HateGuard: Automated Multi-Class Cyberbullying Detection Framework

R.Lahari[1], V.Munni[2]

[1]*Pursuing Computer Science, Andhra Loyola Institute of Engineering and Technology, Vijayawada - 12*

[2]*AssociateProfessor,Department of CSE(AIML),Andhra Loyola Institute of Engineering and Technology, Vijayawada – 12*

ABSTRACT - *The rapid growth of social media platforms has led to a significant increase in cyberbullying, posing serious challenges to user safety and online communication. Detecting and categorizing such harmful content at scale requires efficient and automated Natural Language Processing (NLP) techniques. This work presents a multi-class cyberbullying detection system designed to classify textual data from social media into six categories: age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and non-cyberbullying. The proposed approach follows a structured pipeline involving text preprocessing techniques such as tokenization, stopword removal, and lemmatization using NLTK, followed by feature extraction using Term Frequency–Inverse Document Frequency (TF-IDF). Multiple machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Multinomial Naive Bayes, are trained and evaluated to determine the most effective classifier. Experimental results show that Logistic Regression achieves the best performance, with an accuracy of 81.9% and an F1-score of 0.822, demonstrating reliable classification across multiple categories.*

Keywords- Cyberbullying detection, Natural Language Processing, text classification, TF-IDF, machine learning, Logistic Regression, Support Vector Machine, Random Forest, XGBoost, Naive Bayes, social media analysis, multi-class classification

I. INTRODUCTION

The rapid expansion of social media platforms has transformed the way individuals communicate, share opinions, and interact in digital spaces. While these platforms offer numerous benefits, they have also become a breeding ground for harmful behaviors such as cyberbullying, which can have serious psychological and social consequences for individuals. Cyberbullying manifests in various forms, including harassment based on age, gender, ethnicity, and religion, making its detection a complex and multi-dimensional problem. Traditional moderation techniques, which rely heavily on manual review, are not scalable given the vast volume of user-generated content produced across platforms.

To address these challenges, there is a growing need for automated systems capable of accurately identifying and categorizing cyberbullying content in real time. Natural Language Processing (NLP) provides effective tools for analyzing textual data and extracting meaningful patterns that can be used for classification tasks. In this context, machine learning-based approaches have shown promise in detecting abusive language by learning from labeled datasets. However, many existing systems focus only on binary classification, failing to capture the nuanced differences between various types of cyberbullying.

This work aims to develop a multi-class classification system that can detect and categorize cyberbullying into six distinct classes: age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and non-cyberbullying. By leveraging text preprocessing techniques, feature extraction methods, and multiple machine learning algorithms, the proposed system seeks to improve the accuracy and granularity of cyberbullying detection, thereby contributing to safer and more inclusive online environments.

II. LITERATURE SURVEY

Cyberbullying detection has been an active area of research within Natural Language Processing (NLP), with various approaches proposed to identify abusive and harmful content on online platforms. Early research primarily focused on traditional machine learning techniques combined with handcrafted features such as bag-of-words and Term Frequency–Inverse Document Frequency (TF-IDF) [15]. These approaches provided a foundation for automated text classification and were widely adopted due to their simplicity and computational efficiency [13].

Algorithms such as Multinomial Naive Bayes and Support Vector Machine (SVM) were commonly used in these early systems and demonstrated reasonable performance in detecting offensive language [1], [2]. However, these models relied heavily on surface-level textual features and failed to capture deeper contextual meaning and semantic relationships. As a result, their effectiveness decreased when dealing with subtle,

implicit, or context-dependent forms of cyberbullying [5].

To overcome these limitations, deep learning-based models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks were introduced [3], [9]. These models are capable of capturing sequential dependencies in text, allowing them to better understand context and improve classification performance. By learning representations automatically from data, they reduced the need for manual feature engineering and achieved higher accuracy compared to traditional approaches.

More recently, transformer-based models such as BERT and RoBERTa have significantly advanced the field [11], [12]. These models leverage attention mechanisms to understand contextual relationships within text more effectively and have achieved state-of-the-art results in various NLP tasks, including cyberbullying detection. However, they require large datasets and high computational resources, which can limit their practical deployment in resource-constrained environments [10].

In addition to model advancements, research has also focused on improving dataset quality and handling class imbalance, which is a common issue in cyberbullying datasets [6], [7]. Techniques such as data augmentation, resampling, and cost-sensitive learning have been applied to enhance performance across minority classes. Despite these efforts, many existing systems are limited to binary classification and fail to provide fine-grained categorization of cyberbullying types [8]. This highlights the need for multi-class classification frameworks that can deliver more detailed and actionable insights, which motivates the approach proposed in this work.

III. PROPOSED WORK

This paper proposes a cyberbullying detection system that integrates Natural Language Processing (NLP) techniques with machine learning models and rule-based analysis to identify and classify harmful content in social media text. The framework is designed to process raw textual data, such as tweets, and generate accurate classification results across multiple categories of cyberbullying.

A. System Overview

The proposed system presents an end-to-end framework for detecting and classifying cyberbullying in social media text using Natural Language Processing (NLP) and machine learning techniques. The system is designed as a modular pipeline that processes raw textual input and produces a categorized output indicating the type of cyberbullying. Initially, textual data, such as tweets, is collected and passed through a preprocessing stage to remove noise and standardize the content. This stage includes operations such as tokenization, stopword removal, and lemmatization to improve text quality and consistency [14].

B. Data Collection

The dataset used in this study consists of textual data collected from social media platforms, primarily in the form of tweets. These tweets are labeled into six predefined categories representing different types of cyberbullying, along with a non-cyberbullying class. The dataset is curated to include diverse linguistic patterns, informal language, abbreviations, and slang commonly found in online communication. This diversity ensures that the model learns realistic patterns of cyberbullying behavior [6].

C. Data Preprocessing

Data preprocessing is a critical step in the proposed system, aimed at improving the quality and consistency of textual input. The raw text data undergoes multiple preprocessing operations to remove noise and standardize the content. Initially, tokenization is performed to split the text into individual words or tokens, enabling easier analysis. Stopword removal is then applied to eliminate commonly used words such as “the,” “is,” and “and,” which do not contribute significantly to classification [14].

D. Feature Engineering

Feature engineering is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) technique to convert textual data into numerical form [15]. TF-IDF assigns weights to words based on their importance within a document and across the entire dataset. Words that appear frequently in a specific text but are rare across other documents receive higher importance, making them useful for classification tasks.

E. Model Training

Multiple machine learning models are trained to perform multi-class classification of cyberbullying text. These models include Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Multinomial Naive Bayes. Each model is trained using the TF-IDF feature vectors and labeled data to learn patterns associated with different categories of cyberbullying. The implementation and training of these models are carried out using machine learning libraries such as Scikit-learn [13].

F. Model evaluation

Model evaluation is conducted using standard performance metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how well each model

performs across different classes. Accuracy measures the overall correctness of predictions, while precision and recall evaluate the model's ability to correctly identify cyberbullying instances. The evaluation methodology follows commonly used practices in text classification tasks [8].

G. Prediction Pipeline

The final system is implemented as a prediction pipeline that processes raw input text and generates classification results in real time. The pipeline integrates preprocessing, feature extraction, and the trained machine learning model into a unified workflow. When a user inputs text, it is first cleaned and transformed into TF-IDF features before being passed to the classifier. The system then outputs the predicted category of cyberbullying.
The pipeline supports the following classification outputs:

- Age-based cyberbullying
- Ethnicity-based cyberbullying
- Gender-based cyberbullying
- Religion-based cyberbullying
- Other cyberbullying
- Non-cyberbullying

This structured pipeline ensures efficient and scalable deployment, making it suitable for integration into automated moderation systems.

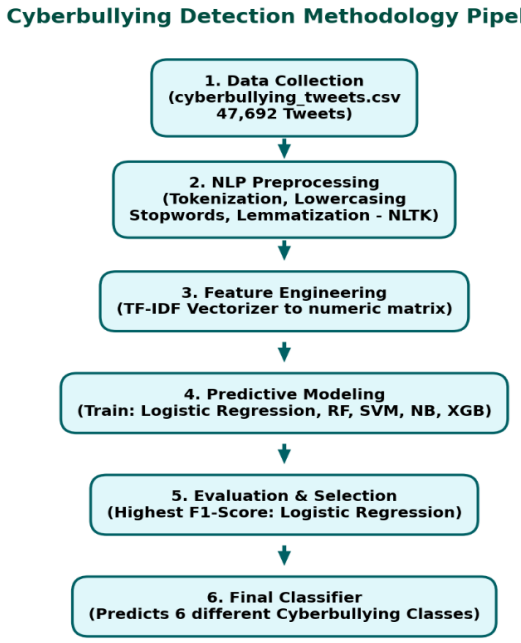

Fig.-1: Methodology Pipeline

IV. REQUIREMENT ANALYSIS

A. Hardware Requirements

The system requires a standard computing environment with a minimum of 8 GB RAM and a multi-core processor such as Intel i5 or equivalent for efficient data processing and model training. While GPU support is not mandatory, it can significantly speed up training for large datasets. Adequate storage is required to handle datasets and model files. The system can be deployed on personal computers or cloud-based platforms for scalability and performance.

B. Software Requirements

The proposed system is implemented using Python as the primary programming language. Key libraries include NLTK for text preprocessing, Scikit-learn for machine learning models, and Pandas and NumPy for data handling and manipulation. TF-IDF vectorization is performed using Scikit-learn utilities. The development environment may include Jupyter Notebook or any Python IDE. Additional tools such as Matplotlib or Seaborn can be used for visualization and performance analysis.

C. Functional Requirements

The system must process raw textual input, perform preprocessing, and convert text into numerical features. It should classify the input into predefined cyberbullying categories using trained machine learning models. The system must support multi-class classification and generate accurate predictions. Additionally, it should allow easy integration into real-time applications for automated content moderation and analysis.

V. RESEARCH AND METHODOLOGY

The proposed system follows a structured machine learning-based methodology for detecting and classifying cyberbullying in textual data. The workflow begins with data collection from social media sources, primarily consisting of tweets labeled into multiple categories of cyberbullying. The collected data is first subjected to preprocessing to remove noise and standardize the text. This includes tokenization, stopword removal, lemmatization, and elimination of irrelevant elements such as URLs, mentions, and special characters. These steps ensure that the textual data is clean and suitable for analysis [14].

Following preprocessing, feature extraction is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) technique [15]. This method converts textual data into numerical feature vectors by assigning importance to words based on their frequency and relevance across documents. The resulting feature vectors provide a structured representation of the text, enabling machine learning models to learn meaningful patterns associated with cyberbullying.

Multiple classification algorithms are then trained using the extracted features, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Multinomial Naive Bayes. The dataset is divided into training and testing sets to evaluate the performance of each model effectively. Hyperparameter tuning is applied to optimize model performance and ensure generalization. The implementation of these models is carried out using machine learning frameworks such as Scikit-learn [13].

The models are evaluated using standard metrics such as accuracy, precision, recall, and F1-score, which are widely used in text classification tasks [8]. Based on comparative analysis, the best-performing model is selected for deployment. The final system integrates preprocessing, feature extraction, and classification into a unified prediction pipeline capable of processing raw text inputs and producing classification outputs in real time. This methodology ensures an efficient, scalable, and reliable approach for multi-class cyberbullying detection.

Dataset

The dataset used in this study consists of labeled textual data collected from social media platforms, primarily in the form of tweets.Each data sample is categorized into one of six classes: age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and non-cyberbullying. The dataset includes a wide variety of linguistic patterns, including informal language, abbreviations, slang, and context-specific expressions commonly found in online communication.

To ensure the quality and effectiveness of the dataset, preprocessing steps are applied to remove duplicates, irrelevant content, and noise such as URLs, user mentions, hashtags, and special characters. This cleaning process helps improve the consistency of the data and enhances model performance. The dataset is carefully analyzed to understand class distribution, as cyberbullying datasets often suffer from class imbalance. Techniques such as balanced sampling or weighting strategies may be considered to address this issue.

VI. RESULTS AND ANALYSIS

The performance of the proposed cyberbullying detection system is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions, while precision evaluates the proportion of correctly identified positive instances among all predicted positives. Recall measures the model's ability to correctly identify actual cyberbullying instances, and the F1-score provides a balanced measure by combining both precision and recall. These metrics are particularly important in multi-class classification tasks, where class imbalance can significantly affect performance.

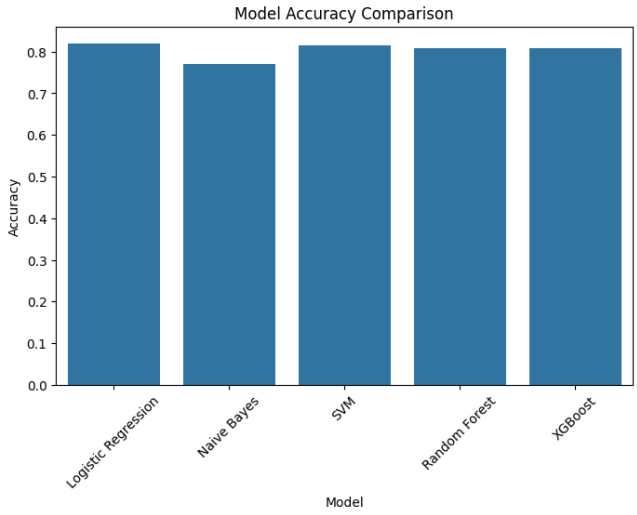

Fig.2 - Accuracy Comparison of Different Machine Learning Models

The experimental evaluation was conducted by training and testing multiple machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Multinomial Naïve Bayes, on the cyberbullying dataset. Each model was assessed using standard classification metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive performance comparison.

Among the evaluated models, Logistic Regression demonstrated the best overall performance, achieving an accuracy of 81.9% and an F1-score of 0.822. This indicates that the model maintains a balanced trade-off between precision and recall, making it more reliable for multi-class classification compared to other models. In contrast, ensemble models such as Random Forest and XGBoost showed comparable accuracy but exhibited slightly lower F1-scores, suggesting less consistent performance across minority classes. Similarly, Multinomial Naïve Bayes, while computationally efficient, struggled with capturing complex contextual relationships in textual data, leading to comparatively weaker classification results.

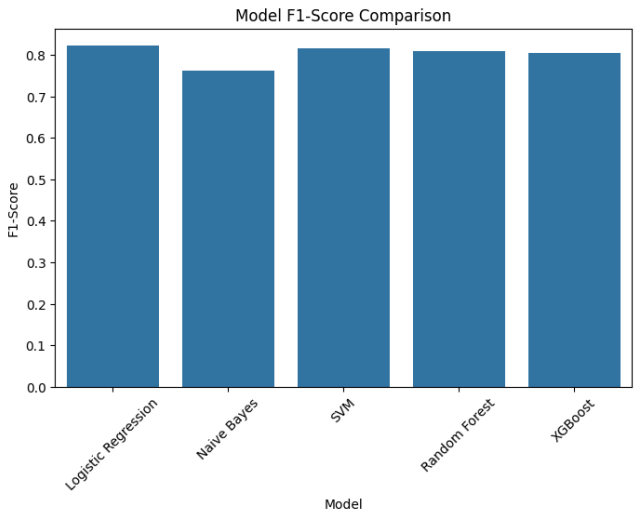

Fig.3 – F1 – Score comparison of different machine learning models

Graphical analysis further supports these findings through performance comparison charts and evaluation curves. The accuracy and F1-score comparison graph illustrates that Logistic Regression outperforms other models in terms of balanced performance. Precision-recall curves indicate that the model maintains a good trade-off between identifying true positives and minimizing false positives across different classes. Additionally, training and validation trends suggest stable learning behavior with minimal overfitting, as performance remains consistent across both datasets.

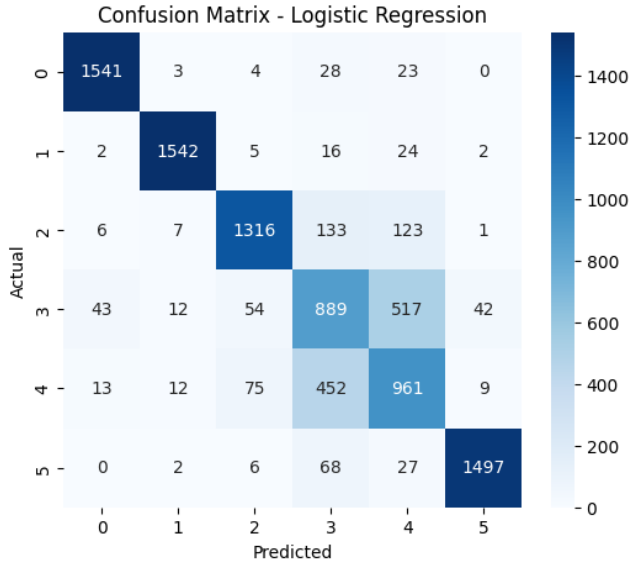

Fig-.4: Confusion Matrix for Logistic Regression Model

VII. CONCLUSION

The study presents a machine learning-based approach for detecting and classifying cyberbullying in social media text using Natural Language Processing techniques. The proposed system follows a structured pipeline that includes text preprocessing, feature extraction using TF-IDF, and multi-class classification using various machine learning algorithms. Among the evaluated models, Logistic Regression demonstrates the best overall performance, achieving an accuracy of 81.9% and an F1-score of 0.822, indicating reliable classification across multiple categories of cyberbullying.

The system effectively categorizes text into six distinct classes, enabling more detailed analysis compared to traditional binary classification approaches. The results highlight that classical machine learning models, when combined with proper preprocessing and feature engineering, can provide efficient and scalable solutions for cyberbullying detection. However, certain limitations remain, particularly in handling context-dependent language and subtle forms of harassment. Future improvements can focus on incorporating advanced deep learning models to enhance contextual understanding and overall performance, making the system more robust for real-world deployment in automated content moderation systems.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their project guide and faculty members for their continuous support, guidance, and valuable suggestions throughout the development of this work. Their expertise and encouragement played a crucial role in shaping the direction and quality of the project. The authors also thank their institution for providing the necessary resources and infrastructure required to carry out this research effectively.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proc. ICWSM, 2017.

[2] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,” in Proc. NAACL-HLT, 2016, pp. 88–93.

[3] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proc. WWW Companion, 2017, pp. 759–760.

[4] I. Kwok and Y. Wang, “Locate the hate: Detecting **tweets** against blacks,” in Proc. AAAI, 2013, pp. 1621–1622.

[5] S. Nobata, J. Tetreault, A. Thomas, Y. Mehdad,

and Y. Chang, "Abusive language detection in online user content," in Proc. WWW, 2016, pp. 145–153.

[6] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, and C. Buntain, "A large labeled corpus for online harassment research," in Proc. CSCW, 2017.

[7] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in Proc. ITASEC, 2017.

[8] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. SocialNLP, 2017, pp. 1–10.

[9] Y. Zhang, B. Wallace, and J. Tetreault, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in Proc. ESWC, 2018.

[10] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in Proc. ACL, 2018.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[12] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[14] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media, 2009.

[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.