IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

Cyber-bullying Prediction using Machine Learning and Transformer-Based Deep Learning Models

Prof. Sudhanshu Tripathi, Prof. Rohan B Kokate, Miss. Shruti Vinod Pande
Master of Computer Applications
J D College of Engineering and Management, Nagpur, Maharashtra

Abstract: *The exponential growth of online communication platforms has led to a significant increase in harmful digital interactions, commonly referred to as cyber-bullying. Identifying such content manually is impractical due to the scale and speed of data generation. This study proposes an intelligent detection framework that combines conventional machine learning techniques with advanced deep learning and transformer-based models. Traditional classifiers, including Naive Bayes, Logistic Regression, and Support Vector Machines, are compared with deep neural architectures such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and the transformer-based **BERT** model developed by Google. Textual inputs are processed through normalization, tokenization, and vectorization techniques, including TF-IDF and contextual embeddings. Experimental findings demonstrate that transformer-based models significantly outperform traditional approaches, achieving an accuracy of 96%. The proposed framework highlights the effectiveness of contextual learning in detecting abusive language and offers a scalable solution for real-world deployment*

Keywords: Cyber-bullying, NLP, Machine Learning, Deep Learning, BERT, Text Analytics

I. INTRODUCTION

The digitalization of communication has transformed how individuals interact, but it has also introduced new forms of online harassment. Cyber-bullying, characterized by repeated hostile behavior through electronic means, has emerged as a serious societal issue. Victims often experience emotional distress, anxiety, and long-term psychological effects.

Given the immense volume of user-generated content across platforms, manual moderation systems are insufficient. Automated detection mechanisms powered by artificial intelligence have become essential. This research focuses on designing a robust predictive system capable of identifying cyber-bullying in textual data by leveraging both statistical learning and deep contextual models.

II. RELATED WORK

Earlier approaches to cyber-bullying detection relied heavily on rule-based systems and keyword matching techniques. While computationally efficient, these methods failed to capture semantic meaning and contextual variations.

Subsequent studies introduced machine learning models, which improved classification accuracy by learning patterns from labeled datasets. However, these approaches were limited in handling complex linguistic structures.

Recent advancements in deep learning enabled the use of neural networks such as CNN and LSTM, which capture hierarchical and sequential features in text. More recently, transformer-based architectures such as BERT have demonstrated superior performance by incorporating bidirectional context and large-scale pre-training.

DOI: 10.48175/IJARSCT-34802

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

III. PROPOSED METHODOLOGY

3.1 Data Acquisition

A labeled dataset consisting of user-generated textual content is utilized for training and evaluation. The dataset includes both abusive and non-abusive samples.

3.2 Text Preprocessing

To ensure consistency and reduce noise, the following preprocessing steps are applied:

- Removal of hyperlinks, special symbols, and redundant characters
- Conversion of text to lowercase
- Tokenization into meaningful units
- Elimination of stop words
- Lemmatization to normalize word forms

3.3 Feature Engineering

Two different representations are used:

- TF-IDF: Captures statistical importance of words
- Contextual Embeddings: Generated using transformer-based models

3.4 Model Development

3.4.1 Traditional Machine Learning Models

- Naive Bayes
- Logistic Regression
- Support Vector Machine

These models are trained on TF-IDF features.

3.4.2 Deep Learning Models

- CNN: Extracts local patterns and discriminative features
- LSTM: Captures sequential dependencies and contextual flow

3.4.3 Transformer-Based Model

The BERT model is fine-tuned for binary classification. Unlike conventional models, it processes text bidirectionally, enabling a deeper understanding of context.

3.5 Evaluation Criteria

The models are evaluated using:

- Accuracy
- Precision
- Recall
- F1-score

IV. EXPERIMENTAL RESULTS

Naive Bayes	85%	83%	82%	82.5%
Logistic Regression	88%	86%	85%	85.5%
SVM	91%	90%	89%	89.5%
CNN	92%	91%	90%	90.5%

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, May 2026
Impact Factor: 8.2

LSTM	93%	92%	91%	91.5%
BERT	96%	95%	94%	94.5%

V. DISCUSSION

The experimental outcomes reveal that transformer-based architectures significantly outperform both traditional and deep learning models. While machine learning models rely on statistical features, deep learning models improve performance by learning hierarchical patterns.

However, the superior performance of BERT is attributed to its ability to incorporate bidirectional context and semantic understanding. This allows it to better interpret subtle linguistic cues such as sarcasm and implicit abuse.

VI. NOVEL CONTRIBUTION

This research makes the following contributions:

- A hybrid evaluation of machine learning, deep learning, and transformer-based approaches
- Demonstration of improved performance using contextual embeddings
- A scalable framework suitable for real-time cyber-bullying detection
- Comparative analysis highlighting limitations of traditional models

VII. CONCLUSION

This study presents a comprehensive framework for detecting cyber-bullying using advanced computational techniques. The results confirm that transformer-based models provide superior accuracy compared to traditional approaches. The findings emphasize the importance of contextual understanding in text classification tasks.

The proposed system can serve as a foundation for intelligent moderation tools aimed at improving online safety.

VIII. FUTURE SCOPE

- Development of multilingual detection systems
- Integration with real-time streaming platforms
- Expansion to multimodal data (text + images)
- Exploration of next-generation transformer models

REFERENCES

[1] N. Dinakar et al., “Modeling the Detection of Textual Cyberbullying,” ICWSM, 2011.
[2] T. Davidson et al., “Automated Hate Speech Detection,” ICWSM, 2017.
[3] Tom M. Mitchell, Machine Learning, McGraw-Hill.
[4] Abraham Silberschatz et al., Operating System Concepts, Wiley.
[5] BERT, https://arxiv.org/abs/1810.04805