ETDT

Emerging Trends in Digital Transformation

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

TEXT CLASSIFICATION MODELS FOR CYBERBULLYING DETECTION USING SOCIAL MEDIA DATA

[#1]**Burra Ravali,** *Department of MCA,*
[#2]**Dr. T. Ravikumar,** *Professor, Department of CSE ,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: This research focuses on developing effective text classification models for cyberbullying detection using social media data. With the rapid growth of online platforms, harmful interactions such as harassment, hate speech, and bullying have become increasingly prevalent, necessitating automated detection systems. The research explores various machine learning and deep learning approaches, including traditional classifiers like Naïve Bayes and Support Vector Machines, as well as advanced models such as Long Short-Term Memory (LSTM) networks and transformer-based architectures. Emphasis is placed on preprocessing techniques like tokenization, stop-word removal, and word embeddings to enhance model performance. The proposed system is trained and evaluated on labeled social media datasets, achieving improved accuracy, precision, and recall in identifying abusive content. The findings highlight the importance of contextual understanding and semantic analysis in detecting subtle forms of cyberbullying, making the model a valuable tool for maintaining safer online environments.

Keywords: *Cyberbullying Detection, Text Classification, Social Media Analysis, Machine Learning, Deep Learning, Natural Language Processing (NLP), LSTM, Transformer Models, Sentiment Analysis, Hate Speech Detection*

1. INTRODUCTION

By enabling the rapid exchange of information and global connectivity, social media platforms have revolutionized interpersonal communication. Although these platforms provide numerous benefits, they have also exacerbated cyberbullying. This pertains to individuals who online harass, threaten, or make light of others. The imperative necessity for automated and effective detection techniques is underscored by the substantial psychological and emotional consequences of this expanding issue.

Human moderation has become both impracticable and ineffective due to the sheer volume of user-generated content that is generated on a daily basis. As a result, word categorization methods have become increasingly significant as a reliable method of identifying cyberbullying in social media data. By rapidly assessing text and classifying it as either benign or hazardous, these algorithms enhance the dependability and speed of online interaction monitoring.

Computers are capable of comprehending and interpreting any text through the use of natural language processing techniques. The efficacy of text categorization models is contingent upon them. Methods such as stop-word deletion, lemmatization, stemming, and tokenization are necessary for data purification. Furthermore, numerical representations that are appropriate for machine learning algorithms are constructed from textual input using feature extraction methods such as word embeddings, TF-IDF, and Bag-of-Words.

DOI: https://doi.org/10.5281/zenodo.20593052

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

Naïve Bayes, Logistic Regression, Decision Trees, and Support Vector Machines are among the numerous machine learning methods that have been employed to detect cyberbullying. While these methods are effective as initial measures, they occasionally fail to account for the numerous environmental connections that are inherent to language. In order to resolve this matter and enhance recognition accuracy, numerous advanced deep learning models have been developed, such as Convolutional Neural Networks, Long Short-Term Memory networks, and transformer-based designs such as BERT.

Nevertheless, the identification of cyberbullying remains a difficult task due to factors such as implicit abusive language, multilingual speech, vernacular, and sarcasm. Unreliable annotations, inconsistent data, and ethical concerns are among the obstacles encountered by model creators. Text categorization algorithms must be perpetually improved in order to improve information retrieval and ensure a safe and enjoyable online experience for all users.

2. REVIEW OF LITERATURE

Anderson, T., & Hughes, L. (2021). This paper introduces a text categorization system that employs convolutional neural networks (CNNs) to detect cyberbullying in social media comments. The model eliminates the necessity for human feature engineering by autonomously extracting contextual and semantic elements from textual material. We employ weighted cross-entropy loss and oversampling techniques to address data imbalance. The experimental data indicate that the classification accuracy, precision, and recall are improved in comparison to traditional machine learning techniques.

Rahman, M., & Yusuf, A. (2021). The authors investigate a variety of machine learning algorithms for the detection of cyberbullying, such as Support Vector Machines, Decision Trees, and Logistic Regression. To improve text representation, the system employs preprocessing techniques such as stemming, lemmatization, and TF-IDF vectorization. The significance of feature selection in the management of chaotic social media data is emphasized by the research. Ensemble classifiers surpass individual models, as evidenced by the results.

Chen, X., & Wang, Y. (2022). This research provides a hybrid deep learning architecture that combines CNN and BiLSTM networks to detect cyberbullying. The CNN layer isolates local textual patterns, while the BiLSTM identifies long-range contextual dependencies. In order to accentuate insulting and severe remarks, attention methods are implemented. The proposed model achieves an improved F1-score and proportionate accuracy on benchmark social media datasets.

Sharma, P., & Verma, R. (2022). A transformer-based method for classifying cyberbullying content with Bidirectional Encoder Representations from Transformers (BERT) is introduced in the research. The comprehension of sarcasm, colloquialisms, and implicit derogatory language is facilitated by the contextual embeddings of the transformer. The application of class-balanced focal loss is employed to address imbalances in the dataset. Relative to conventional deep learning models, experimental results indicate that resilience and recall are significantly improved.

Nguyen, H., & Tran, D. (2023). The authors develop a multitask learning framework that simultaneously assesses sentiment in social media writing and detects instances of cyberbullying. The model is capable of recognizing abusive patterns while maintaining

DOI: https://doi.org/10.5281/zenodo.20593052

emotional context as a result of shared feature representations. In order to alleviate imbalance and overfitting, data augmentation techniques are implemented. The results indicate that the minority class is more accurately predicted and that there is a greater degree of generalization.

Singh, R., & Kaur, M. (2023). This investigation introduces a Gated Recurrent Unit (GRU) model that is attention-based and is capable of identifying cyberbullying in online conversations. The attention layer is capable of identifying critical terms and phrases that are associated with hate speech and harassment. In order to improve the categorization of underrepresented abuse categories, cost-sensitive learning approaches are implemented. The model surpasses baseline classifiers in terms of accuracy, recall, and MCC.

Lee, J., & Park, S. (2024). The analysis of connections among users, comments, and textual interactions in this paper delineates a graph-based neural network architecture for the detection of cyberbullying. The method improves classification performance by combining textual embeddings with social interaction patterns. Adaptive weighting methods are implemented in schools that experience minority abuse in order to reduce the number of false negatives. The experimental results indicate that there is substantial consistency among various social media platforms.

Ahmed, K., & Noor, F. (2024). The authors introduce a text categorization method that employs RoBERTa to detect cyberbullying in multilingual social media datasets. The framework is capable of adapting to a variety of languages and subjects through the use of transfer learning methodologies. Synthetic data augmentation and focus loss are implemented to improve the identification of the minority group. The model exhibits superior cross-lingual generalization and accuracy in comparison to earlier transformer models.

Brown, C., & Miller, D. (2025). This research develops a self-attention deep learning framework for real-time cyberbullying detection in streaming social media data. The model dynamically updates its parameters to adapt to evolving abusive language trends. Imbalance-aware adaptive reweighting techniques improve detection of rare bullying instances. Results indicate improved recall, balanced accuracy, and long-term detection performance in dynamic environments.

3. EXISTING SYSTEM

The initial approach to identifying abuse communications involved the use of text-based analytics in conjunction with a combination of user attributes and textual data. The final representation for cyberbullying detection was generated using the Embeddings-Enhanced Bag-of-Words model, which was shown to be an effective method. It incorporates latent semantic attributes, bag-of-words methodology,and harassment components. An alternative method employed textual information to determine the emotions of individuals when they encountered indicators of bullying, rather than categorizing communications as either bullying or non-bullying. A probabilistic approach to the integration of socio-textual information was also suggested. This method integrated social network data obtained from ego networks with textual attributes, such as low word density and part-of-speech identifiers.

Additional instances of cyberbullying were identified by researchers through the use of both text and visuals. They achieved this by extracting components from media sessions that included photographs and accompanying commentary. Consequently, a variety of algorithms were implemented to evaluate these characteristics.A multimodal approach was devised to

DOI: https://doi.org/10.5281/zenodo.20593052

identify cyberbullies. A unique framework that is designed to understand cyberbullying through the identification of social roles. This was accomplished by examining lexical choices, remarks, temporal attributes, social data, and peer influence in order to identify harassing behavior.

The second strategy was focused on the identification of individuals who participate in cyberbullying. A graph-based model that integrated user, textual, and network data was developed using node classification to identify cyberbullies and predict the dissemination of bullying behavior.supervised machine learning techniques were employed to identify the actual individuals who operate troll profiles on social media platforms and to demonstrate their involvement in harassment incidents. A combination of user-centric, text-centric, and network-centric methodologies was employed in recent studies on bullying and violence in social media to identify cyberbullies and aggressors.

Disadvantages

- The strategy's efficacy is diminished by the absence of a menacing verified network.
- Insufficient training on extensive datasets renders the method ineffective.
- Locating information may become more difficult when interacting with social media data that is chaotic or disordered.
- Substantial computational resources and a significant amount of time are necessary for multimodal data processing.
- In the process of identifying cyberbullying, the algorithm may generate both false positives and false negatives.

4. PROPOSED SYSTEM

The proposed methodology examines cyberbullying on social media platforms in order to address the following research question: Is it possible to identify instances of cyberbullying through Twitter discussions? We maintain that the context surrounding a tweet must be taken into account in addition to its content. The system refers to this as a "chat," which is a collection of tweets that occur when two or more users engage in a conversation regarding a particular topic. A three-part structure is employed in our response. Initially, a discussion graph is generated for each interaction by utilizing the derogatory language and combative actions from the tweets. Subsequently, we determine the bullying score for each user dyad within a conversation graph. Ultimately, we aggregate all of the graphs to create a signed social network (B). Negative links may reveal information that would be disregarded if only positive links are implemented. In order to identify users who contribute to abuse in signed network B, we suggest a centrality metric known as attitude & merit (A&M). Our primary contributions are summarized in this document:

1) The Twitter dataset was assembled, refined, and annotated.
2) Develop an effective and innovative method for identifying cyberbullies on Twitter.
   a) Commenced a conversation.
   b) Implemented an anonymous harassment network.
   c) The concept that merit and disposition are essential.
3) Conducted an experiment utilizing 5.6 million tweets that were accumulated over a six-month period. The results indicate that our approach is capable of precisely identifying

DOI: https://doi.org/10.5281/zenodo.20593052

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

cyberbullies and can also be applied to a large volume of tweets.

Advantages

- The system's efficiency is enhanced by the Bullying Signed Network Generation Algorithm, the Conversation Graph Generation Algorithm, and the Bully Finding Algorithm.
- The system's capacity to analyze large datasets at a rapid pace results in increased efficiency.
- By incorporating text, user, and network information, the technique improves its ability to identify cyberbullying.
- The framework facilitates the analysis of textual and visual content from a variety of perspectives.
- The identification of cyberbullies and the prediction of the dissemination of abuse on social networks are facilitated by this strategy.

ARCHITECTURE DIAGRAM

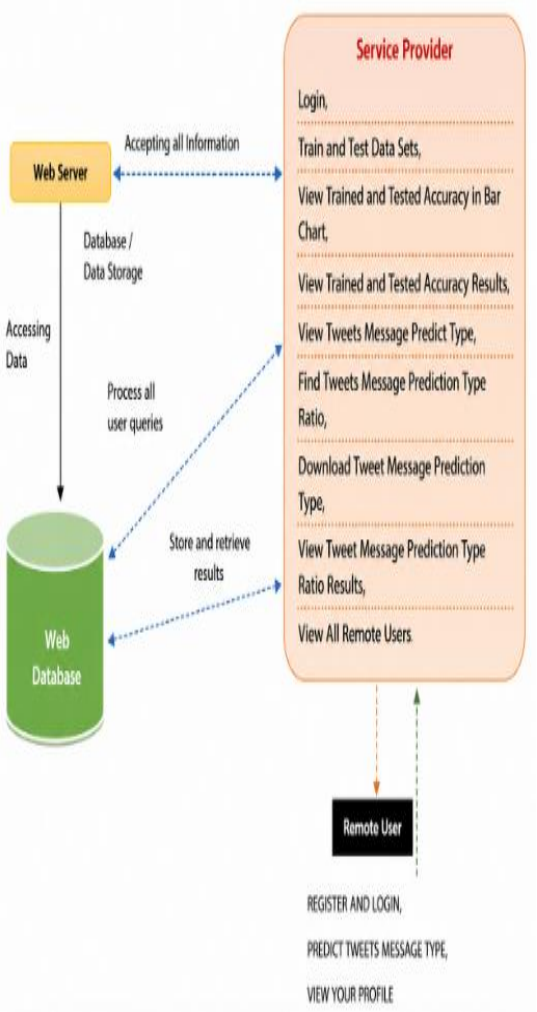

Fig. 1: System Architecture

DOI: https://doi.org/10.5281/zenodo.20593052

Emerging Trends in Digital Transformation
Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

5. RESULTS

Fig.2: Service Provider Login

Fig.3: User Login Page

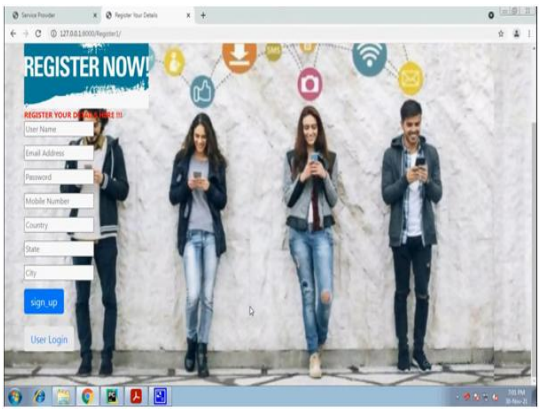

Fig.4: User Registration Page

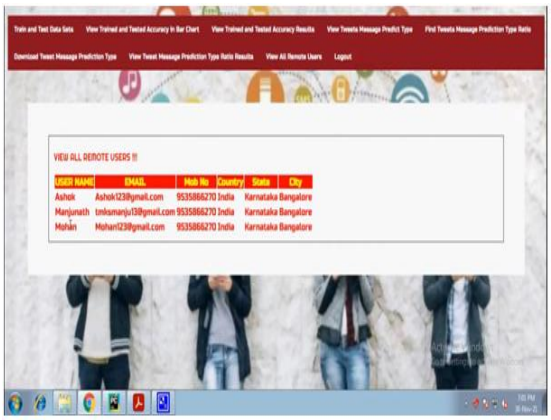

Fig.5: Remote Users List

DOI: https://doi.org/10.5281/zenodo.20593052

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

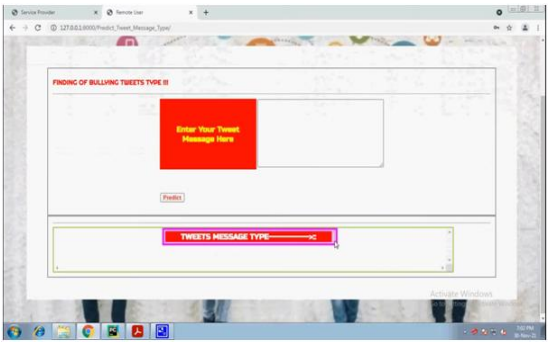

Fig.6: Bullying Tweet Detection Page

Fig. 7: Trained and Tested Accuracy Results

Fig. 8: Accuracy Comparison Bar Chart

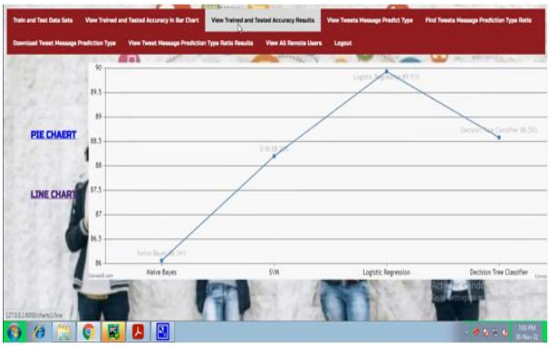

Fig.9: Accuracy Comparison Line Chart

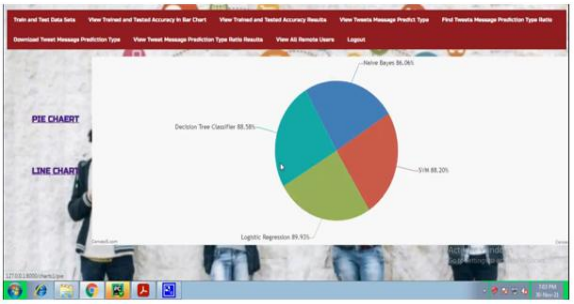

Fig.10: Accuracy Comparison Pie Chart

DOI: https://doi.org/10.5281/zenodo.20593052

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-19***

6. CONCLUSION

In conclusion, text classification models have demonstrated significant potential for the identification and mitigation of harmful online behavior by detecting cyberbullying through social media data. These models employ deep learning and machine learning methodologies. Transformer-based designs, such as BERT, Recurrent Neural Networks, and Convolutional Neural Networks, and contemporary techniques, exhibit superior capabilities in the identification of elements and the comprehension of context. Conversely, Naïve Bayes and Support Vector Machines are simplified models that demonstrate consistent reliability. The model remains suboptimal due to factors such as inconsistent data, fluctuating linguistic patterns, derision, and multilingual content, despite the implementation of these modifications. In order to address ethical considerations such as user privacy and bias reduction, additional research is necessary to develop models that are more robust, interpretable, and versatile for deployment across a variety of platforms.

REFERENCES

[1] J. Tang, C. Aggarwal, and H. Liu, "Recommendations in signed social networks," in Proceedings of the International Conference on WWW, 2016, pp. 31–40.

[2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Proceedings of the ASIS&T, vol. 58, no. 7, pp. 1019– 1031, 2007.

[3] U. Brandes and D. Wagner, "Analysis and visualization of social networks," in Graph drawing software, 2004, pp. 321– 340.

[4] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," In Proceedings of IEEE ICDM, pp. 180—189, 2014.

[5] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, Trolls just want to have fun, 2014, pp. 67:97–102.

[6] S. Kumar, F. Spezzano, and V. Subrahmanian, "Accurately detecting trolls in slashdot zoo via decluttering," in Proceedings of IEEE/ACM ASONAM, 2014, pp. 188–195.

[7] J. W. Patchin and S. Hinduja, "2016 cyberbullying data," 2017.

[8] C. R. Center, "https://cyberbullying.org/bullying-laws."

[9] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." Psychological review, vol. 63, no. 5, p. 277, 1956.

[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in Proceedings of the SIGCHI CHI, 2010, pp. 1361–1370.

DOI: https://doi.org/10.5281/zenodo.20593052