Real-Time Cyberbullying Detection in User-Generated Content Using NLP Techniques and SVM

Pola Rishita, Arragokula Abhinay, Jangidi Ashvith Reddy, Sulthani Shirisha, Rajkumar B Patil, Gagandeep Arora

Department of Computer Science and Engineering, Vardhaman College of Engineering, India

rishithapola567@gmail.com, abhiarragokula123@gmail.com, ashvithreddy05@gmail.com, shirishasulthani@gmail.com, yourrajpatil@gmail.com, Gagandeeparora250379@gmail.com

ABSTRACT

Cyberbullying has become a popular issue in online spaces, seriously affecting people's mental well-being and overall safety on the digital front. As social media and other platforms keep growing, harmful behaviours are popping up more often, and old-school manual checks can't keep up with the volume or the need for instant responses. In this work, we introduce a framework for detecting cyberbullying that's both explainable and emotion-aware. It combines machine learning and natural language processing to recognise toxic text. The setup starts with solid text prep: normalising words, removing common stop words, and lemmatising to get everything into a standard form. Then it converts the text into features using TF-IDF (term frequency-inverse document frequency). But it doesn't stop at basic word counts; the model also pulls in deeper context, such as overall sentiment, hints of sarcasm, mentions of gender, and clear signs of personal attacks, to make predictions more reliable. We train a supervised classifier to separate bullying posts from harmless ones, and it outputs a probability score indicating how damaging a message might be. To address the black-box problem with most classifiers, our system provides straightforward explanations in plain language, breaking down the words and context that led to each decision. On top of that, there's a part that detects emotions like anger, sadness, or fear, which often show up in this kind of nasty exchange. We built the whole thing into a user-friendly web app with Streamlit, so it handles live user input or crunches through large batches of data at once. When we tested it, the framework held its own in terms of accuracy while remaining transparent and easy to use in practice. That makes it a good fit for tasks like moderating social media, teaching digital safety in schools, or monitoring online communities.

Keywords*:* Cyberbullying Detection, TF-IDF, Support Vector Machine (SVM), Explainable Artificial Intelligence (XAI), Sentiment Analysis, Sarcasm Detection

1. Introduction

Cyberbullying stands out as one of the biggest types of online harassment. Places like Twitter (now X), Facebook, Instagram, TikTok, and Snapchat draw in billions of people, letting users crank out tons of content through comments, photos and videos, and even live broadcasts [1] [2]. What sets it apart from old-school bullying, the kind that happens face- to-face in spots like schools or offices, is how cyberbullying stays hidden, spreads everywhere, and hits right away [4]. Those behind it can hide their identities with fake names, location tags, or VPNs, which makes it way tougher to track them down and hold them accountable.

Cyberbullying hits people no matter their age, gender, or where they live. That said, teenagers and young adults appear particularly vulnerable, as they are still developing their social identities and spend significant time online. When it happens, victims often face serious mental and social fallout, like dealing with anxiety or depression, pulling away from friends, feeling worse about themselves, thinking about suicide, or even trying it. If it drags on over time, it can lead to lasting problems with schoolwork, relationships, and jobs.

ISBN: 978-81-993813-1-5

Recent reports really drive home how big this problem is. According to the Cyberbullying Research Centre's 2025 report, more than 37% of kids between 12 and 17 run into abusive content every week. A Pew Research Centre survey from the same year puts it at about 41% of U.S. teens who've dealt with online harassment. At the same time, big tech platforms handle billions of posts every day, making it impossible to rely solely on human moderators. On top of that, tougher regulations are coming into effect, such as the transparency rules in the European Union's Digital Services Act from 2024, which require automated moderation that's clear and accountable. [5]

The usual ways of spotting cyberbullying just don't cut it for real-time use. Those keyword filters, for one, miss the bigger picture and regularly overlook sarcasm, veiled threats, cultural slang, or those subtle emotional cues. Even standard machine learning setups like Naive Bayes tend to let too many sneaky or indirect cases slip through, leading to a bunch of false negatives. On top of that, while deep learning models that act like total black boxes can be pretty spot-on, they don't explain themselves well, eroding user trust and making it hard to meet regulatory standards. To address these shortcomings, this paper proposes a real-time, explainable framework for detecting cyberbullying. It combines natural language processing methods with support vector machine classification [6].

The system relies on organised preprocessing, a mix of feature engineering approaches, and contextual signals to boost its reliability. In particular, it merges TF-IDF for text representation with sentiment polarity ratings, hints of sarcasm, detection of gender references, signs of direct attacks, and emotion identification (such as anger, sadness, fear, or neutral). The system performs binary classification to identify bullying or non-bullying content, and produces probabilistic scores indicating the level of risk. What sets it apart from typical methods is its ability to deliver clear, human-interpretable explanations drawn from identified language patterns and emotional cues, which boosts transparency and helps meet regulatory standards [11]. On top of that, it works with both live user inputs in real time and larger batch processes, making it easy to scale up for tasks like moderating social media, running educational tools, or bolstering online safety.

2. Research Methodology

The cyberbullying detection framework is based on machine learning. This approach uses Natural Language Processing techniques and a Support Vector Machine classifier, shown in Figure 1. The goal of the cyberbullying detection framework is to find content that users create. It does this by looking at the words people use, how they feel when they write something, and the context of what they say. The cyberbullying detection framework uses these factors to determine what is harmful.

A. **Data Collection** - The dataset used in this research consists of user-generated text samples collected from online platforms. Each sample is assigned to one of two categories: cyberbullying or non-cyberbullying. The collected dataset is divided into training and testing sets. The training data is used to develop the classification model, while the testing data is used to evaluate the performance of the proposed framework

B. **Data Preprocessing** - The collected text data is cleaned before the model is applied. The preprocessing steps include converting text into lowercase, removing URLs, special characters,

and unnecessary words. Tokenisation and lemmatisation are performed to prepare the text for analysis as per Fig. 1

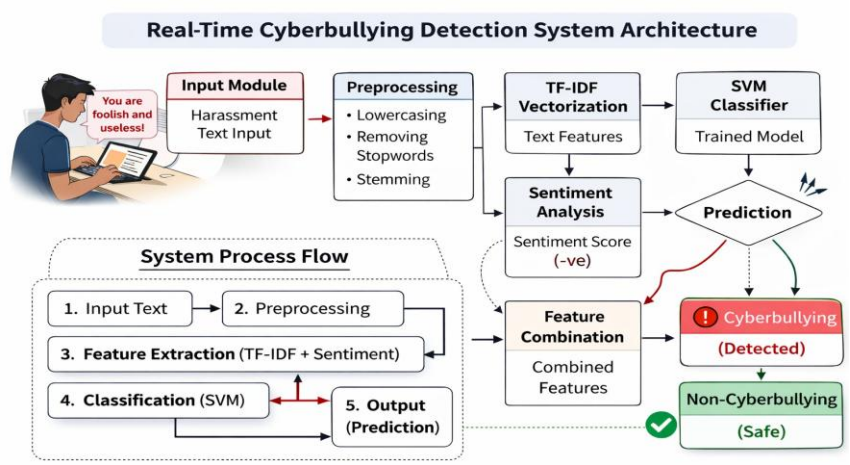

Figure 1: System Architecture of the Proposed Cyberbullying Detection Framework

.

C. **Feature Extraction** - The processed text is converted into numerical features using TF-IDF. Additional features such as sentiment, sarcasm, gender references, and offensive expressions are also considered to improve cyberbullying detection. The features used for identifying cyberbullying patterns are listed in Table 1.

Table 1: Summary of Extracted Features

Feature	Description
TF-IDF	Statistical text representation
Sentiment	Polarity score (-1 to 1)
Sarcasm	Rule-based sarcasm detection
Gender Flag	Gender-related reference detection
Attack Flag	Direct abusive phrase detection

D. **Classification Model** - The extracted features are given to the SVM classifier for training and prediction. The model learns patterns from the data and classifies messages as cyberbullying or non-cyberbullying.

E. **Performance Evaluation** - The performance of the proposed system is measured by comparing the predicted results with the actual labels. Various evaluation metrics, such as accuracy, precision, recall, and F1-score, are used to assess how effectively the model detects cyberbullying content. These measures help analyse the reliability of the proposed approach and assess its ability to classify harmful and non-harmful messages correctly.

3. Theory and Calculation

Given an input message $T_{i2,}$ the system processes the text and generates real-time prediction outputs. Let the dataset be defined as $S = \{M_1, M_2, ..., M_k\}$. Each message undergoes lowercase conversion, URL and special character removal, tokenisation, stopword removal, and lemmatisation. The cleaned dataset is represented as $S_p = g_{\text{preprocess}}(S)$.

To enhance contextual understanding, statistical and rule-based features are combined.

3.1. TF-IDF Representation

The textual representation is computed as:

$$TFIDF(w,m) = TF(w,m) \times \log(N/DF(w) \quad (1)$$

3.2. Additional Linguistic Indicators

The framework extracts the following contextual features:

- Sentiment polarity $S_i \in [-1,1]$
- Sarcasm indicator $Sar_i \in \{0,1\}$
- Gender reference flag $G_i \in \{0,1\}$
- Direct attack flag $A_i \in \{0,1\}$

The final hybrid feature vector is:

$$X_i = [TFIDF_i, S_i, Sar_i, G_i, A_i] \quad (2)$$

3.3. SVM-Based Classification

A Support Vector Machine (SVM) classifier is employed for binary classification. The SVM determines the optimal separating hyperplane:

$$\theta \cdot z + \beta = 0 \quad (3)$$

3.4. Severity and Explainability

The classifier's probability output is transformed into a severity score, computed as Severityi = P(yi = 1) × 100, representing the risk level (0–100). Emotion detection categorises text into Anger, Sadness, Fear, or Neutral using lexicon-based rules. To enhance moderation sensitivity, a safety override assigns Anger if bullying is detected, but no explicit emotion is identified.

3.5. Performance Evaluation

The performance of the proposed hybrid cyberbullying detection framework was evaluated using standard classification metrics. Let P_t, N_t, P_f, and N_f denote true positives, true negatives, false positives, and false negatives, respectively. The evaluation metrics are defined as follows:

$$Accuracy = \frac{Pt+Nt}{Pt+Nt+Pf+Nf}$$

$$precision = \frac{Pt}{Pt+Pf}$$

$$recall = \frac{Pt}{Pt+Nf}$$

$$F1 - score = 2 \times \frac{precision\ X\ recall}{precision+recall}$$

4. Results and Discussion

The results of the proposed real-time cyberbullying detection framework were evaluated using a supervised learning setup with an 80:20 train–test split. The dataset consisted of 8,450 labelled instances, yielding 1,690 test samples. A Support Vector Machine (SVM) classifier was trained using hybrid features that combined TF-IDF representations with contextual rule-based indicators, including sentiment polarity, sarcasm detection, gender reference flags, and

direct attack markers. To assess the impact of contextual augmentation, the proposed hybrid model was compared against a baseline model utilising only TF-IDF features. The baseline model achieved an overall accuracy of 90.12%, while the proposed hybrid model achieved 89.35% on the test set. The detailed performance values obtained from the proposed hybrid SVM model are presented in Table 2.

Table 2: Performance Metrics of the Proposed Hybrid SVM Model

Model	Accuracy	precision	recall	F1-score
Hybrid SVM	89.35%	0.90	0.92	0.91

The confusion matrix for the proposed hybrid SVM classifier is shown in Figure 2, which illustrates the distribution of correctly and incorrectly classified instances.

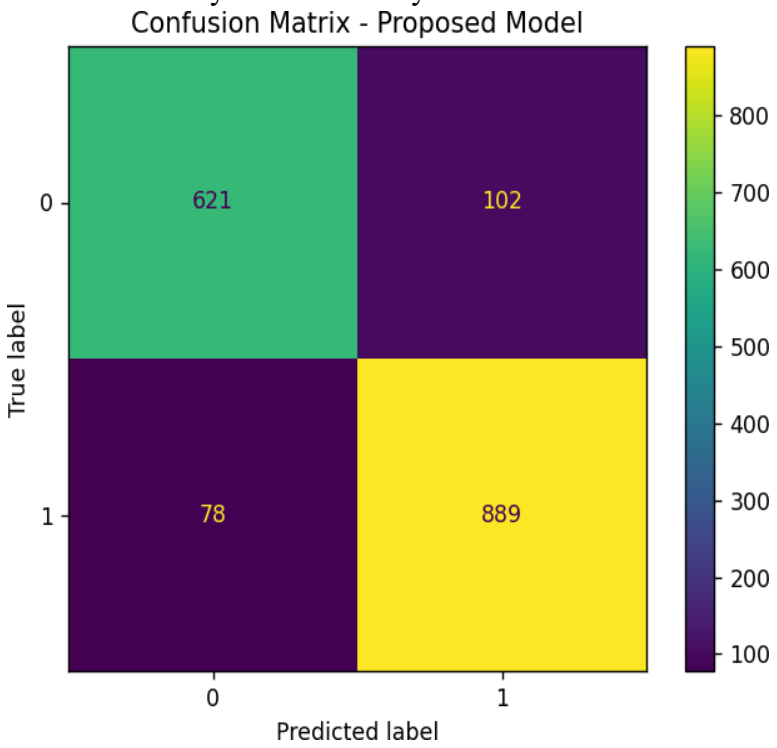

Figure 2: Confusion Matrix of the Proposed Hybrid SVM Classifier

For the cyberbullying category (Class 1), the hybrid model achieved a precision of 0.90, a recall of 0.92, and an F1-score of 0.91 across 967 samples. The higher recall indicates improved detection of abusive content, thereby reducing false negatives, which is critical in real-world moderation systems. For the non-cyberbullying category (Class 0), the model achieved a precision of 0.89, a recall of 0.86, and an F1-score of 0.87 across 723 samples. The slightly lower recall for this class indicates that some non-harmful posts were misclassified as abusive, possibly due to strong emotional expressions. The macro-averaged F1-score of 0.89 and weighted average F1-score of 0.89 demonstrate balanced overall performance across both classes. These results indicate that while the hybrid model does not significantly improve raw accuracy over the baseline, it enhances the reliability of harmful-content detection, making it more suitable for deployment in safety-sensitive social media environments. The classification results for cyberbullying and non-cyberbullying classes are shown in Table 3.

Table 3: Classification Report of the Proposed Hybrid SVM Model

Class	precision	recall	F1-score	support
0 (Non-Cyberbullying)	0.89	0.86	0.87	723
1 (Cyberbullying)	0.90	0.92	0.91	967
Accuracy		0.89		1690
Macro Avg	0.89	0.89	0.89	1690
Weighted Avg	0.89	0.89	0.89	1690

The final output generated by the proposed cyberbullying detection system is shown in Figure 3. The system classifies user-entered text and displays the predicted category along with its confidence score.

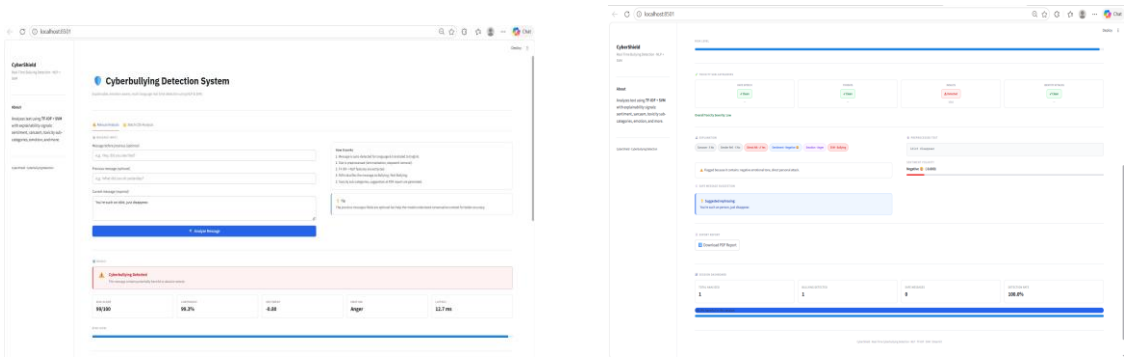

Figure 3: Output of Cyberbullying Detection System

5. Conclusions

The proposed cyberbullying detection framework offers an effective approach to identifying harmful content in user-generated text by combining natural language processing techniques with machine learning. The study focuses on analysing textual patterns and extracting meaningful features to differentiate cyberbullying messages from normal communication. The use of TF-IDF-based feature extraction, along with the Support Vector Machine classifier, improves the system's ability to recognise abusive and harmful online behaviour.

The inclusion of additional factors, such as sentiment information, sarcasm patterns, and offensive expressions, helps the model consider the message's context rather than just individual words. The developed approach shows that automated detection systems can support online platforms by reducing the effort required for manual monitoring and improving user safety. However, the system's performance depends on the quality and diversity of the training dataset, and challenges may arise when dealing with new slang, hidden meanings, or complex forms of cyberbullying. Future improvements can focus on using larger and more diverse datasets, advanced deep learning techniques, and real-time integration with social media platforms to enhance detection accuracy. The framework can also be extended by adding multilingual support and improved explainability features to help users and moderators understand the reasons behind each prediction. Overall, this research demonstrates the potential of combining NLP and machine learning techniques to create reliable tools for identifying cyberbullying and promoting safer digital environments.

Acknowledgements

The authors would like to thank Vardhaman College of Engineering for providing the necessary support and resources to conduct this research.

Funding source

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. F. Almufareh, N. Z. 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, vol. 13, pp. 78348–78359, 2025, doi: 10.1109/ACCESS.2025.3558843.

[2] S. Unnava and S. R. Parasana, "A study of cyberbullying detection and classification techniques: A machine learning approach," Engineering, Technology & Applied Science Research, vol. 14, no. 4, pp. 15607–15613, Aug. 2024, doi: 10.48084/etasr. 7621.

[3] S. Aliyu, K. S. Niksirat, K. Huguenin, and M. Cherubini, "On the role and form of personal information disclosure in cyberbullying incidents," Proceedings on Privacy Enhancing Technologies, vol. 2023, no. 4, pp. 468–483, Aug. 2023, doi: 10.56553/popets-2023-0120.

[4] 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.

[5] E. I. Galyashina and V. D. Nikishin, "The concepts of aggressive information impact through the lens of internet users' worldview security," Journal of Siberian Federal University: Humanities & Social Sciences, vol. 14, no. 11, pp. 1660–1673, Nov. 2021, doi: 10.17516/1997-1370-0848.

[6] R. Bayari and A. Bensefia, "Text mining techniques for cyberbullying detection: State of the art," Advances in Science, Technology and Engineering Systems Journal, vol. 6, no. 1, pp. 783–790, Feb. 2021, doi: 10.25046/aj060187.

[7] S. Joshi, H. G. Nagariya, N. Dhanotiya, and S. Jain, "Identifying fake profile in online social network: An overview and survey," in Communications in Computer and Information Science, vol. 1240, pp. 17–28, 2020, doi: 10.1007/978-981-15-6315-7_2.

[8] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," Social Network Analysis and Mining, vol. 12, no. 1, p. 99, Dec. 2022, doi: 10.1007/s13278-022-00934-4.

[9] R. S. Gün and G. G. Akduman, "What is cyberbullying?" in Bullying in Media and Beyond, Turkey: IGI Global, pp. 473–485, doi: 10.4018/978-1-6684-5426-8.CH028.

[10] S. Alotaibi and A. Mehmood, "A deep analysis of textual features based cyberbullying detection using machine learning," in Proc. IEEE Global Conf. Artificial Intelligence and Internet of Things (GCAIoT), Dubai, UAE, Dec. 2022, doi: 10.1109/GCAIOT57150.2022.10019058.

[11] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT," Information, vol. 14, 2023, doi: 10.3390/infoXXXXX.

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

[13] Z. Hadi, E. Suryadi, A. Akbar, Zaenudin, and R. Muslim, "Cyber bullying sentiment analysis based on social categories using the chi-square test," Journal of Computer Technology, vol. 2, no. 1, pp. 1–9, Jul. 2024.

[14] 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," Journal of Computational Biomedical Informatics, vol. 6, no. 2, pp. 1–12, 2024, doi: 10.56979/602/2024.

[15] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi, "A hybrid transformer-based model for optimizing fake news detection," IEEE Access, vol. 12, pp. 160822–160834, 2024, doi: 10.1109/ACCESS.2024.3476432.

[16] 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.

[17] R. Maskat, M. F. 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, Computing and Artificial Intelligence, Dec. 2020, pp. 1–6, doi: 10.1145/3446132.3446412.

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

[19] Shrivastava, G., Gupta, D., & Sharma, K. (Eds.). (2021), “Cyber crime and forensic computing: Modern principles, practices, and algorithms,” De Gruyter.

[20] P. Parameswaran, A. Trotman, V. Liesaputra, and D. Eyers, "Detecting the target of sarcasm is hard: Really?" Information Processing & Management, vol. 58, no. 4, Jul. 2021, Art. no. 102599.