# Cyber Bullying Detection In Social Media Platforms

Vinay, Department of Artificial Intelligence and Data Science, Don Bosco Institute of Technology, vinaysawalgi8@gmail.com
Chethangiri K, Department of Artificial Intelligence and Data Science, Don Bosco Institute of Technology, chethangiri333@gmail.com
Kushal H K, Department of Artificial Intelligence and Data Science, Don Bosco Institute of Technology, kushalhk023@gmail.com
Dhanush A B, Department of Artificial Intelligence and Data Science, Don Bosco Institute of Technology, dhanushab851@gmail.com
Dr. Nandini K S, Associate Professor, Department of Artificial Intelligence and Data Science, Don Bosco Institute of Technology, nandini.ks@dbit.co.in

**Abstract- One of the unintended effects of the rapid prolifer ation of social media platforms is the spread of cyberbullying, which has become a major problem that severely affects the psychological health of users. Conventional detection methods are often incapable of understanding the contextual subtleties of online abuse, and typical intervention strategies are not sufficiently adaptable. In order to overcome these obstacles, this article presents a sophisticated cyberbullying identification and alleviation system based on a Bidirectional Long Short Term Memory (Bi-LSTM) network.Compared to a unidirectional model, the Bi-LSTM framework reads the text data not only in the forward but also in the backward direction, thus it can keep a larger semantic context is very important for the identification of the less obvious forms of abuse. The proposed method, which uses a comprehensive dataset from Kaggle, categorizes the offensive content into three different degrees of toxicity: Low, Medium, and Intensive. Besides that, the platform features an innovative user-intervention tool that is founded on a fluctuating reputation score; if the reputation score of a user drops below the crucial limit of 10.0, then that user is automatically blocked so that no more damage can be done.The main point of the experimental results is to show that this method is very effective in terms of classification and performance and that it is able to significantly outperform the traditional baselines. Hence, it provides a reliable and scalable tool to make the digital world safer not only through accurate detection but also by proactive user management.**

Index Terms— Cyberbullying Detection, Bidirectional LSTM, Multi-level Toxicity, Reputation Score, User Blocking, Social Media Safety.

## I. INTRODUCTION

The swift global expansion of the internet alongside the nearly universal adoption of social media channels have essentially restructured worldwide communication, allowing for instant connectivity and the quick spreading of information [1]. In the face of these digital innovations, which are packed with novel possibilities for interaction among people, a dreadful and all-pervading evil called cyberbullying has surfaced as a side effect [2]. As per the definition, it is a hostile, deliberately hurtful act, a single person or a group on the one hand, aiming at an individual or a group on the other, and to carry out this act through electronic communication means, performing the act many times against a victim who can hardly defend themselves. Cyberbullying is a form of harassment that surpasses the limitations of the physical world [3]. Unlike traditional bullying, which happens in person, the aggressor can be unknown in an online attack, the haters can spread the word at a great speed and their harassment can last for an indefinite time, and therefore the victim's privacy is violated [1] [5]. This virtual plague's psychological effects are quite substantial and scary as well; thus those who are targeted may become gravely anxious, depressed, socially isolated, and in extreme cases, they may develop suicidal thoughts [4] [5]. Due to the enormous amount of content created by users on such platforms as Twitter, Facebook, and Instagram, the task of human moderators is not only inefficient but also impossible to scale [6]. Thus, there is a strong and urgent call for the creation of automated, smart tools that perform real-time detection and removal of offensive user-generated text. Initial computer-based solutions for this problem mainly involved the use of conventional Machine Learning (ML) methods. Research in this area has found that classifiers like Support Vector Machines (SVM), Random

Forest (RF), Logistic Re gression (LR), and Na¨ ıve Bayes (NB) are very effective in the identification of toxic language [7] [8]. To illustrate, articles note that SVM classifiers along with Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction can result in accuracy as high as 96.14% in certain binary classification scenarios [9]. Likewise, the accuracy of some Random Forest implementations can be as high as 94.2% by combining the outputs of multiple decision trees [8] [10]. Without exception in these achievements, traditional ML mod els encounter major hurdles. They are usually built upon un changeable feature extraction methods such as Bag-of-Words (BoW) or N-grams, which in most cases do not recognize intricate semantic relationships and the linguistic context of human language [11]. Especially with the cases of indirect aggression, sarcasm, and irony, these models prove to be weak, as these are subtle bullying forms that need the model to have a deep understanding of the sequential nature of the context rather than the mere frequency of the words [7] [12]. To take care of these weaknesses, the scholarly works now focus more on the Deep Learning (DL) models. The rise of neural networks, in particular, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, as potent instruments for the task of text classification, has been evidenced [13]. Normal LSTM networks deal with sequential data by memory of the previous inputs, thus are able to handle long sentences better than traditional ML. Furthermore, advanced structures like LSTM-Autoencoders have shown the potential to attain outstanding accuracies, even up to 99% in artificially generated data or specific low-resource languages, datasets [14]. Nevertheless, conventional unidirectional LSTM models have a built-in architectural drawback: they only analyze text from the beginning to the end, thus they can use the past context but are not aware of the future context in a sentence [15]. Their unidirectional processing may at times result in incorrect interpretations as the true implication of a phrase depends most often on the words that come after the sequence. The latest comparison studies clearly point out that bidirectional processing is indispensable for grasping the full connection between the words and the sentences, henceforth, leading to higher pinpointing accuracies [15]. By identifying these shortcomings, the current research presents an efficient, context-aware cyberbullying identifica tion system based on a Bidirectional Long Short-Term Mem ory (Bi-LSTM) network. Unlike the typical networks, the Bi-LSTM takes the input sequences one after another from both the forward and the backward directions. This double processing feature of the model helps in capturing the most detailed context as well as long-term dependencies, thereby ensuring more accurate deciphering of complex and nebulous language [15]. Thanks to an all-encompassing and varied dataset obtained from Kaggle, this research goes far beyond the mere binary detection and is able to perform the fine grained classification. The offensive content is divided into three distinct levels of toxicity: Low, Medium, and Intensive, thus giving the possibility of determining the threat level first for the intervention strategies and then accordingly to the subsequent order of severity [16]. Moreover, efficient cyberbullying prevention cannot be accom plished by just detecting the content without the proactive user management. Based on user accountability and social link analysis, which are the core ideas of the recent literature [5] [16], we propose a new dynamic reputation score mechanism. This system is always on the lookout for user behavior; thus, those who keep on publishing toxic content will see their reputation score lowered. To maintain a network that is secure for all users, the system is designed to automatically hinder any user whose reputation score drops below the critical threshold of 10.0, thus preventing recidivism. The experiment results are the proof of the effectiveness of this combined method, thus, the proposed Bi-LSTM model not only attains classification performance at a higher level but also is a feasible solution in the real world for ensuring digital safety.

## II. PROBLEM STATEMENT

Even social media technologies and automated moderation tools advancing at a rapid pace, effectively identifying and stopping cyberbullying is still the biggest challenge in the realm of digital safety. Online hostility is highly dependent on numerous interconnected linguistic subtleties—such as sarcasm, irony, indirect threats, and changing slang—that determine the meaning of a message in complicated, non-linear ways which traditional statistical methods and unidirectional learning models are unable to comprehend or represent. Be cause of this, platforms frequently find it difficult to differ entiate between innocent joking and actual harassment, which results in incidents not being detected, victims experiencing severe psychological distress, and the creation of a toxic online environment.Moreover, current systems are largely incapable of determining the differing intensities of these assaults or recognizing offenders that commit the acts repeatedly in real time and, therefore, they are unable to provide the option of prioritizing cases with the most severe attacks or blocking those who intend to harm based on their behavioral history. The absence of detailed, context-sensitive, and anticipatory intervention tools underscores the urgent requirement for sophisticated Deep Learning methods that are capable of integrating bidirectional semantic comprehension with variable user reputation metrics to identify toxicity more accurately and guarantee the existence of a safer digital community.

## III. LITERATURE REVIEW

Cyberbullying identification has evolved through various stages, starting from simple keyword filtering and now reach ing advanced computational intelligence systems. This part of the paper presents an exhaustive critique of the existing work, segmented into Traditional Machine Learning methods, Deep Learning advancements, and Novel Severity/Behavioral approaches.

A. Traditional Machine Learning Approaches: Initial works primarily targeted the applications of super vised Machine Learning (ML) algorithms in offensive material classification. A great number of studies have emphasized the effectiveness of Support Vector Machines (SVM) and Random Forest (RF) classifiers when accompanied by powerful feature extraction techniques.
   - Performance of SVM: The experiment of Kadam et al. [9] showed that the accuracy of SVM classifiers can be 2 as high as 96.14%, thus only slightly better than that of Random Forest (96.01%) and Decision Trees (95.39%) in two hate speech datasets. In the same way, Sathya and Fernandez [4] experimented with SVM that combined with the features provided by NLP like Term Frequency Inverse Document Frequency (TF-IDF) and Linguistic Inquiry and Word Count (LIWC2), and they declared the achieved accuracy was 93.15%, thereby proving the efficiency of the model in handling high-dimensional text data. The team of Dhumale et al. [10] supported an idea of adding Word2Vec embeddings (Skip-gram model) to bring SVM to 95% accuracy and to show that semantic representation is important even for traditional models.
   - Ensemble Methods: Ensemble learning has never stopped to show its potential. The work of Sayed et al. [8] was centered around using a Random Forest classifier to label tweets in categories of bullying (Religion, Age, Gender, Ethnicity, and Non-Cyberbullying), and good results were achieved with the accuracy of 94.2%. With a more complicated approach, Benassou et al. [11] described the ”Stacked Model” concept that used a combination of Random Forest, Gradient Boosting, and SVM to achieve a detection rate of 98% thus leading to a very significant performance improvement over the single classifiers.
   - Real-time & Hybrid Systems: In their paper, Mathur et al. [1] introduced a real-time cyberbullying identification system for Twitter using Selenium as a means for web scraping. Their study showed that a carefully Random Forest classifier was the most effective (94.06% accuracy) among Adaboost and Gradient Boosting. In addition, the work of Priyadharshini et al. [6] involved a hybrid model of Na¨ ıve Bayes with TF-IDF where they proclaimed that the discriminative feature extraction and probabilistic classification integration is the most viable way to the detection of bullying texts in the social sphere that is cross-platform.

B. Deep Learning and LSTM Architectures: Machine learning models generally come with high ac curacy; however, researchers have noted that these models frequently fail to adapt to the changing and context-rich nature of social media texts, especially in the case of non English or code-mixed languages [2] [3]. Consequently, the academic research community has redirected its attention to Deep Learning (DL).
   - LSTM & Autoencoders: Cuzzocrea et al. [14] brought into being TLA-NET, a LSTM-Autoencoder network that can be trusted to handle the low-resource languages synthetic data, for example, Hindi and Bangla. Their model went as far as making an accuracy of 99% which is a very high f igure, proving that LSTM architectures are very effec tive for sequential pattern identification where traditional models are usually weak. Correspondingly, Akter et al. [13] experimented with different algorithms on a Bangla dataset and found that LSTM led to an excellent accuracy of 99.80% which is highly notable thus outclassing the traditional algorithms like XGBoost that could only reach 74% by a big margin.
   - Bidirectional Processing: It is true that regular LSTMs are strong; however, due to their unidirectional nature, they cannot fully utilize the information. To overcome this limitation, Berbery et al. [12] converted a standard LSTMinto a Bidirectional LSTM (Bi-LSTM) and trained it on a Twitter dataset that they had freshly collected. The outcomes of their trials indicate without a doubt that Bi-LSTM surpasses CNN, and GRU in performance and thus be capable of achieving the greatest recall (94.54%) and accuracy (92%) measure. This infers that the usage of both forward and

backward directions for text analysis is very important/interchangeable permissive for grasping not only the intricate semantic relations in bullying tweets but also the general.

C. Severity Determination and Behavioral Analysis: Recent innovations have gone beyond a straightforward binary differentiation (bullying vs. non-bullying) to involve the measurement of the intensity of attacks and the analysis of user behavior.

- Severity Classification: Obaid et al. [16] unveiled a pioneering concept that combines an LSTM network with Fuzzy Logic. Their mechanism not only detects bullying but also identifies its severity levels i.e., ”Low,” ”Medium,” and ”High” with a 93.67% accuracy rate. Such a granulated way is indispensable for the reschedul ing of the staff.
- Link Prediction Social Graph: Pal and Shetty [5] in troduced a method of social network analysis based on ”Connection Probability.” Their model estimates the probability of a friendly relationship between users; hence, user interactions with a low connection probability (strangers) are more tightly checked for the presence of hate speech. This behavioral context deepens the understanding of the text.

**D.** Research Gaps and Proposed Solution: Even though the systems of today show great precision, they mostly focus on detection [1] [9] or binary classification [13]. Just a handful have a 3-level detailed toxic content classification (Low, Medium, Intensive) along with a fully automated user management system. Furthermore, while rep utation systems have been envisaged in theory, there are very few instances where users are automatically blocked as a result of a changing score threshold (e.g., ¡ 10.0) within Deep Learning frameworks. This study goes beyond these limitations by implementing a Bi-LSTM model to understand the deep contextual nuances and then integrating it with a reputation-based blocking method to create a safer digital space.

## IV. SYSTEM ARCHITECTURE

The intended cyberbullying identification and alleviation system is architecturally detailed as a modular, end- to step pipeline that takes in unstructured social media text, 3 determines the toxicity severity, and implements automated user management actions. The system design is broken down into four main modules: Data Preprocessing, Bi-Directional Feature Learning, Multi-Level Classification, and Dynamic Reputation Management. Figure 1 shows the detailed data flow of the system model.

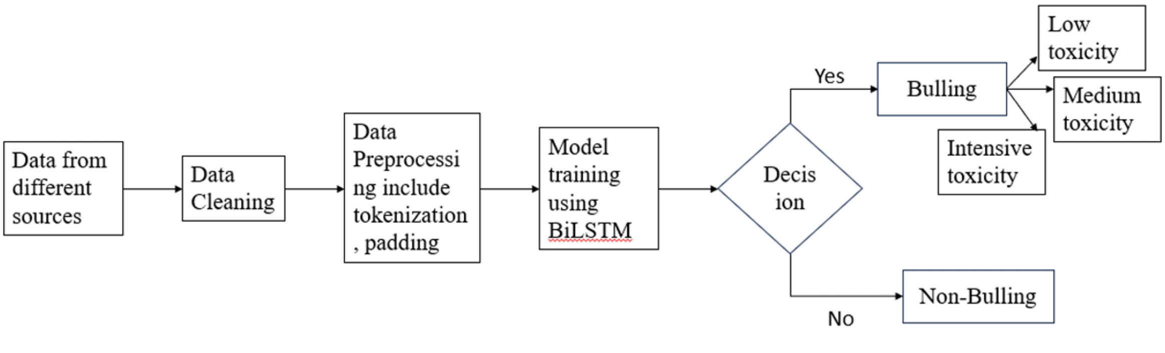
Data from different sources
Data Cleaning
Data Preprocessi ng include tokenization , padding
Model training using BiLSTM
Decis ion
Yes
Bulling
Low toxicity
Medium toxicity
Intensive toxicity
No
Non-Bulling

Figure. 1. Data Flow Diagram of Cyber Bullying Detection in Social Media Platform

## REFERENCS

[1] S. A. Mathur, B. Dharmasivam, S. Isarka, and J. C. D, ”Analysis of tweets for cyberbullying detection,” in 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), 2023, pp. 269–274.

[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, 2024.

[3] A. Toktarova, D. Sultan, and Z. Azhibekova, ”Review of machine learning models in cyberbullying detection problem,” in 2024 IEEE 4th International Conference on Smart Information Systems and Technolo gies (SIST), 2024, pp. 255–260.

[4] J. Sathya and F. M. H. Fernandez, ”Effective automatic cyberbullying detection using a hybrid approach SVM and NLP,” in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024.

[5] V. B. Pal and P. Shetty D, ”Integrating link prediction and comment anal ysis for enhanced cyberbullying detection in online social interactions,” in 2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2024.

[6] M. Priyadharshini, B. Nithya, H. J, S. K, A. F. Banu, and V. Murugesh, ”Advanced cyberbullying detection: A hybrid model integrated with Na¨ ıve Bayes,” in 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2024.

[7] A. Perera and P. Fernando, ”Cyberbullying detection system on social media using supervised machine learning,” Procedia Computer Science, vol. 239, pp. 506–516, 2024.

[8] F. R. Sayed, E. H. Elnashar, and F. A. Omara, ”Cyberbullying detection in social media using natural language processing,” Scientific African, vol. 28, e02713, 2025.

[9] B. Kadam, R. Borhade, A. Bardeskar, V. Shelke, and P. Bhosale, ”Cy berbullying detection using machine learning algorithms,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. V, pp. 1326–1328, May 2023.

[10] R. B. Dhumale, A. K. Dass, A. Umbrajkaar, and P. Mane, ”Enhancing cyberbullying detection with advanced text preprocessing and machine learning,” International Journal of Electrical and Computer Engineering (IJECE), vol. 15, no. 3, pp. 3139–3148, June 2025.

[11] L. F. Benassou, S. Bendaouia, O. Salem, and A. Mehaoua, ”Detection of cyberbullying in online comments: Latest advances and challenges,” in 2023 IEEE International Conference on E-health Networking, Appli cation & Services (Healthcom), 2023.

[12] K. Berbery, K. Samrouth, N. Bakir, and M. Dawood, ”Cyberbullying dataset collection for an enhanced automatic detection on Twitter using deep learning techniques,” in 2025 International Conference on Smart Applications, Communications and Networking (SmartNets), 2025.

[13] F. Akter, M. U. F. Jahangir, R. R. Chowdhury, and M. F. Rabbi, ”Cyber bullying detection on social media platforms utilizing different machine learning approaches,” International Journal of Computer Applications, vol. 186, no. 61, Jan. 2025.

[14] A. Cuzzocrea, M. S. Akter, H. Shahriar, and P. G. Bringas, ”Cy berbullying detection, prevention, and analysis on social media via trustable LSTM-autoencoder networks over synthetic data: The TLA NET approach,” Future Internet, vol. 17, no. 84, 2025.

[15] K. Berbery et al., ”Cyberbullying dataset collection for an enhanced automatic detection on Twitter using deep learning techniques,” in 2025 International Conference on Smart Applications, Communications and Networking (SmartNets), 2025.

[16] M. H. Obaid, S. K. Guirguis, and S. M. Elkaffas, ”Cyberbullying detection and severity determination model,” IEEE Access, vol. 11, pp. 97391–97399, 2023.