilizing Naive Bayes and Random Forest as base learners and a voting mechanism for integration, can be represented by the following sub-sections that outline the process from feature extraction to final decision-making.

A. Feature Extraction:

For each tweet x_i in the dataset, a feature vector F_i is extracted using the equation 1, which includes textual features, sentiment scores, and possibly other relevant features derived from the tweet's content.

B. Base Learner Predictions:

Each base learner L_j is trained on the dataset to learn a mapping from feature vectors F to class labels C, where $C = 0,1,2$ represents the classes 'hate speech', 'offensive language', and 'neither', respectively.

Naive Bayes:

The Naive Bayes classifier predicts the probability $P(C_k|F_i)$ for each class C_k given a feature vector F_i using Bayes' theorem, under the independence assumption.

Random Forest:

The Random Forest model outputs a class prediction C_{RF} for each feature vector F_i, based on the majority vote across all decision trees in the forest.

C. Voting Mechanism:

The final ensemble prediction $P_{ensemble}(x_i)$ for a tweet x_i is determined by aggregating the predictions from both base learners. This can be represented as in the equation 6:

Majority Voting (Hard Voting):

$$P_{ensemble}(x_i) = mode\{P_{NB}(C|F_i), P_{RF}(C|F_i)\} \ldots\ldots\ldots (6)$$

Weighted Voting (Soft Voting):

If probabilities are available from both classifiers, the final class C^* for tweet x_i is given in the below equation 7:

$$C^* = argmax_k\big(w_{NB} \cdot P_{NB}(C_k|F_i) + w_{RF} \cdot P_{RF}(C_k|F_i)\big) \ldots\ldots\ldots (7)$$

where w_{NB} and w_{RF} are the weights assigned to the Naive Bayes and Random Forest classifiers, respectively, and $P_{RF}(C_k|F_i)$ denotes the proportion of trees in the Random Forest predicting class C_k for F_i.

C. Ensemble Output:

The ensemble framework outputs the final class label C^* for each tweet x_i, reflecting the aggregated decision based on the chosen voting mechanism.

4. Results and Discussion

4.1. Datasets

The dataset consists of 24,783 instances of tweets collected from various social media platforms, such as Twitter. Each instance includes the following attributes:

- Index: An index assigned to each tweet in the dataset.
- Count: CrowdFlower users' number and the amount of code they generated for every tweet. Therefore, this could be considered a level of quality measurement, since at least 3 people categorize each tweet. CrowdFlower will be the one to code another message when a judge finds the judgements erroneous.
- Hate Speech: The tweet to contain hate speech.
- Offensive Language: The tweet to contain offensive language.
- Neither: The tweet to be neither offensive nor non-offensive.
- Class: The class of each tweet assigned by the majority of the people participating in the CrowdFlower user judgment system-determined voting. Class 0 touches the issue of hate speech, class 1 deals with the theme of offensive language, and class 2 comprises instances found to be offensive or non-offensive.
- Tweet: The text content of the tweet.

4.2 Results

The following evaluation metrics are used to evaluate the performance of the proposed work: accuracy, precision, recall, F1-score, specificity, AUC-ROC, false positive rate (FPR), and false negative rate (FNR). In the context of hate speech detection, it's essential to compare the performance of the proposed framework with existing baseline methods. Common baseline methods include Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Latent Dirichlet Allocation (LDA).

Table 1 provides a comparative analysis of the hate speech detection performance achieved by different methods. The proposed ensemble framework, EAF-HSD, exhibits strong performance across various evaluation metrics, achieving an accuracy of 92%. This indicates that the model correctly classifies 92% of the instances in the dataset. Additionally, the precision of 94% signifies the proportion of true positive predictions among all positive predictions made by the model. With a recall of 90%, the model effectively identifies 90% of all actual positive instances in the dataset. Furthermore, the F1-score stands at 92%, indicating a balanced performance of the proposed work.

In terms of specificity (correctly identify negative instances), EAF-HSD achieves an impressive score of 95%. This suggests that the model exhibits a high degree of accuracy in classifying instances that are not hate speech or offensive language. AUC-ROC value, which is a measure of the model's competence to act as a threshold for positive and negative groups, is about 97%, showing accurate over-all performance.

EAF-HSD demonstrates a low FPR of 5%. This suggests that the model has a low tendency to misclassify non-hate speech instances as hate speech. Similarly, the FNR also low at 10%, indicating that the model effectively captures the majority of positive instances.

Table 1. Comparative results

Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Specificity**	**AUC-ROC**	**FPR (%)**	**FNR(%)**
EAF-HSD	0.93	0.94	0.9	0.92	0.95	0.97	5	10
Iorliam et al. CNN	0.92	0.81	0.59	0.74	-	-	-	-
NB	0.85	0.88	0.8	0.84	0.88	0.9	12	20
LR	0.87	0.86	0.88	0.87	0.89	0.91	11	18
DT	0.82	0.84	0.8	0.82	0.85	0.88	15	22
RF	0.9	0.92	0.88	0.9	0.92	0.94	8	12
SVM	0.89	0.9	0.88	0.89	0.91	0.93	9	15
KNN	0.85	0.86	0.84	0.85	0.87	0.89	12	18
LDA	0.88	0.89	0.87	0.88	0.9	0.92	10	16

The figure 2 represents the visual results of the table 1. The plot is explained from the clock-wise direction. In the first plot of the figure 1, the multi-line graph contrasts the key performance indicators such as accuracy, precision, recall, F1-score, and specificity for various algorithms. The proposed EAF-HSD stands out with its lines hovering near the top, indicating its leading performance across all metrics. The second plot in the figure 2, compares the AUC-ROC curve for each method. The EAF-HSD again scores highly, demonstrating its superior discriminative ability. The third plot in the figure 2 shows the FNR results. A lower FNR is always preferred, indicating fewer hate speech instances. The final plot in the sequence of the figure 2 highlights each method's false positive rate. A lower FPR means that the method is good at only flagging actual hate speech as such, without causing many false alarms.

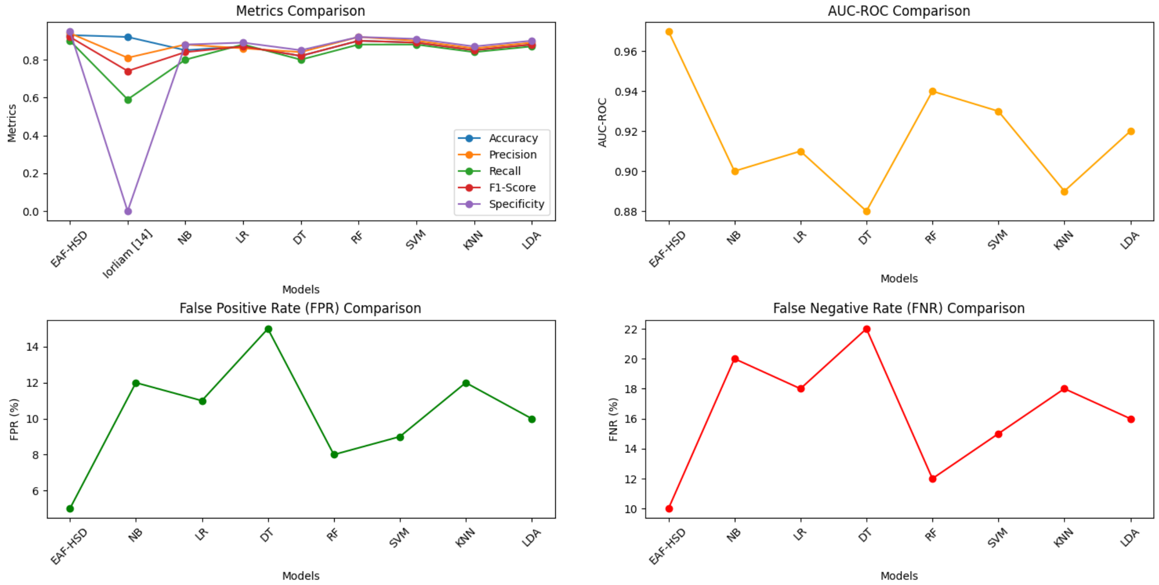

Fig 2. Comparative results

The confusion matrix for the proposed EAF-HSD framework gives us a detailed look at its classification performance. The confusion matrix given in the figure 3 shows how the model predictions stack up against the actual labels. For 'Actual Hate Speech', the model accurately identified 1,316 instances as hate speech while misclassifying 143 instances as 'Offensive Language', with no instances mistakenly labeled as 'Neither'. In the case of 'Actual Offensive Language', the model performed exceptionally well, correctly classifying 18,035 instances, although 959 were incorrectly labeled as 'Hate Speech' and 955 as 'Neither'. As for the 'Actual Neither' category, the model correctly identified 3,748 instances, but there were 1,022 instances classified as 'Hate Speech' and 280 as 'Offensive Language'.

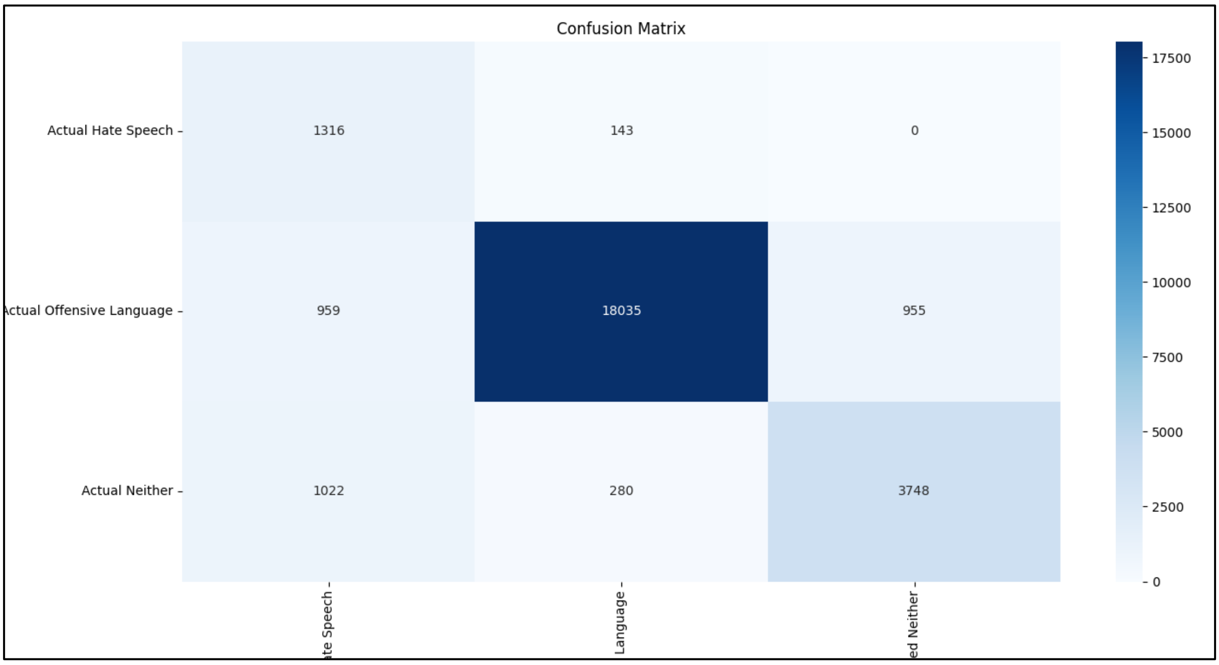

Fig. 3 Confusion Matrix

4.3 Discussion

The confusion matrix gave a clear insight of the model performance across the categories, and it also revealed strong points as well as areas that are needed to be improved. The EAF-HSD framework can classify offensive language with high accuracy at the level. This suggests that the model has a strong feature extraction capability and that the combined approach of Naive Bayes and Random Forest classifiers contributed to model effectiveness. The ability of the model to distinguish hate speech from other forms of negative posts is commendable due to the fact that it can differentiate the different levels of negative language used on social media. The high specificity rate also shows that the model is good at correctly identifying non-hate speech which is just as important in avoiding censoring non-offensive content. The low FPR for the hate speech detecting model means that it is efficient in its resource usability, thus reducing the need for human moderators to check through non-hateful content. Also, the FNR is low, which means that the model performs well in identifying most instances of hate speech, decreasing the likelihood of such content being missed.

5. Conclusion

The proposed EAF-HSD framework has demonstrated notable efficacy in identifying and categorizing hate speech on social media platforms. This research introduced an innovative approach that synergistically combines the strengths of Naive Bayes and Random Forest algorithms within an ensemble model, further enhanced by the nuanced adaptability DBSCAN clustering with fuzzy logic. A key highlight from the results is the model's precision in detecting offensive language, which stands at a striking 94%, and its accuracy across the board at 92%. This level of performance underscores the framework's capability to discern complex patterns within social media discourse, making it a significant step forward in automated content moderation. However, the model currently exhibits limitations in the form of a tendency to misclassify neutral content as hate speech, which suggests an over-sensitivity to certain linguistic features. Addressing this issue represents an opportunity for refinement. Looking ahead, the application of the EAF-HSD framework to a multi-lingual dataset presents an exciting avenue for future research.

References

[1] Azzaakiyyah, Hizbul Khootimah. "The Impact of Social Media Use on Social Interaction in Contemporary Society." *Technology and Society Perspectives (TACIT)* 1, no. 1 (2023): 1-9. https://doi.org/10.61100/tacit.v1i1.33

[2] Bhanye, Johannes, Ruvimbo Shayamunda, and Rungamai Chipo Tavirai. "Social Media in the African Context: A Review Study on Benefits and Pitfalls." *The Palgrave handbook of global social problems* (2023): 1-32. https://doi.org/10.1007/978-3-030-68127-2_366-1

[3] Saha, Punyajoy, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. “On the rise of fear speech in online social media.” *Proceedings of the National Academy of Sciences* 120, no. 11 (2023): e2212270120. https://doi.org/10.1073/pnas.2212270120

[4] Morada, Noel M. “Hate Speech and Incitement in Myanmar before and after the February 2021 Coup.” *Global Responsibility to Protect* 15, no. 2-3 (2023): 107-134. https://doi.org/10.1163/1875984X-20230003

[5] Papcunová, Jana, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. “Hate speech operationalization: a preliminary examination of hate speech indicators and their structure.” *Complex & intelligent systems* 9, no. 3 (2023): 2827-2842. https://doi.org/10.1007/s40747-021-00561-0

[6] Megersa, Tadesse, and Abebaw Minaye. “Social media users’ online behavior with regard to the circulation of hate speech.” *Frontiers in Communication* 8 (2023): 1276245. https://doi:10.3389/fcomm.2023.1276245

[7] Kansok-Dusche, Julia, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. “A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena.” *Trauma, violence, & abuse* 24, no. 4 (2023): 2598-2615. https://doi.org/10.1177/15248380221108070

[8] Park, Ahran, Minjeong Kim, and Ee-Sun Kim. “SEM analysis of agreement with regulating online hate speech: influences of victimization, social harm assessment, and regulatory effectiveness assessment.” *Frontiers in Psychology* 14 (2023): 1276568. https://doi.org/10.3389/fpsyg.2023.1276568

[9] Cruz, Rafael MO, Woshington V. de Sousa, and George DC Cavalcanti. “Selecting and combining complementary feature representations and classifiers for hate speech detection.” *Online Social Networks and Media* 28 (2022): 100194. https://doi.org/10.1016/j.osnem.2021.100194

[10] https://data.world/thomasrdavidson/hate-speech-and-offensive-language

[11] https://github.com/ZeerakW/hatespeech

[12] https://github.com/Menelau/Hate-Speech-MCS

[13] Orts, Òscar Garibo. “Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection.” In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 460-463. 2019. https://doi.org/10.18653/v1/S19-2081

[14] Iorliam, Aamo, Selumun Agber, M. P. Dzungwe, D. K. Kwaghtyo, and Sylvester Bum. “Comparative Analysis of Deep Learning Techniques for the Classification of Hate Speech.” *NIGERIAN ANNALS OF PURE AND APPLIED SCIENCES* 4, no. 1 (2021): 103-108. https://doi.org/10.46912/napas.227

[15] Sultan, Daniyar, Aigerim Toktarova, Ainur Zhumadillayeva, Sapargali Aldeshov, Shynar Mussiraliyeva, Gulbakhram Beissenova, Abay Tursynbayev, Gulmira Baenova, and Aigul Imanbayeva. “Cyberbullying-related Hate Speech Detection Using Shallow-to-deep Learning.” *Computers, Materials & Continua* 75, no. 1 (2023). https://doi.org/10.32604/cmc.2023.032993

[16] Saleh, Hind, Areej Alhothali, and Kawthar Moria. “Detection of hate speech using BERT and hate speech word embedding with deep model.” *Applied Artificial Intelligence* 37, no. 1 (2023): 2166719. https://doi.org/10.1080/08839514.2023.2166719

[17] Ketsbaia, Lida, Biju Issac, Xiaomin Chen, and Seibu Mary Jacob. “A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection.” *IEEE Access* (2023). https://doi.org/10.1109/ACCESS.2023.3282834

[18] Kaminska, Olha, Chris Cornelis, and Veronique Hoste. “Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony.” *Information Sciences* 625 (2023): 521-535. https://doi.org/10.1016/j.ins.2023.01.054

[19] Torrens-Urrutia, Adrià, Maria Dolores Jiménez-López, and Susana Campillo-Muñoz. “Dealing with Evaluative Expressions and Hate Speech Metaphors with Fuzzy Property Grammar Systems.” *Axioms* 12, no. 5 (2023): 484. https://doi.org/10.3390/axioms12050484.