RESEARCH PAPER

Integrating Sentiment Analysis with Machine Learning for Cyberbullying Detection on Social Media

Amgoth Venkatesh Pawar[1*], **SK Mahaboob Basha**[1]

[1]Department of Computer Science and Engineering, Sree Dattha Institute of Engineering and Sciences, Sheriguda, Hyderabad, Telangana, India

[1]Student (Amgoth Venkatesh Pawar) | Email: amgoth.venkatesh@gmail.com

[1]Professor (SK Mahaboob Basha) | Email: email2mahaboob@gmail.com

***Corresponding Author:** Amgoth Venkatesh Pawar, Student, Department of Computer Science and Engineering, Sree Dattha Institute of Engineering and Sciences, Sheriguda, Hyderabad, Telangana, India | Email: amgoth.venkatesh@gmail.com

Abstract— The potential effects of cyberbullying through social media are widespread and have serious psychological and emotional consequences and thus the need to develop effective detection systems. Sentiment analysis using advanced ML offers a valuable and scalable system of identifying harmful interactions and creating safer online space. The report uses the Cyberbullying Tweets dataset of Kaggle, which contains labeled instances of various types on cyberbullying. Preprocessing includes normalization, noise removal such as URLs and HTML tags, tokenization, WordNet lemmatization, VADER sentiment analysis, as well as label encoding. TF-IDF vectorization is used to extract feature and SMOTE oversampling is used to reduce the imbalance in classes. Such visualization techniques as distribution plots and word clouds are used to clarify trends in bullying and the dominance of categories. The suggested structure has multiple classifiers, such as LR, RF, XGBoost, Decision Tree, Naive Bayes, SVM, Extra Tree, Gradient Boost, and AdaBoost, and the hyperparameter optimization is done in connection with the application of GridSearchCV. The accuracy, precision, recall, F1-score, and confusion matrices are used to evaluate the performance of the model in terms of false positives and false negatives. Other enhancements include the SMOTEENN to add data balance and Voting Classifier ensemble (MLP with Bagging and Logistic Regression) to strong classification. Explainable AI approaches such as LIME and SHAP allow interpretability by identifying meaningful characteristics, whereas Flask-based implementation allows real-time predictions with scores of confidence and subject modeling to achieve usability and transparency. The results show that the proposed Voting Classifier outperforms all the baseline models significantly and achieves 97.4% when considered in all evaluation parameters, therefore demonstrating the reliability, scalability, and feasibility of the proposed system in detecting cyberbullying.

"Keywords— Cyberbullying Detection, Sentiment Analysis, Machine Learning, TF-IDF, SMOTE, Ensemble Learning, Explainable AI, Flask Deployment."

How to cite this article: Pawar AV, Basha SKM. Integrating sentiment analysis with machine learning for cyberbullying detection on social media. Int J Drug Deliv Technol. 2026;16(75s): 388-394. DOI: 10.25258/ijddt.16.75s.42

I. INTRODUCTION

The social media has transformed the nature of interpersonal communication, which has brought together individuals, sharing of ideas and real time conversation. Despite the positive aspects, the sites have also become favorable conditions of negative practices, including cyberbullying. Cyberbullying, which is the intentional harm (commonly caused by computer) through the use of digital platforms, such as the sharing of offensive content and threats, fabrication and impersonation, and disclosure of personal data. These acts have devastating effects on people and especially teenagers who often end up experiencing anxiety, depression, social isolation and at worst cases, suicidal tendencies [6], [7]. The growing popularity of social media applications such as Twitter, Facebook, and Instagram increases the need to create advanced detection tools that prevent users of these applications to be affected by mental pain [1], [2].

The traditional type of manual moderation and flagging systems though predominant are ineffective in managing the large amount of content generated daily [8]. To address this challenge, scholars have explored algorithms that employ sentiment analysis and machine learning in order to automatically detect harmful trends [3], [4]. Sentiment analysis can be used to improve understanding of a text by determining the emotional polarity of a text, which is better at classifying a text in a context. Combining the supervised algorithms with linguistic and semantic information allows a detection of abusive interactions to be performed with scalability and efficiency [5]. The introduction of transformer-based architectures and ensemble learning methods have already shown how a cyberbullying detection system can be improved in accuracy [9], [10]. These methods apply contextual embeddings and collaborative decision-making to overcome traditional models as well as offer robustness when unbalanced data are involved.

Focus on the English-language content, which comprises the majority of the online discourse, enhances the generalization of detection algorithms to different platforms [1]. Besides categorization, there is the use of explainability to ensure transparency in that it focuses on the important features

that determine predictions [4]. The goal is to design a clever, reliable and intelligible system which will autonomously recognize harmful interaction, add sentiment indicators and provide instant feedback. This will ensure that guardians, educators and other stakeholders take necessary action in time, ensuring safer and healthier digital environments [2], [5].

II. Related Work

Cyberbullying is a major problem of the digital era, particularly since the use of social media has grown in popularity among the teenagers and the children of young age. The reports conducted on the matter indicate that the level of the online harassment is increasing, and many of the victims report a significant amount of the psychological and emotional distress [11], [12]. It has been found that cyberbullying victims often experience anxiety, depression, and low self-esteem, which have long-term negative consequences on mental health [13]. It has led to increased focus on computation methods to discover, classify and mitigate harmful behavior on the internet.

The literature highlighted the growing importance of machine learning and sentiment analysis techniques in the context of automatic detection of cyberbullying. The traditional ways of using keywords and rule-based systems cannot work well due to the diverse linguistic forms that offenders use [14]. The intricate pattern of harmful communication has been discovered by using supervised and unsupervised methods of learning. Comparison of the different detection and classification strategies has shown that machine learning models such as logistic regression, support vector machines and random forests are able to provide reliable performance when applied to structured data [14]. Preprocessing and feature extraction heavily depend on the accuracy of these systems, which means that complex sentiment analysis algorithms have to be included.

Sentiment analysis is a necessity to detect cyberbullying by trying to detect emotional undertones within the text. Another popular tool of sentiment analysis is VADER, which has been shown to be effective in detecting the polarity in short messages on social media [15]. Studies that were conducted with the help of sentiment attributes report that negative emotional coloring and abusive or bullying behavior is highly correlated [16]. According to Atoum [16], sentiment polarity affects the classification models positively in terms of detection rate compared with the use of textual characteristics. Moreover, the hybrid procedures utilizing TF-IDF vectors and sentiment characteristics have shown a higher precision and robustness of changing online conditions.

Another important point of view is the role of the session-based analysis in recognizing cyberbullying events. Yi and Zubiaga [17] argue that the analysis of posts is only a partial study that fails to capture the contextual development of the conversation. With the incorporation of the session-based method, detection frameworks are able to recognize interaction between numerous exchanges, which improves the ability to distinguish between legitimate and malicious information. The era of modeling is essential in social networks because the circumstances determine the intensity and purpose of a message.

The most recent advances in deep learning and transfer learning will greatly enhance the effectiveness of detecting cyberbullying programs. Teng and Varathan [18] compared the standard machine learning to the transfer learning systems and found that transformer-based systems such as BERT significantly outperform the classical systems, especially when dealing with complex language. The models take advantage of using pre-trained contextual embeddings thus making them generalizable to multiple datasets and meanwhile capturing subtle verbal cues. However, challenges such as data imbalance and high-computing cost still exist and need the hybrid solutions incorporating traditional classifiers and modern deep learning frameworks.

Other negative behaviors such as cyber trolling and hate speech have overlapping traits with cyberbullying. A sentence-level sentiment analysis of cyber trolling tweets with machine learning techniques was conducted by Sharif et al. [19], demonstrating that the addition of sentiment components to the analysis increased the classification accuracy. This highlights a broad applicability of emotion-advanced detection methods in online environments regulation. Similar outcomes have been observed in cases of multilingualism. Almutiry et al. [20] developed an Arabic cyberbullying detector that utilized Arabic sentiment analysis, which demonstrated the importance of models that utilize language-related specifics to meet cultural and linguistic peculiarities. Their results showed that sentiment based methodologies do work in English and in other languages and therefore increased the applicability of the systems to the rest of the world.

III. Materials And Methods

The proposed solution aims to enhance cyberbullying detection on social media with the help of sentiment analysis and advanced machine learning models. Normalization, tokenization, noise, and label-encoding are some of the preprocessing methods that are used on Kaggle Cyberbullying Tweets dataset in order to guarantee structured inputs. TF-IDF is also used in terms of feature extraction i.e., the frequency patterns of various terms in context, and SMOTE in terms of correcting imbalance in classes and supporting minor classes [27], [28]. Many different classifiers are evaluated, including Logistic Regression, Hyperparameter tuning is achieved through GridSearchCV, including RF, XGBoost, Decision Tree, Naive Bayes, SVM, Gradient Boost, and AdaBoost. In order to maximize the reliability of prediction, SMOTEENN is used to reduce noise and to sample the data as even, whereas a Voting Classifier is used to combine multiple models in order to make a robust decision. LIME and SHAP methodologies are explainable AI which makes them more interpretable, as it focuses on important features and therefore offers transparency in the categorization outputs. Such a system will ensure that harmful interactions can be identified using scalable, accurate, and understandable means.

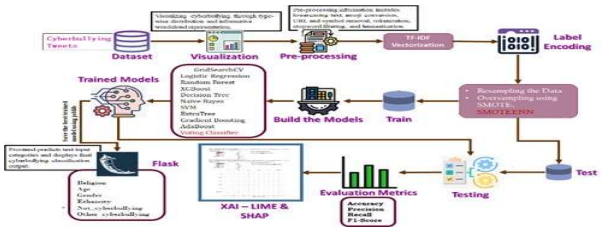


Fig. 1. System Architecture

Figure 1 shows a stepwise data science process of cyberbullying detection. The process starts by having a Cyberbullying dataset, which goes through a detailed Text Preprocessing. These include steps like converting emojis to text, removing stop words, special characters, and URLs to normalize the data to lower case, and clean it up. The resulting text is in turn treated with a model whereby TF-IDF vectorization and label encoding are used to convert the data into a machine learning format. The next phase is Classification, whereby, the Bullying Classification is carried out by SVM, RF, and XGB under the Bullying Classification model. The final evaluation of the results is on the basis of Accuracy and is presented in the form of Word Clouds and a Confusion Matrix.

A) Dataset Collection

In this paper, the data is to be analyzed with the help of Cyberbullying Tweets dataset provided by Kaggle that contains 47,692 tagged records. Every record includes a tweet message and its corresponding cyberbullying category, which consists of six distinct categories: religion, age, gender, ethnicity, other cyberbullying and non-cyberbullying. The dataset is fairly distributed across its categories, and each category has approximately 7,800 samples, thus providing the confidence of good representation of the different types of bullying. This diversity makes it easy to train and evaluate machine learning models effectively, including language, tone, and target demographic diversities. Informal language, hash tags and mentions that tend to occur in social media writing are essential to pre-process the data. Data is a reliable reference point in exploring approaches to cyberbullying based on sentiment detection [23].

	tweet_text	cyberbullying_type
0	In other words #katandandre, your food was cra...	not_cyberbullying
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying
3	@Jason_Gio meh. :P thanks for the heads up, b...	not_cyberbullying
4	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying

Fig.2 Dataset Collection

B) Pre-Processing

The pre-processing stage is a critical step in the development of social media content to go through machine learning, as it is aimed at cleaning and normalizing the information. It involves transforming text into lower-case, changing emojis into textual conversions, removing URLs and special characters and breaking down phrases into single words. Stopwords are removed and lemmatization is used to reduce words to basic forms hence ensuring consistency and improving model performance.

i) Data Processing: Data processing involves a systematic process of preparing the raw twitter text to be analyzed using a machine learning. The tweets will first be lowered to standardize, and the emojis will be translated into the written description to maintain the emotional value. URLs, mentions, hashtags, and special characters are removed in order to minimize noise which can affect model performance. In tokenization, the text is broken into individual words or tokens, and then the stopwords are filtered out of the text and are not involved in categorization. Lemmatization reduces words to the simplest forms and enables the model to be treated as the similar forms of words. Label encoding is used to encode the types of cyberbullying found in the categories used in classification into a number, therefore, improving the interoperability with machine learning algorithms. These actions ensure the structured, unambiguous, and informative data on the input.

ii) Data Visualization: Data visualization explains the distribution and properties of cyberbullying content in the data. By tracing the frequency of each cyberbullying group, we can see the balance or imbalance between classes, which we need to create and evaluate models. Word clouds give a graphic representation of the most frequently used words in each category, highlighting the common words associated with specific forms of bullying, like religion, gender or age. Bar charts and distribution plots will help to quickly compare the categories and may reveal some trends or anomalies in the data. The visualization is useful in the interpretation of sentiment patterns and textual characteristics, therefore, supporting informed decision-making in the feature extraction and preprocessing phases. These preliminary research works are important in understanding the organization of the dataset.

iii) TF-IDF Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is a method to convert textual data into numeric items of features to be used in machine learning models. Term frequency indicates how a word is used in one document, and inverse document frequency removes the weight of the most common words used by all documents, now bringing into focus those words that are more unique as a means of categorization. The application of TF-IDF to the text of the twitter data transforms each tweet into a vector, which indicates the importance of each word relative to the data, and thus, it is easier to use the algorithms to identify patterns associated with the types of cyberbullying. This method maintains contextual importance and reduces the power of too frequently-used but less enlightening words. TF-IDF works well with very short, informal texts like tweets because it balances the number of occurrences of words and distinctive words, therefore, improving the accuracy of the classifier.

iv) Data Sampling: Data sampling is an action taken to counteract the class imbalance that is common to cyberbullying datasets (such as other cyberbullying, religiosity, etc.); these classes might be overrepresented. The oversampling technique (n.d.b.) like SMOTE (Synthetic Minority Oversampling Technique) produces synthetic samples of a minority group by interpolating the available samples. This balances the sample, preventing machine learning models to favor the majority classes and improve the generalization in all types of cyberbullying. Suitable resampling will ensure that the classifiers obtain meaningful

patterns within each category, rather than focusing on the most significant classes. Balanced sampling, preprocessed and encoded data, and integration with vectors offer an opportunity to achieve a complete representation of all types of cyberbullying, and to make accurate and fair predictions of different situations related to cyberbullying on the Internet.

C) Train & Test

The sample is divided into training and test samples in a ratio of 80: 20 to use as much data as possible to train the model and still maintain a representative portion to evaluate the performance of the model. The training set will include 80 percent of the samples in each category of cyberbullying, providing sufficient samples on which the machine can identify patterns, language nuances, and any other means of expressing sentiments that, in turn, are harmful. The remaining 20% is used as the testing set and is not presented during training allowing a fair assessment of the effectiveness of the model. Stratified splitting is applied to maintain the proportional representation of all categories of both sets, therefore, avoiding the impact of class imbalance on evaluation. Such arrangement will ensure effective training and validation of effective cyberbullying detection.

D) Algorithms

Logistic Regression: The Logistic Regression is based on the TF-IDF characteristics used to classify cyber-bullying by predicting it. The simplicity of the model ensures can interpretability, which also highlights the importance of the words correlated with religion, gender, or age-related bullying, as well as it enables quick training and evaluation on large volumes of social media data.

$$\hat{y}_i = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} \qquad (1)$$

Random Forest: Random Forest is built to determine complex cyberbullying patterns using TF-IDF vectors by constructing multiple decision trees. It reduces overfitting, gives more stable predictions, and delivers feature importance scores, and offers strong multi-class detection in the high-noise and high-dimensional social media datasets with interpretable classification scores.

$$Gini = 1 - \sum_{i=1}^{C} (Pi)^2 \qquad (2)$$

XGBoost: XGBoost is an incremental decision tree booster that is regularized, and is therefore capable of handling large TF-IDF feature spaces. It emphasizes misclassified events to enhance accuracy, diagnostic linguistic indicators as well as generalizing throughout underrepresented forms of cyberbullying, thereby giving quick convergence, scalability, and reliable execution on social media data.

Decision Tree: The Decision Tree is a hierarchical splitting feature that uses hierarchical feature splitting to identify cyberbullying and offers understanding frameworks that highlight significant abusive language. It can work across multiple classes to classify by gender, religion, or age-related harassments, providing quick and non-linear predictions without high computational costs that are suitable to set up baseline sentiment analysis of text.

$$I\,(i) = 1 - \sum_{i=1}^{k} p_i^2 \qquad (3)$$

Naive Bayes: Naive Bayes applies both the Bayes theorem and the TF-IDF features to successfully classify cyberbullying, especially in short posts of social media. It is a high-dimensional data processing model that provides high-speed, probabilistic predictions, which reveal word contributions, and maintains scalability, robustness, and stable baseline classification.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad (4)$$

Support Vector Machine (SVM): SVMs make use of hyperplanes to help differentiate between dissimilar groups within TF-IDF feature space to enable maximum marginal classification. Non-linear text relationships are implemented in kernel functions and enhance accuracy in revealing subtle types of abuse. Generalization capabilities and its resilience ensure reliable identification on the noisy social media data.

$$minimize\ \frac{1}{2}\ ||W||^2 + C\sum_{i=1}^{n} \xi_i \qquad (5)$$

Extra Trees: Extra Trees builds randomized decision trees with different splits and thresholds, which reduces the variance and overfitting in the detection of cyberbullying. It is very efficient in processing TF-IDF, which provides some information on the relevance of features and the predictive accuracy is also good when using it in different unstructured social media datasets.

Gradient Boosting: Gradient Boosting progressively upgrades the weak learners to minimize errors, which are useful to detect subtle cases of abuse language in TF-IDF features. The approach of boosting focuses on inaccurate cases, thus encouraging minority class discovery, correcting the data lopses, and improving predictive accuracy in the process of multi-class cyberbullying identification missions.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad (6)$$

AdaBoost: AdaBoost successively combines weak classifiers, and it focuses on difficult misclassified classes of cyberbullying by assigning dynamic weights. It enhances recognition of subtle harassment within high-dimensional TF-IDF features, which ensures fair and accurate inferences on social media datasets, and lowers bias and increases robustness.

$$H(x) = sign\left(\sum_{t=1}^{T} \propto_t h_t(x)\right) \qquad (7)$$

Voting Classifier: The Voting Classifier is an amalgamation of the predictions of a large number of models including MLP, Bagging and Logistic Regression. The application of soft voting to combine strengths enhances prediction accuracy in classifying cyberbullying, balances problematic categories, mitigates biases, and offers stably predictive and interpretable results using diverse social media datasets.

IV. Experimental Results

Accuracy: Accuracy of a test means its ability to help differentiate between patient and healthy cases. To determine the validity of a test, the ratio of true negative and positive of all cases assessed is to be calculated. This can be mathematically stated as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (8)$$

Precision: Precision determines the proportion of correctly identified cases of those identified as positive. As a result, the precision formula can be stated as:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (9)$$

Recall: Recall is a machine learning measurement that determines the ability of a model to identify all relevant examples of a given class. It is the ratio of correctly identified affirmative observations of the total real affirmative observations, which provides information regarding the effectiveness of a model in detecting the events of a certain classification.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

F1-Score: F1 score is a measure used to test the accuracy of a machine learning model. It is a combination of the accuracy and recall score of a model. The metric of accuracy is the number of correct predictions that a model has produced on the whole dataset.

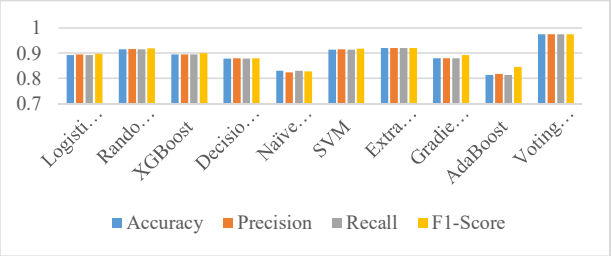
$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 \quad (11)$$

Table.1 Performance Evaluation

ML Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.892	0.894	0.892	0.897
Random Forest	0.915	0.916	0.915	0.918
XGBoost	0.894	0.895	0.894	0.899
Decision Tree	0.878	0.879	0.878	0.880
Naïve Bayes	0.830	0.824	0.830	0.827
SVM	0.914	0.915	0.914	0.917
Extra Trees	0.919	0.919	0.919	0.921
Gradient Boosting	0.879	0.879	0.879	0.892
AdaBoost	0.813	0.817	0.813	0.846
Voting Classifier	0.974	0.974	0.974	0.974

Table 1 shows the performance of the models with Voting Classifier getting the highest accuracy, Precision, recall and F1-score.

Fig.3 Comparison Graph



In figure 3, a bar chart is used to compare the performance of different machine learning models across four metrics; Accuracy, Precision, Recall, and F1-Score.

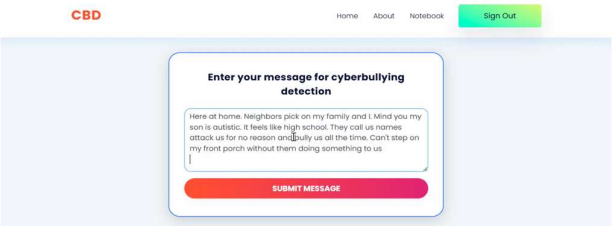


Fig.4 Upload Input Text Message

Figure 4 shows a user interface of a cyberbullying detecting system, which has a text input field.

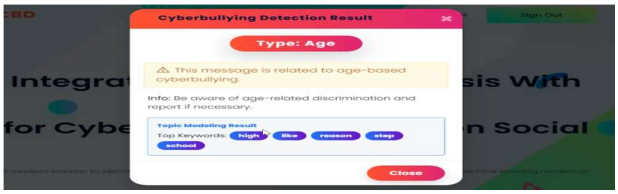


Fig.5 Predict Result

As shown in Figure 5, the results of a cyberbullying detection system show that a message received is considered as age-based cyberbullying.

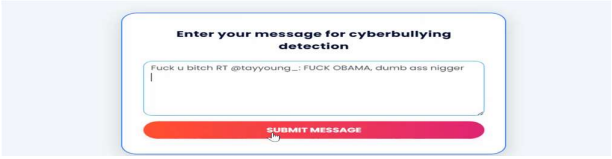


Fig.6 Upload another Input Text

In fig. 6, one can see a user typing a message with a hashtag and a potentially offensive word into a cyberbullying detection program.

Fig.7 Final Outcome

The result of a cyberbullying detection system is shown in figure 7, which proves that the message provided is not cyberbullying.

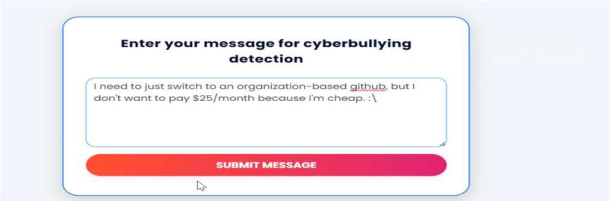


Fig.8 Upload Input Text Message

Figure 8 also shows how a user can input a message that is not related to cyberbullying into the detection system.

Fig.9 Result

The result of a message recognized as being part of Other Cyberbullying is shown in Figure 9, where the message is considered to be a victim of cyberbullying.

V. Conclusion

The technology has the ability to effectively integrate sentiment analysis with advanced machine learning techniques in detecting cyberbullying in social media. The Kaggle Cyberbullying Tweets dataset has gone through normalization, noise removal, tokenization, lemmatization, vaderSentiment analysis, and label encoding preprocessing steps to generate clean data to be used in modeling. TF-IDF vectorization internalized semantic features, but both SMOTE and SMOTEENN balanced the classes to enhance the performance in minor groups. Different algorithms, including LR, RF, XGBoost, SVM, and ensemble methods, have been optimized on the grid search by using the grid search control variable and evaluated by measuring the accuracy, precision, recall, F1-score, and confusion matrices, with good results of detection and a low level of false prediction. An ensemble of Voting Classifier (MLP, Bagging, and Logistic Regression) was outstanding as it scored 97.4 on all the metrics. Interpretability was done through LIME and SHAP, which included determining the important factors in prediction. Finally, flask implementation with topic modeling allowed real-time forecasting with a confidence score, which is a reliable, transparent, and scalable model of efficient cyberbullying detection in social media platforms.

Deep learning models such as LSTM, BiLSTM, or Transformers can be added to the system, which allows one to capture the surrounding semantics and improve the detection accuracy. This can be achieved by introducing multilingual support to detect cyber bullying in various languages and different regional accents. Preemptive response may also entail real-time monitoring of the social media platforms. Intense natural language processing systems are better able to determine sarcasm, slang, and new patterns of abuse. Mobile communication and integration with chat service can enable instant feedback and inform the consumers. Additionally, the combination of sentiment analysis and user behavioral patterns might be used to identify potential threats before they increase. Visual and multimedia content analysis expansion can lead to detection comprehensiveness and strength.

References

[1] Alhejaili, R. (2025). Machine learning approaches for sentiment analysis on social media. In AI-Driven: Social Media Analytics and Cybersecurity (pp. 21-43). Cham: Springer Nature Switzerland.

[2] Pranith, B. Y. K. G. (2025). Machine Learning Solutions for Cyberbullying Detection and Prevention on Social Media.

[3] Akter, F., Jahangir, M. U. F., Chowdhury, R. R., & Rabbi, M. F. Cyberbullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches. International Journal of Computer Applications, 975, 8887.

[4] Mubeen, M., Muskan, A., Akram, A., Rashid, J., Alshalali, T. A. N., & Sarwar, N. (2025). Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method. International Journal of Computational Intelligence Systems, 18(1), 1-24.

[5] Syafiq, A. M., Ab Razak, M. F., Abidin, A. F. Z., Mohamad, S., & Kamaruddin, N. K. (2025). Social network approach for cyberbullying detection using machine learning. Journal of Governance and Integrity, 8(1), 874-880.

[6] Hu, Y., Clancy, E. M., & Klettke, B. (2023). Understanding the vicious cycle: Relationships between nonconsensual sexting behaviours and cyberbullying perpetration. Sexes, 4(1), 155–166. doi: 10.3390/sexes4010013.

[7] Aliyu, S., Niksirat, K. S., Huguenin, K., & Cherubini, M. (2023). On the role and form of personal information disclosure in cyberbullying incidents. Proc. Privacy Enhancing Technol., 2023(4), 468–483. doi: 10.56553/popets-2023-0120.

[8] Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. Online Social Netw. Media, 36, 100250. doi: 10.1016/j.osnem.2023.100250.

[9] Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., & Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. Social Netw. Anal. Mining, 12(1), 99. doi: 10.1007/s13278-022-00934-4.

[10] Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., & Hasan, K. (2021). Am I being bullied on social media? An ensemble approach to categorize cyberbullying. Proc. IEEE Int. Conf. Big Data (Big Data), 2442–2453. doi: 10.1109/BIGDATA52589.2021.9671594.

[11] Vogels, E. A. (2022). Teens and cyberbullying 2022. Pew Research Center. Retrieved from https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/.

[12] Cook, S. (2024). Cyberbullying statistics and facts for 2024. Comparitech. Retrieved from https://www.comparitech.com/internet-providers/cyberbullying-statistics/.

[13] Collantes, L. H., Martafian, Y., Khofifah, S. N., Fajarwati, T. K., Lassela, N. T., & Khairunnisa, M. (2020). The impact of cyberbullying on mental health of the victims. Proc. 4th Int. Conf. Vocational Educ. Training (ICOVET), 30–35. doi: 10.1109/ICOVET50258.2020.9230008.

[14] Unnava, S., & Parasana, S. R. (2024). A study of cyberbullying detection and classification techniques: A machine learning approach. Eng. Technol. Appl. Sci. Res., 14(4), 15607–15613. doi: 10.48084/etasr.7621.

[15] Endsuy, R. (2021). Sentiment analysis between VADER and EDA for the U.S. presidential election 2020 on Twitter datasets. J. Appl. Data Sci., 2(1), 8–18. doi: 10.47738/jads.v2i1.17.

[16] Atoum, J. O. (2020). Cyberbullying detection through sentiment analysis. Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), 292–297. doi: 10.1109/CSCI51800.2020.00056.

[17] Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. Online Social Netw. Media, 36, 100250. doi: 10.1016/j.osnem.2023.100250.

[18] Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11, 55533–55560. doi: 10.1109/ACCESS.2023.3275130.

[19] Sharif, W., Ashraf, M., Shifa, A., Shahid, M., Mumtaz, Q., Ijaz, U., Anwar, M., & Ikram, M. (2024). Sentence level sentiment analysis of cyber trolling tweets using machine learning technique. J. Comput. Biomed. Informat., 6(2), 1–12. doi: 10.56979/602/2024.

[20] Almutiry, S., Fattah, M. A., & Almunawarah, S. A.-A. (2021). Arabic cyberbullying detection using Arabic sentiment analysis. Egyptian J. Lang. Eng., 8(1), 39–50. doi: 10.21608/ejle.2021.50240.1017.

[21] Halim, L. R., & Suryadibrata, A. (2021). Cyberbullying sentiment analysis with word2Vec and one-against-all support vector machine. Int. J. New Media Technol., 8(1), 57–64. doi: 10.31937/ijnmt.v8i1.2047.

[22] Atoum, J. O. (2021). Cyberbullying detection neural networks using sentiment analysis. Proc. Int. Conf. Comput. Sci. Comput. Intell., 158–164. doi: 10.1109/CSCI54926.2021.00098.

[23] Perdana, R. A., Widodo, C. E., & Santoso, R. (2024). Sentiment analysis of Naïve Bayes, decision tree, and K-nearest neighbor (K-NN) algorithms for cyberbullying comments on Instagram accounts. Proc. 11th Int. Conf. Inf. Technol., Comput., Electr. Eng. (ICITACEE), 253–258. doi: 10.1109/icitacee62763.2024.10762775.

[24] Ahmad, I. S., Darmawan, M. F., & Talib, C. A. (2023). Cyberbullying awareness through sentiment analysis based on Twitter. Stud. Comput. Intell., 1080, 195–211. doi: 10.1007/978-3-031-21199-7_14.

[25] Wang, J., Fu, K., & Lu, C.-T. (2020). SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection. Proc. IEEE Int. Conf. Big Data (Big Data), 1699–1708. doi: 10.1109/BIGDATA50022.2020.9378065.

[26] Maskat, R., Zainal, M. F., Ismail, N., Ardi, N., Ahmad, A., & Daud, N. (2020). Automatic labelling of Malay cyberbullying Twitter corpus using combinations of sentiment, emotion and toxicity polarities. Proc. 3rd Int. Conf. Algorithms, Comput. Artif. Intell., 1–6. doi: 10.1145/3446132.3446412.

[27] Almuayqil, S. N., Humayun, M., Jhanjhi, N. Z., Almufareh, M. F., & Javed, D. (2022). Framework for improved sentiment analysis via random minority oversampling for user tweet review classification. Electronics, 11(19), 3058. doi: 10.3390/electronics11193058.

[28] Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artif. Intell., 61, 863–905. doi: 10.1613/jair.1.11192.

[29] Wu, F., Gao, B., Pan, X., Su, Z., Ji, Y., Liu, S., & Liu, Z. (2023). FACapsnet: A fusion capsule network with congruent attention for cyberbullying detection. Neurocomputing, 542, 126253. doi: 10.1016/j.neucom.2023.126253.

[30] Mahmud, M. I., Mamun, M., & Abdelgawad, A. (2022). A deep analysis of textual features based cyberbullying detection using machine learning. Proc. IEEE Global Conf. Artif. Intell. Internet Things (GCAIoT), 166–170. doi: 10.1109/GCAIOT57150.2022.10019058.