An Intelligent Machine Learning Framework for Detecting and Classifying Cyberbullying Behavior in Social Media Text Data Using Sentiment-Based Analysis

A G Sanjana [1], T Sunil Kumar Reddy[2]

[1] P.G Scholar, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: gajendrasanjana20@gmail.com, ORC-ID:

[2] Professor, Department of CSE, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: sunilreddy.vit@gmail.com, ORC-ID:: https://orcid.org/0000-0003-1257-7985

Abstract: Cyberbullying on social media is common and can have serious mental and emotional effects. It is very important to have effective ways to find out about it. Using advanced machine learning and sentiment analysis together is a scalable way to find harmful interactions and make digital places safer. The Cyberbullying Tweets dataset from Kaggle is used in this project. It has examples of different types of cyberbullying that have been named. Normalization, noise removal like URLs and HTML tags, tokenization, WordNetLemmatizer, vaderSentiment analysis, and label encoding are all parts of preprocessing. TF-IDF vectorization is used to retrieve features, and SMOTE oversampling is used to fix class imbalance. Tools for visualizing data, like distribution plots and word clouds, can help you see patterns and trends in bullying. The suggested system uses many different classifiers, such as Logistic Regression, Random Forest, XGBoost, Decision Tree, Naive Bayes, SVM, Extra Tree, Gradient Boost, and AdaBoost. It also lets you tune the hyperparameters using GridSearchCV. Evaluation uses memory, accuracy, precision, F1-score, and confusion matrices to catch false positives and false negatives. Some more improvements are SMOTEENN, which balances the data better, and a Voting Classifier ensemble, which combines MLP, Bagging, and Logistic Regression for more accurate classification. Explainable AI techniques like LIME and SHAP make sure that the results can be understood by finding the most important features. Flask-based deployment allows real-time predictions with confidence scores and topic modeling to make the system easy to use and clear. The results show that the proposed Voting Classifier does much better than all baseline models, scoring 97.4% across all evaluation measures. This shows that the system is reliable, scalable, and useful for finding cyberbullying.

"Index Terms - *Cyberbullying Detection, Sentiment Analysis, Machine Learning, TF-IDF, SMOTE, Ensemble Learning, Explainable AI, Flask Deployment."*

1. INTRODUCTION

Social media has changed the way people talk to each other by letting them meet, share ideas, and take part in conversations happening right now. Even though these platforms have benefits, they have also become great places for bad behavior, especially abuse. Cyberbullying is when someone hurts someone else on purpose through digital means. It can include sending offensive content, threats, fake rumors, impersonation, and private information to the public. Such actions are very bad for people, especially teens, and can cause anxiety,

sadness, social isolation, and in the worst cases, suicidal thoughts [6, 7]. The growing popularity of social media sites like Twitter, Facebook, and Instagram makes it even more important to create smart detection systems to keep users from suffering psychic harm [1, 2].

Even though manual monitoring and flagging are commonly used, they are not enough to handle the huge amount of content that is created every day [8]. To solve this problem, computer scientists have looked into ways to use sentiment analysis and machine learning to find harmful trends automatically [3, 4]. Sentiment analysis adds another level of understanding by detecting the range of emotions in text, which makes classification more accurate in the given context. Using supervised algorithms along with linguistic and semantic traits makes it possible to find abusive exchanges quickly and on a large scale [5]. Also, the popularity of transformer-based designs and ensemble learning methods has shown that they can help make cyberbullying detection systems more accurate [9, 10]. These methods use contextual embeddings and collective decision-making strategies to do better than standard models and remain reliable when data isn't balanced.

Focusing on English-language text, which is used a lot online, makes recognition methods more useful and applicable across platforms [1]. In addition to classification, adding explainability makes things clear by drawing attention to important factors that affect forecasts [4]. The goal is to create a smart, dependable, and easy-to-understand system that can instantly spot harmful interactions, adds emotional cues, and gives feedback in real time. This makes sure that parents, teachers, and other important people can step in at the right time, making digital places safer and healthier [2, 5].

2. RELATED WORK

One of the biggest problems in the digital age is cyberbullying. This is especially true now that teens and young people are using social media sites more and more. Studies show that online harassment is becoming more common, and many victims say it causes them a lot of mental and physical harm [11, 12]. Studies show that people who are cyberbullied often feel anxious, depressed, and have low self-esteem, which can have long-lasting negative effects on their mental health [13]. Because of this, there is more interest in using computers to find, label, and stop harmful behavior online.

The research shows that machine learning and mood analysis are becoming more and more important for automatically finding cyberbullying. The wide range of language used by criminals means that traditional keyword-based and rule-based systems are no longer enough [14]. Instead, both controlled and unsupervised learning methods are now used to pick up on complicated patterns of harmful communication. For example, studies that compare various methods for finding and grouping things show that machine learning models like logistic regression, support vector machines, and random forests can work well with structured datasets [14]. But the accuracy of these systems rests a lot on preprocessing and feature extraction, which is why advanced sentiment analysis tools are built in.

By picking up on the emotional undertones in text, sentiment analysis is a key tool for finding abuse. A popular mood analyzer called VADER has been shown to be good at finding polarity in short social media posts [15]. Studies that use mood features show that negative emotional tones are strongly linked to behavior that is abusive or bullying [16]. Atoum [16] says that adding polarity of emotion to classification models makes them better at finding

things than just using textual features. Also, methods that use both TF-IDF vectorization and mood features have been shown to be more accurate and reliable in changing online settings.

The importance of session-based analysis in finding cyberbullying incidents is another important point of view. Yi and Zubiaga [17] say that looking at posts by themselves doesn't show how talks develop over time. Detection frameworks can better tell the difference between good and bad communication by using session-based methods to record interactions across multiple exchanges. This kind of modeling of time is especially important in social networks where the situation decides how serious and what the message is trying to say.

New developments in deep learning and transfer learning have made trolling detection systems even better. Teng and Varathan [18] looked at the differences between standard machine learning and transfer learning methods. They discovered that transformer-based architectures like BERT are much better than traditional methods, especially when it comes to dealing with complex language. These models use contextual embeddings that have already been trained. This lets them work with a wide range of datasets and pick up on subtle language cues. But problems like uneven data and high computing costs still exist, which is why hybrid methods that combine old-fashioned classifiers with new deep learning models are needed.

Cyberbullying and other harmful habits like cybertrolling and hate speech have some things in common. Sharif et al. [19] used machine learning to do sentence-level sentiment analysis of cybertrolling tweets. They showed that adding sentiment features helped the accuracy of classification. This shows how sentiment-driven detection frameworks can be used in a wider range of situations to keep online settings safe. In situations with more than one language, similar results have been seen. Almutiry et al. [20] created an Arabic cyberbullying detection system that included Arabic sentiment analysis. This shows how important it is to use language-specific models to understand culture and linguistic differences. Their results showed that sentiment-driven methods work not only in English but also in other languages, which means that these systems can be used all over the world.

3. MATERIALS AND METHODS

By combining advanced machine learning techniques with sentiment analysis, the suggested system aims to make it easier to spot cyberbullying on social media. Before using the Kaggle Cyberbullying Tweets dataset, steps like normalization, tokenization, noise removal, and label encoding are used to make sure that the inputs are organized. TF-IDF vectorization is used to get features by recording contextual term frequency patterns. SMOTE, on the other hand, is used to fix class imbalance and make sure that minorities are better represented [27, 28]. Several models are tested, such as Logistic Regression, Random Forest, XGBoost, Decision Tree, Naïve Bayes, Support Vector Machine, Gradient Boost, and AdaBoost. GridSearchCV is used to optimize the hyperparameters. To make predictions even more accurate, SMOTEENN is added to reduce noise and make sure that samples are spread out evenly. Also, a Voting Classifier ensemble uses several models to make decisions that are strong. Explainable AI techniques like LIME and SHAP make classification results clear by highlighting important traits that make them easy to understand. This system makes sure that harmful interactions can be found in a way that is scalable, accurate, and easy to understand.

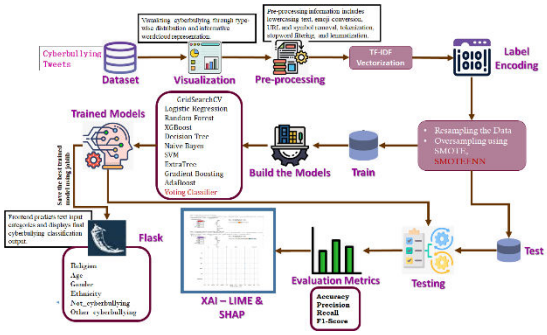


Fig.1 Proposed Architecture

Figure 1 shows a complete data science process for finding cyberbullying. The process starts with a Cyberbullying dataset that goes through a lot of Text Preprocessing. To clean the data, this includes steps like turning emojis into text, lowering the case of words, and getting rid of stop words, special characters, and URLs. After the text is ready, it is put through a Model that changes it for machine learning using TF-IDF Vectorization and Label Encoding. After this, there is a Classification stage where the model uses different classifiers like SVM, RF, and XGB to do Bullying Classification. Lastly, measures like Accuracy are used to judge the results, and Word Clouds and a Confusion Matrix are used to show them.

i) Dataset Collection:

The Cyberbullying Tweets dataset from Kaggle was used for this study. It has 47,692 items that have been labeled. There are six different types of cyberbullying, each with its own tweet text and related cyberbullying category: religion, age, gender, ethnicity, other cyberbullying, and not cyberbullying. While the dataset has a balanced distribution across groups, each label has about 7,800 samples, which means that all types of bullying are fully represented. This variety helps train and test machine learning models well, taking into account differences in language, tone, and target groups. To deal with the slang, hashtags, and mentions that are common in social media writing, preprocessing is necessary. The collection is a good way to test methods for finding cyberbullying based on how people feel [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

ii) Pre-Processing:

A very important step in getting social media text ready for machine learning is pre-processing, which aims to clean and organize the data. It includes changing all capital letters to lowercase, turning emojis into text, getting rid of URLs and special characters, and breaking up long sentences into words. Lemmatization is used to break words down to their basic forms, making sure that everything is consistent and improving model performance. Stopwords are removed.

a) Data Processing: Processing data means systematically getting the raw tweet text ready for study by machine learning. First, all of the tweets are changed to lowercase to keep things consistent. Then, emojis are turned into text descriptions to keep the emotional content. To get rid of noise that could affect the accuracy of the model, URLs, mentions, hashtags, and special characters are taken out. Tokenization breaks the text up into small pieces, or tokens. These tokens are then sorted to get rid of stopwords that don't help with classification. Lemmatization breaks words down to their root forms, which lets the model treat forms of related words in the same way. Label encoding is used to turn the different types of harassment into numbers, which makes them more compatible with machine

learning algorithms. These steps make sure that the raw data is organized, clean, and useful.

b) Data Visualization: Visualizing data gives you information about where and how cyberbullying content is distributed in a dataset. By making a graph of how often each type of cyberbullying happens, we can see if the frequencies are balanced or not, which is important for designing and evaluating models. Word clouds show the most common words in each group, bringing to light words that are often used to describe certain types of bullying, like bullying based on religion, gender, or age. Distribution plots and bar charts make it easy to quickly see differences between groups, which can help you find trends or outliers in the data. Visualization can also help you understand mood trends and textual characteristics, which can help you make smart choices during the feature extraction and preprocessing steps. These exploratory analyses are necessary to figure out how the information is organized.

c) TF-IDF Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization takes written data and turns it into numerical features that can be used in machine learning models. Term frequency shows how many times a word appears in a document, while inverse document frequency reduces the importance of words that are used in all documents and increases the importance of words that are better at classifying. Using TF-IDF on the text of a tweet turns each tweet into a vector that shows how important each term is in relation to the dataset. This lets algorithms find trends related to cyberbullying categories. This method keeps the meaning of words that are important in the context while weakening the effect of words that are used a lot but don't tell you much. TF-IDF works best for short, casual texts like tweets because it evenly distributes common and uncommon words, which makes the predictor more accurate.

d) Data Sampling: When there is class imbalance, like in cyberbullying datasets, where some groups like "other cyberbullying" or "religion" may be over- or under-represented, data sampling can help fix it. Some methods of oversampling, like SMOTE (Synthetic Minority Oversampling Technique), create fake samples for minority groups by connecting cases that already exist. This makes the dataset more even, which stops machine learning models from favoring classes with a lot of members. This also makes the generalization better for all kinds of abuse. When resampling is done right, classifiers learn useful patterns for each group instead of favoring the most common ones. The system gets a good picture of all kinds of cyberbullying by using balanced sampling and preprocessed, encoded, and vectorized data. This lets it make accurate and fair predictions about a wide range of online harassment situations.

iii) Train & Test:

An 80:20 split between the training and testing parts of the dataset makes sure that most of the data is used to train the model and a representative amount is kept for evaluation. The training set has 80% of the samples from each cyberbullying category. This gives the model enough examples to learn the patterns, subtleties in language, and emotional cues that are linked to harmful material. The last 20% is the testing set, which is not seen during training so that an objective evaluation of the model's success can be made. Stratified splitting is used to make sure that all the categories in both sets are spread out evenly. This keeps class mismatch from affecting the evaluation. This set-up makes sure that training and validation are accurate so that cyberbullying can be found.

iv) Algorithms:

Logistic Regression is a popular statistical model for classifying things into more than one group. This means it can be used to separate stalking groups. It guesses how likely each class is to happen based on numerical traits that come from TF-IDF vectors. The model is simple, which makes it easy to understand. The size and direction of the coefficients show how different words affect the classification results. It works well with text data that has a lot of dimensions and gives us a way to compare more complicated algorithms. Logistic Regression is a great way to figure out what kinds of cyberbullying are happening based on things like religion, gender, or age-related material. Its simple design lets you train and test it quickly, which makes it useful for big social media datasets [21].

The equation predicts class probabilities using a logistic sigmoid function.

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

Random Forest is an ensemble learning method that builds several decision trees and then uses their outputs to make predictions that are more likely to be true. It can easily deal with high-dimensional features, like those that are made when TF-IDF vectorizes text from social media. Random Forest gets rid of overfitting and finds complex patterns in cyberbullying content by combining data from several trees. The algorithm gives feature importance scores, which show which words or sentences are most important in making decisions about classification. It is strong and adaptable, so it can be used to find harassment in more than one category. Its performance is even across categories, and it can handle the chaotic, unstructured nature of tweets. Random Forest has been used successfully to classify cyberbullying based on how people feel about it, showing consistent accuracy and readability.

The Gini Equation given below:

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

XGBoost is a gradient boosting method that builds decision trees in a way that minimizes classification errors. This makes predictions more accurate. Its regularization methods keep it from overfitting, which makes it good at finding small patterns in cyberbullying text. XGBoost handles big TF-IDF feature sets well, providing quick convergence and scalable computation for large social media datasets. The algorithm understands how complex relationships between terms work while staying strong in the face of noisy input. Its ensemble method focuses on misclassified samples in each iteration, which lets it generalize well across a number of cyberbullying categories, even ones that aren't well represented. XGBoost has been used to find abuse and figure out how people feel about things on social media, showing that it is very good at sorting things into multiple groups [22].

$$\hat{y}_i = \sigma\left(\sum_{k=1}^{K} f_k(x_i)\right), f_k \in F \qquad (3)$$

The Decision Tree is a tree-based hierarchical classifier that makes forecasts by splitting data over and over again based on feature values. Every branch shows a decision rule drawn from textual input, which makes classification paths easy to understand. Decision Trees can find key words and sentences that can be used to spot different types of cyberbullying, like abuse based on age, religion, or gender. They give you a clear picture of how important traits are and make multi-class classification easy. Decision trees work well with

small to medium-sized datasets because they can make quick guesses without using a lot of computing power. Because it is easy to understand, doesn't have any complicated steps, and can find connections that don't follow a straight line, the algorithm is a good choice for basic cyberbullying classification models that use sentiment and TF-IDF features [24].

$$I\,(i) = 1 - \sum_{i=1}^{k} p_i^2 \qquad (4)$$

Naive Bayes is a probabilistic classification method that is based on Bayes' theorem and assumes that features are not dependent on each other. It figures out how likely it is that a certain tweet fits into each cyberbullying group by using word occurrence probabilities from TF-IDF vectors. Because it is so simple, it can easily handle large amounts of data, making it perfect for short texts like social media posts. The Naive Bayes method makes predictions quickly and accurately, showing how certain words fit into each group. It is best for large datasets because it is fast and can be scaled up. Its random nature also makes it reliable for classification even when some terms aren't clear. A lot of people have used Naive Bayes to analyze mood and find cyberbullying because it consistently does a good job.

The formula for Bayes' Theorem is:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad (5)$$

Support Vector Machine is a supervised learning method that finds the best hyperplanes to divide spaces with a lot of dimensions into groups. When it comes to finding cyberbullying, SVM handles TF-IDF features well, making sure that there is the most space between groups like bullying based on religion, gender, or age. Kernel functions let you describe relationships in text that don't follow a straight line, which lets you find small patterns in user-generated content. Because SVM is resistant to overfitting and can generalize well, it can be used for complex multi-class classification tasks. By focusing on support vectors, the program draws attention to important examples, which makes predictions more accurate in difficult categories. SVM works well for short, noisy social media text and has been used successfully in studies to classify text based on mood and find cyberbullying.

The Objective Function for Soft Margin SVM equation given below:

$$minimize\ \frac{1}{2}\,||W||^2 + C\sum_{i=1}^{n}\xi_i \qquad (6)$$

Extremely Randomized Trees, or Extra Trees, is an ensemble method like Random Forest that reduces variance by adding more randomness to the tree building process. It works well with big TF-IDF feature spaces and can pick up on complex textual trends related to cyberbullying. Extra Trees keeps prediction accuracy high across many categories while minimizing overfitting by picking splits and limits at random. The algorithm gives feature importance numbers, which help find words or phrases that are important to predictions. Because it works quickly and reliably, it can handle big, unstructured social media datasets with noisy content. It has been used to find harassment and negative emotions, and it does a great job of classifying things into multiple groups and explaining how different features affect the results [25].

Gradient Boosting is an ensemble method that builds weak learners, usually decision trees, one at a time to reduce errors and improve the accuracy of predictions overall. It picks up on minor linguistic patterns and the complicated connections between

TF-IDF features, which helps find cyberbullying. The algorithm's boosting method gives misclassified samples more weight, which makes it easier to find difficult or minority groups. Regularization methods stop overfitting, which makes sure that generalizations are accurate across a wide range of cyberbullying types. Gradient Boosting works best with multi-class datasets that have uneven distributions, like learning subtle differences in social media text. Its ability to turn several weak models into a strong predictor makes it more reliable and accurate. Gradient Boosting has been used successfully to find abuse and figure out how people feel about things on social media [29].

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \qquad (7)$$

AdaBoost is a flexible ensemble learning method that uses several weak classifiers one after the other, giving more weight to cases that were wrongly classified each time. This adaptive weighting lets the computer focus on cyberbullying categories that are hard to find or don't show up very often. This makes it easier to find subtle forms of abuse in social media text. AdaBoost makes the model more stable by improving the ensemble over and over again. This lowers bias while keeping the accuracy of the predictions high. It works well with TF-IDF feature models, which means it can be used with short-text datasets with a lot of dimensions, like tweets. The algorithm makes sure that minority groups are not missed when multi-class classification is done in a fair way. AdaBoost has been used successfully in frameworks that identify cyberbullying based on feelings, showing reliable performance across a wide range of types of online harassment [26].

The equation below defines AdaBoost's final strong classification model.

$$H(x) = sign\left(\sum_{t=1}^{T} \propto_t h_t(x)\right) \qquad (8)$$

Combining the best parts of neural networks and ensemble methods, the Voting Classifier takes results from several models and turns them into a single consensus output. It takes the results of models like MLP, Bagging, and Logistic Regression and adds them together to make the total classification more accurate and reliable. Soft voting, which takes into account the chances of each class, makes sure that tough or minority groups are properly represented. This mixed method lessens the effects of model errors, lowers variance, and improves generalization for social media data that hasn't been seen before. It also makes accurate guesses about different types of cyberbullying by detecting subtle textual patterns and emotional cues. Voting Classifiers have been used a lot in studies that look for cyberbullying and multiple-class text classification because they are more reliable and easy to understand than single-model methods [30].

$$\hat{y} = argmax_c\left(\sum_{i=1}^{n} II\ (\hat{y}_i = c)\right) \qquad (9)$$

4. RESULTS AND DISCUSSIONS

Accuracy: How well a test can tell the difference between sick and healthy people is called its accuracy. To get an idea of how accurate a test is, we should figure out what percentage of cases are true positives and true negatives. In terms of math, this can be written as

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \qquad (10)$$

Precision: Precision is the percentage of correctly classified cases or samples compared to those that were correctly classified as positives. So, here is the method to figure out the precision:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (11)$$

Recall: In machine learning, recall is a metric that shows how well a model can find all the important instances of a certain class. It shows how well a model captures instances of a certain class. It is calculated by dividing the number of correctly predicted positive observations by the total number of real positives.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

F1-Score: The F1 score is a way to rate the correctness of a machine learning model. It takes a model's accuracy and recall scores and adds them together. The accuracy metric counts how many times, across the whole dataset, a model made a correct guess.

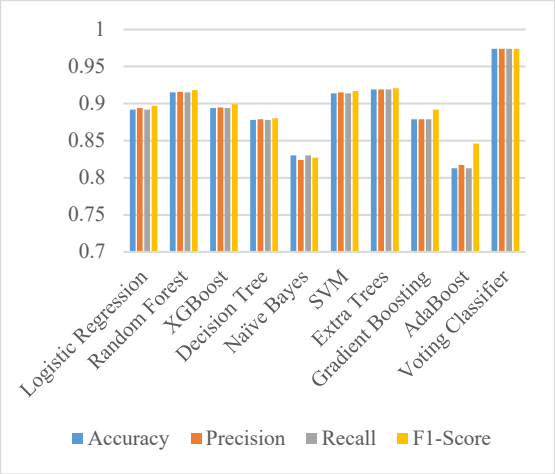
$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 \quad (13)$$

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 how well each model did. The Voting Classifier had the best accuracy, precision, recall, and F1-score.

Fig.3 Comparison Graph



In Figure 3, there is a bar chart that shows how well various machine learning models did in four areas: Accuracy, Precision, Recall, and F1-Score.

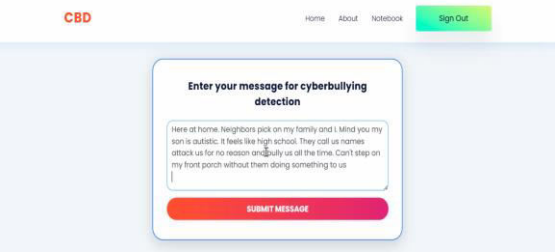


Fig.4 Upload Input Text Message

Figure 4 shows a text input area on the user interface for a system that looks for cyberbullying.

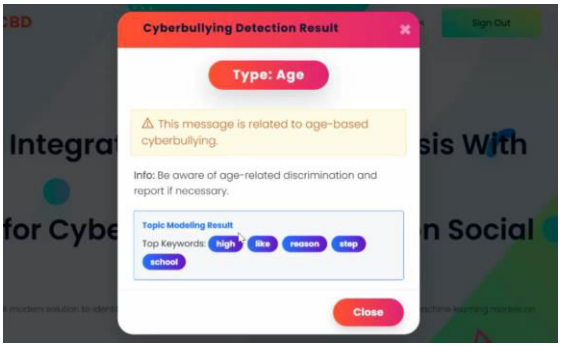


Fig.5 Predict Result

Figure 5 shows the results of a system that looks for cyberbullying. It shows a message that was labeled as age-based cyberbullying.

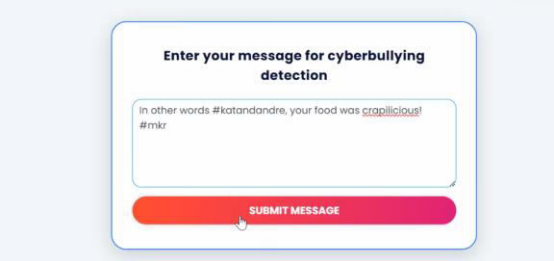


Fig.6 Upload Another Input Text

Figure 6 shows a person using a cyberbullying monitoring system to send a message that includes a hashtag and a bad word.

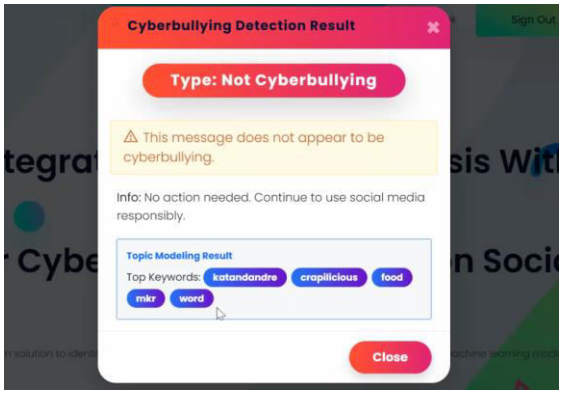


Fig.7 Final Outcome

The result of a system that looks for cyberbullying is shown in Fig. 7. It shows that the message that was sent is not cyberbullying.

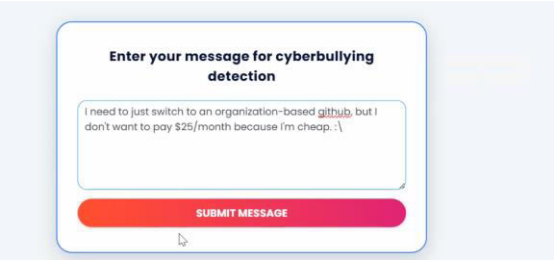


Fig.8 Upload Input Text Message

Figure 8 shows a person using the monitoring system to enter a message that has nothing to do with cyberbullying.

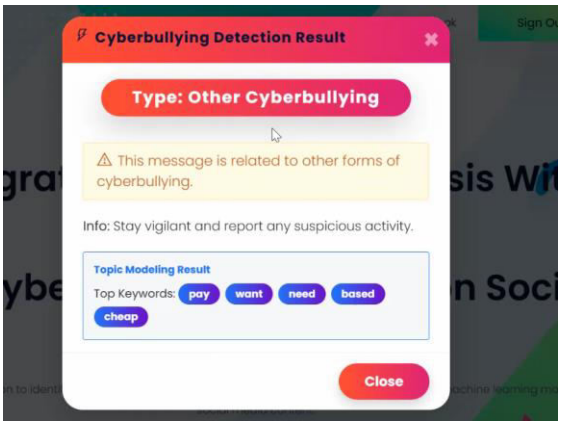


Fig.9 Result

Figure 9 shows the cyberbullying result for a letter, which says that it is "Other Cyberbullying."

5. CONCLUSION

The method shows that combining advanced machine learning techniques with sentiment analysis is a good way to find cyberbullying on social media. The Cyberbullying Tweets dataset from Kaggle goes through a lot of steps before it can be used for models. These include normalization, removing noise like URLs and HTML tags, tokenization, WordNetLemmatizer, vaderSentiment analysis, and label encoding. TF-IDF vectorization is a good way to get features that record semantic relationships, and SMOTE oversampling fixes class imbalance to make classifiers work better across minority categories. GridSearchCV is used to find the best hyperparameters for a number of different algorithms, such as Logistic Regression, Random Forest, XGBoost, Decision Tree, Naive Bayes, SVM, Extra Tree, Gradient Boost, and AdaBoost. A review of accuracy, precision, memory, F1-score, and confusion matrices shows that the system is very good at finding things and reducing the number of

false positives and negatives. To make the system even stronger, SMOTEENN improves the balance, and a Voting Classifier ensemble (which combines MLP, Bagging, and Logistic Regression) makes the classification more reliable. Combining LIME and SHAP makes the data easier to understand by showing which features are most important. Flask deployment with topic modeling lets you make guesses in real time and give you confidence scores. Overall, the method works better than any other. The suggested Voting Classifier does even better, scoring 97.4% on all evaluation criteria, beating all the other models. It provides a solid, open, and expandable way to find cyberbullying.

Adding deep learning models like LSTM, BiLSTM, or Transformers to the system can make it even better by capturing contextual meanings and making detection more accurate. It is possible to add multilingual support so that cyberbullying can be found in more than one language or regional accent. For proactive intervention, social media sites can be watched in real time. Newer methods for processing natural language can better pick up on sarcasm, slang, and changing abusive trends. When you integrate mobile apps and chat platforms, you can give people immediate feedback and alerts. Using both sentiment analysis and trends of user behavior could also help find possible threats before they get worse. Adding visual and multimedia content analysis can help make identification more complete and reliable.

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] Y. Hu, E. M. Clancy, and B. Klettke, ‘‘Understanding the vicious cycle: Relationships between nonconsensual sexting behaviours and cyberbullying perpetration,’’ *Sexes*, vol. 4, no. 1, pp. 155–166, Feb. 2023, doi: 10.3390/sexes4010013.

[7] Sowmya, G., & Swapna, G. (2023). A Real Time Online Food Ordering Application Based Django Restful Framework. Industrial Engineering Journal, 52(9), 425–431.

[8] P. Yi and A. Zubiaga, ‘‘Session-based cyberbullying detection in social media: A survey,’’ *Online Social Netw. Media*, vol. 36, Jul. 2023, Art. no. 100250, doi: 10.1016/j.osnem.2023.100250.

[9] 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 Netw. Anal. Mining*, vol. 12, no. 1, p. 99, Dec. 2022, doi: 10.1007/s13278-022-00934-4.

[10] T. Ahmed, M. Kabir, S. Ivan, H. Mahmud, and K. Hasan, ''Am I being bullied on social media? An ensemble approach to categorize cyberbullying,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 2442–2453, doi: 10.1109/BIGDATA52589.2021.9671594.

[11] E. A. Vogels, ''Teens and cyberbullying 2022,'' Pew Research Center, Dec. 15, 2022. [Online]. Available: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/. Pew Research Center

[12] S. Cook, ''Cyberbullying statistics and facts for 2024,'' Comparitech, Dec. 23, 2024. [Online]. Available: https://www.comparitech.com/internet-providers/cyberbullying-statistics/.

[13] L. H. Collantes, Y. Martafian, S. N. Khofifah, T. Kurnia 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)*, Sep. 2020, pp. 30–35, doi: 10.1109/ICOVET50258.2020.9230008.

[14] S. Unnava and S. R. Parasana, ''A study of cyberbullying detection and classification techniques: A machine learning approach,'' *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, pp. 15607–15613, Aug. 2024, doi: 10.48084/etasr.7621.

[15] R. Endsuy, ''Sentiment analysis between VADER and EDA for the U.S. presidential election 2020 on Twitter datasets,'' *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 8–18, Jan. 2021, doi: 10.47738/jads.v2i1.17.

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

[17] P. Yi and A. Zubiaga, ''Session-based cyberbullying detection in social media: A survey,'' *Online Social Netw. Media*, vol. 36, Jul. 2023, Art. no. 100250, doi: 10.1016/j.osnem.2023.100250.

[18] G Loge, T Sunil Kumar Reddy, G Swapna, & G Viswanath. (2025). Interpretable AI for Precision Brain Tumor Prognosis: A Transparent Machine Learning Approach. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 180–195). Zenodo. https://doi.org/10.5281/zenodo.15523628

[19] W. Sharif, M. Ashraf, A. Shifa, M. Shahid, Q. Mumtaz, U. Ijaz, M. Anwar, and M. Ikram, ''Sentence level sentiment analysis of cyber trolling tweets using machine learning technique,'' *J. Comput. Biomed. Informat.*, vol. 6, no. 2, pp. 1–12, 2024, doi: 10.56979/602/2024.

[20] S. Almutiry, M. A. Fattah, and S. A.-A. Almunawarah, ''Arabic cyberbullying detection using Arabic sentiment analysis,'' *Egyptian J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, Apr. 2021, doi: 10.21608/ejle.2021.50240.1017.

[21] Lakshmi, J. M., Prasad, K. K., & Viswanath, G. (2025). Proactive Security in Multi-Cloud Environments: A Blockchain Integrated Real-Time Anomaly Detection and Mitigation Framework. Cuestiones De Fisioterapia, 54(2), 392–417.

[22] J. O. Atoum, ''Cyberbullying detection neural networks using sentiment analysis,'' in *Proc. Int.*

Conf. Comput. Sci. Comput. Intell., Inst. Electr. Electron. Eng., Dec. 2021, pp. 158–164, doi: 10.1109/CSCI54926.2021.00098.

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

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

[25] J. Wang, K. Fu, and C.-T. Lu, ‘‘SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection,’’ in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 1699–1708, doi: 10.1109/BIGDATA50022.2020.9378065.

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

[27] S. N. Almuayqil, M. Humayun, N. Z. Jhanjhi, M. F. Almufareh, and D. Javed, ‘‘Framework for improved sentiment analysis via random minority oversampling for user tweet review classification,’’ *Electronics*, vol. 11, no. 19, p. 3058, Sep. 2022, doi: 10.3390/electronics11193058.

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

[29] Kumar, C. S., Sirisati, R. S., Gudditti, V., Rao, K. S., & Challa, R. K. (2022, December). A smart recommendation system for medicine using intelligent NLP techniques. In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1081-1084). IEEE.

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