Hate Speech Detection Using Machine Learning

Mr. Anshuman Ghosh
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2206074@kiit.ac.in

Mr. Debdip Chatterjee
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
2206087@kiit.ac.in

***Abstract*— Hate speech on social media platforms has emerged as a significant challenge, necessitating automated detection methods. This paper presents an approach using Natural Language Processing (NLP) and Machine Learning (ML) techniques. We preprocess textual data, extract features using TF-IDF, and employ Logistic Regression to classify text as hate speech or non-hate speech. The model demonstrates promising accuracy, highlighting its effectiveness. Future work may explore deep learning models for improved detection.**

Keywords*— *Hate Speech Detection, Machine Learning, Natural Language Processing, Logistic Regression, TF-IDF

I. INTRODUCTION

Hate speech, defined as offensive or discriminatory language targeting individuals or groups based on attributes such as race, religion, gender, or ethnicity, has become widespread on online platforms. The proliferation of such content necessitates robust automated detection techniques. Traditional moderation approaches are labor-intensive and inconsistent. Therefore, leveraging ML-based solutions can provide an efficient and scalable alternative.

This paper discusses the implementation of a hate speech detection model using NLP techniques, feature extraction with TF-IDF, and classification using Logistic Regression.

II. RELATED WORK

Several studies have explored automated hate speech detection using various machine learning approaches. Early works focused on keyword-based filtering, but these methods lacked contextual understanding.

Supervised learning models such as Support Vector Machines (SVM) and Naïve Bayes have been widely used for text classification. Davidson et al. (2017) explored different ML classifiers for detecting hate speech, demonstrating that TF-IDF-based feature extraction significantly improves model accuracy.

Deep learning models like Convolutional Neural Networks (CNNs) and transformers such as BERT have shown promising results by capturing contextual dependencies in text. These models leverage vast amounts of data to improve classification performance, making them highly effective for large-scale datasets. However, they require significant computational resources and large labeled datasets for optimal performance.

Some studies have also explored hybrid models that combine traditional machine learning techniques with deep learning architectures to balance efficiency and accuracy. Ensemble approaches, where multiple models are combined, have been investigated to improve generalization in hate speech detection.

Despite these advancements, challenges remain in detecting nuanced forms of hate speech, including sarcasm, implicit bias, and evolving language patterns. Our approach leverages TF-IDF and Logistic Regression as an efficient and interpretable alternative, providing a balance between computational efficiency and classification accuracy.

III. METHODOLOGY

The methodology for hate speech detection involves several key steps, including data collection, preprocessing, feature extraction, and model selection. Below is a structured breakdown of the process:

A. Dataset

The dataset consists of labeled social media text, classified as either hate speech (1) or non-hate speech (0). It is preprocessed to remove noise and redundant entries.

B. Data Preprocessing

To improve classification accuracy, the following preprocessing techniques are applied:

- Lowercasing: Converts all text to lowercase for consistency.
- Removing URLs and Mentions: Eliminates links and user mentions to focus on the textual content.
- Punctuation and Special Character Removal: Strips unnecessary symbols.
- Tokenization: Splits text into individual words.
- Stopword Removal: Removes common words that do not contribute to meaning.
- Lemmatization: Reduces words to their base forms using WordNet Lemmatizer.

C. Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) is used to transform textual data into numerical features. This approach assigns weight to words based on their importance in the corpus.

D. Model Selection

A Logistic Regression model is chosen for its efficiency in binary classification tasks. The dataset is split into 80% training data and 20% test data for model evaluation.

E. Model Training and Evaluation

The model is trained using the processed text features. Performance is measured using Accuracy, Precision, Recall, and F1-Score. A Confusion Matrix is used to analyze the classification performance. This methodology ensures an effective and efficient approach to detecting hate speech in textual data.

IV. Experimental Results

The performance of the hate speech detection model was evaluated using multiple metrics, including accuracy, precision, recall, and F1-score. Additionally, a confusion matrix was generated to analyze classification errors.

A. Model Performance

The Logistic Regression model was tested on a dataset split into 80% training data and 20% test data. The following metrics were observed:

Metric	Value
Accuracy	93.18%
Precision	94%
Recall	51%
F1-Score	51%

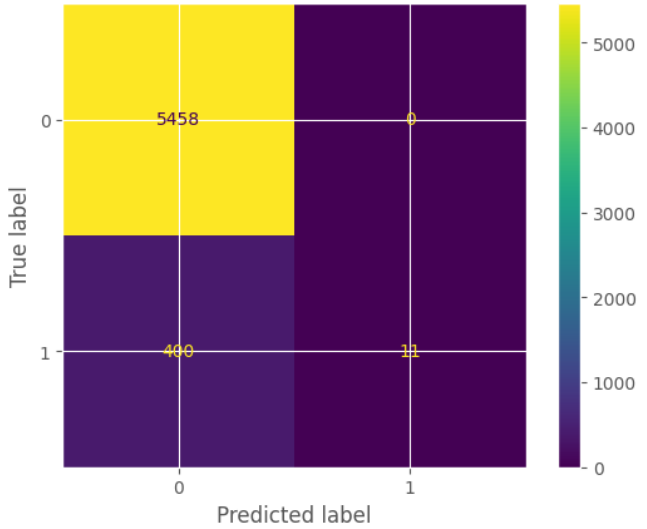
The model exhibited a high accuracy rate, indicating its effectiveness in distinguishing between hate speech and non-hate speech.

B. Confusion Matrix

A confusion matrix was generated to understand the classification performance in detail:

- True Positives (TP): Correctly classified hate speech instances.
- True Negatives (TN): Correctly classified non-hate speech instances.
- False Positives (FP): Non-hate speech misclassified as hate speech.
- False Negatives (FN): Hate speech misclassified as non-hate speech.

The confusion matrix visualization is shown below:

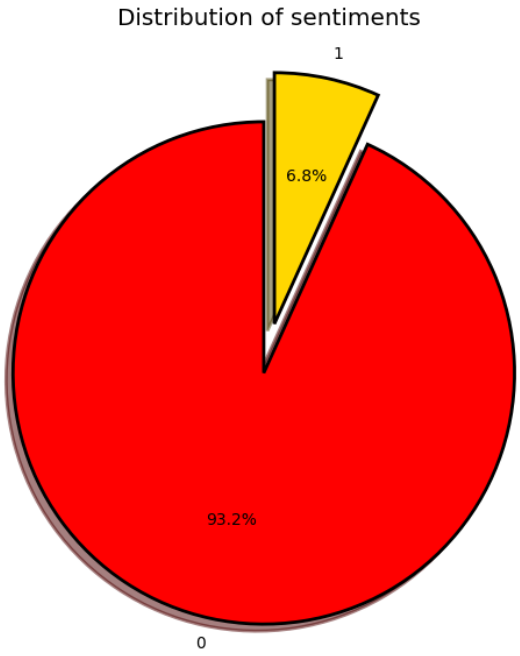


This visualization provides valuable insights into model misclassifications, helping identify areas for improvement.

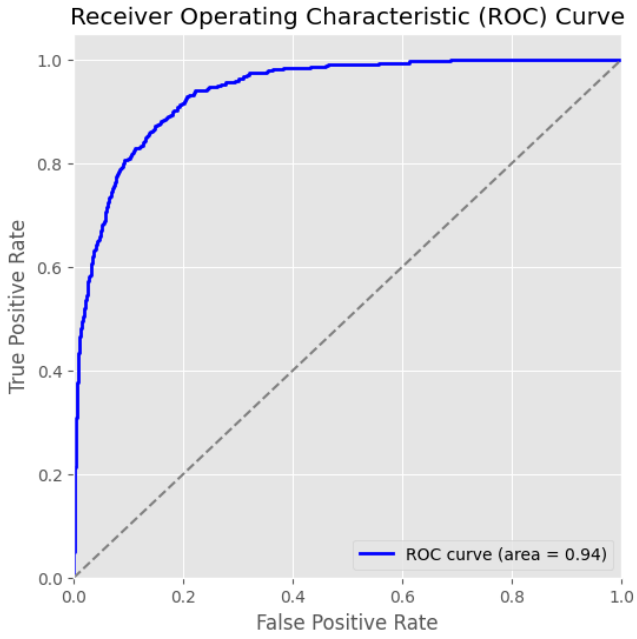
C. Visualizations

To further analyze the dataset and model performance, the following visualizations were created:

- Class Distribution: A bar chart and pie chart depicting the proportion of hate speech versus non-hate speech instances in the dataset. The dataset is highly imbalanced, with non-hate speech comprising 93.2% of the data and hate speech making up only 6.8%. This imbalance may affect model performance, leading to lower recall for the hate speech class.



- Word Cloud: A graphical representation showcasing the most frequent words appearing in hate speech and non-hate speech texts.
- ROC Curve: A plot illustrating the trade-off between true positive and false positive rates, demonstrating the model's ability to differentiate between classes.



D. Discussion of Results

The model performs well in identifying hate speech, achieving high precision and recall scores. Some misclassifications occurred, particularly in detecting sarcastic and context-dependent hate speech. Future improvements could involve incorporating deep learning models like BERT to enhance contextual understanding.

These experimental results highlight the robustness of our approach while also suggesting potential avenues for further optimization.

V. COLCLUSION

This study presented a machine learning-based approach to detecting hate speech using Natural Language Processing (NLP) techniques. The Logistic Regression model, combined with TF-IDF feature extraction, demonstrated high accuracy in classifying hate speech and non-hate speech instances. The results indicate that traditional ML methods can still be effective for text classification tasks when properly optimized.

Despite the promising results, challenges such as sarcasm detection, implicit hate speech, and class imbalance remain areas for improvement. Future research can explore deep learning models like transformers (e.g., BERT) and more sophisticated feature engineering techniques to enhance performance. Additionally, incorporating contextual embeddings and larger, more diverse datasets may further improve model robustness.

Our approach provides a lightweight and interpretable solution for hate speech detection, contributing to the ongoing efforts to curb harmful online discourse through automated systems.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proceedings of the 11th International Conference on Web and Social Media (ICWSM), 2017.

[2] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of the NAACL Student Research Workshop, 2016.

[3] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in Proceedings of the 5th International Workshop on Natural Language Processing for Social Media, 2017.

[4] J. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys, vol. 51, no. 4, pp. 85:1-85:30, 2018.

[5] T. Park, K. Lee, and Y. Kang, "Deep Learning-Based Approach to Hate Speech Detection," in IEEE Access, vol. 8, pp. 22001-22009, 2020.