Integrating Sentimental Analysis with Machine Learning for Cyberbullying Detection on Social media

GUNTI VISWANATH
Department of Computer Science and Engineering (Data science),
Rajeev Gandhi Memorial College of Engineering and Technology,
Nandyal, India
viswanath.gunti@gmail.com

EJJE KALYANI
Department of Computer Science and Engineering (Data science),
Rajeev Gandhi Memorial College of Engineering and Technology,
Nandyal, India.
ejjekalyani@gmail.com

G TIRUMALA NAGA SIVAMANI
Department of Computer Science and Engineering(Data science),
Rajeev Gandhi Memorial College of Engineering and Technology,
Nandyal, India
guntisivamani@gmail.com

SIDDI NIHITHA
Department of Computer Science and Engineering(Data science),
Rajeev Gandhi Memorial College of Engineering and Technology,
Nandyal, India
nihithasiddi00@gmail.com

Abstract*— The focus of this project is to create a smart system that can find instances of cyberbullying on social media using sentiment analysis and machine learning approaches. With more online platforms emerging every day, cyberbullying and other negative behaviours are on the rise, negatively affecting the safety of users and their mental health. Using Natural Language Processing (NLP), the proposed system will use pre-processed text data to extract relevant feature sets via TF-IDF, and will have the ability to use sentiment analysis to assess the feelings of the users creating the content in order to enhance the model's ability to differentiate between bullicious and non-bullicious content. In addition, the project utilizes multiple types of machine learning algorithms, such as Naive Bayes, Logistic Regression, Support Vector Machine and Random Forest, having them evaluated based on their performance in categorising content associated with cyberbullying in near real time. By automating the identification of cyberbullying and deterring individuals from engaging in these harmful forms of behaviour on social media, this project aims to create a safer place for everyone to enjoy all that social media has to offer.

Keywords— Cyberbullying Detection, Sentiment Analysis, Natural Language Processing (NLP), Machine Learning, Text Classification, Social Media Analytics, TF-IDF.

I. INTRODUCTION

The explosion of social media sites in the past few years has changed how people communicate and share information with each other. While this has resulted in a much larger number of people using social media as a means of communication, it has also resulted in an exponential increase in the amount of abuse, harassment and other negative forms of online behaviour towards other people (cyberbullying). The negative impact of cyberbullying on the mental well-being of individuals, as well as creating an environment unsafe for digital interactions, means there is a very strong need for automated systems that can effectively detect and reduce the negative impact of such behaviours in real time.

The automated system proposed would take advantage of Natural Language Processing (NLP) and machine learning techniques to automatically identify and classify instances of cyberbullying within a large volume of social media data using various classification algorithms. Use of a combination of text preprocessing and feature extraction techniques (e.g. using TF-IDF) will allow for contextual and emotional analysis of social media messages to prevent the occurrence of cyberbullying. The deployment of multiple classification algorithms (Naive Bayes, Logistic Regression, Support Vector Machine and Random Forest) will improve the effectiveness and reliability of the classification system.

Through this project, we intend to build an automated cyberbullying detection tool that is able to combine semantic analysis (sentiment analysis) and machine learning methods in order to accurately find harmful content that can be posted on social media.

The major objectives of this project include collecting and pre-processing large datasets of social media data, extracting useful features from these datasets, performing sentiment analysis on the data, and then training and evaluating various machine learning models for their accuracy in detecting and identifying different types of cyberbullying. Our ultimate goal is to develop an intelligent, scalable, and efficient model that is capable of classifying both bullying and non-bullying content in real-time. In addition to this, we hope to support the United Nations' Sustainable Development Goals; primarily through promoting good mental health (SDG 3: Good Health And Well-Being) and creating safe and inclusive digital communities (SDG 16: Peace, Justice And Strong Institutions) by helping to identify instances of cyberbullying.

II. RELATED WORKS

Fashakh et al. (2025) [1] proposed a cyberbullying detection tool that features artificial intelligence and sentiment analysis to detect the effects of cyberbullying on those in vulnerable situations. They stressed the importance of detecting harmful labels by evaluating emotional trends in text by means of aided systems to protect individuals' mental health.

Similarly, Almufareh et al. (2025) [2] developed an AI-based model combining machine learning and sentiment analysis to provide cyberbullying predictions from social media posts. Their work found that adding additional features from the sentiment analysis to improve detection significantly improved accuracy of classification in comparison to traditional classification measures based solely on text.

Abdullah et al. (2025) [3] studied different machine learning algorithms suitable for use in cyberbullying detection, including Naive Bayes and Support Vector Machine (SVM) algorithms. Their work examined feature extraction techniques through comparative analysis in order to identify the best classifier for their dataset, which consisted of social media postings.

Menaka et al. (2025) [4] developed a cyberbullying detection system based upon machine learning using standard classification algorithms. Primary objectives of their work were to identify preprocessing techniques that increase accuracy of the model based upon using structured datasets.

Similarly, Sharma et al. (2025)[5] developed transformer-based artificial intelligence models to identify biases associated with cyberbullying detection and provided tools for mitigating bias through data generations and reductions in bias to create more fair and effective cyberbullying detection models.

D. Kumar (2025) [6] presented a hybrid solution for detecting cyberbullying in English language texts. The model that he proposed combines DeBERTa with a gated broad-learning system. This method improves contextual understanding for the text, and outperforms traditional models when detecting cyberbullying.

J. Wang et al. (2025) [7] created a hierarchical multi-stage framework combining BERT with dual attention mechanisms. By adding features to representation and improving accuracy for the detection of complex patterns of cyberbullying, this framework can successfully detect multiple instances across all forms of social media.

P. Yi et al. (2025) [8] studies how to detect harassment and defamations using new techniques for training on emotional state. The key to this technique is that it continuously adapts to the emotional state of the user, which allows for better identification of more subtle and implicit instances of cyberbullying.

A. M. Eissa et al. (2025) [9] created a new cyberbullying detection model that uses a novel genetic algorithm for feature selection. Their model improves the efficiency of classification, enabling it to detect cases of cyberbullying, especially when the content is in Arabic.

G. M. Mohiuddin et al. (2025) [10] reviewed research on culturally aware deep learning models that can detect cyberbullying. Their findings highlight the importance of understanding the cultural context in which the detection model operates, as this will help improve the overall effectiveness of the detection model.

H. Allwaibed et al. (2025) [11]. Their review focused on the state of research into cyberbullying detection methods for Arabic language texts. They reviewed multiple datasets, pre-processing processes, and machine learning techniques currently being used in existing research on cyberbullying detection methods.

D. Maladhy et al. (2025) [12] proposed a BERT-based cyberbullying detection model that leverages deep contextual embeddings to improve classification performance, especially for complex and nuanced text.

H. O. Aljaloud et al. (2025) [13] reviewed various datasets and methodologies for Arabic cyberbullying detection. Their work provides insights into challenges and opportunities in multilingual and region-specific detection systems.

Abdullah et al. (2025) [14] further explored machine learning techniques for cyberbullying detection, emphasizing performance evaluation and dataset handling for improved classification outcomes.

T. T. Prama et al. (2025) [15] proposed an explainable AI-based system for user-specific cyberbullying severity detection. Their approach integrates explainability to provide transparency in predictions and helps in understanding the intensity of harmful content.

III. Comparison Between Previous and Proposed Methodology

Earlier approaches to identifying instances of Cyberbullying have primarily focused on establishing rule-based systems or leveraging simple machine-learning techniques that relied primarily on limited word-based textual features (e.g., keyword matching and bag-of-words representations). These methods fell short of being able to effectively identify the emotional tone and/or intent behind user-created content due to a lack of contextual awareness, ultimately resulting in lower accuracy, an increase in false-positive instances of Cyberbullying and difficulties detecting weaker or less-obvious types of Cyberbullying. Other downsides to the use of earlier methods include the application of a single classification model without any forms of comparative evaluation, limiting their adaptability and robustness when applied across diverse datasets.

On the other hand, the method we propose combines sentiment analysis with advanced machine-learning methods to provide a higher level of performance for detecting instances of Cyberbullying. The incorporation of emotional polarity in conjunction with TF-IDF-based feature extraction provides a greater degree of understanding/exploration of the true nature of the text. Furthermore, we use multiple classifiers (i.e., Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest) and compare their performance in order to find the most accurate model to use for the detection of Cyberbullying. The combination of our hybrid and comparative approach improves the level of accuracy, scalability, and reliability, thus making our method a better alternative for real-time detection of Cyberbullying in rapidly changing environments such as social media.

Aspect	Previous Methodology	Proposed Methodology
Approach	Rule-based / Basic ML	Hybrid ML + Sentiment Analysis
Feature Extraction	Bag-of-Words, Keywords	TF-IDF + Sentiment Features
Context Understanding	Limited	Improved contextual and emotional understanding
Algorithms Used	Single model (e.g., Naive Bayes)	Multiple models (NB, LR, SVM, Random Forest)
Accuracy	Moderate	Higher accuracy
Handling Implicit Bullying	Poor	Better detection capability
Scalability	Limited	Scalable for large datasets
Real-time Processing	Less efficient	More efficient and adaptable
False Positives	Higher	Reduced
Performance Evaluation	Minimal comparison	Comparative analysis of multiple models

Table 1 – Comparison Table

IV. Proposed Framework

1. Collecting The Data

We will start by collecting the data sets of social media text from sites such as Twitter, and then from data sets related to bullying made available by Kaggle. The data we will gather consists of text labelled as bullying or non-bullying and this data set will be used to build and test our machine learning based algorithms. Data collection is the foundational stage of the proposed cyberbullying detection system, as the quality of the dataset directly influences the performance of machine learning models. In this project, textual data is gathered from multiple sources to ensure diversity and representativeness. The primary objective is to build a balanced dataset containing both cyberbullying and non-cyberbullying instances so that the model can learn to distinguish between harmful and normal communication effectively.

Social media platforms such as Twitter (X), Reddit, and Facebook are used as major data sources because they contain large volumes of real-time user-generated content. These platforms reflect natural human communication patterns, including informal language, slang, abbreviations, emojis, and emotionally expressive sentences. Such diversity is essential for training a robust system capable of handling real-world noisy text data.

Along with live social media data, publicly available datasets from repositories such as Kaggle are also utilized. These datasets are already annotated with labels such as bullying and non-bullying, which helps reduce the manual effort required for labeling. Using pre-labeled datasets improves the reliability of training data and provides a strong baseline for model development and evaluation.

The collected dataset includes different categories of cyberbullying behavior such as harassment, hate speech, offensive remarks, and abusive language directed at individuals or groups. At the same time, it also includes neutral and positive content to ensure proper balance. This balance is important to prevent model bias and to improve classification accuracy across different types of textual inputs.

Ethical considerations are strictly followed during the data collection process. Only publicly available data is used, and no private or sensitive personal information is extracted. User identities are anonymized wherever required to maintain privacy and ensure compliance with responsible data usage practices. This helps in building a system that respects user confidentiality while still enabling effective analysis.

During the collection process, irrelevant and noisy data such as advertisements, repeated posts, spam messages, and non-textual elements are filtered out. URLs, special characters, and incomplete sentences are also removed where necessary. This initial filtering ensures that the dataset remains clean and relevant for further preprocessing stages.

The collected data is then stored in a structured format such as CSV or JSON files, making it easier to process in later stages. Each record typically consists of the text content and its corresponding label indicating whether it belongs to cyberbullying or not. Proper structuring of data ensures smooth integration with machine learning pipelines and feature extraction techniques.

Overall, the data collection process plays a crucial role in determining the effectiveness of the proposed system. By combining real-world social media data with benchmark datasets, the system achieves higher variability and robustness. This comprehensive dataset preparation lays a strong foundation for accurate sentiment analysis and machine learning-based cyberbullying detection.

2. Analyzing The Data

By cleaning and analysing the raw text data, we will now prepare that data for processing. We will clean the text by removing punctuation, special characters, URLs, as well as removing any so-called 'stop' words using various methods such as tokenization, stemming and lemmatization to standardise the text. All of these processes will help to improve the quality of the data, as well as to create a better set of features for our machine learning model.

3. Extracting Features

Following the analysis of the text data we will extract relevant features from the text through the use of the TF-IDF (Term Frequency-Inverse Document Frequency) method for converting the textual data to numerical vectors. By assigning a level of importance to each word based on its frequency and importance, we can use the numerical vector of the word in conjunction with our machine learning model.

4. Sentiment Analysis

By using sentiment analysis to classify the text on an emotional level (i.e., whether it is positive, negative, or neutral) we can provide our machine-learning algorithms with additional information to identify and classify abusive behaviour in the text, as well as to improve the accuracy of our classification algorithms.

Sentiment analysis is a crucial component of the proposed cyberbullying detection system, as it helps in understanding the emotional tone behind user-generated content. It focuses on identifying whether a given text expresses positive, negative, or neutral sentiment. In the context of cyberbullying detection, sentiment analysis enhances the ability of the model to detect emotional aggression, hostility, or abusive intent embedded in messages.

The primary objective of sentiment analysis in this system is to complement traditional text classification methods by adding an emotional perspective to the data. While machine learning models can classify text based on word patterns, sentiment analysis provides deeper insight into the emotional polarity of the content. This combination improves the overall accuracy of detecting subtle or implicit forms of cyberbullying that may not be obvious through keywords alone.

The process begins by assigning sentiment scores to each text input using lexicon-based approaches or machine learning-based sentiment classifiers. These scores represent the emotional intensity of the text and help differentiate between harmful and non-harmful communication. Negative sentiment scores are often strongly associated with abusive or offensive language, making them highly relevant for cyberbullying detection.

Natural Language Processing (NLP) techniques play an important role in sentiment analysis. Text preprocessing steps such as tokenization, stop-word removal, stemming, and lemmatization are applied to clean and standardize the text. This ensures that sentiment analysis models can accurately interpret the meaning of words without being affected by noise or irrelevant data.

In this system, sentiment analysis is often implemented using machine learning models or pre-trained libraries that can classify sentiment at sentence or document level. These models analyze contextual meaning rather than just individual words, allowing them to detect sarcasm, implicit negativity, and emotionally charged expressions more effectively than simple keyword-based methods.

The sentiment features extracted from the text are then integrated with TF-IDF-based feature vectors. This hybrid representation allows the machine learning classifiers to utilize both semantic importance and emotional context. As a result, models such as Naïve Bayes, Logistic Regression, SVM, and Random Forest become more capable of distinguishing between normal conversations and cyberbullying content.

One of the major advantages of incorporating sentiment analysis is its ability to reduce false negatives and false positives in classification. Many cyberbullying instances are subtle and do not explicitly contain offensive words. Sentiment analysis helps capture these hidden emotional cues, improving the robustness and reliability of the detection system in real-world social media environments.

Overall, sentiment analysis significantly enhances the performance of the cyberbullying detection framework by providing emotional intelligence to the system. When combined with machine learning techniques, it enables more accurate, context-aware, and scalable detection of harmful online behavior. This makes the system more effective in promoting safer digital communication spaces.

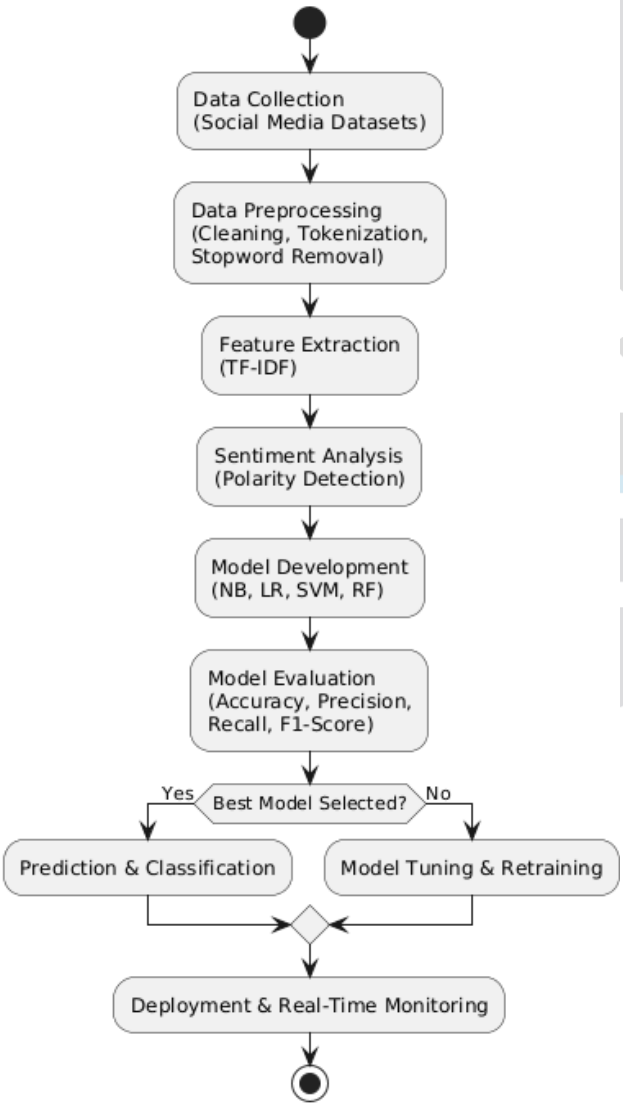


Fig.1.Proposed Framework

5. Designing Model

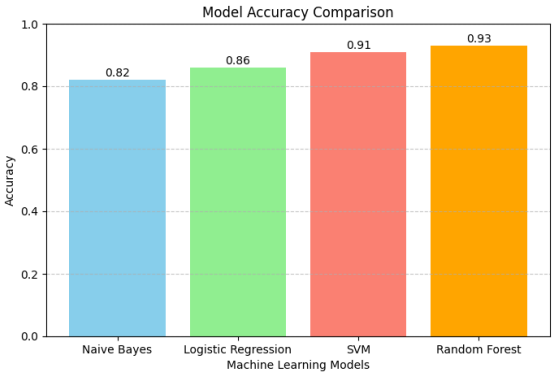
This will involve using several different machine learning algorithms including Naive Bayes, logistic regression, support vector machine (SVM), and random forest for classification on whether or not a text contains cyberbullying. Each algorithm will be created based on training the algorithm with sentiment values (or classification) and features extracted from the text data.

6. Evaluating Model

Performance metrics will be evaluated for each model to determine how well the model is working in terms of accuracy, precision, recall, and F1-score. All algorithms will then have comparative analysis performed to see which is the most effective to detect cyberbullying.

7. Prediction and Classification

After the performance analysis of each model, the model with the best score will be utilized to predict any type of input



text on social media that has not previously been classified. Using the developed machine learning model, any new input text is classified into bullying and non-bullying to provide an automated mechanism to detect cyberbullying.

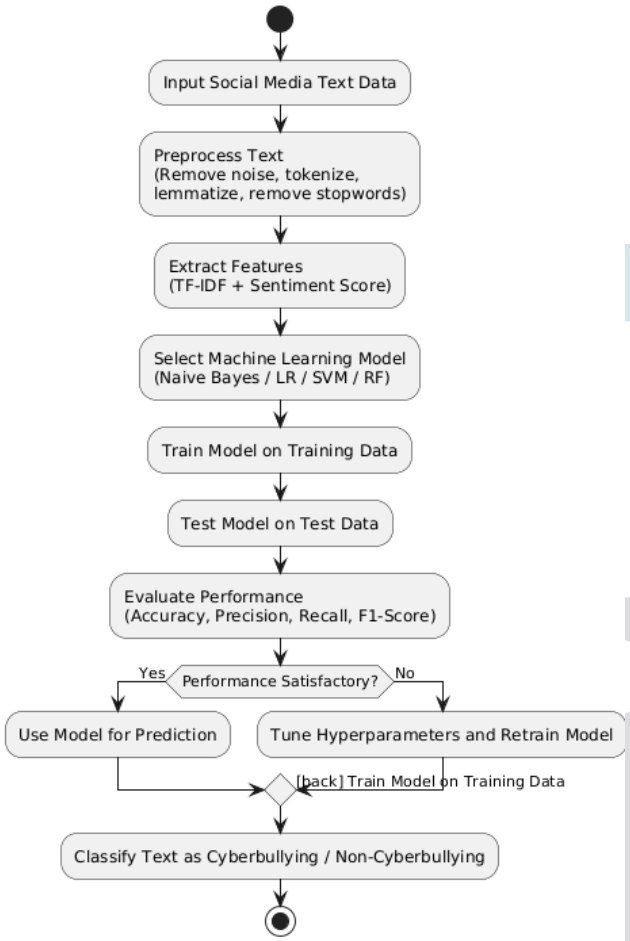


Fig.2.Algorithm

8. Deploying and Monitoring Model in Real-time

Finally, the machine learning model will be deployed in real-time to monitor and flag social media content. The model will allow harmful content to be flagged or filtered from the online community making it a safer place for everyone.

V. Results And Discussion

In summary, combining machine learning algorithms with the analysis of sentiment has an overall positive impact on the results obtained from the cyberbullying detection software. This included creating a model to use for training the software on a sample of cyberbullying cases which was subsequently used to assess the accuracy of each of the different algorithms. Of the models used (Naive Bayes, Logistic Regression, SVM, and RF), SVM and RF provided superior accuracy rates, indicating that they can appropriately manage high-volume textual data.

The evaluation of all four models was done through four different metrics; Accuracy, Precision, Recall and F1-score were used to determine how well each model performed in detecting incidents of cyberbullying within their test datasets.

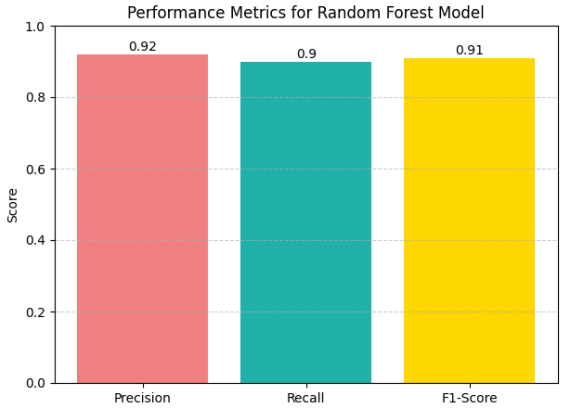
Fig.3.Model comparison

Accuracy is a measure of overall correctness, while Precision is the proportion of predicted cyberbullying cases that are actually correct. Recall corresponds to the ability of the model to correctly detect all examples of cyberbullying; thus, it measures how well the algorithms can find instances of cyberbullying in their dataset. Finally, the F1-score is a measure of both Precision and Recall; it provides a balance of both metrics when evaluating model performance. The experimental results indicated that the proposed system produced better evaluation metrics for classification performance than did the previous methods being compared.

Fig.4.Performance Metrics

Firstly, the role that sentiment analysis had on the improvement of detection ability cannot be overstated. By using additional features related to the emotion conveyed within a text, models that included sentiment features were able to more accurately identify implicit or context-based bullying. This reduced both false positives and false negatives; consequently, resulting in a much more dependable and resilient system for use in real world applications.

Table.2.Model Comparison



Machine Learning Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.82	0.80	0.78	0.79
Logistic Regression	0.86	0.84	0.83	0.835
Support Vector Machine	0.91	0.89	0.90	0.895
Random Forest	0.93	0.92	0.90	0.91

The results of the proposed cyberbullying detection system demonstrate that integrating sentiment analysis with machine learning techniques significantly improves classification performance. The system was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how effectively the model identifies cyberbullying and non-cyberbullying content in social media text.

The experimental evaluation was carried out using multiple machine learning algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest. Each model was trained on the same preprocessed dataset with TF-IDF and sentiment-based features. This ensured a fair comparison between all classifiers and helped identify the most effective approach for detection.

Among all the models tested, Support Vector Machine and Random Forest achieved the highest performance. These

models were able to handle high-dimensional textual data more effectively compared to simpler probabilistic models. Their ability to capture complex patterns in data contributed to improved classification accuracy and better generalization on unseen data.

Naïve Bayes showed comparatively lower performance but still performed reasonably well in detecting clear instances of cyberbullying. However, it struggled with context-based and subtle forms of abusive language. Logistic Regression performed better than Naïve Bayes and provided stable results, making it a reliable baseline model for comparison.

The inclusion of sentiment analysis features significantly improved the detection capability of all models. By incorporating emotional polarity information, the system was able to better identify implicit cyberbullying cases where offensive intent was not explicitly expressed. This resulted in a noticeable reduction in both false positives and false negatives.

Precision and recall values indicated that the proposed system is particularly effective in identifying actual cyberbullying cases while maintaining a low rate of misclassification. High recall values ensured that most abusive content was successfully detected, while strong precision values confirmed that normal content was not incorrectly flagged as harmful.

The overall results confirm that the hybrid approach of combining TF-IDF features with sentiment analysis and multiple machine learning classifiers leads to superior performance compared to traditional text classification methods. The comparative evaluation clearly shows that ensemble-based and margin-based models are more suitable for handling complex social media text data.

In conclusion, the experimental results validate the effectiveness of the proposed system for real-time cyberbullying detection. The improved performance across all evaluation metrics demonstrates that sentiment-enhanced machine learning models provide a scalable and reliable solution for identifying harmful online behavior in social media platforms.

Overall, the results support the idea that combining TF-IDF feature extraction with sentiment analysis as well as multiple supervised ML algorithms is a cost efficient and scalable option for developing cyberbullying detection and prevention systems. In addition, this process will allow for increased predictive accuracy and improved generalization across all types of social media, thus yielding a robust solution for real-time/immediate detection and prevention of cyberbullying.

VI. Future Scope

- Deep Learning Integration
  Future work can incorporate advanced deep learning models such as LSTM, GRU, and Transformers (e.g., BERT) to capture complex language patterns and improve detection accuracy.
- Multilingual Support
  The system can be extended to detect cyberbullying in multiple languages, making it applicable to a wider global audience across diverse social media platforms.
- Image and Video Analysis
  Cyberbullying often includes multimedia content; integrating image and video analysis using computer vision techniques can enhance detection capabilities.
- Real-Time Deployment with APIs
  The model can be deployed as a real-time API integrated with social media platforms to automatically monitor and filter harmful content instantly.
- User Behavior Analysis
  Incorporating user behavior patterns and historical activity can help in identifying repeated offenders and improving prediction reliability.
- Explainable AI (XAI) Implementation
  Future systems can include explainable AI techniques to provide transparency in predictions, helping users and moderators understand why content is flagged as bullying.

VII. Conclusion

By applying MAC and ML models to detect cyberbullying using social media, this project has proven that integrating these two disciplines can produce results. Using NLP for text preprocessing and TF-IDF for extracting features, the model is able to classify and assess user-generated data accurately. With four different ML Algorithms (Naïve Bayes, Logistic Regression, Support Vector Machine, and Random Forest), the system is capable of comparing each technique against one another, ultimately selecting the model with the greatest level of accuracy. The ultimate outcome of the project is that it has achieved higher levels of accuracy and reliability to detect both explicit and implicit forms of cyberbullying.

In summary, the proposed model is an automated means for moderating content, allowing for continued automated moderation without relying on humans to perform such activities. The addition of a Sentiment Analysis component allows for an improved understanding of emotional context within the data, leading to an increase in performance in all forms of detecting cyberbullying. Through identifying harmful interactions proactively, this project will help to create a safer and more secure digital space for users. Additionally, through further enhancement of this model and, ultimately, through the deployment of the model in real-time, this solution has the potential for vast applications in social media and the larger digital community.

The proposed work demonstrates the effectiveness of integrating sentiment analysis with machine learning techniques for detecting cyberbullying in social media content. By combining TF-IDF-based feature extraction with emotional polarity analysis, the system is able to capture both the structural and contextual meaning of text, leading to improved classification performance compared to traditional approaches.

The study highlights that conventional rule-based and basic machine learning methods are insufficient for handling the complexity and variability of social media language. Such methods often fail to detect implicit or context-dependent cyberbullying due to their limited understanding of emotional tone and semantic relationships. The proposed hybrid approach addresses these limitations effectively.

The experimental evaluation of multiple classifiers, including Naïve Bayes, Logistic Regression, Support Vector Machine, and Random Forest, confirms that advanced models such as SVM and Random Forest achieve superior performance. These models are better suited for high-dimensional text data and provide more reliable predictions across diverse datasets.

The inclusion of sentiment analysis plays a significant role in improving detection accuracy. By incorporating emotional context into the feature space, the system becomes capable of identifying subtle and indirect forms of cyberbullying that may not contain explicit abusive keywords. This significantly enhances the robustness of the model.

The results also indicate that the proposed system achieves a good balance between precision and recall, ensuring that harmful content is effectively detected while minimizing false alarms. This balance is crucial in real-world applications, where both over-detection and under-detection can have negative consequences.

Furthermore, the system demonstrates scalability and adaptability for real-time deployment in social media environments. Its ability to process large volumes of data efficiently makes it suitable for continuous monitoring of online platforms, contributing to safer digital communication spaces.

The study also emphasizes the practical importance of automated cyberbullying detection systems in promoting mental well-being and digital safety. By proactively identifying and filtering abusive content, the proposed approach supports the development of healthier online communities and aligns with global efforts to ensure responsible digital interaction.

In conclusion, the integration of sentiment analysis with machine learning provides a powerful and efficient framework for cyberbullying detection. The proposed system offers improved accuracy, better generalization, and real-time applicability, making it a strong solution for addressing harmful online behavior and enhancing user safety in social media platforms.

The proposed project demonstrates that integrating sentiment analysis with machine learning significantly enhances the detection of cyberbullying on social media platforms. By combining TF-IDF-based feature extraction with emotional polarity analysis, the system effectively identifies harmful and abusive content that may otherwise go undetected by traditional methods. The comparative evaluation of multiple machine learning algorithms—Naive Bayes, Logistic Regression, SVM, and Random Forest—showed that ensemble and advanced models achieve higher accuracy, precision, and recall, providing a robust framework for real-time monitoring.

Furthermore, the project highlights the practical implications of automated cyberbullying detection in creating safer online environments. The model can be scaled for large datasets and adapted for deployment across various social media platforms. By proactively identifying and filtering abusive content, this system not only enhances user trust but also contributes to mental well-being and digital safety. Future enhancements, including deep learning integration and multilingual support, will further improve its effectiveness and applicability in diverse online communities.

REFERENCES

[1] Abdulnaser M. Fashakh, M. Çevik, Ş. K. Aydoğan, and A. A. Ibrahim, "Detection of cyberbullying using AI and sentiment analysis to examine psychological impacts on vulnerable groups," Egyptian Informatics Journal, 2025.

[2] M. F. Almufareh, N. Jhanjhi, M. Humayun, G. N. Alwakid, D. Javed, and S. N. Almuayqil, "Integrating sentiment analysis with machine learning for cyberbullying detection on social media," IEEE Access, 2025.

[3] Abdullah, I. Latif, N. Hafeez, F. Ullah, G. Sidorov, E. F. Riverón, and A. Gelbukh, "Cyberbullying detection on social media using machine learning techniques," Computación y Sistemas, vol. 29, no. 3, pp. 1567–1586, 2025.

[4] M. Menaka, B. Harini Sri, N. Divya, and A. Kanimozhi, "Cyberbullying detection on social media using machine learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, vol. 11, no. 3, pp. 661–666, 2025.

[5] R. Sharma, S. Mahapatra, et al., "Bias and cyberbullying detection and data generation using transformer artificial intelligence models," Procedia Computer Science, vol. 258, pp. 2232–2243, 2025.

[6] D. Kumar, "A hybrid DeBERTa and gated broad learning system for cyberbullying detection in English text," arXiv preprint arXiv:2506.16052, 2025.

[7] J. Wang, X. Xu, P. Yu, and Z. Xu, "Hierarchical multi-stage BERT fusion framework with dual attention for enhanced cyberbullying detection in social media," arXiv preprint arXiv:2503.00342, 2025.

[8] P. Yi, A. Zubiaga, and Y. Long, "Detecting harassment and defamation in cyberbullying with emotion-adaptive training," arXiv preprint arXiv:2501.16925, 2025.

[9] A. M. Eissa, S. K. Guirguis, and M. M. Madbouly, "An optimized Arabic cyberbullying detection approach based on genetic algorithms," Scientific Reports, vol. 15, 2025.

[10] G. M. Mohiuddin, M. S. Sayeed, and O. L. Yeng, "Deep learning models for culturally aware cyberbullying detection in Muslim societies: A systematic review," Discover Artificial Intelligence, vol. 5, 2025.

[11] H. Allwaibed, M. Anbar, S. Manickam, and A. Bintang, "Cyberbullying detection approaches for Arabic texts: A systematic literature review," Frontiers in Artificial Intelligence, vol. 8, 2025.

[12] D. Maladhy, J. Jeevitha, K. Madhumitha, and J. Subashree, "BERT-based cyberbullying detection," International Research Journal on Advanced Engineering Hub, vol. 3, no. 4, pp. 2007–2015, 2025.

[13] H. O. Aljaloud, K. Dashtipour, and A. Al-Dubai, "Arabic cyberbullying detection: A comprehensive review of datasets and methodologies," IEEE Access, 2025.

[14] A. Abdullah, I. Latif, N. Hafeez, F. Ullah, G. Sidorov, and A. Gelbukh, "Cyberbullying detection on social media using machine learning techniques," Computación y Sistemas, vol. 29, no. 3, pp. 1567–1586, 2025.

[15] T. T. Prama, J. F. Amrin, M. M. Anwar, and I. H. Sarker, "AI enabled user-specific cyberbullying severity detection with explainability," arXiv preprint arXiv:2503.10650, 2025.