Journal of IoT in Social, Mobile, Analytics, and Cloud (ISSN: 2582-1369)
www.irojournals.com/iroismac/

VeriSphere: An NLP and XGBoost-Based Framework for Early Detection of Cyberbullying in Social Media

Aririthissh A M.[1], Prasanth J.[2], Yayathiraja N.[3], Preethi Monika K.[4]

[4]Assistant Professor, [1,2,3]Student, Department of Artificial Intelligence and Data Science, NPR College of Engineering and Technology, lndia

E-mail: [1]aririthissham920822243007@nprcolleges.org, [2]prasanthj920822243039@nprcolleges.org, [3]yayathirajan920822243061@nprcolleges.org, [4]preethimonikak@nprcolleges.org

Abstract

The increasing adoption of social media platforms has increased the incidences of cyberbullying which poses serious problems in terms of user security, psychological safety and effective content moderation on the internet. This paper presents VeriSphere which is an intelligent cyberbullying detection framework that incorporates NLP with the use of the XGBoost classifier for identifying cyberbullying at an early stage. The proposed system uses a robust preprocessing pipeline that consists of text normalization, tokenization, stop-word removal, lemmatization and TF–IDF feature extraction to derive discriminating features for classification purposes. XGBoost classifier is used to classify abusive content from non-abusive texts in an efficient way and with good generalization through the use of regularization. In addition, automatic notification system sends emails to administrators as soon as cyberbullying is identified. Experiments have shown that the proposed system can classify cyberbullying cases with 92.84% accuracy, 91.76% precision, 92.31% recall, 92.03% F1-score and 93.90% ROC-AUC which indicate high performance and good discrimination ability.

Keywords: Cyberbullying Detection, Natural Language Processing (NLP), XGBoost Classification, TF–IDF Feature Representation, Automated Content Moderation, Real-Time Text Analytics.

DOI: https://doi.org/10.36548/jismac.2026.3.006
Extended version of article received from **International Conference on Smart Innovations in Engineering & Technology (ICSIET 2026)**
Received:23.05.2026, received in revised form: 28.06.2026, accepted: 15.07.2026, published: 27.07.2026

1. Introduction

Due to rapid growth of social media, digital communication was revolutionized in a way that now allows millions of users to share their views, ideas and experiences by using posts, comments and messages. Though these platforms provide global connections and information sharing opportunities, but unfortunately cyberbullying, hate speech and harassment through these platforms are becoming very common. Spread of such abusive and offensive content is posing many challenges to user safety, mental well-being and quality of digital community. The victims of cyberbullying usually suffer from anxiety, depression and social isolation, while growing number of user generated content makes it increasingly difficult to perform moderation. Therefore, creation of intelligent systems that are able to detect harmful content automatically is becoming highly important.

Several researchers have explored the topic of automatic detection of cyberbullying through rule-based filtering, traditional machine learning techniques, deep learning systems, and transformer language models. Traditional machine learning techniques have shown adequate results in combination with properly constructed textual features, while deep learning models and contextual language models have shown significant improvement in semantic analysis and classification. However, the current methods continue facing a number of difficulties, which include inability to comprehend the nuances of the context, sarcasm, slang, and newly emerging online languages. Moreover, many advanced deep learning systems demand substantial computational power, huge amounts of labeled data, and lengthy training time, thus making the implementation of such systems impractical for social media monitoring on a large scale.

In view of these challenges, the work describes VeriSphere, an intelligent system for early detection of cyberbullying in social media based on NLP and XGBoost classifier. The described system uses a systematic text preprocessing pipeline to convert unstructured information from social media into feature representation and, thus, perform efficient classification of bullying and non-bullying texts. XGBoost integration enables efficient learning, high generalization capability, and good predictive accuracy at low computational costs. Apart from identifying instances of cyberbullying, the described system also allows for automated moderation of social media based on monitoring and notifications about harmful actions in online environment. The described system offers a scalable and computationally

efficient approach for practical social media applications and its efficiency is demonstrated by experimental evaluation.

2. Related Works

The emergence of social media sites has greatly contributed to the increase in textual content created by users, thus making the automated identification of cyberbullying and hate speech an interesting area of research within NLP. Several recent studies have widely covered machine learning and deep learning techniques that are used for identifying abusive language, offensive text, and cyber harassment. Feature engineering-based techniques like Bag of Words, TF IDF, and lexically based features together with classifiers such as SVM, Naïve Bayes, and Logistic Regression were used traditionally for this purpose. Even though these techniques have shown promising classification results, their effectiveness was highly hindered by the lack of contextual information and failure to understand semantic relations of textual data [1], [2], [10].

Deep learning has seen several developments that include the utilization of convolutional neural networks, recurrent neural networks, and attention mechanisms for better feature representation and contextual language modeling. There have been various research works in combining convolutional neural networks with bidirectional recurrent neural networks to enable the capturing of both local and sequential features of texts in order to enhance the detection of implicit hate speech and abusive language. Contextual language models and word embeddings have also made contributions to the increased performance of classification through enhanced semantic representation of textual data. Furthermore, explainable artificial intelligence has been employed in hate speech detection algorithms to ensure that the process becomes more transparent. However, deep learning methods still rely on significant computing power, large-scale labeled data sets, and more time for inference [3], [4], [5], [6].

Multilingual analysis, contextual understanding of language and multimodal analysis of content have been incorporated in the recent research efforts in order to cater to the diversity in the nature of cyberbullying on various social media platforms. Detection systems that are able to process multiple languages and mixed languages have shown greater resilience against regional language variations. Additionally, the use of multimodal data has enabled greater recognition of abusive content present within memes and multimedia posts. Even though the above approaches have provided increased efficiency in the detection process, they have also

added complexity to the task. Thus, there is a requirement for an approach that can be highly computationally efficient, scalable and easily interpretable so that cyberbullying can be detected from text content efficiently. This research problem has motivated the development of the proposed VeriSphere framework which combines natural language processing techniques with the xgboost classifier for efficient and reliable detection of cyberbullying [7], [8], [9], [10].

3. Methodology

The proposed VeriSphere framework aims at identifying the cyberbullying posts on social media text using the combination of NLP and XGBoost Learning Framework. This framework works in transforming the unstructured textual data into useful features to achieve effective discrimination between bullying and non-bullying text with computational efficiency. The process can be better understood from the four interrelated steps involved as shown in Figure 1. These four steps include initialization of the framework, preparing the dataset and preprocessing, feature transformation and classification using XGBoost learning framework and finally automated prediction.

3.1 Proposed Framework

This suggested framework develops a processing pipeline comprising data acquisition, linguistic preprocessing, feature extraction, machine learning based classification and automatic moderation of the process. In the first place, textual data from social media websites is gathered and provided to the preprocessing component, wherein unnecessary symbols, duplicate data, and linguistic errors are eliminated for better data processing. Then, feature vectors for the processed text are created by applying NLP techniques, thus providing a semantic interpretation of the textual data. Then, these feature vectors are provided to the XGBoost classifier for learning discriminative decision boundaries between bullying and non-bullying texts. Lastly, the prediction results are passed on to the notification module for the purpose of moderating offensive posts. The interaction between all these components is depicted in Figure 1.

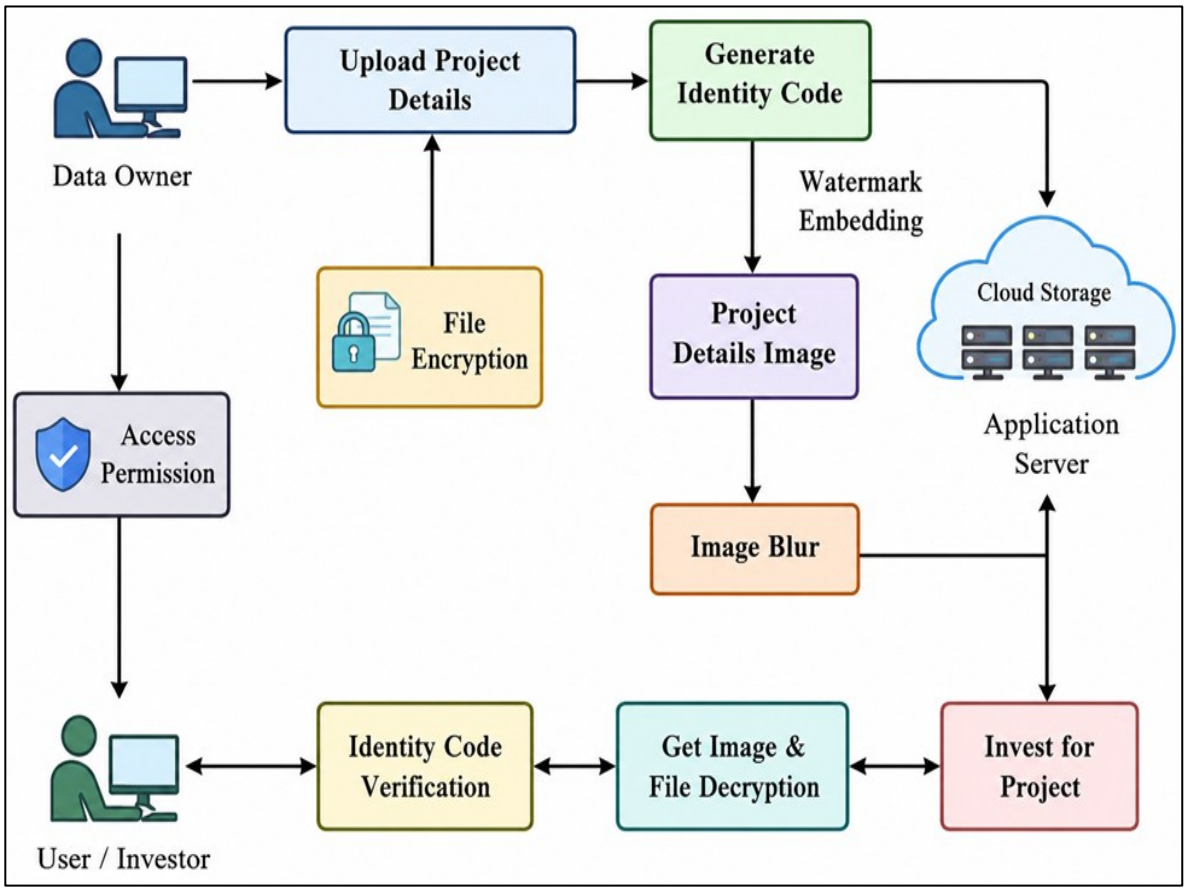

Figure 1. Overall architecture of the proposed VeriSphere framework

3.2 Dataset Description

The VeriSphere approach was formulated and assessed using the Toxicity Parsed Dataset, which is sourced from the freely accessible Kaggle Cyberbullying Dataset archive [11]. This dataset has a collection of 159,686 labelled social media posts in the form of text, and therefore, it provides appropriate input data for training the supervised cyberbullying model. Each of these social media posts is labelled with a class label of either non-bullying (0) or bullying (1) nature, allowing the suggested approach to develop linguistic features for classification of binary texts. The dataset contains a wide variety of online communication types, namely, toxic, abusive, offensive, and normal text conversations.

Table 1. Characteristics of the cyberbullying dataset

Parameter	Description
Dataset Name	Toxicity Parsed Dataset
Dataset Source	Kaggle Cyberbullying Dataset Repository [11]

Domain	Social Media Text
Data Format	CSV
Language	English
Total Samples	159,686
Non-Bullying Samples	144,324
Bullying Samples	15,362
Classification Type	Binary Classification
Input Feature	Text
Train–Test Split	80%: 20%
Training Samples	127,749
Testing Samples	31,937
Feature Representation	TF–IDF
Classification Algorithm	XGBoost

The dataset was analyzed before building a model to make sure that it is consistent and of high quality. The duplicates and missing records were eliminated from the dataset, but the textual information was left in its original form to be used for further Natural Language Processing-based (NLP-based) pre-processing. In order to provide an unbiased estimation of the model's performance, the dataset was split into training set (80%) and test set (20%), which would enable XGBoost Classifier to learn linguistic features. Table 1 describes the characteristics of the datasets used.

The text preprocessing stage involves cleaning up the raw text data from the social media platforms by eliminating noise and retaining valuable language data:

$$T_p = f(T_r) \tag{1}$$

where,

- T_r = Raw textual content collected from social media platforms.
- T_p = Preprocessed textual data after cleaning, tokenization, stop-word removal, and lemmatization.
- $f(\cdot)$ = Text preprocessing function that transforms noisy input into normalized textual representations.

3.3 NLP Feature Representation and XGBoost Classification

Post data preprocessing stage, the cleaned-up text data is converted into numeric feature vectors using Natural Language Processing based feature representation method. In the chosen representation method, the emphasis is given to each term based on its frequency in the document as well as the entire corpus, hence focusing on discriminative terms related to abusive behavior.

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \tag{2}$$

where,

- $TFIDF(t,d)$ = Weight of term t in document d.
- $TF(t,d)$ = Frequency of term t in document d.
- $DF(t)$ = Number of documents containing term t.
- N = Total number of documents in the corpus.
- t = Individual textual term.
- d = Social media document or post.

The feature vectors are then passed into the XGBoost classifier, which trains decision trees through gradient boosting. The algorithm improves its predictions iteratively while maintaining the complexity of the model using regularization. Equation (3) gives the optimization objective of the algorithm, whereas Equations (4) - (5) define the regularization and prediction functions for the algorithm.

The objective function formulation considers minimizing the prediction loss while keeping the model complexity under control through regularization:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where,

- $\mathcal{L}$ = Overall objective function.
- n = Total number of training samples.
- y_i = Actual class label of the i_{th} sample.
- $\hat{y}_i$ = Predicted class label of the i_{th} sample.

- $l(y_i, \hat{y}_i)$ = Classification loss function.
- K = Total number of boosted decision trees.
- f_k = Decision tree generated during the k_{th} boosting iteration.
- $\Omega(f_k)$ = Regularization term controlling model complexity.

The regularization term is introduced in order to lower model complexity and increase the generalization ability of the classifier:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

where,

- $\Omega(f)$ = Regularization penalty.
- γ = Penalty coefficient controlling tree complexity.
- T = Number of terminal (leaf) nodes in the decision tree.
- λ = L2 regularization coefficient.
- w_j = Weight assigned to the j_{th} leaf node.

The final output is derived by averaging the results computed by all of the boosted decision tree:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

where,

- $\hat{y}_i$ = Predicted output for the i_{th} social media post.
- x_i = Feature vector of the i_{th} input sample.
- K = Total number of boosted trees.
- $f_k(x_i)$ = Prediction generated by the k_{th} decision tree.

The use of TF–IDF feature representation along with XGBoost classification represents an efficient way to detect cyberbullying in textual data. Feature extraction allows capturing important linguistic information in the text, whereas XGBoost is capable of learning the data dependencies due to the ensemble learning algorithm implemented in the model. The described approach can be used for practical purposes, as it provides a highly accurate and stable way to analyze large amounts of unstructured data from social networks.

Algorithm 1. Proposed VeriSphere framework for early cyberbullying detection

Input: Social media textual posts (D)

Output: Cyberbullying prediction (Bullying / Non-Bullying)

Step 1. Load the cyberbullying dataset D.

Step 2. Split D into training and testing sets.

Step 3. For each text sample in D:

a. Perform text cleaning and normalization.

b. Remove URLs, punctuation, and special characters.

c. Convert text to lowercase.

d. Apply tokenization.

e. Remove stop words.

f. Perform lemmatization.

Step 4. Transform the preprocessed text into TF-IDF feature vectors.

Step 5. Train the XGBoost classifier using the generated feature vectors.

Step 6. Optimize the objective function with regularization.

Step 7. Predict the class label for each testing sample.

Step 8. Classify the text as Bullying or Non-Bullying.

Step 9. Trigger notification/moderation for detected cyberbullying content.

Step 10. Return the final prediction results.

4. Results and Discussion

4.1 Automated Cyberbullying Detection and Email Notification

The suggested VeriSphere system analyzes the social media messages in real-time through the combination of Natural Language Processing with the XGBoost classifier to detect any cases of cyberbullying. Once the message is submitted in the form of the text, the preprocessing step gets rid of all unnecessary symbols, standardizes the text, and creates TF-IDF feature vectors for discrimination purposes. Depending on the prediction results, the system decides whether the post contains bullying or not and initiates the necessary moderation process instantly.

After classification, the notification process is initiated in case the detected information turns out to be cyberbullying in nature. The notification module generates an email notification of the abusive content and forwards it to the authorized administrator. This helps in taking any necessary action as soon as possible. The inclusion of intelligent text classification along with automated notifications increases the practical implementation of the proposed system in social media networks. Figure 2 shows the email notification produced by the proposed framework when cyberbullying is detected. This notification will contain the detected message and proof that cyberbullying was successfully detected. Through such notifications, social media administrators will be able to monitor abusive messages and moderate them accordingly. The automated notification system will enhance the effectiveness of the whole framework.

The detailed execution process of the proposed VeriSphere framework is depicted in Algorithm 1. The algorithm includes all the steps involved in data set acquisition, pre-processing, features generation, training of the XGBoost model, predicting cyberbullying, and sending notifications. It serves as a complement to the above mathematical formulation through provision of an algorithmic representation of the execution process of the proposed framework.

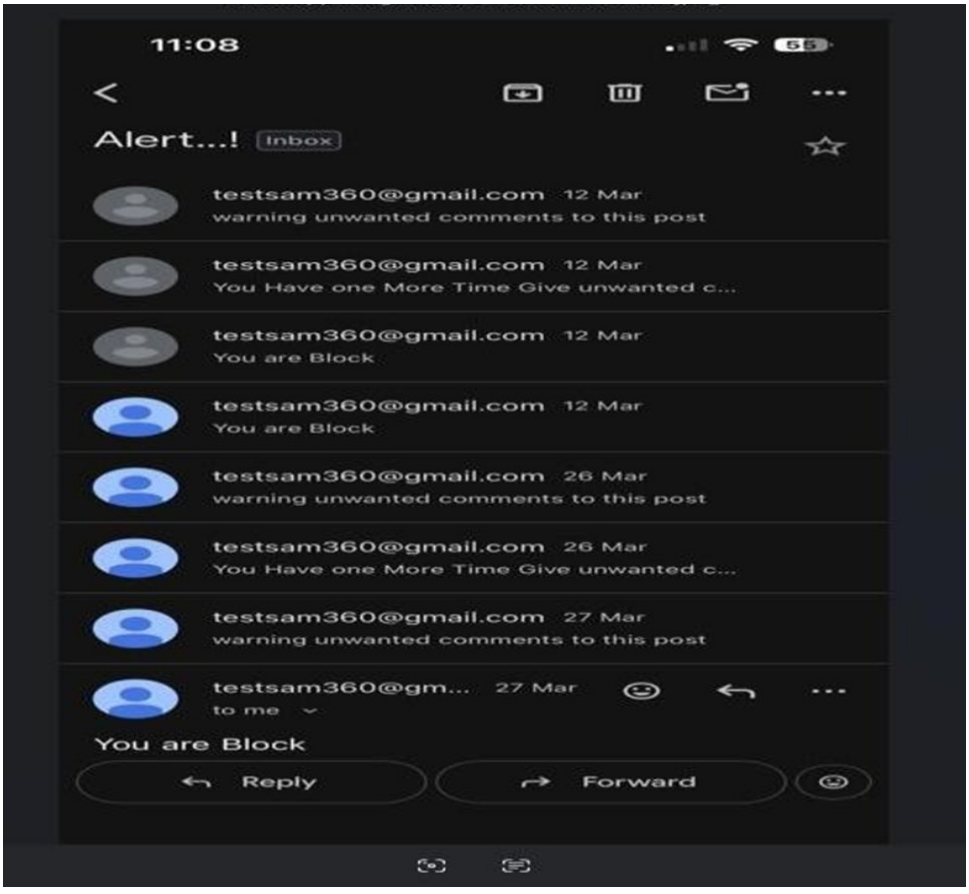

Figure 2. Automated cyberbullying detection and email notification

4.2 Quantitative Performance Evaluation

The performance of the suggested VeriSphere framework was evaluated through quantitative metrics of the binary classification, such as Accuracy, Precision, Recall, F1-score, and ROC-AUC (Receiver Operating Characteristic – Area Under the Curve). This is because these metrics evaluate the effectiveness of the framework concerning its predictive capabilities from the point of view of overall accuracy, quality of positive predictions, detection rate of cyberbullying events, balance of Precision and Recall, and discrimination power of the classifier at various decision thresholds. The results of the evaluation are presented in Table 2.

Table 2. Performance evaluation of the proposed VeriSphere framework

Performance Metric	Value (%)
Accuracy	92.84
Precision	91.76
Recall	92.31
F1-score	92.03
ROC–AUC	93.90

As shown in Table 2, the suggested VeriSphere model performs remarkably well when evaluated against all the considered measures. The results show a consistent ability to discriminate between posts containing cyberbullying and those free of such behavior with low false positives, high sensitivity to abusive content, and an appropriate balance between precision and recall measures. The ROC–AUC score further corroborates the excellent discriminative ability of the model as a result of applying the NLP preprocessing, TF–IDF features, and XGBoost classifier.

4.3 Receiver Operating Characteristic (ROC) Analysis

Additionally, the classification ability of the proposed VeriSphere method was further analyzed using the ROC curve. The ROC curve analysis assesses the classification performance of the NLP and XGBoost algorithm at different decision threshold values and offers an overall evaluation of its capacity to discriminate between cyberbullying and normal social media posts. The ROC analysis serves as an additional assessment method for the classification output, which shows how the true positive rate is traded against the false positive rate during the prediction process.

Figure 3 reveals that the ROC curve stays very close to the upper left corner of the graph, showing very good class separability. High changes in the true positive rate with minimal changes in the false positive rate is a result of the good discrimination of the textual features by the preprocessing step of the NLP and TF-IDF approach. In addition to this, gradient boosting algorithm used by the XGBoost classifier helps the model learn complex boundaries within the language.

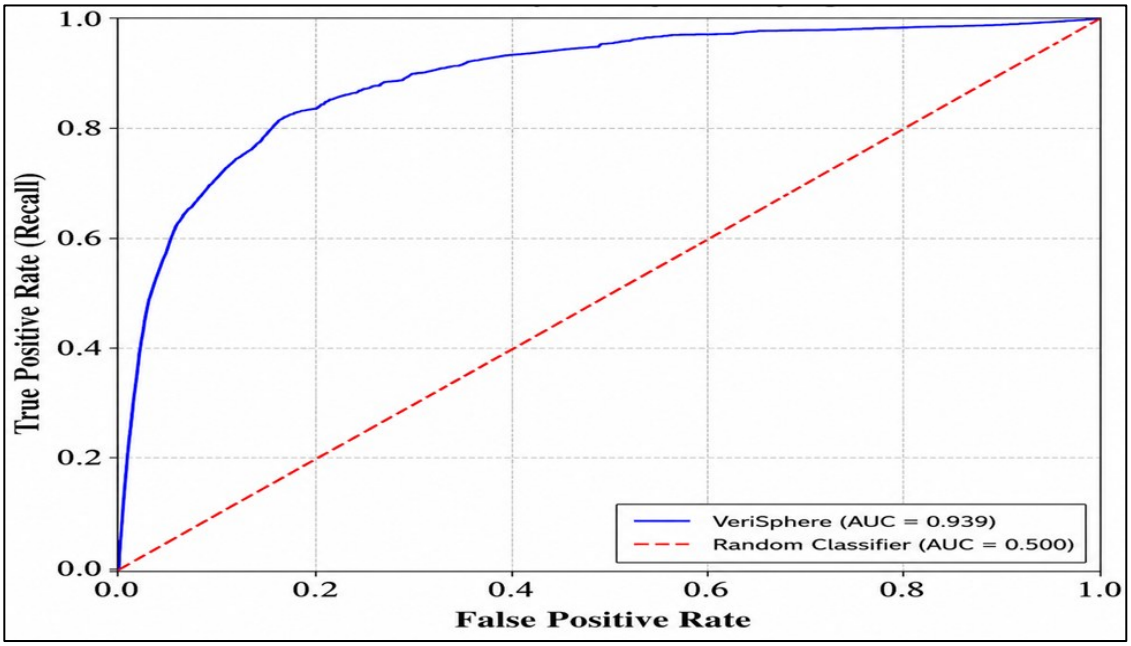

Figure 3. ROC curve of the proposed VeriSphere framework

These results show that the suggested approach possesses reliable predictive behavior at various threshold values and does not experience degradation of classification consistency. The smooth behavior of the graph shows that the algorithm demonstrates a good balance between the sensitivity and specificity in identifying cyberbullying. Consequently, the proposed approach will be able to minimize unnecessary labeling of the content while providing dependable detection of abusive patterns in texts. Thus, these results prove that the proposed VeriSphere system is a computationally effective approach for cyberbullying detection and social media content moderation.

4.4 Discussion

Based on the results of the experiment, it is possible to conclude that the VeriSphere approach suggested for detecting cyberbullying is successfully combining NLP techniques with the XGBoost classifier providing an effective and computationally efficient solution to the

problem. With the help of the pipeline of text preprocessing steps (text normalization, tokenization, stop-words filtering, lemmatization, TF–IDF features), the level of noise in texts is reduced, while linguistic information remains unchanged. As a result, the XGBoost classifier is able to detect patterns corresponding to abusive language with high generalization capability due to ensemble learning and regularization techniques.

Moreover, the ROC analysis clearly indicates that the consistency of the classification process is maintained across various decision levels, and hence the efficiency and reliability of the proposed model can be observed. While comparing with the more complex methods of deep learning, the proposed approach provides a relatively simpler framework in terms of computational complexity, fast inference, and implementation. The incorporation of automatic email notification also makes the framework more practical for the detection of cyberbullying because of its ability to moderate such posts in time.

5. Conclusion and Future Work

VeriSphere, an intelligent cyberbullying detection system, was developed in this study. VeriSphere consists of a combination of NLP methods with XGBoost classifier for early detection of abusive contents on social networks. The proposed system consists of comprehensive text preprocessing, TF–IDF feature representation, and ensembles to precisely differentiate between bullying and non-bullying content with good predictive performance and high computation efficiency. VeriSphere achieved 92.84% accuracy, 91.76% precision, 92.31% recall, 92.03% F1-score, and 93.90% ROC–AUC which clearly shows the effectiveness and efficiency of the proposed system. In addition to this, the inclusion of an automated email alert mechanism increases the practicality of the system for timely moderation of online toxic behavior without putting much manual effort into it. Thus, it can be said that the proposed system provides an effective and reliable solution for real-time cyberbullying detection on social media platforms. The future research efforts are aimed at extending the framework for supporting multiple languages and code-mixed text using transformer-based language models for better context comprehension, multimodal cyberbullying detection using visual text, and developing adaptive learning algorithms that identify evolving cyberbullying patterns with computational efficiency for large scale real-time deployment.

References

[1] Subramanian, Malliga, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, and G. Manikandan. "A Survey on Hate Speech Detection and Sentiment Analysis using Machine Learning and Deep Learning Models." Alexandria Engineering Journal 2023, vol. 80: 110-121.

[2] Alkomah, Fatimah, and Xiaogang Ma. "A Literature Review of Textual Hate Speech Detection Methods and Datasets." Information 2022, vol. 13, no. 6: 273.

[3] Khan, Shakir, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, and Abdul Rauf Baig. "BiCHAT: BiLSTM with Deep CNN and Hierarchical Attention for Hate Speech Detection." Journal of King Saud University-Computer and Information Sciences 2022, vol. 34, no. 7: 4335-4344.

[4] Mehta, Harshkumar, and Kalpdrum Passi. "Social Media Hate Speech Detection using Explainable Artificial Intelligence (XAI)." Algorithms 2022, vol.15, no. 8: 291.

[5] Saleh, Hind, Areej Alhothali, and Kawthar Moria. "Detection of Hate Speech Using BERT and Hate Speech Word Embedding with Deep Model." Applied Artificial Intelligence 2023, vol 37, no. 1: 2166719.

[6] Pérez, Juan Manuel, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac et al. "Assessing the Impact of Contextual Information in Hate Speech Detection." IEEE Access 2023, vol. 11: 30575-30590.

[7] Bilal, Muhammad, Atif Khan, Salman Jan, and Shahrulniza Musa. "Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform." IEEE Access 2022, vol.10: 121133-121151.

[8] Raza Ur Rehman, Hafiz Muhammad, Mahpara Saleem, Muhammad Zeeshan Jhandir, Eduardo Silva Alvarado, Helena Garay, and Imran Ashraf. "Detecting Hate in Diversity: A Survey of Multilingual Code-Mixed Image and Video Analysis." Journal of Big Data 2025, vol.12, no. 1: 109.

[9] Karim, Md Rezaul, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. "Multimodal Hate Speech Detection from Bengali Memes and Texts."

in International Conference on Speech and Language Technologies for Low-resource Languages, Cham: Springer International Publishing, 2022: 293-308.

[10] Jahan, Md Saroar, and Mourad Oussalah. "A Systematic Review of Hate Speech Automatic Detection using Natural Language Processing." Neurocomputing 2023, vol. 546: 126232.

[11] F. Elsafoury, Cyberbullying Datasets, Mendeley Data, vol. 1, 2020. Available: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset