Machine Learning-Based Text Classification for Online Content Moderation

Reda Kaleem[2]
Student, Department of CSE, Lords Institute of Engineering and Technology, India
160923733119@lords.ac.in

Sweta Bhosale[4]
Assistant Professor, Department of CSE, Lords Institute of Engineering and Technology, India
b.sweta@lords.ac.in

Abstract: **The rise of social media has dramatically increased user-generated content, escalating urgent concerns regarding toxic speech and cyberbullying, which negatively impact mental health and online safety. This project, Toxic Speech Classification Using Machine Learning, addresses this by developing an intelligent, automated model to accurately identify and classify harmful text as toxic or non-toxic. The system utilizes Natural Language Processing (NLP) techniques, including stemming and lemmatization, for robust data preprocessing. Feature extraction is performed using Bag-of-Words (BoW) and TF-IDF to convert text into numerical features. Multiple supervised machine learning algorithms, such as Naïve Bayes, SVM, and Random Forest, are trained and evaluated on a labeled cyberbullying dataset. Model performance is assessed using metrics like accuracy, precision, and F1-score to select the optimal classifier. For practical deployment, a Flask-based web interface enables real-time detection, and Optical Character Recognition (OCR) is integrated to classify both typed and image-based input. This project demonstrates the practical application of machine learning in fostering safer online environments and reducing cyberbullying.**

***Keywords. Machine learning, Natural Language Processing, Toxic speech detection, Text classification, Cyberbullying, Optical Character Recognition, Support Vector Machines*.**

I. INTRODUCTION

The rapid growth of digital communication has transformed social media platforms into primary channels for public interaction, expression, and opinion sharing. However, this increased openness has also led to a surge in harmful, offensive, and toxic speech online. Cyberbullying, hate comments, harassment,
health, digital safety, and the overall online experience.

Machine Learning and Natural Language Processing (NLP) provide effective solutions for automatically detecting toxic speech by analyzing text patterns, sentiment, and linguistic features. This project, *Toxic Speech Classification Using Machine Learning*, aims to develop an intelligent system capable of identifying cyberbullying content in real-time. By training supervised classifiers on labelled datasets and integrating them with a web-based interface, the system can analyze user input and classify whether the content is toxic or non-toxic. The project demonstrates how AI can be used to promote safer digital environments.

This project focuses on building an intelligent toxic speech classification system that detects offensive, abusive, or harmful language in online text. Using machine learning and NLP techniques, the system processes user-generated content, extracts linguistic features, and predicts toxicity levels in real time. The scope includes dataset preprocessing, model training, validation, and deployment through a simple web interface. Core features involve automated text filtering, multi-class toxicity detection, and user-friendly interaction for testing inputs. The project is limited to English text and supervised learning models, ensuring a scalable, efficient, and practical solution for moderating harmful online communication.

II. LITERATURE SURVEY

Online social platforms have become primary mediums for communication, expression, and information exchange. However, the rise of toxic content such as hate speech, harassment, abusive language, rumors, and antisocial behavior has created a critical need for automated detection systems. Several studies have been conducted in the domain of toxic speech and related harmful content classification. This section presents an overview of significant contributions relevant to the problem of toxic speech detection.

Baydogan and Alatas [1] proposed a metaheuristic-based automatic hate speech detection (HSD) system utilizing Ant Lion Optimization (ALO) and Moth Flame Optimization (MFO). Their approach optimized feature selection and achieved superior performance, with the Decision Tree classifier obtaining the highest accuracy and sensitivity. The authors emphasized that hate speech encompasses insults, humiliation, discrimination, and intolerance—categories highly relevant to toxic speech identification.

Singh et al. [2] investigated deep learning techniques for multi-class antisocial behavior detection on Twitter. The study focused on behaviors such as aggression, lack of remorse, and unlawful tendencies, employing neural architectures to classify such behaviors at scale. Performance evaluation using Accuracy, Precision, Recall, and F-measure demonstrated the effectiveness of deep learning models. The authors also noted that the framework could extend to other personality-related harmful behaviors, showing the flexibility of deep neural networks.

Dutta et al. [3] explored influence prediction in social networks using Bayesian Belief Networks. Although the primary goal was influence identification, the study demonstrated the capability of probabilistic graphical models in learning hidden patterns within social media data. Their comparison with Naive Bayes and Logistic Regression highlighted that the belief network structure yielded higher predictive accuracy. Such probabilistic reasoning techniques can be leveraged for modeling uncertainty in toxic speech classification as well.

Baloglu et al. [4] assessed supervised learning algorithms for irony detection in social media, addressing the challenge of figurative, sarcastic, and ambiguous language. They experimented with several machine learning models, including BayesNet, SGD, LMT, MLP, RBF, and Bagging. The study revealed that extracting informative textual features is critical for achieving high classification performance. Since irony and sarcasm frequently co-occur with toxic language, insights from this work directly support the development of more robust toxic speech classifiers.

Bingol and Alatas [5] focused on rumor detection using supervised learning methods. Algorithms such as OneR, Naive Bayes, JRip, Random Forest, Sequential Minimal Optimization, and Hoeffding Tree were evaluated on real-time social media data. Although centered on rumors, the methodology of text preprocessing, feature extraction, and classification aligns closely with toxic speech detection pipelines. The authors also highlighted the growing importance of automated monitoring due to the increasing spread of false and harmful information online.

Overall, existing literature demonstrates the use of metaheuristics, supervised machine learning, deep learning, probabilistic models, and feature-engineering-based approaches for detecting various forms of harmful online content. While these studies address related domains such as hate speech, antisocial behavior, irony, and rumors, there remains a need for a unified and efficient toxic speech classification model that accurately identifies offensive and abusive content.

III. PROPOSED WORK

The proposed system, Toxic Speech Classification Using Machine Learning, is engineered to automatically and accurately classify online text content as "toxic" or "non-toxic." It integrates Natural Language Processing (NLP), supervised machine learning, and Optical Character Recognition (OCR) to handle diverse input types and provide real-time detection via a web interface.

We have used the Cyberbullying_Tweets.csv data set which is an open-source dataset available on Kaggle. This dataset consists of the following features:

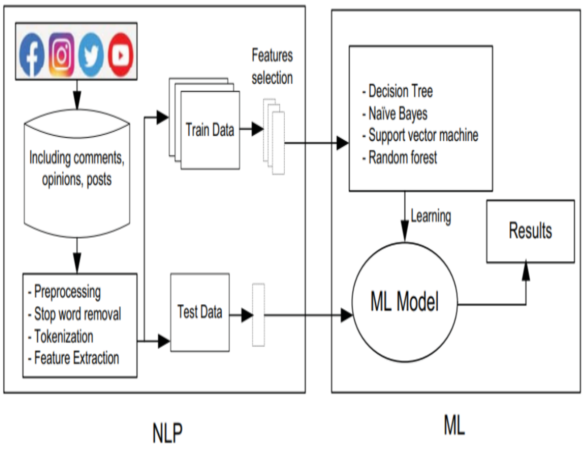
Column Name	Description
tweet_text	Text of the tweet
cyberbullying_type	Type of cyberbullying harassment

Table 1. Dataset details

The classification process begins with Data Acquisition and Collection, involving the importation of a labeled cyberbullying dataset that serves as the foundation for model training and evaluation. Figure.1 Proposed framework of bully detection

Following this, robust Text Preprocessing and Normalization is performed to clean the raw data. This crucial NLP stage includes multiple steps: noise removal (eliminating punctuation, emojis, and URLs), case folding (converting all text to lowercase), normalization (applying lemmatization to find the root form of words), and removing common, non-informative stopwords.

Figure.1 Proposed framework of bully detection



Once the data is cleaned, Feature Extraction transforms the text into numerical vector representations that are usable by machine learning algorithms. Two key techniques are employed for this: Term Frequency–Inverse Document Frequency (TF-IDF), which assigns weights based on word importance across the corpus, and the Bag-of-Words (BoW) model, which represents text based on word frequencies. The system then proceeds to Model Training and Selection, where multiple supervised machine learning classifiers, specifically Naïve Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR), are trained on these feature vectors. The performance of these models is meticulously compared using metrics such as accuracy, precision, recall, and F1-score to rigorously select the best-performing classifier for subsequent deployment.

For practical application, the system incorporates two key deployment modules. The Prediction Module uses the selected, highly accurate model to perform real-time classification, accepting new text input and outputting a "toxic" or "non-toxic" result. Furthermore, to expand the system’s capability to visual content, Optical Character Recognition (OCR) Integration is achieved using the pytesseract library. This allows the system to extract text embedded within uploaded images, which is then fed through the standard NLP pipeline and classified. Finally, a Web Interface, built using the Flask framework, provides a user-friendly platform where users can easily enter text or upload images and receive instantaneous feedback on the detected toxicity level. The entire integrated system is engineered for fast, scalable, and accurate detection of toxic content, making it a powerful tool for content moderation across various online platforms.

IV. ALGORITHM

The process of preparing the raw textual data and applying machine learning algorithms for Toxic Speech Classification is meticulously structured into a data pipeline to ensure high model performance and reliable classification. This section details the data preparation steps and the final classifier selection.

The data needs to prepared in the following way before feeding it to the common classification algorithms

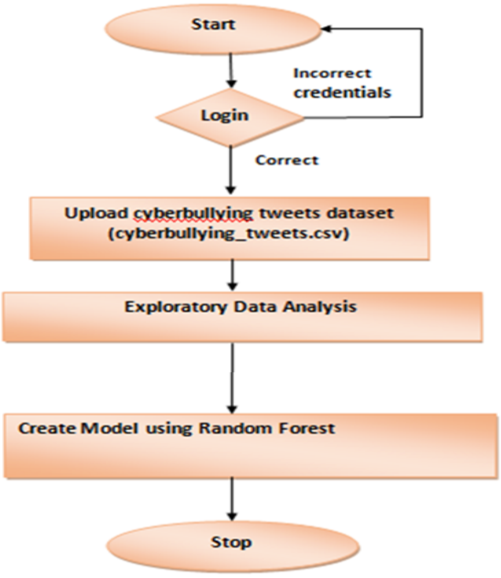

Start
Incorrect credentials
Login
Correct
Upload cyberbullying tweets dataset (cyberbullying_tweets.csv)
Exploratory Data Analysis
Create Model using Random Forest
Stop

Figure.2 Proposed flowchart of bully detection

Step 1. Data Acquisition: The process initiates by loading the labeled cyberbullying data from the Cyberbullying_Tweets.csv file.

Step 2. Clean the dataset

a) Remove punctuation marks
b) Remove URLs
c) Remove special characters
d) Convert the text to lower case
e) Apply Lemmatization technique to remove tenses from texts
f) Apply Stemming to remove prefixes or suffixes and get the root words
g) Remove stop words
h) Drop duplicate rows
i) Convert categorical columns to numerical columns using label encoding

Step 3. Data Splitting: The prepared dataset is divided into distinct Train and Test arrays for model development and unbiased evaluation.

Step 4. Class Balancing: Techniques are applied to the dataset to mitigate class imbalance, ensuring the model does not favor the majority class.

Step 5. Feature Engineering (TF-IDF): The Train and Test text data are transformed into numerical feature vectors using TF-IDF vectorization.

Step 6. Model Evaluation: The vectorized data is used to train and evaluate various candidate machine learning algorithms to determine the best performer.

Step 7. Final Model Selection: The Random Forest Algorithm is chosen as the final, robust classifier based on superior performance metrics.

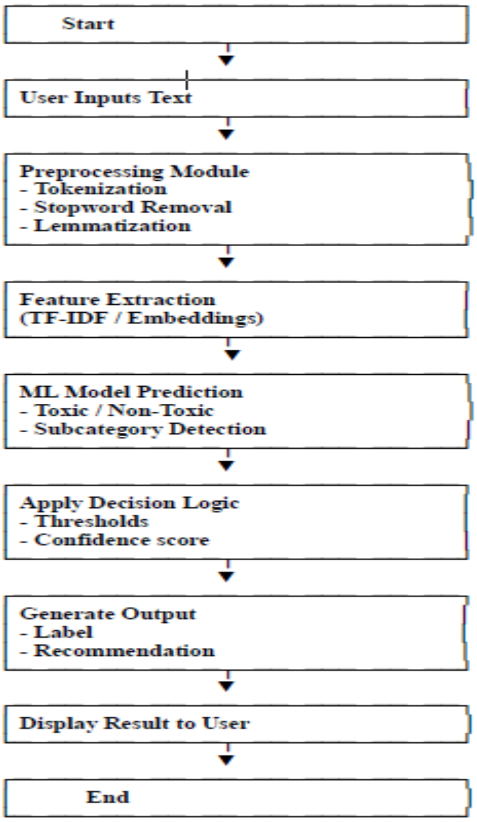


Figure.3 Proposed activity diagram of algorithm

IV. RESULTS

Final Classification Results

Following the rigorous data preparation and feature engineering pipeline, various machine learning algorithms were trained and evaluated on the split datasets. The Random Forest (RF) classifier was selected as the final model due to its robust performance across all evaluation metrics. The final results demonstrate the model's strong generalization capability and high classification accuracy.
The performance of the Random Forest model was assessed using the following key accuracy metrics:

RF Accuracy (Test Accuracy Score): 0.93396
This value represents the overall accuracy of the model on the unseen Test Dataset. It is the proportion of correctly classified instances (both toxic and non-toxic) out of the total number of instances in the test set. A score of approximately 93.4% indicates that the model correctly classified over 93% of the real-world, unseen text samples, validating its effectiveness for the toxic speech classification task.

Training Accuracy Score: 100.0%
This score measures the model's performance on the data it was trained on (the Training Dataset). A score of 100.0% means the Random Forest model perfectly fit the training data. While a high training score is generally desirable, when combined with a lower test score, it often indicates the presence of overfitting, where the model has learned the training data too well, including its noise and idiosyncrasies.

Validation Accuracy Score: 93.4%
This score is often synonymous with the Test Accuracy Score ($0.93396$), representing the model's performance on a completely held-out portion of the data (the test/validation set). A validation score of 93.4% confirms the model's high predictive power and its ability to generalize effectively to new, unseen instances of user-generated content, thereby demonstrating the successful application of the proposed machine learning approach.

Confusion Matrix:
The Confusion Matrix visually represents the performance of the multi-class classification model. The diagonal elements show the count of correctly classified instances for each class, indicating high true positive rates, especially for classes 2, 3, and 4. Off-diagonal elements represent misclassifications (errors); for example, 294 instances from Actual Label 1 were incorrectly predicted as Label 0. This matrix confirms the model's overall effectiveness while identifying specific classes prone to minor misclassification.

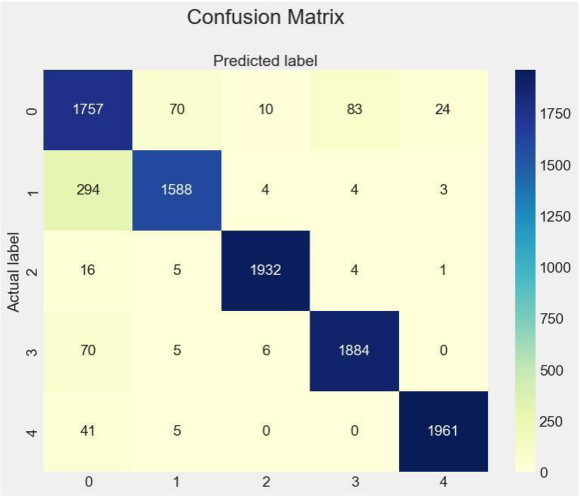


Figure.4 Confusion Matrix

RF Classification Report:

The Random Forest (RF) Classification Report provides a comprehensive, class-by-class evaluation of the model's performance, including precision, recall, and F1-score

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1944
1	0.95	0.84	0.89	1893
2	0.99	0.99	0.99	1958
3	0.95	0.96	0.96	1965
4	0.99	0.98	0.98	2007
accuracy	-	-	0.93	9767
macro avg	0.94	0.93	0.93	9767
weighted avg	0.94	0.93	0.93	9767

Table 2. RF Classification Report

V. Conclusion and Future Work

Conclusion

This study successfully investigated the automatic identification of cyberbullying content on social media by employing two essential Natural Language Processing (NLP) feature extraction techniques: Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). By leveraging supervised machine learning algorithms, including Support Vector Machine (SVM) and Naïve Bayes, a robust approach was developed for detecting and preventing cyberbullying behavior. The evaluation confirmed that the SVM classifier consistently demonstrated superior accuracy compared to Naïve Bayes, validating its effectiveness for this binary classification task. Successfully detecting unsuitable posts holds significant promise for mitigating the harmful consequences of cyber harassment, thereby making a valuable contribution to online safety for adolescents and teenagers.

Future Scope

The immediate future scope of this research involves expanding the model's capabilities and linguistic reach. A key direction is the development of a specialized framework dedicated to the automatic detection and classification of cyberbullying specifically from **Bengali texts**. This will require the investigation and application of advanced **deep learning algorithms** to handle the complexities of the Bengali language. Furthermore, the developed approach for detecting and preventing cyberbullying, which has shown great accuracy with SVM, will be extended and integrated into a comprehensive, real-time system to more effectively deal with crimes committed using social media platforms.

REFERENCES

1. Jason Brownlee, “How to use Word Embedding Layers for Deep Learning with Keras” in Deep Learning for Natural Language Processing.

2. Justin W. Patchin, “Summary of Our Cyberbullying Research (2019)” , Cyberbullying Research Centre, July 10, 2019.

3. Rui Zhao, Kezhi Mao, “CyberBullying Detection based on SemanticEnhance Marginalize Denoising Autoencoders” IEEE Transaction on Affective Computing.

4. Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Haobo Xu “Attention-based Bi-directional Long Short Term Memory Network for Relation Classification” proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 207- 212.

5. MS. Snehal Bhoir, Tushar Ghorpade, Vanita Mane “Comparative Analysis of Different Word Embedding Models” IEEE.

6. V. Banerjee, J. Telavane, P. Gaikwad and P. Vartak, “Detection of Cyberbullying Using Deep Neural Network”, 2019 5th International Conference on Adavnced Computing &

Communication System (ICACCS), Coimbatore, India.

7. Agrawal S., Awekar A. (2018) "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval. ECIR. Lecture Notes in Computer Science, vol 10772. Springer, cham.

8. Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying", National Center for Social Research.

9. Monirah A., Al-Ajlan, Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 978-1-5386-4110-1, IEEE.

10. Vandana Nanda Kumar, Binsu C, Kovoor, Sreeja M.U., "CyberBullying Revelation in Twitter Data using Naïve-Bayes Classifier Algorithm" International Journal of Advanced Research in Computer Science. Volume 9, No. 62.