International Journal of Communication Networks and Information Security
2024, 16(3), 6703
ISSN: 2073-607X,2076-0930
https://ijcnis.org/

Research Article

Context-Aware Hate Speech Detection: A Comparative Study of Machine Learning Models

Subrata Paul[1][0000-0001-5920-0654], Anirban Mitra[2][0000-0002-6639-4407], Shivnath Ghosh[3], Amitava Podder[4][0009-0004-9268-3781]

[1, 3,4]Department of CSE, School of Engineering, Brainware University, Ramkrishnapur Road, Barasat, Kolkata, West Bengal, 700125, India.
[2]Department of Computer Science and Engineering, ASET, Amity University, Kolkata,Kadampukur, New Town, 700135, West Bengal, India.
maitysubhadip77@gmail.com[1], mitra.anirban@gmail.com[2],shivghosh.cs@gmail.com[3], amitavapodder24@gmail.com[4]

ARTICLE INFO

Received: 3 Aug 2024
Accepted: 8 Sep 2024

ABSTRACT

Hate speech detection has grown into an essential obligation in today's social media-driven world, where adverse and discriminatory material spreads quickly. This paper conveys an in-depth investigation of recognising hateful speech employing machine learning approaches, with a focus on Logistic Regression and Random Forest models. Sentiment analysis is also investigated as an important factor in differentiating among hateful and non-hateful material, especially on social media, in which jokes and memes can occasionally lack malicious intent. The study contrasts the efficacy of the two models, with a focus on the balance of precision and recall in hate speech recognition. Although the Random Forest model surpassed Logistic Regression, both models struggled to precisely recognise hate speech, particularly in sophisticated and contextualised scenarios. Sentiment analysis is proposed as a future approach for enhancing the identification of potentially hazardous material. The results help to advance the research and development of more precise, context-aware detection of hateful speech systems.

Keywords: Hate Speech Detection, Machine Learning, Random Forest, Logistic Regression, Sentiment Analysis, Social Media, Context Awareness, Precision, Recall.

1. Introduction

Hate speech is a pervasive issue in online communities, social media, and other digital platforms. The anonymity provided by the internet has enabled the rapid dissemination of hate speech, contributing to an environment of toxicity and harm. The need for automated tools to detect and combat hate speech is crucial to create safer online spaces. Hate speech on online platforms has become a pressing societal concern, reflecting the intersection of technology, communication, and social dynamics. The term "online platforms" encompasses a wide range of digital spaces, including social media, forums, comment sections, and other interactive websites. Online platforms often provide users with a degree of anonymity, allowing them to express opinions without direct accountability. This anonymity can embolden individuals to engage in hate speech they might not express offline. Hate speech on online platforms can reach a global audience within seconds. This rapid dissemination contributes to the swift spread of harmful narratives, making it a formidable challenge for moderation efforts.

Algorithms on social media platforms are designed to maximize user engagement. As a result, content that elicits strong emotional responses, including hate speech, may be algorithmically amplified. This contributes to the creation of online echo chambers where like-minded individuals reinforce and intensify each other's beliefs. Hate speech takes various forms, including explicit language, derogatory comments, memes, images, and even subtle forms of discrimination. Its diverse nature makes it challenging to develop a one-size-fits-all solution for detection and moderation.

There exist several wide ranges of impacts of Online Hate Speech. Hate speech on online platforms has been linked to real-world consequences, including hate crimes, discrimination, and violence. The online environment can serve as a breeding ground for extremist ideologies that may manifest in offline actions. Besides, individuals targeted by hate speech may experience psychological and emotional harm. The constant exposure to discriminatory content can lead to stress, anxiety, and a sense of exclusion, affecting users' mental well-being. Hate speech can contribute to the erosion of social cohesion by fostering divisions and animosities among different groups. It may hinder constructive dialogue and impede efforts to build understanding and tolerance.Persistent exposure to hate speech can contribute to the normalization of discriminatory attitudes and behaviors. It may desensitize individuals to prejudiced language, making it more socially acceptable.Victims of hate speech may feel silenced or intimidated, leading to self-censorship and reluctance to express opinions or participate in online discussions. This can undermine the principles of free speech and open discourse. Online hate speech can spill over into the offline world, potentially leading to real-world harassment, discrimination, or violence. There have been instances where online threats escalated to physical harm.Members of marginalized communities are often disproportionately affected by hate speech. It can perpetuate systemic inequalities and contribute to the marginalization of already vulnerable groups.Hate speech can erode trust within online communities and platforms, making users hesitant to engage in open and respectful dialogue. This erosion of trust undermines the potential for constructive conversations.Online hate speech poses challenges for legal systems and regulators. Determining the boundaries between freedom of expression and harmful speech requires careful consideration, and jurisdictions may vary in their approaches.Hate speech can deepen existing social divides and contribute to the polarization of communities. It fosters an "us vs. them" mentality, creating tensions between different groups and hindering social cohesion.

The objective of this paper is to provide a review of the existing approaches for hate speech detection and additionally propose a model that would overcome the research gaps of the existing approaches. The problem of hate speech detection using machine learning (ML) typically revolves around mitigating the harmful impact of hate speech in online environments. This paper discusses and contrasts two machine learning models (Logistic Regression and Random Forest) for detecting offensive language in social media content. It also examines the effectiveness of sentiment analysis in differentiating non-hateful from possibly damaging speech. Furthermore, model performance was assessed using accuracy, precision, recall, and F1-score, revealing difficulties with the categorization of nuanced data. Future enhancements to hate speech detection include sentiment analysis and contextual awareness in machine learning models.

The paper is divided into several sections. Section 2 provides the existing literature related to the study and provides a comparative study of the approaches alongside the study of the research gaps. Section 3 illustrates the essential considerations in the design of the proposed model. Further the proposed model has been elaborately discussed in section 4. Section 5 discussed the simulation results while implementing the proposed model. The paper finally ends with the discussion on ethical considerations in section 6 and a general conclusion in section 7 respectively.

2. Literature Review

The proliferation of inappropriate and harmful content on social media has made hate speech detection an essential field of research.The systematic review follows a structured methodology, encompassing an extensive literature search spanning recent years. Various academic databases, conference proceedings, and reputable journals have been scrutinized to compile a comprehensive dataset of studies related to hate speech detection in social networks. The inclusion criteria prioritize relevance, novelty, and the utilization of diverse methodologies.Several studies have explored hate speech detection using various techniques, including Support Vector Machines (SVMs) and neural networks. R. Saravanan et al [1] conducted a survey focusing on SVM-based approaches, highlighting their effectiveness in natural language processing tasks but also discussing challenges like imbalanced datasets. Shanita Biere et al [2] utilized Convolutional Neural Networks (CNNs) to analyze tweets, achieving good performance but noting misclassifications of non-hate speech. They suggested that larger datasets could enhance CNN performance.Deep learning techniques have also been extensively studied for hate speech detection. M. U. Akram et al [3] provided a comprehensive overview of neural network architectures, addressing challenges such as data sparsity. Reviews by [4] and [5] emphasized the superiority of deep learning over traditional methods, showcasing various models and strategies for enhancement, including pre-training and transfer learning.

Moreover, some studies, like [6], investigated hate speech detection in social media using SVMs and Random Forests, achieving high accuracies. Additionally, a review article by [7] discussed challenges in multilingual settings and proposed techniques like cross-lingual transfer learning. Overall, these

studies contribute to the advancement of hate speech detection, highlighting both the progress made and the avenues for future research.Studies on hate speech detection have benefited greatly from large annotated corpora and powerful classification approaches. Florio et al. [8] investigated the temporal robustness of a BERT model for Italian, focusing on hate speech against immigrants in the Italian Twittersphere. They found that the performance of the model varied with the temporal distance of the fine-tuning set, suggesting the importance of considering temporal factors in hate speech detection. Mossie& Wang [9] proposed a hate speech detection approach using Spark distributed processing and deep learning algorithms like GRU. Their experiments on Amharic texts showcased the efficacy of word embedding techniques and deep learning for hate speech detection, emphasizing the significance of identifying vulnerable minority groups.

Kovács et al. [10] introduced a deep NLP model combining convolutional and recurrent layers for hate speech detection in social media data. Their model achieved promising results on the HASOC2019 corpus, highlighting the potential of deep learning in addressing hate speech detection challenges. Mullah &Zainon [11] reviewed ML algorithms and techniques for hate speech detection, offering insights into the critical steps involved and identifying research gaps and challenges. They provided a comprehensive overview of classical ML, ensemble approaches, and deep learning methods, aiding researchers and professionals in algorithm selection.Toktarova et al. [12] conducted a comprehensive comparison of traditional ML and deep learning techniques for hate speech detection on Twitter. Their findings indicated the superiority of deep learning, particularly BiLSTM, in accurately identifying hate speech. They also explored the integration of word embeddings to enhance model performance, contributing to a nuanced understanding of hate speech detection methods. Mossie& Wang [13] addressed the challenge of hate speech identification in Amharic Facebook posts using Apache Spark, Random Forest, and Naïve Bayes. Their model achieved promising accuracy, leveraging Spark's capabilities for big data processing.Alrehili [14] provided a survey of NLP techniques used in automatic hate speech detection on online social networks (OSNs), highlighting approaches such as dictionaries, bag-of-words, and N-grams. This survey underscores the importance of employing sophisticated NLP techniques to combat the spread of hateful and offensive speech on OSNs, emphasizing the role of technology in addressing societal challenges posed by hate speech.

2.1. Comparative Study

The provided literature in the previous section presents a diverse range of approaches for hate speech detection in social networks, employing various techniques from traditional machine learning to deep learning methods.

Paper Name	Approach Used	Advantages	Disadvantages
A Survey on Hate Speech Detection using Support Vector Machines	Support Vector Machines (SVMs)	Feature engineering, feature selection, and kernel selection.	imbalanced datasets and model interpretability
Hate Speech Detection Using Natural Language Processing Techniques	Natural Language Processing, Convolutional Neural Network (CNN)	Analyze tweets annotated with three labels: hate, offensive language and neither. potential of CNNs is acknowledged, particularly with larger and higher-quality datasets	Misclassifying non-hate speech**.**
A Comprehensive Survey on Hate Speech Detection using Neural Networks	Neural networks	This model will provide various natural language processing tasks, including text classification and sentiment analysis	Data sparsity model interpretability
Hate Speech Detection using Convolutional Neural Network: A Literature Review.	Convolutional Neural Networks (CNNs)	high accuracy in identifying hate speech in multiple languages, including English, Spanish, and Arabic	needs to various modifications to CNN architectures, such as using pre-trained word embedding and attention mechanisms, to improve the performance
Deep Learning Techniques for	Deep learning,	Efficacy in capturing complex linguistic patterns	narrow scope biased selection of studies

Hate Speech Detection: A Review.	Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs)	providing valuable insights for researchers and practitioners in the field.	insufficient critical analysis, affecting the overall completeness and applicability of the model.
Hate Speech Detection in Social Media using Support Vector Machines and Random Forests	Support Vector Machines (SVM) and Random Forests (RF) algorithms	robust approach with high interpretability and accuracy.	limited scalability to large datasets challenges in capturing nuanced linguistic contexts compared to more complex neural network approaches.
Hate Speech Detection in Multilingual Settings: A Review of Techniques and Challenges	Cross-lingual transfer learning, multilingual embedding	comprehensive analysis of the current state-of-the-art techniques, providing valuable insights for researchers and practitioners in the field.	Potential bias in the selection of reviewed techniques insufficient coverage of emerging methodologies challenges in providing concrete solutions for the complex issue of hate speech detection in diverse linguistic environments.
Time of your hate: The challenge of time in hate speech detection on social media	Diachronic Perspective with BERT Model	sensitivity to temporal distance	improved performance with an adequate time window
Vulnerable community identification using hate speech detection on social media	Spark-based Model for Amharic Hate Speech Detection	utilize deep learning algorithms, including GRU, and employ Word2Vec for feature extraction. identifies a highly vulnerable ethnic group, emphasizing cultural considerations.	
Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources	Deep Natural Language Processing Model	combining convolutional and recurrent layers methods for resource expansion leveraging unlabeled data and similarly labeled corpora to mitigate overfitting	
Advances in machine learning algorithms for hate speech detection in social media: a review	Machine Learning Algorithms	analyze baseline components, improvements over time, new datasets, and performance metrics	
Hate speech detection in social networks using machine learning and deep learning methods	Traditional and Deep Learning Techniques on Twitter	evaluates LSTM, BiLSTM, and CNN models, with BiLSTM emerging as the most accurate.	

Social network hate speech detection for Amharic language	Random Forest and Naïve Bayes	Apache Spark for hate speech detection in Amharic Facebook posts. employ, with Word2Vec and TF-IDF for feature selection. achieves promising results	
Automatic hate speech detection on social media: A brief survey.	NLP techniques	Various approaches like dictionaries, bag-of-words, and N-grams are discussed, outlining their advantages and disadvantages.	

To sum up, the literature covers a wide array of approaches, each with its strengths and challenges. SVMs, deep learning, ensemble methods, and linguistic considerations in different languages contribute to the evolving landscape of hate speech detection in social networks. Consideration of cultural context, effective feature selection, and advancements in neural network architectures emerge as critical factors for improving model performance.

2.2 Research Gaps in Existing Machine Learning Algorithms

Based on the literature survey presented in the previous section, there are several notable research gaps and areas for further exploration in hate speech detection:

- Interpretability and Explainability: While SVMs and traditional machine learning algorithms have been effective in hate speech detection, there's a lack of focus on model interpretability. Research should delve deeper into understanding how these models make decisions, especially when dealing with sensitive issues like hate speech, to ensure transparency and accountability.
- Cross-lingual and Multilingual Hate Speech Detection: Despite efforts to address hate speech detection in multilingual settings, there's still a need for robust approaches that can effectively identify hate speech across different languages and cultures. Current methods often face challenges related to language identification, code-switching, and cultural nuances, indicating a need for further research in this area.
- Contextual Understanding: Hate speech detection models, especially those based on deep learning techniques like CNNs and LSTMs, struggle with understanding context and distinguishing between hate speech and non-hate speech accurately. There's a need to develop models that can better capture semantic nuances, sarcasm, and cultural references to improve detection accuracy.
- Mitigating Bias and Fairness: Hate speech detection models can inherit biases present in the training data, leading to unfair outcomes, especially for marginalized communities. Future research should focus on developing methods to mitigate bias and ensure fairness in hate speech detection models, including data augmentation techniques, adversarial training, and fairness-aware algorithms.
- Real-time Detection and Response: Hate speech spreads rapidly on social media platforms, making real-time detection and response crucial. There's a need for scalable and efficient hate speech detection systems that can operate in real-time, allowing platforms to take immediate action to curb the spread of hate speech and protect users.
- Ethical Considerations: Hate speech detection research must consider ethical implications, including privacy concerns, freedom of speech, and potential unintended consequences of automated moderation. Future studies should explore ethical frameworks and guidelines for developing and deploying hate speech detection systems responsibly.
- Evaluation Metrics and Benchmark Datasets: Standardized evaluation metrics and benchmark datasets are essential for comparing the performance of hate speech detection models accurately. There's a need for consensus on evaluation metrics and the development of benchmark datasets that cover diverse languages, cultures, and social media platforms.

Addressing these research gaps will not only advance the field of hate speech detection but also contribute to creating safer and more inclusive online environments. Additionally, interdisciplinary collaboration between researchers, policymakers, and industry stakeholders is crucial for developing effective and ethical hate speech detection solutions [1-15].

3. Essential Considerations in Design

Natural Language Processing (NLP) constitutes a pivotal field in modern computing, enabling machines to comprehend, interpret, and generate human language. Tokenization, the initial step in

NLP, involves breaking down text into manageable units called tokens, forming the basis for subsequent analysis. Text preprocessing further refines raw data by eliminating noise, handling special characters, and standardizing text for enhanced analysis, thus optimizing the quality of extracted information. Part-of-Speech Tagging assigns grammatical labels to words, facilitating syntactic and semantic analysis by elucidating the grammatical structure of sentences. Named Entity Recognition (NER) identifies and categorizes entities within text, aiding in the extraction of structured information crucial for information retrieval tasks. Semantic Analysis delves into the meaning and intent behind words and phrases, supporting applications such as sentiment analysis and relationship extraction. Syntactic Analysis, or parsing, involves dissecting sentence structure to discern relationships between words, enabling the creation of parse trees for syntactic representation. Machine Translation enables the translation of text between languages, benefiting applications such as online language services and cross-language information retrieval. Text Generation employs language models to produce human-like text, facilitating the creation of chatbots, content generation, and automatic summarization. Sentiment Analysis involves determining the emotional tone of text, useful for analyzing customer reviews, social media sentiments, and public opinion. When considering NLP over Support Vector Machines (SVM) for hate speech detection, several factors come into play. NLP's suitability for large and complex datasets, leveraging deep learning architectures to capture nuanced hate speech characteristics, contrasts with SVM's effectiveness in scenarios with smaller datasets or high-dimensional feature spaces. While NLP may offer less interpretability due to the complexity of deep learning models, SVMs provide interpretable decision boundaries crucial for understanding hate speech features. Additionally, NLP's computational intensity and resource requirements may pose challenges, contrasting with SVM's generally lower resource demands, making it more viable for scenarios with limited computational resources. Moreover, NLP's capability for multimodal integration makes it well-suited for tasks involving the analysis of diverse data types alongside text, while SVMs may necessitate additional techniques for such integration. Ultimately, the choice between NLP and SVM depends on factors such as data characteristics, interpretability needs, resource constraints, and multimodal requirements, shaping the selection process for hate speech detection applications [16-20].

4. Methodology

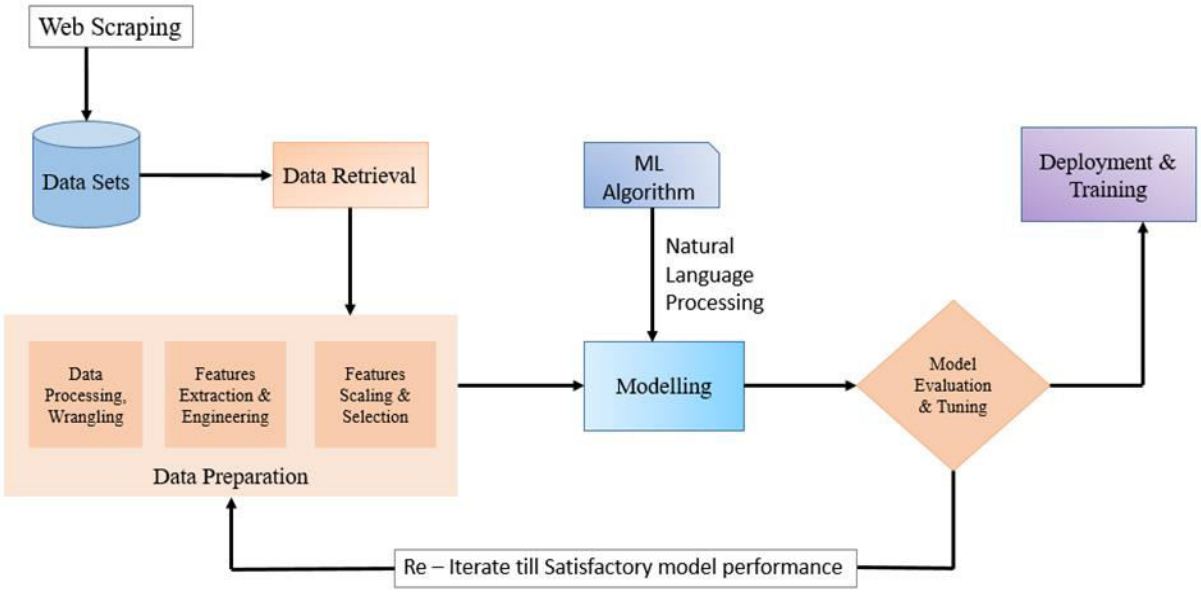


Fig 1. Proposed Methodology in Hate Speech detection in social media

The figure 1 presented above presents a proposed methodology for hate speech detection in social media. Preparing a machine learning (ML) model involves a systematic process to ensure its effectiveness and reliability in addressing the defined problem. The first step, Collect and Prepare Data, entails gathering relevant data sources and performing data cleaning and preprocessing tasks to handle missing values, outliers, and inconsistencies. Subsequently, the dataset is split into training, validation, and test sets to facilitate model training, tuning, and evaluation. Following data collection, the Explore and Analyze Data phase involves conducting exploratory data analysis (EDA) to understand dataset characteristics, visualize distributions, correlations, and patterns, and identify potential features and relationships impacting the model. Feature Engineering then focuses on selecting relevant features and creating new ones to enhance model performance. Data Encoding

converts categorical variables into numerical representations using techniques like one-hot encoding or label encoding. The dataset is then divided into input features (X) and the target variable (y) in the Split Data step, with subsequent division into training, validation, and test sets. Choosing a suitable model architecture is crucial in the Choose a model phase, considering factors such as complexity, interpretability, and scalability. Model Training involves training the model on the training dataset, adjusting hyperparameters, and utilizing techniques like cross-validation for generalization assessment. Model Evaluation assesses performance using the validation dataset, employing evaluation metrics tailored to the problem nature. Model Tuning fine-tunes hyperparameters and considers regularization and optimization techniques to enhance generalization. Validation and Testing validate the model on the validation set and evaluate its performance on the test set. Interpretation of Results involves understanding the model's effectiveness and insights gained from feature importance. If the model meets desired criteria, it is Deployed for real-world use, integrating it into production environments. Finally, the Model is Monitored and Maintained, with mechanisms established to track performance over time and update the model as needed based on new data or domain changes. This systematic approach ensures the development of robust and reliable ML models for various applications.

5. Result and Discussion

5.1 Description of the dataset

The hate speech detection dataset consists of labelled comments from multiple sources that includes Social Network (like Facebook, Twitter (X), LinkedIn) and Online News Portals like (Times of India, India Today etc) and have been classified into two categories: hateful and non-hateful comments. It contains a wide range of data types, including examples of hate speech and normal speech, as well as some inappropriate vocabulary. To guarantee adequate instruction, the dataset includes a diverse range of sentiments and expressions commonly encountered in social media interactions. Before processing, the dataset was cleaned to remove irrelevant fields, special characters, and irrelevant spaces, leaving only the 'text' and 'label' fields for further analysis. This diverse and large dataset provides a solid foundation for training machine learning models to differentiate between hate speech and benign content.

5.2 Evaluation Parameters

In machine learning, different evaluation parameters are utilised for evaluating the effectiveness of models, particularly for tasks involving classification such as hate speech detection. These parameters are derived from the confusion matrix, that comprises True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Here are the key evaluation parameters:

Accuracy: It symbolises the percentage of instances correctly categorised (both hate and non-hate speech) out of the overall number of instances. It is calculated by the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is an overall indicator of the the model's altogether correctness, but it might not be reliable in imbalanced data sets where a single group dominates.

Precision: Precision is defined as the proportion of positive observations that were correctly predicted (true positives) to all anticipated positive observations. It demonstrates how accurate the model is at forecasting hate speech. It is given as:

$$Precision = \frac{TP}{TP + FP}$$

Precision is especially useful while the cost of false positives is high, making hate speech projections as precise as possible.

Recall (Sensitivity or True Positive Rate): Recall is defined as the proportion of positively predicted observations (true positives) to total positive observations. It demonstrates how accurately the model detects actual hate speech. Recall is represented as:

$$Recall = \frac{TP}{TP + FN}$$

High recall is critical because missing real-life instances of hate speech (false negatives) can have serious consequences.

F1-Score: The F1-Score is the harmonic mean of precision and recall, which balances them. It is beneficial while you need to strike a balance between precision and recall, particularly when the classes are imbalanced. The F1-Score is given by the formula:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1-Score is a single measure that balances the trade-off between false positives and false negatives, which makes it perfect for unbalanced datasets.

Specificity (True Negative Rate): Specificity is the ratio of true negatives (non-hate speech correctly classified) to total actual negatives. It is complementary to recall and represented by the formula:

$$Specificity = \frac{TN}{TN + FP}$$

Specificity is critical for correctly identifying non-hate speech, especially when false positive predictions must be minimised.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): It assesses the model's ability to differentiate among classes (hate speech and non-hate speech) at various threshold values. It is given by:

$$AUC - ROC = \int_0^1 TPR(x) dFPR(x)$$

AUC-ROC is a thorough indicator of how well a model distinguishes between classes, with higher values suggesting better classification performance.Each of these metrics offers a unique perspective on the hate speech detection model's performance. Various indicators can be prioritised based on specific goals (for example, reducing false positives or false negatives).

5.3 Simulation Results

In this study, a basic machine learning model for hate speech detection was implemented using Logistic Regression. The process started with the import of the required libraries and machine learning models. A labelled dataset with examples of both hate speech and normal speech was then loaded. The text data was preprocessed to remove noise, tokenise it, and convert it into a format appropriate for model input. The dataset was divided into training and testing subsets to objectively assess the model's performance. The Logistic Regression model was trained on the processed data via the training subset. Following training, the model's performance was evaluated by comparing its accuracy to test data. Key performance metrics such as precision, recall, and F1-score were also computed to provide a thorough evaluation of the model's ability to classify hate speech. These metrics provided valuable insights into areas where the model's performance could be improved and refined.

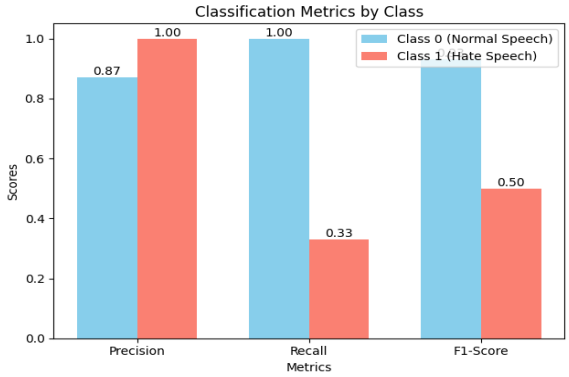
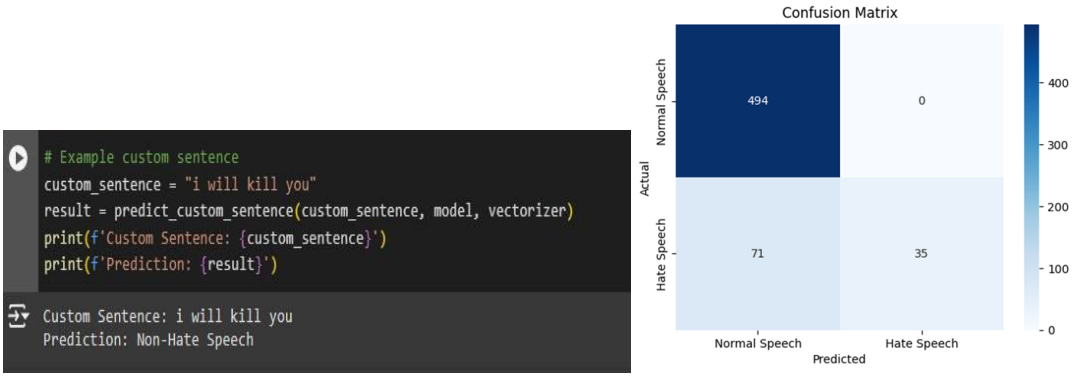


Fig 2. Performance Metrics Comparison for Hate Speech Detection Model using Logistic Regression

The accuracy of the Logistic Regression-based hate speech detection model is 88.17%. The classification report reveals key insights: the model had a precision of 0.87 for class '0' (normal speech) and 1.00 for class '1' (hate speech), indicating that it was extremely accurate in detecting true hate speech cases. The recall, a measure of the model's ability to find all relevant situations, was 1.00 for class '0' but significantly lower (0.33 for class '1'), indicating a difficulty in correctly detecting hate speech. The F1-score, which combines precision and recall, was 0.93 for class '0' and 0.50 for class '1', indicating a trade-off in which normal speech was accurately classified but hate speech detection required enhancements.The macro averages for precision, recall, and F1-score were 0.94, 0.67, and 0.71, respectively, whereas the weighted averages, accounting for class distribution, were 0.90, 0.88, and 0.86. This indicates that the model is extremely accurate but requires further improvement in recording true positive hate speech cases, as demonstrated by the low recall for class '1'.




```
# Example custom sentence
custom_sentence = "i will kill you"
result = predict_custom_sentence(custom_sentence, model, vectorizer)
print(f'Custom Sentence: {custom_sentence}')
print(f'Prediction: {result}')

Custom Sentence: i will kill you
Prediction: Non-Hate Speech
```

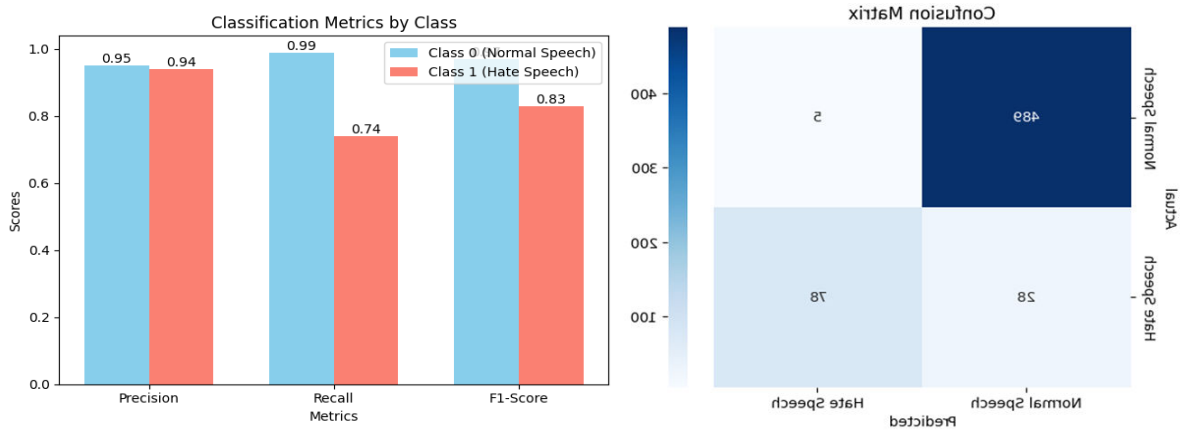
(a) (b)

Figure 3: Model Performance Evaluation and Generalization Issues (a) Custom Sentence Evaluation and (b) Confusion Matrix.

Figure 3 provides useful insights into the effectiveness of the hate speech detection model. Figure 3 (a) shows an instance of the custom sentence, "I will kill you," which was tested using the trained model. Unexpectedly, the model classified the phrase as "Non-Hate Speech," despite the obvious negative and offensive nature of the content. This demonstrates a significant limitation in the model's ability to generalise well to new or unknown data. Figure 3(b) depicts a confusion matrix, which further demonstrates the model's performance. On the test dataset, the model performed admirably, correctly identifying 494 occasions of normal speech and 35 instances of hate speech. However, 71 instances of hate speech were incorrectly classified as normal speech, indicating that the model is extremely successful at detecting normal speech but struggles with correctly recognising hate speech. This classification imbalance indicates overfitting, in which the model executes well on accustomed or training data but fails to generalise effectively to new or unknown inputs. Although the model performs well on the test set, as evidenced by the confusion matrix, its ability to deal with unknown data, as demonstrated by the misclassification of the custom sentence, requires improvement. This suggests that further refinement of the training data, as well as enhancements to the pre-processing and feature extraction techniques, may improve the model's generalisability and hate speech detection accuracy.

The datasets, which included three labelled datasets with both hate and non-hate comments, had been loaded. This diversity in datasets stipulated an extensive variety of data types and a large volume of data, which enabled more effective model training with various sentiments and hateful statements. Additional attributes have been eliminated, leaving only two fields, text and label. The first dataset included three types of data: hate speech, offensive speech, and normal speech. To simplify the analysis, the comments were divided into two categories: hateful and normal. A procedure was written to clean up the dataset by removing numbers, special characters, extra spaces, and other unnecessary elements. This function was subsequent applied to all datasets to get them ready for further processing. Next, the text data was vectorised, converting all text to numerical formats because the model procedures numbers more efficiently than words. The data was divided into training and testing sets, with 80% used for training and 20% for testing, to ensure an accurate assessment of the model's efficiency. The Random Forest model was chosen and trained using the training data. Following training, the model's performance was evaluated on test data, and it achieved an accuracy of around 95%, representing an enhancement over the previous model. Furthermore, precision and

recall values have been established to provide a more detailed assessment of the model's reliability and overall effectiveness.



(a) (b)

Fig 4. (a) Performance Metrics Comparison and (b) Confusion Matrix for Hate Speech Detection Model using Random Forest.

Figure 4 shows the performance metrics and confusion matrix for the Hate Speech Detection Model, which was created using a Random Forest classifier. Figure 4(a) depicts a bar chart contrasting the precision, recall, and F1-scores of normal speech (Class 0) and hate speech (Class 1). The precision values for both categories are high, with Class 0 (normal speech) at 0.95 and Class 1 (hate speech) at 0.94. However, the recall score for Class 1 is significantly lower (0.74) than Class 0, suggesting that the model is unable to detect hate speech as effectively as normal speech. This disparity is also reflected in the F1-scores, which are 0.83, slightly lower than Class 0's 0.97, highlighting the difficulties in balancing precision and recall when detecting hate speech. Figure 4(b) provides additional information about the model's classification performance. Out of 494 instances of normal speech, the model correctly classified 489, misclassifying only five as hate speech. However, of the 106 hate speech instances, the model correctly recognised 78 while misclassifying 28 as normal speech. This supports the metrics' observation that the model does a better job understanding normal speech but challenges to capture every instance of hate speech. The figures illustrate that while the Random Forest model performs well in terms of precision, there is room for improvement in enhancing the recall of hate speech detection.

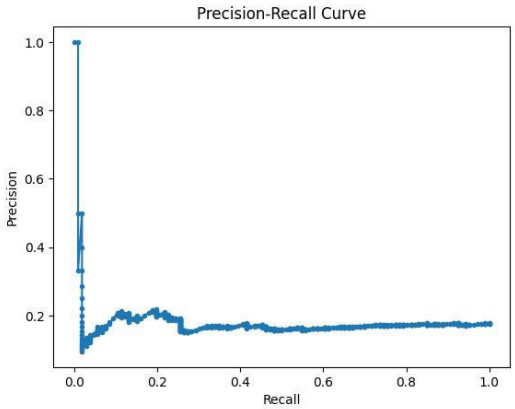


Fig 5. Precision Recall Curve for Random Forest

Figure 5 for the Random Forest model shows the relationship between precision and recall at different thresholds. Precision is defined as the proportion of true positive predictions out of all favourable predictions, whereas recall is the proportion of true positives discovered out of all actual positives. In this graph, the curve begins with a very high precision for low recall values, suggesting that the model is extremely precise when making predictions. However, as recall increases, precision decreases dramatically and eventually stabilise at a lower level. This significant decrease indicates that the Random Forest model is struggling with preserving high precision as it attempts to capture more true

positives (increase recall). The model's performance might be less effective in scenarios where both high precision and recall are needed simultaneously.

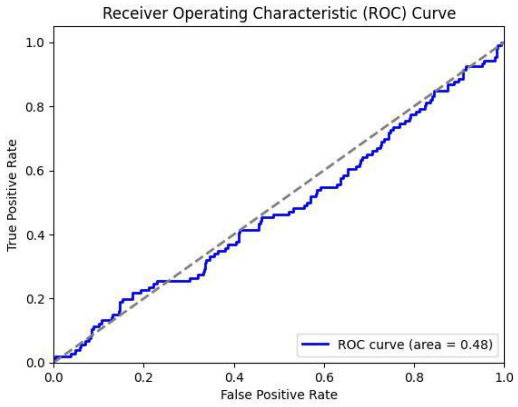


Fig 6. ROC Curve for Random Forest

Figure 6 shows the ROC Curve with an AUC (Area Under the Curve) of 0.48, which is close to 0.5, indicating that the model's ability to distinguish between classes is nearly equivalent to random guessing. The ROC (Receiver Operating Characteristic) curve typically compares the True Positive Rate (TPR) to the False Positive Rate (FPR) at different thresholds. An ideal model would have an AUC close to one, indicating perfect distinction between classes. However, an AUC of 0.48 indicates that the Random Forest model underperforms in this case, and its predictions are unreliable for differentiating hate speech from normal speech.

When contrasting the results of the Logistic Regression and Random Forest models for hate speech detection, the Random Forest classifier performs significantly better. The Logistic Regression model, while effective in basic cases of hate and non-hate speech, had limitations, particularly in its ability to generalize to previously unseen data. It struggled to accurately identify situations of hate speech, as evidenced by lower recall and F1-scores, indicating a trade-off between precision and recall. In contrast, the Random Forest model performed better in terms of accuracy, precision, and overall efficacy when differentiating between hate and normal speech. Despite its superior performance, the Random Forest model faced some challenges in detecting hate speech, as evidenced by the confusion matrix and lower recall for hate speech instances. While both models have advantages, Random Forest provided a more robust solution, though additional optimization and balancing of precision and recall are still required for detecting hate speech with greater accuracy in real-world applications.

The study discovered that sentiment analysis plays an important role in hate speech detection because understanding the context is required before drawing conclusions. Jokes, memes, and humorous comments are common on social media today, and not all of them are meant to be offensive. As a result, a thorough investigation of the sentiment underlying such statements is required. For example, the sentences "Hahaha, dude, you are so dumb that you did this kind of mistake" and "Dude, you are dumb enough to deserve this kind of mistake" may use similar words but convey different meanings. The first sentence could be interpreted as a light-hearted, humorous remark, whereas the second might be regarded as offensive or hateful. This distinction highlights the importance of sentiment analysis in accurately identifying the intention behind seemingly similar statements, thereby preventing misclassification of non-hateful comments as hate speech.

6. Ethical Considerations

Detecting hate speech through automated models necessitates a vigilant approach towards ethical considerations, especially concerning biases inherent in the models. The concern arises from the potential propagation of biases present in training data, which can result in discriminatory outcomes in hate speech detection. To mitigate this, it's imperative to conduct regular audits of training datasets, ensuring representation from diverse demographics. Additionally, diversifying training data sources and implementing debiasing techniques can help minimize the impact of biases. Continuously assessing model performance on diverse datasets allows for the detection and correction of discriminatory patterns, promoting fairness and inclusivity in hate speech detection systems. Furthermore, fostering transparency in the model development process and involving stakeholders

from diverse backgrounds can aid in identifying and addressing biases effectively. By prioritizing fairness and inclusivity, hate speech detection tools can uphold ethical standards and mitigate the risk of perpetuating societal biases.

Another significant ethical concern in hate speech detection revolves around balancing the imperative to prevent harm with the fundamental right to freedom of speech. Hate speech detection tools must navigate this delicate balance, as overly aggressive moderation measures may inadvertently suppress legitimate discourse and curtail free expression. To address this concern, it's essential to establish clear and transparent criteria for hate speech detection, ensuring consistency and accountability in moderation decisions. Providing users with avenues to appeal decisions and incorporating human moderators to handle nuanced cases can help mitigate the risk of over-moderation. Moreover, implementing transparent and accountable content moderation policies, coupled with robust user engagement mechanisms, fosters user trust and confidence in the hate speech detection process. By striking a balance between preventing harm and preserving free expression, hate speech detection tools can uphold fundamental rights while effectively combatting online hate speech. Additionally, fostering open dialogue and collaboration with stakeholders, including civil society organizations and human rights advocates, can provide valuable insights into the ethical implications of hate speech detection and inform the development of responsible moderation practices. Ultimately, by adopting a proactive and inclusive approach to addressing ethical concerns, hate speech detection tools can fulfill their societal role effectively while respecting user rights and freedoms [21, 22].

7. Conclusion

Hate speech detection in online communities presents complex challenges necessitating interdisciplinary approaches and ethical considerations. Through a systematic review of existing literature and identification of research gaps, this paper proposes a methodological framework leveraging NLP techniques and machine learning algorithms for hate speech detection. This study found that both Logistic Regression and Random Forest models are promising for detecting hate speech, with the Random Forest model outperforming the Logistic Regression model in terms of accuracy and precision. However, the difficulties encountered, particularly in detecting nuanced hate speech, emphasize the limitations of these models in their current state. Sentiment analysis is critical in distinguishing between offensive and humorous content, and its incorporation into hate speech detection systems is proposed as a significant improvement for future research. The findings indicate that future research should focus on integrating context-aware models and improving comprehension of the underlying sentiments in social media posts.This increases the accuracy of detecting hate speech in real-world scenarios, where content can vary significantly in tone and intent. Furthermore, the use of deep learning techniques and a wider range of datasets may improve the generalization capabilities of hate speech detection systems.

References

[1] Jagadeesan, M., Saravanan, T. M., Selvaraj, P. A., Ali, U. A., Arunsivaraj, J., & Balasubramanian, S. (2022). Twitter Sentiment Analysis with Machine Learning. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 681-686). IEEE.
[2] Biere, S., Bhulai, S., & Analytics, M. B. (2018). Hate speech detection using natural language processing techniques. Master Business AnalyticsDepartment of Mathematics Faculty of Science.
[3] Rafique, A., Rustam, F., Narra, M., Mehmood, A., Lee, E., & Ashraf, I. (2022). Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus. PeerJ Computer Science, 8, e1004.
[4] Gambäck, B., &Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Proceedings of the first workshop on abusive language online (pp. 85-90).
[5] Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. IEEE Access, 8, 128923-128929.
[6] Wijaya, D. I., &Arifudin, R. (2022). Detecting Hate Speech Tweets and Abusive Tweets In Indonesian Languange Using Random Forest and Support Vector Machine with Voting Classifier Technique. Journal of Advances in Information Systems and Technology, 4(1), 24-32.
[7] Al-Hassan, A., & Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In 6th international conference on computer science and information technology (Vol. 10, pp. 10-5121).
[8] Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences, 10(12), 4180.

[9] Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. Information Processing & Management, 57(3), 102087.
[10] Kovács, G., Alonso, P., & Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. SN Computer Science, 2, 1-15.
[11] Mullah, N. S., &Zainon, W. M. N. W. (2021). Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access, 9, 88364-88376.
[12] Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., ... &Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5).
[13] Mossie, Z., & Wang, J. H. (2018). Social network hate speech detection for Amharic language. Computer Science & Information Technology, 41-55.
[14] Alrehili, A. (2019, November). Automatic hate speech detection on social media: A brief survey. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-6). IEEE.
[15] Ebube, S. (2023). The Role of Legal Frameworks in Addressing Online Hate Speech and Cyberbullying. American Journal of Law and Policy, 1(1), 13-24.
[16] Chetty, N., &Alathur, S. (2018). Hate speech review in the context of online social networks. Aggression and violent behavior, 40, 108-118.
[17] Keller, N., &Askanius, T. (2020). Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. SCM Studies in Communication and Media, 9(4), 540-572.
[18] Kaur, K., & Gupta, S. (2023). Towards dissemination, detection and combating misinformation on social media: a literature review. Journal of business & industrial marketing, 38(8), 1656-1674.
[19] Balayn, A., Yang, J., Szlavik, Z., &Bozzon, A. (2021). Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. ACM Transactions on Social Computing (TSC), 4(3), 1-56.
[20] Wanniarachchi, V. U., Scogings, C., Susnjak, T., &Mathrani, A. (2023). Hate Speech Patterns in Social Media: A Methodological Framework and Fat Stigma Investigation Incorporating Sentiment Analysis, Topic Modelling and Discourse Analysis. Australasian Journal of Information Systems, 27.
[21] Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. Sensors, 19(21), 4654.
[22] Parker, S., &Ruths, D. (2023). Is hate speech detection the solution the world wants?. Proceedings of the National Academy of Sciences, 120(10), e2209384120.