

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

Madasu Sai Nikhita¹, Ravi Kumar Tirandasu^{2*}

¹Master's Student, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India – 522502

Email: sainikhitamadasu@gmail.com

²Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India – 522302

Email: ravi.tirandasu@gmail.com, travikumar@kluniversity.in

Phone: 7507976799

ORCID: 0000-0002-8233-7816

***Corresponding Author:** Ravi Kumar Tirandasu, Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India – 522302

Email: travikumar@kluniversity.in

a) ABSTRACT

Cyberbullying using social networking sites is an extensive network issue that can have significant psychological and emotional impacts, thus, reliable detection systems play an important role. With the power of cutting-edge ML, sentiment analysis can help identify harmful interactions on a scale that is scalable, thereby creating safer online environments. This work uses the Cyberbullying Tweets dataset from Kaggle that contains labeled cyberbullying events of various categories. Pre-processing consists of normalisation, removing noise (like urls and html tags), tokenisation, WordNetLemmatizer, vaderSentiment analysis and label encoding. TF-IDF vectorisation and SMOTE oversampling methods are used for feature extraction and class imbalance, respectively. Distributions plots, word clouds etc. are used to help us appreciate the trends in bullying, and the frequency of different types of bullying. We have applied several classifiers in the suggested framework including LR, RF, XGBoost, DT, NB, SVM, ET, GB and AdaBoost with hyperparameter tuning with GridSearchCV. Evaluation based on: accuracy, precision, recall, F1-score and efficient tracking of false positives and false negatives is done using confusion matrices. The further improvements are the SMOTEENN to balance the data and a VC ensemble (MLP and Bagging combined with LR) for robust classification. Methods like LIME and SHAP, used in XAI, emphasize the most significant factors, making the AI's decisions interpretable. The Flask deployment offers real-time predictions and confidence scores with subject modelling for usability and transparency. The results show that the proposed VC algorithm outperforms all the baseline classifiers with an accuracy of 97.4% in all the evaluation parameters, and the system can be considered reliable, scalable and feasible for the detection of cyberbullying.

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

How to cite this article: Sai Nikhita M, Tirandasu RK., Integrating Sentiment Analysis with Machine Learning for Cyberbullying Detection on Social Media. Int J Drug Deliv Technol. 2026;16(86s): 1937-1944; DOI: 10.25258/ijddt.16.86s.175

Source of support: Nil.

Conflict of interest: None

II. INTRODUCTION

Social media has revolutionized the way people communicate with one another, enabling them to connect, share thoughts and participate in conversations in real time. With these benefits, however, negative behaviors have sprung up, particularly cyberbullying. Cyberbullying is intentionally causing harm to someone with a computer or other technological devices. This includes the sending of bad language, threats, false rumours, false identity or the sharing of private information. These acts have disastrous effects on people, particularly teens, often leaving them feeling anxious, hopeless and lonely, and in some cases, suicidal [6, 7]. As

social media sites such as Twitter, Facebook and Instagram continue to rise in popularity, the demand for developing

advanced detection methods to safeguard social media users from psychological damage grows [1,2].

There are many manual approaches such as manual moderation and flagging, but these cannot cope with the volume of content created daily [8]. Researchers have developed computational methods using sentiment analysis and ML techniques for automatic identification of dangerous patterns that address the above mentioned problem [3, 4].

Sentiment analysis is an additional layer of consciousness of the text's emotional polarity, and therefore more contextual classification. Complementing supervised algorithms with language and semantic data, there is a possibility to detect abusive interactions on a scalable and efficient manner [5]. Moreover, this is the first time the transformer design approach and ensemble learning have been introduced to achieve attractive performance improvement in the accuracy of the cyberbullying detection system [9, 10]. These strategies perform better than conventional models and robustness in imbalanced datasets, with the help of context embedding and ensemble based approach.

When detection systems are focused on English language text—the dominant language of online discourse [1]—the applicability and effectiveness of the detection system to several platforms is enhanced. Explainability provides transparency of the identification of influential features in predictions [4] beyond classification. The aim is to create an intelligent, reliable and interpretable system which can automatically detect detrimental interactions, include sentiment signals and offer real-time feedback. This helps guardians, educators and stakeholders make quick decisions resulting in safer and healthier digital environments [2, 5].

III. RELATED WORK

Cyberbullying is one of the most pressing issues facing our society in today's digital age, especially given the prevalence of social media amongst teenagers and young people. According to surveys [11, 12], online harassment is getting worse and many victims are experiencing severe psychological and emotional harm. The victims of cyberbullying are shown to be anxious, depressed and have low self-esteem which leads to long term detrimental implications on mental health [13]. This has spurred increasing research into computer-based methods for detecting, classifying and blocking harmful activities online.

Some literature shows that ML and sentiment analysis techniques are being employed to identify cyberbullying automatically, especially with an increase in the number of instances. The conventional keyword-based methods and rule-based systems are insufficient today because of the different verbal utterances of the perpetrators [14]. Rather, supervised and unsupervised learning approaches have been used to model intricate patterns of damaging communication. For instance, a paper comparing the various detection and classification approaches shows that ML algorithms, like logistic regression, support vector machines and random forests, can be effectively applied to structured datasets [14]. The performance of such systems is, however, highly dependent on the pre-processing and feature extraction that would be required, hence the need for combining advanced sentiment analysis methods.

Sentiment analysis is an important tool in detecting cyberbullying since it is able to capture the emotional tones in the text. VADER is a commonly utilised sentiment analyser

which has been demonstrated to be successful in detecting polarity of short social media posts [15]. The analysis of the negative emotional tone of the text has been shown to be significantly associated with abusive or bullying behaviour in studies which employ sentiment features [16]. The classification models detection rate is better if the sentiment polarity is added to the model than if the textual features are added to the model only, according to Atoum [16]. Moreover, hybrid approaches with both TF-IDF vectorisation and sentiment features have proven to be more accurate and robust in dynamic online environments.

The involvement of session analysis in the cyberbullying incident detection is another important factor. Yi and Zubiaga [17] argue that their analysis of the posts does not take into account the context in which posts are used or conversed. Session-based techniques can be leveraged to allow detection frameworks to capture interactions over multiple exchanges that can help better identify safe and unsafe communications. This temporal modelling is significant in social networks, because it's the context that determines the impact and nature of a message.

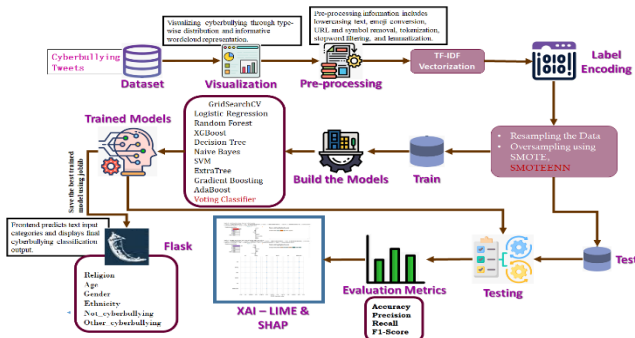
DL and transfer learning have made significant advances in the past few years, greatly enhancing the accuracy of cyberbullying detection systems. Teng and Varathan [18] explored and studied traditional ML algorithms for transferring learning methods and found that transformer models such as BERT significantly outperform traditional methods, particularly in handling language nuances. They rely on pre-trained contextual embeddings and are able to generalise between datasets and capture nuances in the language. But there are still some challenges, such as data asymmetry and expensive computing costs, and hybrid solutions are needed that combine classic classifiers with current DL classifiers.

Cyberbullying has characteristics in common with other negative behaviours on the internet, including internet trolling and hate speech. The sentence level sentiment analysis of cyber trolling tweets was carried out by Sharif et.al. [19] and demonstrated the effectiveness of the sentiment features to boost the classification accuracy by using ML approaches. This shows that sentiment-based detection techniques can be more widely applied to the online more moderate environment. The same has been found in multilingual contexts. Language-specific models are important to understand the cultural and linguistic nuances, as was done by Almutiry et al. [20] who created an Arabic cyberbullying detection system that leverages Arabic sentiment analysis. They demonstrate that sentiment analysis systems work in English, and in other languages, opening the way to a worldwide application.

IV. MATERIALS AND METHODS

The suggested approach aims to enhance the detection of cyberbullying in social media, incorporating sentiment analysis and advanced ML methods. Preprocessing

techniques like normalisation, tokenisation, noise removal, and label encoding are applied to ensure structured inputs using the Kaggle Cyberbullying Tweets dataset. It employs TF-IDF vectorisation to capture the context word frequency patterns, and SMOTE to address class imbalance and boost the minority class [27], [28]. Various classifiers like Logistic Regression, Random Forest, XGBoost, Decision Tree, Naïve Bayes, Support Vector Machine, Gradient Boost and AdaBoost are explored and their hyperparameters are tuned using GridSearchCV. For increasing the prediction confidence, SMOTEENN is used to make a noise reduction and balanced sampling, and a Voting Classifier ensemble is used to combine several models for robust decision making. Explainable AI techniques like LIME and SHAP offer interpretability as well, by pinpointing the most significant features and furnishing transparency in the classification outcomes. It is possible to detect hazardous interactions in an accurate, interpretable and scalable way on the platform.



“Fig. 1. System Architecture”

Fig. 1 depicts a whole data science pipeline for cyberbullying detection. The process begins with a Cyberbullying dataset which is subjected to a careful Text Preprocessing. This means changing emoji into text, translating emoji to lower case and removing stop words, special characters, urls to clean the data. This generated text is then fed into a Model which uses TF-IDF Vectorisation and Label Encoding to transform the text into a machine-learning suitable format. The next stage is the classification where the model does the Bullying Classification by using several classifiers such as SVM, RF and XGB. Lastly, the outcomes are analysed by key indicators such as Accuracy and visualised on Word Clouds and a Confusion Matrix.

A) Dataset Collection

The dataset used in this work is the Cyberbullying Tweets dataset from Kaggle, which has 47,692 labelled items. The records are made up of the text of a tweet and its cyberbullying classification. There are six such classes: religion, age, gender, ethnicity, other cyberbullying and non cyberbullying. The distribution of samples is fairly even with each category having approximately 7,800, thus providing a good representation of various types of bullying. This diversity creates a good training and evaluation set for ML algorithms and can help to capture variance in language, tone, and target groups. Preprocessing is important to deal with

informal language, hashtags and mentions commonly seen in social media writing. The dataset is a good baseline for studying the algorithms of sentiment-based cyberbullying 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 of social media text is a crucial step that enables it to be utilized in ML. It's a data cleansing and normalisation process. It contains text to lower case, emoji to text, remove urls, remove special characters and tokenisation of sentences to words. The stop words are removed, and lemmatisation is applied to reduce words to their roots, which is consistent with the uniformity of the word model and improves the performance of the model.

i) *Data Processing*: Data processing refers to the systematic transformation of the raw content from the twitter feed to make them suitable for analysis by ML. First, we lowercase each tweet in order to have a uniform text. Next, we translate the emojis into their written explanations, so that we can preserve the emotional content . URLS, mentions, hashtags and weird characters are stripped out to reduce noise that may affect model accuracy. The tokenisation process breaks the text into individual words/tokens and the stopwords are filtered out as they do not add anything to the categorisation. Lemmatisation is a process whereby words are converted to their base or root form, enabling the model to recognise other forms of the same word as being equivalent. The data kinds are converted to numerical values that can be used in the ML algorithms via label encoding. The processes ensure that the input data is organized, tidy and informative.

ii) *Data Visualization*: Visualising data provides the understanding of property and distribution of cyberbullying material in a dataset. We can simply show the frequency of each cyberbullying category to view the balance or imbalance in classes, which is very important in constructing models and evaluating the models. The word clouds show the most common words used in each category . There are common words for certain sorts of bullying , like bullying due to religion , gender , or age . The use of distribution plots and bar charts can enable you to compare categories rapidly, and also determine likely trends or irregularities in the data. Visualisation is also useful for gaining insights into the sentiment patterns of the text and the textual features that can be useful in making informed judgments when processing and extracting features from the text. It is important to conduct exploratory investigations in order to understand the structure of this dataset.

iii) *TF-IDF Vectorization*: TF-IDF vectorisation converts text information into a numerical representation of the features that can be utilized by the ML algorithm. Term frequency measures the frequency of a word in a document, and inverse document frequency down-weights common words that appear in all papers, giving more importance to words that are more discriminative for classification. Next, TF-IDF is used to process the text of the tweets and each tweet is converted into a vector, which indicates the relevance of each phrase about the data set. This way, algorithms can find the patterns related to the cyberbullying categories. This strategy helps to keep the context meaningful and reduces the frequency of very common, but less meaningful words. TF-IDF is good for short, informal data such as tweets and balances between common and rare words, so that the classifier will be more accurate.

iv) *Data Sampling*: Data sampling is the issue of class imbalance that often occurs in cyberbullying data sets, such as under-representation or over-representation of classes like "other cyberbullying" or "religion". The SMOTE and other oversampling methods synthesize minority class samples from the existing minority class samples by interpolation. This even distributes data, preventing algorithms from learning only the majority classes and enhancing the generalisation of the learning to all types of cyberbullying. Proper sampling will allow classifiers to learn important patterns for each class, rather than just being swayed by the majority classes. This approach incorporates balanced sampling and preprocessed (encoded and vectorised) data to yield a comprehensive representation of all forms of cyberbullying and generate accurate and equitable predictions across diverse cyberbullying categories.

C) Train & Test

The data is split into a training and a testing subset with an 80:20 ratio, the biggest part of the data is used to train the model, while the representative part is used to evaluate it. 80% of the samples in the training set are from each cyberbullying category. In this way the model can learn patterns, grammatical nuances and sentiment signals that can be used to detect harmful content, because there are enough examples in the training set. The remaining 20% is the testing set, which is not observed during the training process and is used to make an independent assessment of the performance of the model. To avoid class imbalance bias, stratified splitting is used to ensure that the proportions of all categories is equal across both sets. This setup ensures robust training and validation of the system for effective cyberbullying detection.

D) Algorithms

Logistic Regression: It is a ML model used to estimate probabilities to classify the types of cyberbullying using TF-IDF information. Its simplicity ensures interpretability, highlights the importance of words related to bullying for

religion, gender or age and enables rapid training and evaluation with large-scale social media datasets.

$$y_i = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (1)$$

Random Forest: Several DT models are created by RF to learn complex behavioural patterns from TF-IDF vectors. It reduces over-fitting, increases the stability of prediction and provides weight on features which ensures good multi-class detection in noisy high-dimensional social media data with interpretable classification results.

$$Gini = 1 - \sum_{i=1}^C (P_i)^2 \quad (2)$$

XGBoost: A sequential approach to building boosted decision tree models with regularisation, suitable for large TF-IDF feature sets. It aims to classify misclassified cases to improve accuracy, learns nuances of the language and generalizes across under-represented groups of cyberbullying to ensure a rapid convergence, scalability, and reliable performance on social media datasets.

Decision Tree: DT uses hierarchical feature splits for cyberbullying detection, which provides interpretable paths highlighting significant phrases of cyberbullying. It enables quick, non-linear predictions of multi-class categories of gender, religion or age-related harassment with little processing for baseline sentiment-driven text analysis.

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

Naive Bayes: NB is a classification algorithm based on Bayes' theorem and the features of TF-IDF that is able to detect cyberbullying efficiently, particularly at short messages on social media. It handles high-dimensional data, giving quick, probabilistic predictions showing word contributions, while yet being scalable, robust and giving constant baseline classification results.

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

Support Vector Machine (SVM): SVM uses hyperplane to divide the TF-IDF feature space in the optimal way to achieve the maximum margin classification. These non-linear linkages in the text are represented by kernel functions, which produce greater accuracy for subtle abuse patterns. It is also reliable, generalisable and can be used to identify reliably in noisy social media data.

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

Extra Trees: ET is a technique which generates multiple models of random decision trees with various splits and thresholds, reducing the variance and overfitting in cyberbullying detection. It is efficient for the computation of TF-IDF vectors, allows insights into feature relevance and yields very high prediction accuracy on heterogeneous and unstructured social media data.

Gradient Boosting: GB is a method that gradually tweaks weak learners to minimize errors, and that's how it detects subtle abusive language in the TF-IDF features. Its boosting approach builds on misclassified instances to enhance the detection of minority class, balance the imbalanced data and enhance the predictive accuracy for multi-class cyberbullying detection tasks.

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

AdaBoost: AdaBoost iteratively combines weak classifiers by providing adaptive weights to hard misclassified cyberbullying categories. It features a balanced and accurate prediction performance on social media data sets, robust resistance and reduced bias in the identification of small-scale harassment in high-dimensional TF-IDF features.

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right) \quad (7)$$

Voting Classifier: The VC takes the predictions of many models such as MLP, Bagging, Logistic Regression, etc. It enhances the accuracy of cyberbullying categorisation, equalises the hard categories, mitigates biases and provides reliable and interpretable predictions on various social media datasets by capitalising on the strengths of soft voting.

V. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is the ability to distinguish between the patient and healthy cases appropriately. The accuracy of a test should be measured as the proportion of true positive and true negative results of all tests performed. This mathematically can be written as:

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

Precision: Precision is a ratio of the number of true positive instances/samples to the number of occurrences/samples that are labeled as positive. The formula for precision then is:

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

Recall: Recall is a metric used for ML that measures the ability of a model to identify all the examples of a class. It is the ratio of the number of the positives that are correctly predicted as positives and provides insight into the ability of the model to capture positive instances of a specific class.

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

F1-Score: F1 score is a measure to evaluate the correctness of a ML model. It combines the precision and recall scores of the models. Accuracy measure is the number of samples that a model has predicted correctly in the whole data set.

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

“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 best accuracy, precision, recall and F1-score results for the Voting Classifier model.

Fig.3 Graph of comparison
“Fig.3 Comparison Graph”

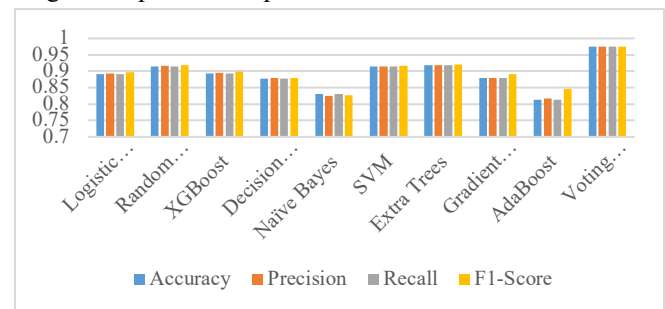
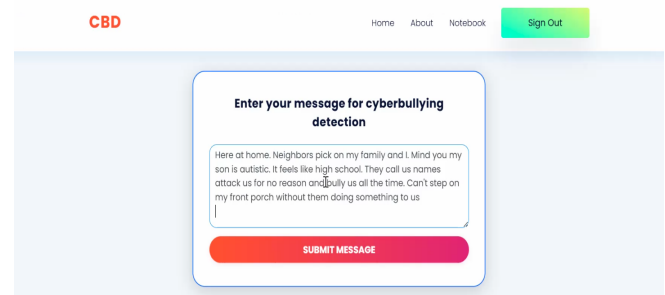
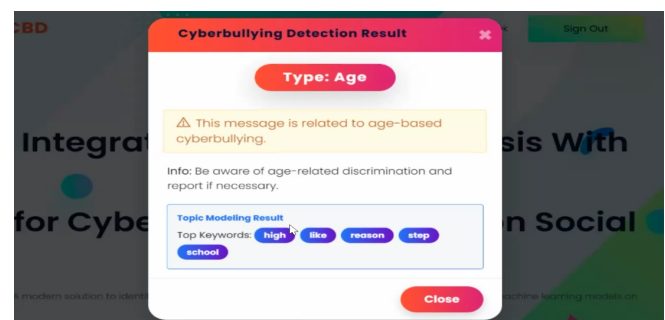


Figure 3. The bar chart of evaluating the performance of several ML algorithms on four metrics: Accuracy, Precision, Recall and F1-Score.



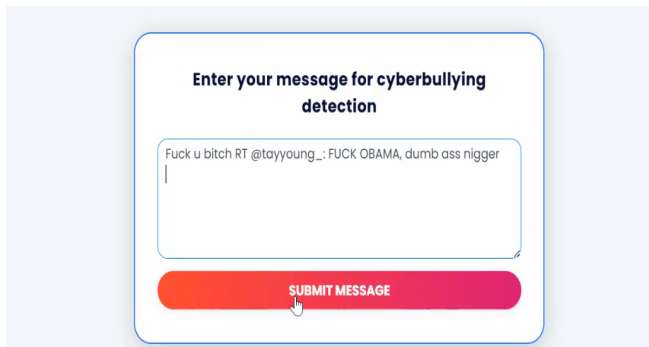
“Fig.4 Upload Input Text Message”

In the user interface from a cyberbullying detection system illustrated in Figure 4, a text entry field is provided for users to enter text.



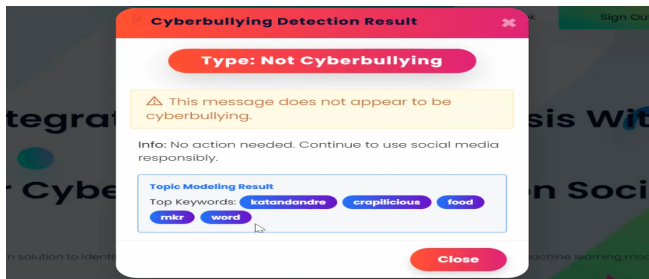
“Fig.5 Predict Result”

Figure 5 shows the results of a cyberbullying detection system that successfully detects a communication as an age-based cyber-bullying incident.



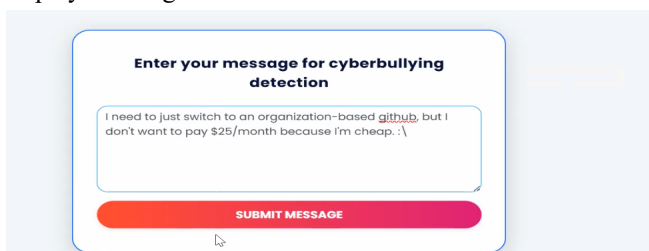
“Fig.6 Upload another Input Text”

Figure 6: User enters an hashtag and an inappropriate word into a cyberbullying detection system.



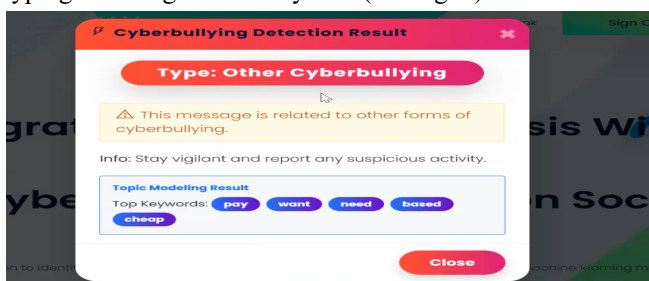
“Fig.7 Final Outcome”

The outcome of the cyberbullying detection system provided that the provided message is not a cyberbullying is displayed in Figure 7.



“Fig.8 Upload Input Text Message”

In the non-bullying context detection system, a user is typing a message into the system (see Fig. 8).



“Fig.9 Result”

The result of the detection of cyberbullying using the figure: Fig. Figure 9 is shown as “Other Cyberbullying.”

VI. CONCLUSION

The approach is able to integrate sentiment analysis together with the latest ML techniques to detect cyberbullying on social media. Kaggle Cyberbullying Tweets data set was utilized and the preprocessing techniques like normalisation, noise removal, tokenisation, lemmatisation, vaderSentiment analysis and label encoding were carried out to get the clean data for modeling. TF-IDF vectorisation used

Semantic features SMOTE and SMOTEENN for better performance for minority class. The models were optimized using GridSearchCV and tested with accuracy, precision, recall, F1 score and confusion matrix. The following ML techniques were employed: LR, RF, XGBoost, SVM and ensemble techniques. They showed good detection with few false predictions . The best overall performance in terms of all the criteria was obtained using a VC with the combination MLP, Bagging and LR which gave 97.4% accuracy. Interpretability was provided by the use of LIME and SHAP in determining the main features that influence the prediction. In conclusion, the Flask deployment with topic modelling provided predictions in real-time, along with confidence scores, making it a powerful, transparent, and scalable solution for effective and efficient cyberbullying detection on social media platforms.

For better improvement of the system, DL models like LSTM, BiLSTM or Transformers can be added to model the contextual semantics and further enhance the detection accuracy. There should be multi-lingual support to detect cyberbullying in many languages and regional dialects. Proactive intervention can be achieved using real-time social media monitoring. More sophisticated natural language processing systems can better recognise sarcasm, slang and new abusive tendencies. Mobile application and chat integration can send notifications and give immediate feedback to users. It can also be used to monitor the sentiment and user trends, which could indicate potential risks and help prevent them from worsening. Detection can be expanded and strengthened if the extension to visual and multimedia content analysis is desired.

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.