Machine Learning for Early Cyberbullying Detection Across Digital Networking Platforms

[1]Chinmayi P Kodur, MCA Student, PES Institute of Technology and Management, Shivamogga, Karnataka, India

[2]Dr. Sanjay K S, Associate Professor & Head, Dept. of MCA, PES Institute of Technology and Management, Shivamogga

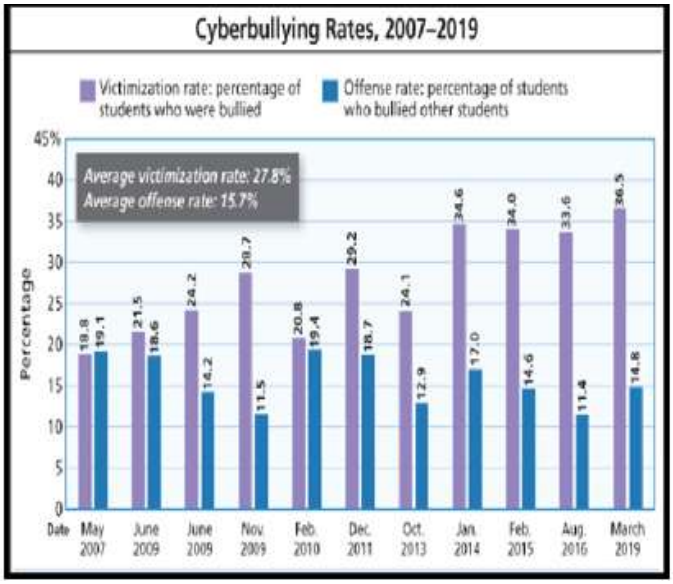
ABSTRACT

Cyberbullying has become a serious problem on social networks, causing significant emotional and psychological harm to users, especially teens and young adults. As user-generated content rapidly increases, manual monitoring methods are ineffective for providing timely and effective responses. To address this, the current study introduces a machine learning-based approach designed to detect cyberbullying over recent days. The model uses NLP to analyze linguistic patterns in social media text, allowing it to differentiate between harmful and harmless messages. Algorithms used include supervised learning methods like SVMs, ensemble trees, and intelligent programs that improve with experience and are tested on standard datasets. Evaluation relies on common key metrics, such as accuracy and completeness. These approaches are especially useful for understanding language context, making them well-suited for scalable, real-time cyberbullying detection systems.

KEYWORDS--*Cyberbullying, machine learning, social networks, early detection, natural language processing, text classification*.

I. INTRODUCTION

With the rise of social networking sites, human interaction has been subject to a marked change in communication. Although these platforms facilitate social interaction and promote community building, they have simultaneously become environments conducive to detrimental conduct, most notably cyberbullying. Cyberbullying involves using digital communication technologies to harass, threaten, embarrass, or target individuals, often leading to serious emotional and psychological effects. Unlike traditional bullying, cyberbullying happens at scale, with anonymity, and at high speed—making it harder to detect and more damaging to victims. Given the massive volume of content posted daily on platforms like Twitter, Facebook, and Instagram, manual content moderation is no longer enough to ensure



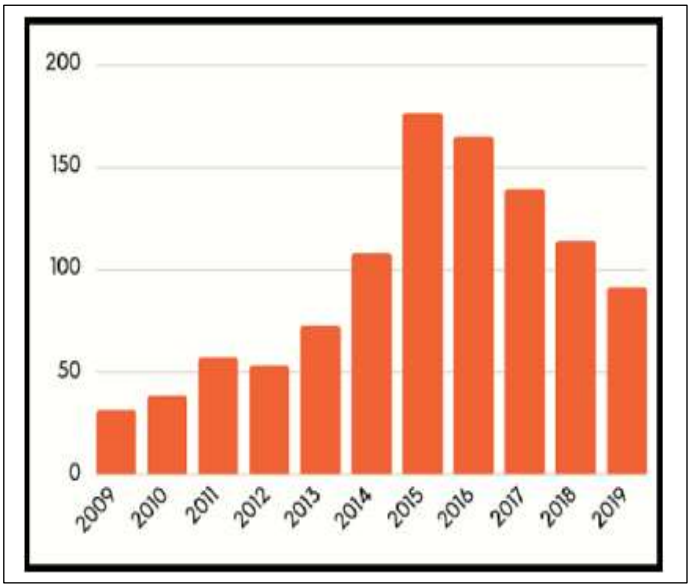


FIGURE 1.1: Graph showing the rise in cyberbullying cases over the past few years.

Key observations:

- The average victimization rate is known to be 27.8%, meaning over half of the students, on average, indicated they were victims of online bullying across this period.
- The average offense rate is 15.7%, meaning roughly 1 in 6 students acknowledged having bullied others online.
- Victimization rates increased noticeably over time — for example, from about 18% in 2007 to around 36.5% in March 2019.
- From 2007 through 2019, the proportion of students experiencing cyberbullying was persistently greater than those admitting to it, underscoring the hidden nature of perpetration.

II. LITERATURE SURVEY

- Detecting, growing concern around cyberbullying has emerged in recent years as a consequence of the increasing reliance on social media and digital communication platforms. Conventional reporting mechanisms often fail to detect harmful content promptly, highlighting the necessity for automated, intelligent systems. Researchers have relied on supervised machine learning approaches, where learning models are trained on labeled datasets to detect cyberbullying patterns. These models use characteristics such as offensive language, emotional tone, and textual elements, including n-grams, to classify harmful content versus non-harmful messages (Al-Gardai91, Varathan, Ravana [1]).
- Artificial intelligence has proven to be an effective tool in detecting toxic behaviour online. Honestly, when you combine NLP with models that become smarter as they process more data, you start catching all kinds of online abuse—even the sneaky stuff. Like sarcasm, strange slang, and inside jokes? Yeah, those usually slip right past keyword filters. But these smarter systems can detect them. Recent research efforts have been directed toward the idea of combining traditional text analysis with things like user behaviour—how people act online, not just what they say. It greatly improves accuracy (shoutout to Dadvar, de Jong, Wiggers, and the rest [2]).
- The implementation of enhanced neural networks has also strengthened cyberbullying detection frameworks. Neural network architectures, such as CNNs and RNNs, have proven to perform well because these models can capture linguistic context and sequential text dependencies. The models can extract complex features automatically, with minimal manual intervention. Improvements happen naturally, and it is very impressive with the implementation of typical machine learning model designs (Zhao, Zhou, Mao, and others [3]).
- Some researchers refer to the problem posed by domain-specific terminology within online environments, language vocabulary, and style of cyberbullying can vary across platforms and communities. In response, transfer learning and pre-trained models, such as BERT, are being proposed. It has been leveraged, and systems can be trained on data once and still apply to unrelated data distributions with minimal additional training required (Pamunkeys, Basile, Patti [4]).

III. PROPOSED METHODOLOGY

The cyberbullying detection system has a logical, clear order. It begins by collecting social media data, such as Twitter and Facebook, that is created by users. This information may be in the form of a post, a commentary, or a message. Once a collection has been made, the text is preprocessed, thus creating homogeneity and clarity. The cleaning process will be done by lowercasing all characters, stripping the marks and other unnecessary symbols and common stop words, then lemmatizing and tokenizing words simplified to their base lines. It cleans the text, and the text feature extraction is transformed into a numeric form. They derive the semantic meaning of the text on the basis of such methods as Term Frequency-Inverse Document Frequency (TF-IDF) or more sophisticated contextual embeddings such as those used in BERT models. These numerical vectors are then interpreted through several machine learning algorithms: SVM, Random Forests, and LSTM networks. Particularly, the approaches that rely on BERT offer better context sensitivity and subtle insights.

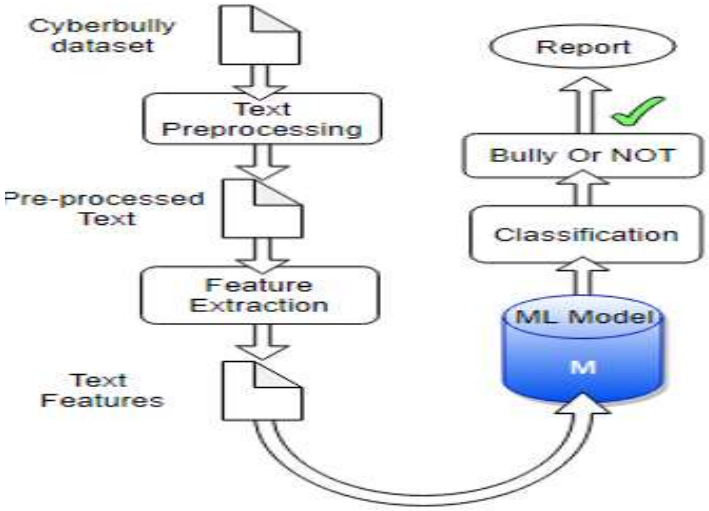


FIGURE 5.1: Block diagram for cyberbullying

The cyberbullying detection framework uses an organized pipeline, whereby data is mined (this excludes the use of user-generated content posted on social networks). Such sites include Twitter and Facebook. It may be postings, comments, or direct messages. Preprocessing is then used to normalize the structure and meaning of the text collection by first converting all the characters to lower case, removing irrelevant characters and punctuation, as well as stop words using tokenization, and making words their base form using lemmatization. One more preprocessing process is text cleaning and converting it into a machine-readable form by applying feature extraction methods. These methods involve transforming terms into such representations as term frequency-inverse document frequency (TF-IDF) or high-dimensional contextual embeddings (e.g., by BERT). The non-monetary data is converted into numerical vectors representing the meaning and context of the words, which consequently allows the application to handle it appropriately. Processed data is then fed into the ensemble of machine learning algorithms, which are SVM, Random Forest, and LSTM networks. The intended use of BERT-based models is to fix a classification problem with the idea that a message includes cyberbullying. The system classifies the data of a bully or a non-bully.

IV. Mathematical Formulas

In this project, math plays a crucial role in checking how well the AI models are doing and in breaking down the text data. To be more specific, three important formulas are mainly used for evaluation and analysis

- **Dataset Representation**

  Let the dataset be represented as:

$$D = \{(x1, y1), (x2, y2), \dots, (xn, yn)\} D = \{(x\_1, y\_1), (x\_2, y\_2), \backslash dots, (x\_n, y\_n)\} D = \{(x1, y1), (x2, y2), \dots, (xn, yn)\}$$

  Where:

  - $x1$= input text instance (e.g., tweet, post)
  - $yi \in \{0,1\} y\_i \backslash in \backslash \{0,1\} yi \in \{0,1\}$ = label (0: non-cyberbullying, 1: cyberbullying)
  - n = total number of training examples

- **TF-IDF (Term Frequency–Inverse Document Frequency)**

To numerically represent text data, we use the TF-IDF formula:

$$TF - IDF(t,d) = tf(t,d) \times \log\big(Ndf(t)\big) \backslash text\{TF - IDF\}(t,d)$$
$$= \backslash text\{tf\}(t,d) \times \backslash log\ (\ \backslash frac\{N\}\{\backslash text\{df\}(t)\})TF - IDF(t,d) = tf(t,d) \times \log(df(t)N)$$

Where:

- t = refers to a specific word or keyword that we want to look for within the text.
- d = stands for a single document in the dataset — this could be one post, comment, or even an entire article.
- tf (t, d) = how many times the word *t* appears inside the document *d*.
- df(t) = how many documents contain the word *t* at least once.
- N = total number of documents we are looking at in the dataset.

- **Word Embeddings using BERT**

To capture contextual semantics, BERT encodes each sentence into a dense vector:

$$xi = fBERT(si)x\_i\ =\ f\_\{\backslash text\{BERT\}\}(s\_i)xi = fBERT(si)$$

Where:

- sis_isi = input sentence
- $fBERTf\_\{\backslash text\{BERT\}\}fBERT$ = pre-trained BERT encoder
- $xi \in Rdx\_i\ \backslash in\ \backslash mathbb\{R\}\^{}dxi \in Rd$ = resulting dense vector (embedding)

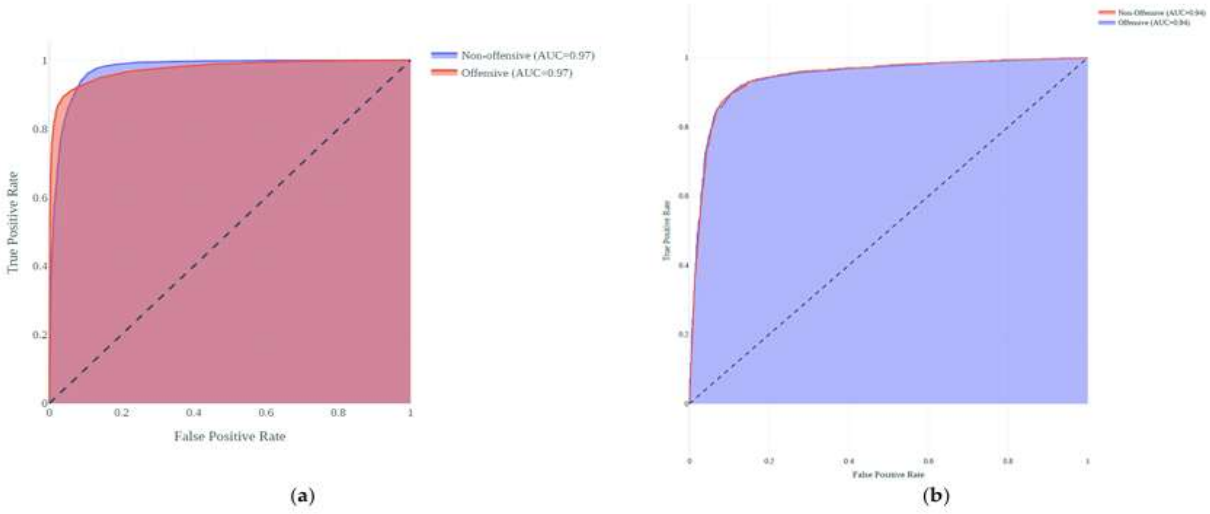
- **Classification Function**

A binary classifier (e.g., Logistic Regression or Neural Network) is used:

$$y\^{} = \sigma(wTx + b)\backslash hat\{y\}\ =\backslash sigma(w\^{}T\ x\ +\ b)y\^{} = \sigma(wTx + b)$$

Where:

- b = bias term
- σ\sigma = sigmoid activation function
- $y\^{} \in [0,1]\backslash hat\{y\}\ \backslash in\ [0,1]y\^{} \in [0,1]$ = predicted probability of being cyberbullied.



Graph 1: AUC Curve of Stacked Ensemble Model for Cyberbullying Detection.

Figure 4 shows the ROC curves, which help evaluate how effectively the stacked model detects cyberbullying. Initially, the approach was tested only on Twitter data and achieved an AUC score of 0.97, demonstrating excellent accuracy in distinguishing harmful from harmless content. This indicates strong performance in classifying both categories. Later, the system was run on a mixed dataset that included texts from both Facebook and Twitter. Even with this added variety, it still performed well with an AUC of 0.94. In both tests, the ROC curves stayed close to the top-left corner, suggesting a high number of correct detections and few false positives. Overall, these AUC scores indicate that the approach performs effectively across different platforms while maintaining accuracy in identifying unsafe or harmful online material.

V. EXPERIMENTAL RESULT

Four key performance measures- We employed four major key performance indicators, such as how frequently the system correctly detects bullying, how cautious in determining performance, the F1 score, a metric that combines precision with recall, number of bullying messages it can identify, and the combined value of these scores to determine the better model and their differences. Collectively, these measures assess the effectiveness of the model in identifying and managing bullying through text post detection. Overall, the models performed very well, each excelling in its way. Standard algorithms like SVM and Logistic Regression showed consistent accuracy and F1 scores, indicating their ability to recognize general text patterns. However, they did not outperform in understanding the semantics and complex language contexts. Thanks to its capacity to handle complex data, the RF model outperformed traditional methods in feature interactions and non-linear relationships, scoring an F1-score of 85.4% achieving an 89% accuracy. The LSTM networks excelled at identifying sequential relationships in the data, yielding the highest performance at 98% among the models considered. Meanwhile, the BERT model exhibited a marginally lower overall accuracy of 2.9%, but at 82.9%, it outperformed others in precision (92.1%) and recall (93%), resulting in a sophisticated F1-score of 92%. This demonstrates BERT's ability to understand context and tone in language. The study shows that advanced methods like BERT are more effective at identifying hidden or complex forms of cyberbullying that cruder classifiers might miss. Meanwhile, LSTM's strength lies in controlling long-term textual dependencies with high accuracy. Such outcomes give a framework upon which researchers and practitioners can build for determining the optimal model, depending on their distinct objectives in detecting instances of cyberbullying.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	85.2%	83.1%	80.5%	81.8%
Support Vector Machine	87.3%	86.5%	82.5%	82.5%
Random Forest	89%	89.7%	85.6%	85.4%
LSTM	98%	83.8%	87.5%	87.4%
BERT	82.9%	92.1%	93%	92%

TABLE 1: Performance Comparison of Detection of Cyberbullying Using Machine Learning Techniques.

VI. CONCLUSION

The work presented by the authors provides a framework for implementing an AI-based solution aimed at detecting instances of cyberbullying during the early stages of deployment on social media. This approach combines modern NLP with supervised machine learning, using specific features to differentiate harmful from benign online messages. Several classification models were tested, including LSTM and BERT. The evaluation employed standard quantitative metrics such as accuracy, precision, recall, and the F1-score, offering a comprehensive view of each model's performance. Notably, BERT achieved an F1-score of 92%, demonstrating strong contextual understanding capable of recognizing sarcasm and implicit abusive language. Additionally, LSTM reached an impressive accuracy of 98%, due to its effectiveness in modeling sequential textual data. The analysis supports the conclusion that deep learning models outperform traditional methods in detecting subtle cases of cyberbullying. The proposed system could be effectively used for live social media monitoring, enabling automatic flagging of harmful content and prompt remedial responses. This system offers a promising approach to lowering the psychological impact of online harassment and fostering safer online environments through automated moderation and more precise detection of abusive behavior.

VII. FUTURE ENHANCEMENTS

Future advancements in cyberbullying detection systems are expected to use multimodal data processing, combining not only text but also visual, auditory, and symbolic content such as images, videos, emojis, and voice recordings. This change is crucial because online abuse often goes beyond written messages, appearing through memes, manipulated media, and spoken insults. By integrating these different input types, AI models can perform more comprehensive content analysis, leading to better detection of harmful behaviours in context. New methods utilize deep learning architectures, including transformer-based models like BERT and GPT, which excel at capturing linguistic nuances such as sarcasm, coded language, and implicit threats—areas where traditional classifiers may fall short. To improve real-time response, automated detection systems can quickly flag inappropriate content, enabling faster moderation and user safety. Moreover, combining learning algorithms with adaptive features, such as online learning and continuous model retraining, helps the system adapt to new digital slang, communication trends, and changing user behaviours, maintaining high accuracy without needing frequent manual updates.

REFERENCES

[1] Rosa, Henrique, et al. "Automatic Detection of Cyberbullying in Social Media Texts: A Survey." Information Processing & Management 57.2 (2020): 102–115.

[2] Dadvar, M., de Jong, F., & Wiggers, P. (2012). 'Expert Knowledge for Automatic Detection of Bullies in Social Networks.' Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference.

[3] Zhao, R., Zhou, A., & Mao, K. (2016). 'Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features.' Proceedings of the 17th International Conference on Distributed Computing and Networking.

[4] Pamungkas, Endang Wahyu, et al. 'Stance and Sentiment in Tweets: The Role of Argumentation.' Computer Speech & Language 60 (2020): 101032.

[5] Zhao, R., Zhou, A., & Mao, K. (2019). 'Automatic Detection of Cyberbullying on Social Networks Based on Deep Learning and Natural Language Processing.' IEEE Access, 7, 114,688–114,698.

[6] Cheng, L., Guo, R., & Hu, B. (2020). 'A Neural Network-Based Approach to Detect Cyberbullying in Social Media Texts.' Journal of Intelligent & Fuzzy Systems, 38(2), 1,697–1,707.

[7] Potha, N., & Maragoudakis, M. (2014). 'Cyberbullying Detection Using Time Series Modeling.' Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics.

[8] Van Hee, C., et al. (2015). 'Automatic Detection of Cyberbullying in Social Media Text.' Proceedings of the International Conference Recent Advances in Natural Language Processing.

[9] Xu, J.-M., Bruckman, A., & Park, H. (2012). 'Mean Comments: Machine Learning for Cyberbullying Detection.' Proceedings of the 2012 ASE/IEEE International Conference on Social Computing.

[10] Dinakar, K., Reichart, R., & Lieberman, H. (2011). 'Modeling the Detection of Textual Cyberbullying.' Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

[11] Agrawal, S., Awekar, A. (2018). 'Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms.' European Conference on Information Retrieval (ECIR).

[12] Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). 'Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in Twitter.' Computers in Human Behavior, 63, 433–443.

[13] Nahar, V., Al-Maskari, S., & Abozinadah, E. A. (2014). 'Detecting Cyberbullying on Social Networks Using Machine Learning Techniques.' International Journal of Distributed Sensor Networks.

[14] Zhang, Z., Robinson, D., & Tepper, J. (2018). 'Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network.' European Semantic Web Conference.

[15] Zhong, H., Li, H., Squicciarini, A., et al. (2016). 'Content-Driven Detection of Cyberbullying on the Instagram Social Network.' Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI).

[16] Gambäck, B., & Sikdar, U. K. (2017). 'Using Convolutional Neural Networks to Classify Hate-Speech.' Proceedings of the First Workshop on Abusive Language Online.

[17] Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). 'Deeper Attention to Abusive User Content Moderation.' Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[18] Yin, D., Xue, Z., Hong, L., et al. (2009). 'Detection of Harassment on Web 2.0.' Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW.

[19] Burnap, P., & Williams, M. L. (2015). 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making.' Policy & Internet, 7(2), 223–242.

[20] Sood, S. O., Antin, J., & Churchill, E. F. (2012). 'Profanity Use in Online Communities.' Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[21] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). 'Using Machine Learning to Detect Cyberbullying.' Proceedings of the 10th International Conference on Machine Learning and Applications.

[22] Nandhini, D., & Sheeba, J. M. (2015). 'Online Social Network Bullying Detection Using Intelligence Techniques.' Procedia Computer Science, 45, 485–492.

[23] Chatzakou, D., Kourtellis, N., Blackburn, J., et al. (2017). 'Mean Birds: Detecting Aggression and Bullying on Twitter.' Proceedings of the 2017 ACM on Web Science Conference.

[24] Sanchez, J., Kumar, S., & Gopalakrishnan, V. (2019). 'Cyberbullying Detection on Twitter Using Deep Learning and Text Mining.' International Journal of Computer Applications, 975, 8887.