International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com
Volume 3 Issue 8 , August 2026
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

A Machine Learning Framework for Real-Time Cyberbullying Detection, Monitoring, and Explainable Analytics on Social Media Platforms

Krupa Rani[1], Heeba A[2]
[1] Assistant Professor, GM University, Davangere
[2] Student, GM University, Davangere.

Article Info

Article History:

Published: 21 August 2026

Publication Issue:
Volume 3, Issue 8
August-2026

Page Number:
495-508

Corresponding Author:
Heeba A

Abstract:

The emergence of social media has not only simplified communication but has brought about more cyberbullying cases. It is not easy to manually detect any cyberbullying case due to the vast numbers of posts generated each day in social media platforms. The current study aims at developing a machine learning framework that identifies cyberbullying cases and monitors the content of the social media through a web-based application. Text data will be processed using TF-IDF vectorization with unigram and bigram features while Logistic Regression, Multinomial Naïve Bayes and Linear SVM classifiers will be used to classify the data. Classification will be done by training and testing 14,602 cleaned and manually labeled social media posts. From the different classifiers considered, Logistic Regression classifier performed the best in terms of predicting the bullying class with an accuracy of 74.70% and an F1-score of 0.4703. Other than classification, the proposed framework offers other functionalities such as predictions explanations, severity level, prediction history, user feedback, administrative analytics and real time Reddit comment analysis.

Keywords: Cyberbullying detection, machine learning, NLP, TF-IDF, logistic regression, text classification, Explainable AI, social media analytics

1. INTRODUCTION

Nowadays social media becomes an inevitable part of human interaction that enables users to express their views, communicate and exchange any data instantly. But at the same time social media could be used to post offensive or abusive content. One of the most critical problems connected to the usage of online communication is cyberbullying, which includes harassment through insults, threats, humiliation, or other actions targeted at some individual.

The significant amount of content produced by users of social media platforms makes it impossible to manually review each of the posts by human moderators. Thus, automated methods should be used to recognize potentially threatening content. ML and NLP enable efficient analysis of the textual content and classification of posts based on their characteristics. Previously developed cyberbullying detection systems have relied on classical machine learning techniques, while modern approaches use deep learning architectures, such as CNN, LSTM, Bi-LSTM, and transformers like BERT.

Though there are many studies providing the evaluation results of their models in terms of performance on different datasets and classification metrics, there is no practical implementation of this issue. Thus, the detection system not only should be able to perform classification but also provide information for users, keep track of predictions, enable monitoring and give opportunity to get human feedback. These requirements are essential for the development of the practical social media-oriented application.

Thus, this research proposes a cyberbullying detection machine learning system and a web application. In this framework, we use TF-IDF features and compare Logistic Regression, Multinomial Naïve Bayes, and Linear SVM

classifiers. Along with the prediction, the proposed framework gives explanation for it, defines the severity level of a message, keeps user history, provides feedback possibility, admin analytics and scanning of messages in real-time from Reddit.

The major contributions of the present work can be listed as follows:

- It uses identical text processing procedures during the stages of training and prediction, thus guaranteeing consistency and equal treatment of training and deployment phases.
- The use of three Machine Learning algorithms—Logistic Regression, Multinomial Naïve Bayes, and Linear SVM—is examined with the help of TF-IDF features. The performance of the models is compared using adequate metrics.
- Furthermore, the model generates explanations to its decisions through identifying the most influential words. The technique doesn't require the existence of an initial list of offensive words.
- A full-featured web application has been developed with such functionalities as prediction history, users' ratings, export to CSV, administration panel, and real-time scanning of Reddit comments.

The remainder of this paper is organized in the following manner. Section II introduces relevant works from the literature. The proposed methodology is described in detail in Section III. Section IV considers the data set and system architecture. Experimental results are presented in Section V and discussed in Section VI, while the conclusions are drawn in Section VII.

2. RELATED WORK

Development of automated cyberbullying detection has become much more advanced during recent years compared to previous efforts. Early research was mostly oriented on classical machine learning approaches and manual text feature extraction. Recent developments focus on deep learning approaches and more advanced transformer-based models, which are capable of capturing more complex patterns in social media language. One can classify the existing research into three major groups: classical machine learning methods, deep learning and transformers, and survey-based papers.

A. Classical Machine Learning Methods

Classical machine learning models are frequently used for cyberbullying detection in social media text. Van Hee et al. [7] performed research on cyberbullying detection with the usage of Support Vector Machine (SVM) on English and Dutch texts. They reported F1-scores equal to 64% for English and 61% for Dutch, demonstrating that languages can affect the performance of text classification model.

Prabowo and Azizah [8] use TF-IDF features along with SVM for cyberbullying detection. The reported accuracy is 93%, with precision equal to 95% and recall being 97%. At the same time, the researchers developed browser-based interface, which is capable of performing text classification in real time.

Almufareh et al. [9] investigate cyberbullying using several machine learning classifiers: Extra Trees, SVM, Logistic Regression, Random Forest, and Naïve Bayes. The authors also consider class imbalance problem using SMOTE sampling. According to their research, Extra Trees classifier showed the highest performance with accuracy of 95.38% and F1-score of 0.95. Thus, handling of imbalanced dataset can influence the performance.

Haidar et al. [10] develop multilingual approach for cyberbullying detection for Arabic and English content using SVM. Accuracy of English part of their dataset is 93.4% and F1-score is 92.7%. Moreover, Almutiry et al. [11] perform cyberbullying detection for Arabic content using sentiment analysis and SVM-based classifier. Thus, one needs to take care of language-specific features when developing cyberbullying detection tool.

Classical machine learning approaches remain popular due to simplicity, efficiency, and interpretability of models. However, performance of such algorithms can be affected negatively when the meaning of message depends heavily on context, sarcasm, word order, or any other social media language feature.

B. Deep Learning and Transformer Models

The area of deep learning has also opened up a new way of cyberbullying detection as they enable models to learn useful representations from text without relying heavily on manually created features. Hasan et al. [2] evaluated multiple deep learning models such as CNN, LSTM, Bi-LSTM, and CNN-BiLSTM and found out that they can outperform traditional models when there is enough labelled data.

Lu et al. [12] proposed character-level CNN model for detecting cyberbullying in text messages on social media. This study also dealt with spelling variations and other features of online language that make it challenging to work with text on social media platforms. Raj et al. [13] tried out machine learning as well as deep learning techniques in one detection pipeline and found both successful outcomes.

Transformer-based language models are the next step in improving NLP models' ability to understand context of text. Devlin et al. [14] presented BERT – bidirectional transformer model that has become widely used backbone for modern text classification models. BERT based models, including those developed specifically for cyberbullying detection, can find contextual relations between words that are too hard to find with simple lexical approaches.

Words' representation approaches like Word2Vec [15] and GloVe [16] were used in different deep learning pipelines. With these approaches, words could be represented by vectors of numbers that contain more information about relations between words than frequency-based representation.

Despite the fact that deep learning and transformer models can provide deeper understanding of context of text, they need more computing power and larger labelled dataset to be used effectively.

C. Surveys and System-Level Studies

Several scholars have conducted reviews concerning the general development of automated cyberbullying detection systems. Salawu et al. [17] have examined various approaches and classified the features utilized in cyberbullying detection as lexical, semantic, and social network-based features. The study shows a variety of information that may be taken into consideration during analysis of online behaviour.

Rosa et al. [18] performed a systematic review of automatic cyberbullying detection and highlighted the lack of standardized datasets and evaluation methods among the existing researches. Al-Garadi et al. [19] reviewed machine learning-based cyberbullying prediction within big data of social media and outlined the problems such as class imbalance, multi-lingualism, and real-time cyberbullying detection.

Teng and Varathan [20] examined traditional machine learning approaches and transfer learning techniques concerning their use in cyberbullying detection. The results of the study show that transfer learning may bring performance improvement although at a cost of increased computational complexity.

D. Research Gap

It is evident from the existing literature that traditional machine learning and modern deep learning technologies may be used for cyberbullying detection. However, most of the studies pay little attention to the implementation of the model as part of a complete system rather than focus on its performance.

Moreover, there are several issues such as class imbalance, real-time processing, explainability, user feedback, and monitoring which should be considered when implementing a model to practice. The current research pays attention not only to the comparison of machine learning classifiers but also to the design of a complete web-based framework

which includes prediction, explanation, severity assessment, user history, feedback, administrative analytics, and social media monitoring.

TABLE I

Summary of Representative Related Work

Study	Technique	Dataset	Reported Result
Van Hee et al. [7]	SVM	EN/NL social media posts	F1 = 64% (EN), 61% (NL)
Prabowo & Azizah [8]	TF-IDF + SVM	Web-crawled comments (1,500)	Acc = 93%, Prec = 95%
Haidar et al. [10]	SVM	Arabic–English Twitter corpus	Acc = 93.4%, F1 = 92.7%
Almufareh et al. [9]	Extra Trees + SMOTE + SA	Penta-class Twitter dataset	Acc = 95.38%, F1 = 0.95
Lu et al. [12]	Character-level CNN	EN Tweets + Chinese Weibo	Acc ≈ 98%
Teng & Varathan [20]	ML vs. Transfer Learning	Social network corpus	Transfer learning > classical ML
This work	TF-IDF + LR/NB/SVM + deployed app	14,602 annotated posts	Acc = 74.70%, F1 = 0.4703 (LR)

3. PROPOSED METHODOLOGY

The system that has been proposed is structured to be an end-to-end system used for cyberbullying detection in social media text. Text preprocessing, feature extraction, machine learning classifier, prediction explanation, severity

analysis, and a web application developed on the Django framework make up the entire system architecture. The entire process involves four stages, including text preprocessing, feature extraction, modeling, and prediction with monitoring.

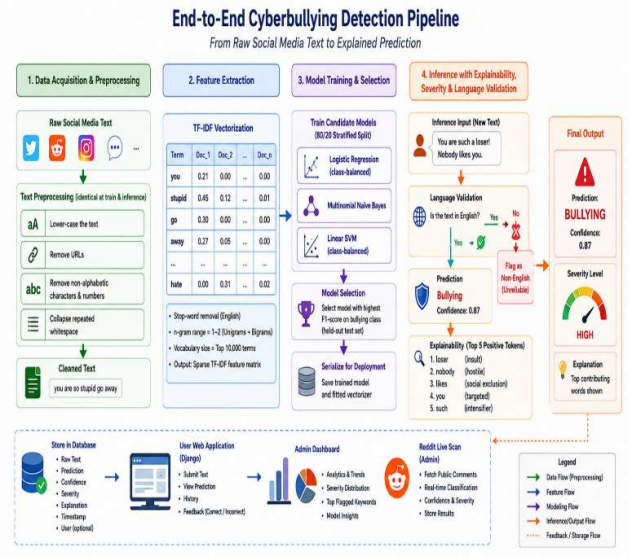

Fig.1 End-to-end cyberbullying detection pipeline from raw social media text to explanation of prediction.

A. Text Preprocessing

Prior to passing the message to the machine learning model, the text is preprocessed by removing the unnecessary parts which could influence the classification process. The same preprocessing technique is used both in the training phase and during prediction to provide consistent text format for the model.

Preprocessing comprises text to lower case conversion, URL removal, keeping only lower case alphabetic letters and spaces, and extra whitespace removal. Using the same technique in both phases avoids inconsistencies between training and deployment stages.

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

B. Feature Extraction

The text is then represented numerically through the use of the Term Frequency-Inverse Document Frequency (TF-IDF) approach after preprocessing. The system takes into account unigrams and bigrams since this will allow it to capture not only single words but also pairs of words that appear consecutively. Only the highest scoring 10,000 terms form the vocabulary.

The reason for choosing TF-IDF is the computational efficiency it offers and the interpretability of its representation of the words appearing in a message. It also enables the system to make predictions without using GPU computation.

C. Model Training and Selection

Three supervised machine learning classifiers are trained and compared: Logistic Regression, Multinomial Naïve Bayes, and Linear Support Vector Machine (SVM). Dataset is split into training and testing subsets using 80:20 ratio.

Class weights are applied to Logistic Regression and Linear SVM in order to balance the impact of the inequality between number of bullying and non-bullying classes. Models are tested on the unseen data subset and special focus is put on the F1-score for the bullying class. The most appropriate classifier is selected.

D. Explainability, Severity and Language Validation

The proposed approach provides an explanation together with the classification result. Coefficients from the linear models show the importance of each word for prediction. While predicting, the system finds top 5 positively weighted words of the message under consideration and returns the output in human-friendly format.

The system also categorizes the prediction confidence or decision score into one of the three severity levels: low, medium and high. As the model was trained using English text, language validation is included in order to filter out the messages which seem to be not in English.

E. Django-Based Web Application

Machine learning algorithms are incorporated into the web application built with Django. ML component includes training script, text cleaning module, trained models and vectorizers, prediction and explanation services.

User part includes account registration and login page, one message prediction, prediction history, feedback mechanism, activity graph and exporting the user history into CSV file. Administrator part includes analytics and monitoring tools such as counts of bullying and non-bullying messages, severity statistics, activity trends, flagged keywords, model insights, dataset exporting and real time scanning of Reddit.

Generally, proposed approach combines the machine learning pipeline and the web application in a way that the user's message can be pre-processed, classified, explained and stored as part of the system's monitoring process.

4. DATASET AND SYSTEM DESIGN

A. Dataset Description

There are 20,002 raw posts of social media posts that come with an explicitly annotated ground truth label (`0` for non-bullying, `1` for bullying) along with the text of the post and the user handle. After the exclusion of all raw posts that have null text or null annotation value, duplicate raw posts, and raw posts that turn into an empty string during the pre-processing step, there are 14,602 posts left to use for training. The class distribution of the posts after applying the exclusion step is shown in Table II, which indicates that the distribution of classes is obviously imbalanced toward the majority class, a characteristic commonly observed in practice as well as in the literature [9], [19]. The data after cleaning can be used for training and testing the proposed models for cyberbullying detection since the data does not include irrelevant and noisy examples anymore.

TABLE II

Class Distribution After Cleaning (N = 14,602)

Class	Count	Percentage
Not Bullying (0)	11,815	80.9%
Bullying (1)	2,787	19.1%

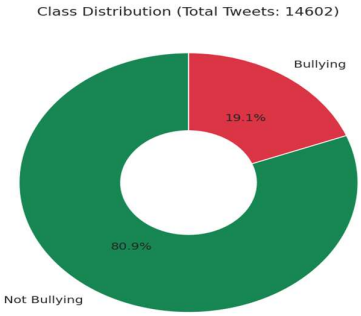

Fig. 2. Class distribution of the cleaned dataset.

The system is designed to be built as a Django web application comprising three main components. The ml component comprises the offline training script, the text-cleaning module, the saved models and vectorizers, the prediction service, and the explanation service. The User component deals with user account registration and authentication, the one-message prediction functionality, the user-specific history display component (which offers the option for confirming or flagging comments, the seven-day activity graph, and the CSV file download of a user's history). The Admin component allows for a session-based admin login, the overall analytics dashboard (bullying vs safe comments, their severity distribution, the fourteen-day activity graph, and the top flagged keywords extracted on-demand), the model insights (which include offline metrics of the current model), the CSV export of the entire dataset, and the live Reddit scanner (which downloads comments from a subreddit of interest in an anonymous way from the Reddit JSON read-only API and classifies them in real-time).

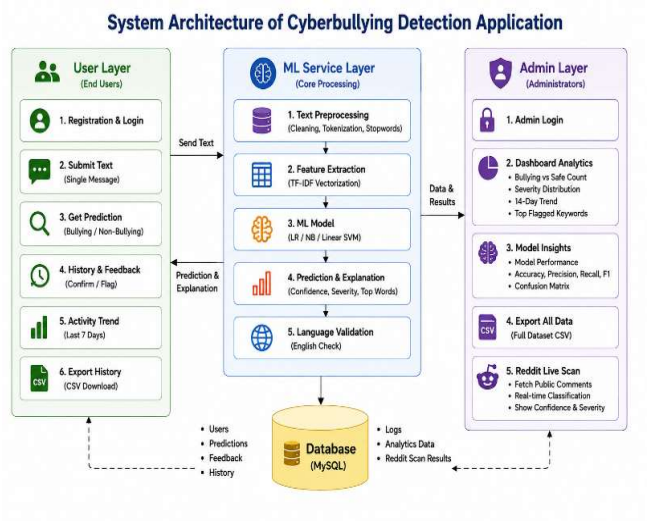

Fig. 3. System architecture showing the User-facing, Administrative, and ML service layers of the Django application.

C. Data Model

Each classification request is stored in the CyberTweets database entry along with the originating user (nullable for anonymous/system scans), the actual text of the message, the predicted label, the confidence score value, the severity value, the time of creation, and an additional user flag that could be correct, incorrect, or not yet reviewed. This last element creates a feedback loop between the automatic prediction and the human assessment and serves as the foundation for the active learning extension described in Section VI.

5. EXPERIMENTAL RESULTS

All model candidates were tested on the same test set (n=2921) that constitutes 20% of the total corpus with accuracy and F1-scores in the class of bullying as the key metrics for comparison since accuracy may not be a good metric on its own due to class imbalance (Table II).

TABLE III

Model Comparison on Held-Out Test Set

Model	Accuracy	F1 (Bullying)	Precision (Bullying)	Recall (Bullying)
Logistic Regression	0.7470	0.4703	0.39	0.59
Multinomial Naïve Bayes	0.8100	0.0246	0.64	0.01
Linear SVM	0.7467	0.4228	0.37	0.49

Although the Naive Bayes algorithm yields the maximum accuracy (81.00%), the F1-Score yielded by the same for Bullying class drops drastically to 0.0246. This is a classic illustration of the fact that accuracy cannot be trusted as a metric in case of class imbalance and the selection of model must take into consideration the F1-Score. Therefore, Logistic Regression was selected as the deployment model.

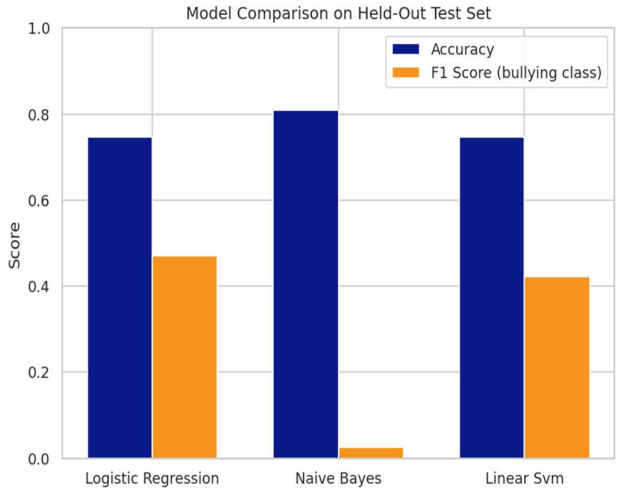

Fig. 4. Accuracy and F1-score comparison across the three candidate classifiers.

TABLE IV

Confusion Matrix — Logistic Regression (Deployed Model)

	Predicted: Not Bullying	**Predicted: Bullying**
Actual: Not Bullying	1,854	509
Actual: Bullying	230	328

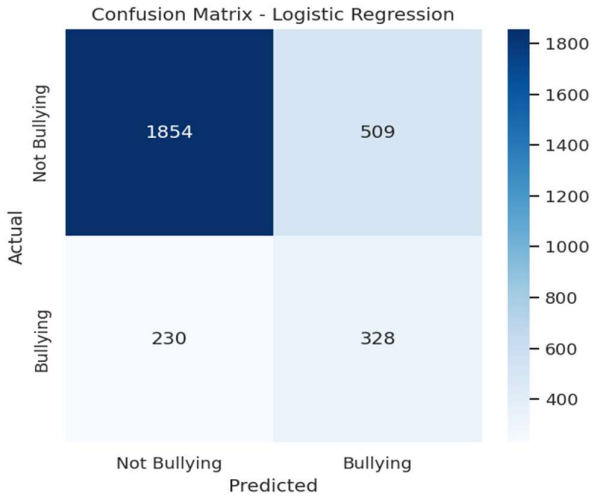

Fig. 5. Confusion matrix heat-map for the deployed Logistic Regression model.

Based on the confusion matrix above, it is clear that our model is able to detect 328 instances out of 558 cases of bullying (recall = 0.59) but gives 509 false positives for 2,363 cases of non-bullying (false-positive rate = 0.22). This is by design as it is a consequence of the class-balancing weights, and it works very well as in the case of a tool that is used to help in moderation, false negatives have more costs than false positives.

Figure 6 shows the distribution of message length in terms of words. The messages in the corpus classified as bullying messages tend to be relatively shorter than those classified as not being bullying messages. This is in agreement with observations made in other corpora [7], [17].

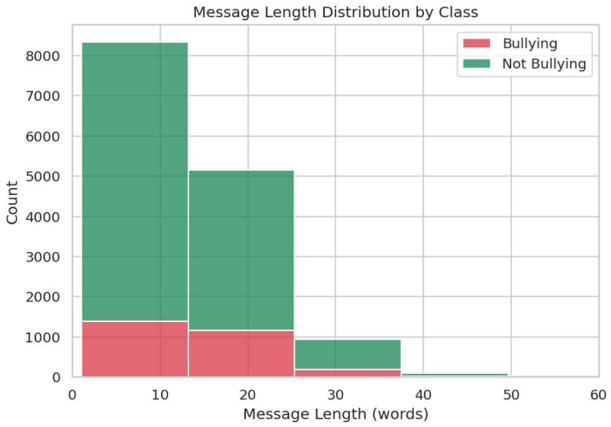

Fig. 6. Message-length distribution by class.

Table V benchmarks the present system's results against the representative prior studies introduced in Section II.

TABLE V

Comparative Results Against Prior Published Work

Study	Best Model	Accuracy	F1-Score
Van Hee et al. [7]	SVM	—	0.64 (EN)
Prabowo & Azizah [8]	SVM + TF-IDF	0.93	—
Haidar et al. [10]	SVM	0.934	0.927
Almufareh et al. [9]	Extra Trees + SMOTE	0.9538	0.95
This work	Logistic Regression	0.7470	0.4703

The relatively lower F1-score in this work compared to Almufareh et al. [9], for instance, is due to the two differences in methodology highlighted in Section VI as follows: (i) the current corpus is a binary corpus and more imbalanced (80.9% vs. 19.1%) compared to the resampled penta-class corpus in [9], and (ii) no synthetic minority oversampling technique such as SMOTE was used in this version of the system because it was designed to ensure that the system was trained only on real data. Regardless of the limitations identified, the proposed approach proves that a relatively simple and easily interpretable machine learning model is capable of performing reasonably well on a real-life cyberbullying detection task. In turn, the current results highlight the potential of applying more computationally intensive solutions, such as resampling algorithms or transformer-based architectures, to the problem.

A. Application Screenshots

The application interface in this work can be demonstrated using the figures below which contain placeholders for screenshot of application interface.

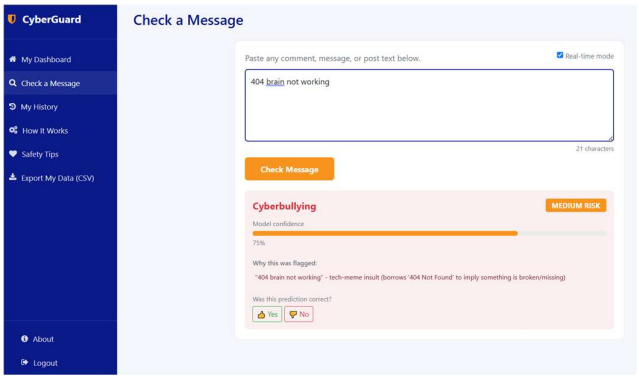

Fig. 7. Single-message cyberbullying check interface, showing prediction label, confidence, severity, and word-level explanation

Fig. 8. Administrator analytics dashboard showing bullying/safe counts, severity breakdown, and fourteen-day trend.

6. DISCUSSION

Based on the results of the developed model, it is clear that a lightweight machine learning algorithm can be effective in identifying cyberbullying in cases where computational resources are limited. Out of all three models, Logistic Regression was chosen to be used due to its best F1-score in relation to the bullying class. Nonetheless, there are some limitations related to the results that should be taken into account before applying the proposed system in practice.

A. Class Imbalance

Firstly, it should be noted that there is a problem of class imbalance in the used dataset, since after preprocessing 80.9% of the messages belong to non-bullying class and only 19.1% – to the bullying one.

Therefore, the metric of accuracy is not sufficient for evaluating the performance of the model since even Naïve Bayes reached an accuracy of 81.00%, while having F1-score of 0.0246 for the minority class, whereas Logistic Regression managed to reach the same accuracy level with F1-score of 0.4703 and therefore was considered more useful for identification of the class.

The currently developed system uses weights for classes in Logistic Regression and Linear SVM without altering the initial data distribution, which can be solved by such methods as SMOTE in the future.

B. Lexical Features and Context

As was mentioned above, the usage of TF-IDF makes the proposed system quite straightforward and effective, however, it also means that it will have several shortcomings. First of all, TF-IDF focuses mostly on representation of importance of certain words and phrases and does not take into account the context of their usage. That is why cases when there is sarcasm or negation, special order of words or implicit forms of cyberbullying can be hard to classify correctly.

There can be used other models like contextual models (BERT) or models specific for cyberbullying (transformers). They might be quite helpful in terms of dealing with these cases. But, at the same time, these models will demand more computing power and may diminish certain qualities of the current system (efficiency, simplicity). For this reason, the future version of the system can become hybrid.

C. Explainability and Accuracy

One of the key features of the proposed system is the capability of providing the explanation of the prediction. Thanks to the fact that the models being used are linear, their coefficients can be easily used in order to find out the words influencing the prediction and making it easier for user/administrator to understand.

At the same time, the explainable approach described above works well only in case of linear models. If it is decided to upgrade the models and make them more advanced and non-linear, then the explanation via coefficients will not work anymore.

D. Deployment and Human-in-the-Loop Design

In contrast to approaches, which only consider offline model evaluation, our proposed design involves the connection of the classifier and full-fledged web application. It keeps track of the prediction and provides a feedback channel for indicating whether the prediction was correct or not. This provides the ground for the participation of human reviewers in the process of detection.

The system is also designed to provide real-time predictions within a single web request by using CPU-only calculations. Such an approach is appropriate in the context of lightweight applications requiring faster and less computationally expensive responses. At the moment, the system does not have any mechanisms to automatically re-train or update the model based on feedback. The use of active learning might be considered as a possible future direction of research to enable the use of feedback for further model improvement.

E. LIMITATIONS

Although the proposed system provides an efficient way for identifying cyberbullying, there are certain limitations that must be considered.

• Limitation in the language: Currently, the system is able to handle only the English language. Although there is language detection, the system neither process nor classify the messages written in other languages.

• Limitation in the dataset: The dataset was collected from a single source in a particular period of time. It may cause a problem in performing tasks on newer datasets or with different social media language.

• Class imbalance: In the cleaned dataset, there is 80.9% of non-bullying messages and 19.1% bullying messages. This causes a problem since the dataset is unbalanced.

• Limitation in the authentication: The administrator login and permissions are provided using the credentials configured through application environment and not using the authentication and permissions system of Django.

Therefore, there are a number of limitations in the existing system framework; however, improvements can be made in the system.

7. CONCLUSION

This paper presented the full-fledged cyberbullying detection system, which comprises the extensively benchmarked TF-IDF and classical ML classification pipeline together with the web-based application able to conduct real-time monitoring, feedback collection, and administrative analytics. Among three available classifiers, Logistic Regression is proved to be the most efficient regarding precision and recall of the minority class of the cyberbullying case; the F1 score of 0.4703 and 74.70% of accuracy were obtained for a test sample, consisting of 2,921 messages extracted from a 14,602 annotated messages corpus. The comparison with the performance of the previously studied systems shows that resampling based ensemble techniques might lead to much more effective results in terms of F1 scores with differently structured datasets. In the current case, however, the lightweight approach presents an acceptable compromise between interpretability and deployment capabilities, including unique public scanning on social media networks capability. Further studies will include resampling by means of SMOTE, transformer models, and active learning of the feedback loop.

References

[1] L. H. Collantes, Y. Martafian, S. N. Khofifah, T. K. Fajarwati, N. T. Lassela, and M. Khairunnisa, "The impact of cyberbullying on mental health of the victims," in Proc. 4th Int. Conf. Vocational Educ. Training (ICOVET), 2020, pp. 30-35.

[2] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," Future Internet, vol. 15, no. 5, p. 179, 2023.

[3] D. Nikolaou, "Does cyberbullying impact youth suicidal behaviors?" Journal of Health Economics, vol. 56, pp. 30-46, 2017.

[4] S. Unnava and S. R. Parasana, "A study of cyberbullying detection and classification techniques: A machine learning approach," Engineering, Technology & Applied Science Research, vol. 14, no. 4, pp. 15607-15613, 2024.

[5] C. N. Kamath, S. S. Bukhari, and A. Deng

[6] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," Social Network Analysis and Mining, vol. 12, no. 1, p. 99, 2022.

[7] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," PLoS ONE, vol. 13, no. 10, 2018.

[8] W. A. Prabowo and F. Azizah, "Sentiment analysis for cyberbullying detection using TF-IDF and SVM," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 6, pp. 1142–1148, 2020.

[9] M. F. Almufareh, N. Z. Jhanjhi, M. Humayun, G. N. Alwakid, D. Javed, and S. N. Almuayqil, "..."

[10] B. Haidar, M. Chamoun

[11] S. Almutiry, M. A. Fattah, and S. A.-A. Almunawarah, "Arabic cyberbullying detection through Arabic sentiment analysis," ...

[12] N. Lu, G. Wu, Z. Zhang, Y. Zheng, Y. Ren, and K.-K.-R. Choo, "Cyberbullying detection in social media text via character-level convolutional neural network with shortcut connections," Concurrency and Computation: Practice and Experience, vol. 32, no. 23, 2020.

[13] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application for detecting cyberbullying via machine learning and deep learning algorithms," SN Computer Science, vol. 3, no. 5, 2022.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.0480

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532-1543.

[16] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," IEEE Transactions on Affective Computing, vol. 11, no. 1, pp. 3-24, 2020.

[17] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," Computers in Human Behavior, vol. 93, pp. 333-345, 2019.

[18] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," IEEE Access, vol. 7, pp. 70701-70718, 2019.

[19] T. H. Teng and K. D. Varathan, "Detection of cyberbullying on social networks using machine learning vs. transfer learning methods," IEEE Access, vol. 11, pp. 55533-55560, 2023.

[20] F. Pedregosa, G. Varoquaux, V. Michel,