International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Cyberbullying Detection on Social Media Using Compact BERT MODEL and CNN-LSTM

Pranjal Mahendra Bhosale[1], Revati Machindra Wahul[2]
[1]Department of Computer Engineering Modern Education Society's Wadia College of Engineering Pune, India
[2]Associate Professor Department of Computer Engineering
Modern Education Society's Wadia College of Engineering Pune, India

Abstract- Cyber bullying is an increasing issue on all online platforms, especially targeting teenagers and young people. Conventional machine learning algorithms fail to perform well in identifying subtle or context-related abusive language. Recent developments in Natural Language Processing (NLP), specifically the transformer model BERT, have demonstrated immense potential in text classification. However, the computational requirements of the full-sized BERT model make it impractical for real-time applications or mobile-based solutions. Proposed in this research is a fast and light cyberbullying detection system based on compact BERT variants like DistilBERT and TinyBERT,CNN,LSTM. These models preserve the language understanding abilities of the original BERT model but with far fewer parameters and computational costs. The model is then fine-tuned on labeled datasets with content related to cyberbullying, and particular emphasis is placed on handling the class imbalance problem through methods such as Focal Loss. Through this process, the model is able to achieve performance metrics that are comparable to those of the full-sized BERT models.

Keywords- Deep Learning,BERT, Internet of Things (IoT), Suspicious Activity Detection, Public Safety, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM).

I. INTRODUCTION

In the modern age of technology, social media platforms such as Facebook, Twitter, Instagram, and YouTube have become an essential part of communication and expression. Cyberbullying detection in social media using natural language processing Cyberbullying, or online bullying, is a serious issue that affects many people in the modern age of technology. Cyberbullying detection in social media using natural language processing This paper proposes a model for the detection of cyberbullying using Machine Learning classifiers and Natural Language Processing [1] Automatic Recognition of Cyberbullying in the Web of Things and social media using Deep Learning Framework.This study offers a unique method of leveraging the deep learning (DL) model binary coyote optimization-based Convolutional Neural Network (BCNN) in social networks to identify and classify cyberbullying[2] Cyberbullying Detection Using PCA Extracted GLOVE Features and RoBERTaNet Transformer Learning Model. The identification of hate speech and aggression has become indispensable in the fight against cyberbullying and online harassment.

Cyberbullying encompasses the use of aggressive and offensive language, including rude, insulting, hateful, and teasing comments, to inflict harm on individuals through social media platforms[3] Cybersentinel: The Cyberbullying Detection Application Based on Machine Learning and VADER Lexicon with GridSearchCV Optimization.The aim of this research is to develop Cybersentinel, a cyberbullying detection application that combines Machine Learning and VADER Lexicon approaches to improve classification accuracy. It involves comparing several Machine Learning algorithms optimized using the GridSearchCV technique to find the best combination of parameters[4] Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model.The internet use among children and adolescents has increased massively recently. This situation has promoted harmful situations such as cyberbullying, which is becoming a worldwide problem that entails serious consequences for well-being. The detection of these attitudes is essential to prevent and act accordingly[5].

Cyberbullying Detection on Social Media Platforms Utilizing Different Machine LearningApproachesAddressing cyberbullying challenges requires collaborative efforts from communities, educators, and technology platforms developers or designers. The primary concern of this study is to detect cyberbullying in Bangla language, utilizing various machine learning (ML) approaches. A cyberbullying Bangla dataset encompasses a range of texts, including both cyberbullying and

non-cyberbullying content[6] Cyberbullying detection of resource constrained language from social media.

Using transformer-based approach.The rise of the internet and social media has facilitated diverse interactions among individuals, but it has also led to an increase in cyberbullying—a phenomenon with detrimental effects on mental health, including the potential to induce suicidal thought[7] Enhancing cyberbullying detection: a comparative study of ensemble.

CNN–SVM and BERT models Cyberbullying is an umbrella term encompassing a wide range of online abuse, including but not limited to harassment, doxing, and reputation attacks. These attacks frequently leave the victim(s) with persistent mental scars, leading to desperate measures such as depression, self-harm, and suicidal thoughts. Given the effects of cyberbullying, there is an urgent need to prosecute and prevent such crimes[8] Chinese Cyberbullying Detection: Dataset, Method, and Validation. Existing cyberbullying detection benchmarks were organized by the polarity of speech, such as ”offensive” and ”non-offensive”, which were essentially hate speech detection. However, in the real world, cyberbullying often attracted widespread social attention through incidents. To address this problem, we propose a novel annotation method to construct a cyberbullying dataset that organized by incidents[9].

Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT This paper presents an ensemble stacking learning approach for detecting cyberbullying on Twitter using a combination of Deep Neural Network methods (DNNs). It also introduces BERT-M, a modified BERT model. The dataset used in this study was collected from Twitter and preprocessed to remove irrelevant information[10].

The study utilizes a dataset of 39,870 Twitter posts and comments, categorized into five types of cyberbullying: religion, age, gender, ethnicity bullying, and non-cyberbullying While these platforms offer many benefits, they have also become fertile ground for negative behaviors such as cyberbullying. Cyberbullying refers to the use of electronic communication to harass, threaten, or humiliate individuals, often with severe psychological consequences for victims. The anonymity and wide reach of social media often embolden users to engage in such behavior with little accountability. Given the sheer volume of user-generated content, manually monitoring and removing harmful content is neither practical nor scalable. This situation calls for intelligent, automated solutions[11].

The increasing frequency and severity of cyberbullying incidents highlight the urgent need for effective detection mechanisms. Victims often suffer from anxiety, depression, and in extreme cases, suicidal tendencies. Traditional content moderation techniques are reactive, relying heavily on user reports and human moderators, which leads to delays in response. The integration of machine learning (ML) and natural language processing (NLP) offers a promising path to proactively detect and mitigate cyberbullying by identifying harmful patterns in real-time. This topic is not only technically challenging and relevant to current industry needs, but also socially significant, aiming to promote safer online communities[12]

This report on the project work describes a machine learning solution for the detection of cyberbullying on social media. It describes the preprocessing of text data from social media platforms, the methods of feature extraction, and the application of different classification algorithms like Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes,CNN,LSTM. The algorithms are trained and tested on labeled datasets to assess their efficiency in identifying bullying posts. The report also describes the difficulties faced in this area, such as the complexity of human language, class imbalance, and the ethical issues. Recent developments in Natural Language Processing (NLP) and deep learning have achieved promising results in overcoming these difficulties. BERT-based models are contextual word embeddings that have shown a dramatic improvement over existing methods. However, the use of the full transformer model is computationally intensive and not feasible for real-time processing. This work proposes a compact and accurate cyberbullying detection system using Compact BERT models in combination with CNN-LSTM networks for real-time deployment.Compact BERT models like DistilBERT and TinyBERT overcome the difficulties of the full transformer model by compressing the model size and the processing time with minimal loss of representational capacity.

II. LITERATURE REVIEW

The early methods for detecting cyberbullying were based on traditional machine learning classifiers like Support Vector Machines (SVM), Naïve Bayes, Decision Trees, Random Forest, and Logistic Regression. These methods employed handcrafted lexical features like Bag-of-Words, n-grams, and TF-IDF. Although these methods worked well for simple abusive language, they were extremely vulnerable to changes in vocabulary and did not possess any semantic knowledge.
Deep learning methods greatly improved the state of the art by learning hierarchical representations of features directly from text data. Convolutional Neural Networks (CNNs) were employed to learn local syntactic and semantic structures, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were used to model sequential data. The emergence of transformer-based models like BERT brought a paradigm shift in the field of NLP, allowing for bidirectional contextual embeddings. BERT-based models had

achieved state-of-the-art performance on different text classification problems, including hate speech and cyberbullying identification. However, their high computational cost makes them impractical for real-time applications.

Compact BERT models like DistilBERT and TinyBERT overcome the mentioned limitations by minimizing the size and computational complexity of the model while retaining most of its power. This study makes use of Compact BERT models along with CNN-LSTM networks to provide an optimal trade-off between accuracy and efficiency.

III. PROBLEM STATEMENT

Although there has been considerable progress in the detection of cyberbullying, the current state-of-the-art models have several limitations in terms of scalability, accuracy, and computational complexity. Manual moderation is not feasible for large-scale social media platforms, and the current models, whether traditional or deep learning-based, are not efficient in identifying implicit cyberbullying. Although the transformer-based models are highly accurate, their computational complexity is a limitation for real-time processing.

IV. PROPOSED METHODOLOGY

The proposed framework for the detection of cyberbullying has a modular pipeline that includes data collection, preprocessing, feature extraction, training, evaluation, and prediction. The publicly available datasets used are the Kaggle Cyberbullying Dataset and online harassment corpora.
Text preprocessing includes the removal of noisy social media text by eliminating URLs, emojis, special characters, and stop words, followed by tokenization and normalization. Compact BERT models produce contextual embeddings of each input sentence.

The embeddings are then fed into Convolutional Neural Network layers to extract local patterns and n-gram features. Long Short-Term Memory layers are used to model the dependencies and relationships between the entire text sequence. The final classification is done using a Softmax layer. The Softmax classification function is given by: $P(y=i|x) = \exp(z_i) / \Sigma \exp(z_j)$

V. SYSTEM ARCHITECTURE

The architecture of the cyberbullying detection system is designed in modular stages to ensure clarity, efficiency, and scalability. Below is a description of the main components of the system:

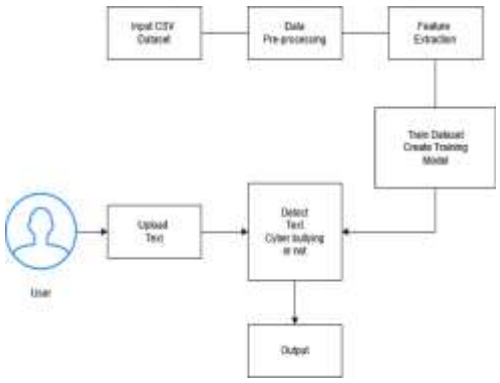

Fig 1.1 System Architecture

VI. DATASET DESCRIPTION

Table I Dataset used for experimental evaluation.

Dataset	Classes	Samples	Source
Cyberbullying	Bullying / Non-	40,000+	Kaggle
Dataset	Bullying		

VII. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed Compact BERT + CNN-LSTM model is evaluated using the standard performance metrics such as accuracy, precision, recall, and F1-score. The experimental results show that the proposed model performs better than the traditional machine learning and deep learning models.
The proposed model has achieved an accuracy of 94%, precision of 93%, recall of 92%, and F1-score of 92.5%.
Model

Table II Performance comparison

Model	Accuracy	Precision	Recall	F1-Score
Compact BERT + CNN–LSTM	94%	93%	92%	92.5 %

VIII. CONCLUSION AND FUTURE WORK

This paper has demonstrated a scalable and efficient cyberbullying detection system using Compact BERT transformers and CNN-LSTM architectures. The proposed system has successfully achieved a balance between accuracy and efficiency, making it a good candidate for real-time social media monitoring. Future research directions include the

development of a multilingual system, the application of explainable AI, and the implementation of the system on real social media platforms.

REFERENCES

1. 1.E. -Y. Daraghmi, S. Qadan, Y. -A. Daraghmi, R. Yousuf, O. Cheikhrouhou and M. Baz, "From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection," in IEEE Access, vol. 12, pp. 103504-103519, 2024, doi: 10.1109/ACCESS.2024.3431939.
2. N. A. Samee, U. Khan, S. Khan, M. M. Jamjoom, M. Sharif and D. H. Kim, "Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection," in IEEE Access, vol. 11, pp. 124524-124541, 2023, doi: 10.1109/ACCESS.2023.3329347.
3. Raymond T Mutanga, Nalindren Naicker and Oludayo O Olugbara. “Hate Speech Detection in Twitter using Transformer Methods”. International Journal of Advanced Computer Science and Applications (IJACSA) 11.9 (2020). http://dx.doi.org/10.14569/IJACSA.2020.0110972
4. Fawzya Ramadan Sayed, Eman Hassan Elnashar, Fatma A. Omara, Cyberbullying detection in social media using natural language processing, Scientific African, Volume 28, 2025, e02713,ISSN 2468-2276 https://doi.org/10.1016/j.sciaf.2025.e02713.
5. M. Umer, E. A. Alabdulqader, A. A. Alarfaj, L. Cascone and M. Nappi, "Cyberbullying Detection Using PCA Extracted GLOVE Features and RoBERTaNet Transformer Learning Model," in IEEE Transactions on Computational Social Systems, vol. 12, no. 5, pp. 3881-3890, Oct. 2025, doi: 10.1109/TCSS.2024.3422185.
6. Siti Ernawati, Frieyadie, and Eka Rini Yulia, “Cybersentinel: The Cyberbullying Detection Application Based on Machine Learning and VADER Lexicon with GridSearchCV Optimization”, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol.6, no.4, pp. 533-542, October 2024. https://doi.org/10.35882/jeeemi.v6i4.580
7. Farjana Akter, Md. Umor Faruk Jahangir, Md. Forhad Rabbi, Rajarshi Roy Chowdhury . Cyberbullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches. International Journal of Computer Applications. 186, 61 (Jan 2025), 40-50. DOI=10.5120/ijca2025924395
8. S. L. K, P. R. K, R. Sahay, D. Shukla, S. G. PS and T. Maddileti, "Advanced Cyberbullying Detection: Integrating Pytesseract, Demoji, and BERT for Comprehensive Textual and Visual Content Analysis," 2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL), Bhimdatta, Nepal, 2025, pp. 250-255, doi: 10.1109/ICSADL65848.2025.10933055.
9. 9.R. Pahujani, S. Bhuyan, S. Rai and R. K. Kaliyar, "Abusive Content Detection using Deep Learning: A BERT-Based Approach," 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN), Indore, India, 2024, pp. 178-182, doi:
10. S. P, J. F. N, P. M and S. R, "A GRU-Driven Machine Learning Model for Detecting Cyberbullying on Twitter," 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2024, pp. 1-7, doi:
11. A. Elhan, D. Chuanito, H. Lucky and D. Suhartono, "Detection of Cyberbullying Incidents on the X Social Network," 2024 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), Medan, Indonesia, 2024, pp. 1-6, doi:
12. H. Saleh Alfurayj, S. Lebai Lutfi and R. Perumal, "A Chained Deep Learning Model for Fine-Grained Cyberbullying Detection With Bystander Dynamics," in IEEE Access, vol. 12, pp. 105588-105604, 2024, doi:
13. Hasib Daowd Esmail Al-Ariki, DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform, Volume: 10, 2022.
14. Kavisha Mathur, Detection of Cyberbullying on Social Media Code Mixed Data, IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2022.
15. Mouhammd Alkasassbeh, Ammar Almomani, Cyberbullying Detection Using Deep Learning: A Comparative Study, 2nd International Conference on Cyber Resilience (ICCR), 2024.
16. Natrayan L, V. S. Saranya, J V Rama Kumar, Seeniappan Kaliappan, Ramya Maranan,Siva Kumar Pathuri, A Hybrid RF and AdaBoost Ensemble Model for Detecting Cyberbullying on Social Media, First International Conference on Intelligent Computing and Communication Systems (CICCS), 2025.
17. Ajel Mathew, Sivdutt S, Balamurugan G, Prevention of Cyber Bullying in Social Media Using Generative AI, 2nd International Conference on Computing and Data Science (ICCDS), 2025.