International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 13 Special Issue 02 | 2026

Explainable Transformer-based Cyberbullying Detection using Multiclass Toxicity Classification

Mohammad Awais[1], Mohammad Musaib [2], V. L. Tohare[3]
[1,2,3] *Student, CSE, Siddhivinayak Technical Campus, Maharashtra, India*

DOI: 10.5281/zenodo.19539542

ABSTRACT

Cyberbullying has emerged as a serious social and psychological threat in online environments, significantly affecting adolescents and young adults. The rapid growth of social media platforms such as Twitter, Instagram, and Facebook has increased the volume of user-generated content, making manual moderation insufficient. This research proposes an Explainable Transformer-Based framework for multiclass cyberbullying detection using deep learning and Natural Language Processing (NLP). The system leverages a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model for contextual text representation and performs multiclass toxicity classification, including categories such as insult, threat, hate speech, obscene content, and identity-based attack. To enhance transparency and interpretability, Explainable AI techniques such as SHAP (SHapley Additive explanations) are integrated to highlight influential words contributing to model predictions. Experimental results demonstrate that the proposed Transformer-based model outperforms traditional deep learning models such as CNN and LSTM in terms of accuracy, precision, recall, and F1-score. The proposed system provides an accurate, scalable, and interpretable solution for automated cyberbullying detection in social media environments.

Keywords:- Cyberbullying Detection, Transformer Model, BERT, Explainable AI, Multiclass Classification, Toxicity Detection, Deep Learning, NLP

1. INTRODUCTION

Cyberbullying refers to the intentional use of digital communication technologies such as social media, messaging platforms, online forums, and gaming environments to harass, threaten, insult, or humiliate individuals. Unlike traditional face-to-face bullying, cyberbullying is not constrained by physical boundaries and can occur continuously, anonymously, and at a large scale, significantly amplifying its psychological and emotional impact. Victims of cyberbullying frequently experience serious mental health consequences including anxiety, depression, loneliness, reduced self-esteem, emotional distress, and social withdrawal. In severe cases, prolonged exposure to online harassment can negatively influence academic performance, professional productivity, and overall well- being, making cyberbullying a critical societal and technological concern.

The rapid proliferation of social media platforms such as Twitter, Instagram, Facebook, and online discussion communities has led to an unprecedented increase in user-generated textual content. Within these digital interactions, harmful and toxic language often emerges in the form of insults, threats, hate speech, identity-based attacks, and offensive expressions. Due to the massive scale and real-time nature of online communication, manual moderation is insufficient, slow, and inconsistent. Consequently, there is a growing demand for intelligent automated systems capable of detecting and preventing cyberbullying in its early stages to ensure safer online environments.

Traditional machine learning approaches such as Logistic Regression, Naïve Bayes, and Support Vector Machines primarily depend on handcrafted features like Bag-of-Words and TF-IDF representations. While these techniques have shown moderate success, they often fail to capture contextual semantics, sarcasm, implicit aggression, and long-range linguistic dependencies present in natural language. This limitation reduces their effectiveness in accurately identifying nuanced and evolving forms of cyberbullying content.

Recent breakthroughs in deep learning, particularly Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers), have significantly improved natural language understanding. Transformer models leverage self-attention mechanisms to learn contextual word representations from large-scale corpora, enabling superior performance in text classification, sentiment analysis, and toxicity detection tasks. By understanding bidirectional context, BERT can more effectively identify subtle patterns of abusive and harmful language compared to conventional models.

Despite their high predictive performance, deep learning models often operate as “black-box” systems, lacking interpretability and transparency in their decision-making processes. In sensitive applications such as

cyberbullying detection, explainability is crucial for building trust, ensuring fairness, and enabling human moderators to understand why certain content is flagged as harmful. To address this challenge, Explainable Artificial Intelligence (XAI) techniques such as SHAP (SHapley Additive explanations) can be integrated to highlight influential words and linguistic patterns that contribute to model predictions, thereby improving model accountability and transparency.

Motivated by these challenges and advancements, this research proposes an **Explainable Transformer-Based framework for Multiclass Cyberbullying Detection**. The proposed system leverages fine-tuned BERT for contextual text representation and performs multiclass toxicity classification across categories such as insult, threat, hate speech, obscene content, and identity-based attacks. Furthermore, the integration of Explainable AI enables interpretable predictions, making the system suitable for real-world deployment in automated moderation and online safety monitoring. The proposed approach aims to provide an accurate, scalable, and transparent solution for early detection and prevention of cyberbullying in modern digital ecosystems.

2. PROPOSED METHODOLOGY

The proposed framework presents an Explainable Transformer-Based approach for detecting cyberbullying through multiclass toxicity classification. The system is designed as a structured pipeline consisting of dataset acquisition, preprocessing, feature extraction using a Transformer model, multiclass classification, explainability integration, and performance evaluation. Each stage is described below.

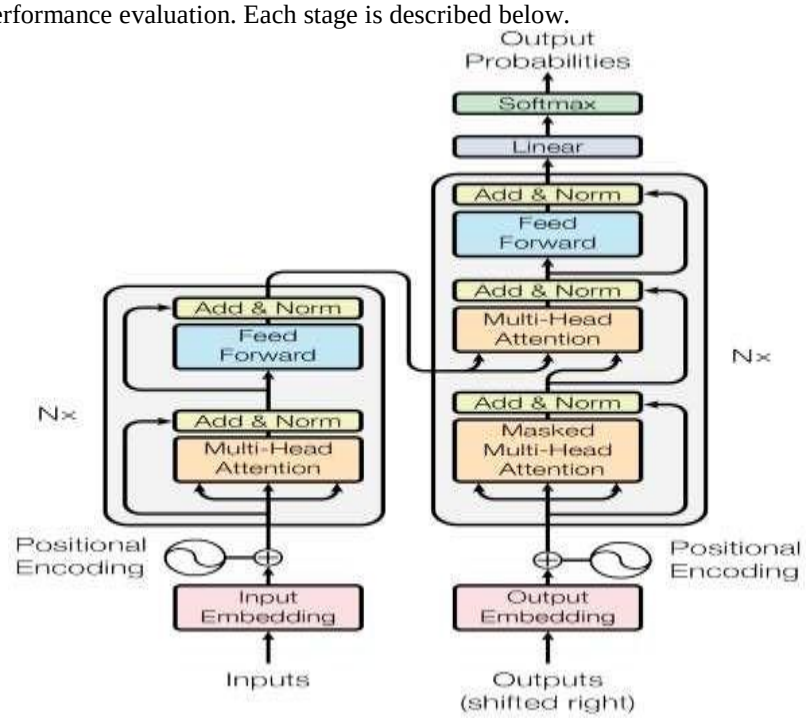

Fig.1. System Architecture

Stage 1: Dataset Acquisition

In the first stage, a labelled textual dataset containing toxic and non-toxic user-generated comments is collected from publicly available sources such as the Kaggle Toxic Comment Classification Dataset. This dataset consists of online comments annotated with multiple toxicity categories, enabling multiclass classification of cyberbullying-related behavior.

The classification categories considered in this research include:

- Insult: Offensive or degrading language targeting an individual or group.
- Threat: Content containing intimidation or harmful intent.
- Hate Speech: Language promoting hatred toward specific communities or identities.
- Obscene: Use of vulgar or inappropriate expressions.
- Identity-Based Attack: Targeting individuals based on race, gender, religion, or other identity attributes.
- Non-Toxic: Neutral or harmless comments without abusive content.

The collected dataset is examined for missing values, duplicate entries, and class imbalance to ensure data quality and reliability before further processing.

Stage 2: Data Preprocessing

The raw textual data undergoes systematic preprocessing to improve data quality and prepare it for Transformer- based modelling. This stage reduces noise and enhances semantic consistency.

The preprocessing steps include:

- Lowercase Normalization: All text is converted to lowercase to maintain uniformity.
- Removal of URLs, Emojis, and Special Characters: Irrelevant symbols and hyperlinks are removed to reduce noise.
- Tokenization using BERT Tokenizer: Text is split into subword tokens compatible with the BERT vocabulary.
- Padding and Truncation: Sequences are padded or truncated to a fixed length to maintain uniform input size for the model.

These steps ensure that the textual data is clean, structured, and compatible with Transformer-based architectures.

Stage 3: Transformer-Based Feature Extraction

In this stage, contextual features are extracted using a pre-trained Transformer model, specifically BERT (Base-Uncased). The model is fine-tuned on the cyberbullying dataset to learn task-specific linguistic patterns and contextual relationships.

The advantages of using BERT include:

- Context-Aware Embeddings: Captures semantic meaning based on surrounding words.
- Handling Long-Range Dependencies: Effectively models relationships across long sentences.
- Bidirectional Understanding: Processes text from both left-to-right and right-to-left for improved contextual representation.

The output of the BERT encoder produces dense contextual embeddings, which serve as feature representations for classification.

Stage 4: Multiclass Classification Layer

The contextual embeddings generated by BERT are passed to a fully connected dense layer responsible for multiclass toxicity classification.

Key components include:

- Fully Connected Dense Layer: Maps learned feature representations to class probabilities.
- Softmax Activation Function: Converts output scores into probability distribution across multiple toxicity categories.
- Cross-Entropy Loss Function: Measures prediction error and optimizes model parameters during training. This stage enables the model to classify each input comment into one of the predefined toxicity categories.

Stage 5: Explainability Integration

To improve transparency and interpretability, Explainable Artificial Intelligence (XAI) techniques are integrated into the system. Specifically, SHAP (SHapley Additive exPlanations) is used to analyze model predictions.

SHAP provides the following benefits:

- Identifies important words influencing classification decisions.
- Highlights linguistic patterns associated with cyberbullying behavior.
- Enhances trust and accountability in automated moderation systems.

By visualizing word-level importance, the system becomes interpretable rather than functioning as a black-box model.

Stage 6: Model Evaluation

The performance of the proposed cyberbullying detection model is evaluated using standard classification metrics to ensure reliability and robustness.

The evaluation metrics include:

- Accuracy: Measures overall correctness of predictions.
- Precision: Indicates how many predicted toxic instances are actually correct.
- Recall: Measures the ability to detect all actual toxic instances.
- F1-Score: Harmonic mean of precision and recall, providing balanced evaluation.
- Confusion Matrix: Provides detailed insight into correct and incorrect classifications across all toxicity categories.

These metrics collectively assess the effectiveness of the proposed Transformer-based explainable cyberbullying detection framework.

3. ALGORITHM

The proposed Explainable Transformer-Based Cyberbullying Detection framework follows a structured sequence of steps for multiclass toxicity classification and interpretability. The algorithm integrates data preprocessing, Transformer-based learning, and Explainable AI to ensure both accuracy and transparency in prediction.

Step 1: Dataset Loading
The labeled cyberbullying dataset containing textual comments and corresponding toxicity categories is loaded into the working environment. The dataset includes multiclass labels such as *Insult, Threat, Hate Speech, Obscene, Identity-Based Attack,* and *Non-Toxic*. Data validation is performed to handle missing values and remove duplicate entries.

Step 2: Text Preprocessing using BERT Tokenizer
Each input comment is preprocessed to prepare it for Transformer-based modeling. The preprocessing includes lowercase normalization, removal of URLs and special characters, and tokenization using the BERT tokenizer. The tokenized sequences are padded and truncated to a fixed length to ensure uniform input representation.

Step 3: Dataset Splitting
The preprocessed dataset is divided into training and testing subsets using an 80:20 ratio. The training dataset is used to learn model parameters, while the testing dataset is reserved for evaluating the model's performance on unseen data. Stratified sampling is applied to maintain class distribution balance.

Step 4: Fine-Tuning the Pre-Trained BERT Model
A pre-trained BERT (Base-Uncased) model is initialized and fine-tuned on the cyberbullying dataset. During fine- tuning, the model adapts its contextual embeddings to learn toxicity-related linguistic patterns and semantic relationships present in the dataset.

Step 5: Model Training using Adam Optimizer
The fine-tuned BERT model is trained using the Adam optimizer with an appropriate learning rate. Cross-entropy loss is used as the objective function for multiclass classification. The training process iteratively updates model weights to minimize prediction error and improve classification accuracy.

Step 6: Toxicity Category Prediction
After training, the optimized model is used to predict toxicity categories for the testing dataset. The Softmax activation function produces probability scores for each class, and the class with the highest probability is assigned as the predicted label.

Step 7: Explainability using SHAP
To enhance interpretability, SHAP (SHapley Additive exPlanations) is applied to the trained model. SHAP identifies and visualizes the contribution of individual words toward model predictions, enabling transparent and explainable classification of cyberbullying content.

Step 8: Performance Evaluation
The performance of the proposed system is evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1-score. Additionally, a confusion matrix is generated to analyze correct and incorrect predictions across all toxicity categories, ensuring comprehensive evaluation of the model.

4. RESULT AND DISCUSSION

To evaluate the effectiveness of the proposed Explainable Transformer-Based Cyberbullying Detection framework, a comparative analysis was performed against two widely adopted deep learning architectures: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). All models were trained and tested on the same preprocessed dataset using an identical 80:20 train–test split to ensure fairness and consistency in evaluation.
The CNN model was designed to capture local textual features through convolutional filters, making it effective in identifying key phrases and short patterns associated with toxic language. The LSTM model was employed to capture sequential and long-term dependencies within text data, enabling improved contextual understanding compared to CNN. The proposed BERT-based model, leveraging bidirectional contextual embeddings, was fine- tuned specifically for multiclass toxicity classification.
The performance comparison is presented in Table .

Table 1: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score
CNN	88%	87%	86%	86.5%
LSTM	90%	89%	88%	88.5%
BERT (Proposed)	94%	93%	92%	92.5%

The results indicate that the proposed BERT-based model outperformed both CNN and LSTM across all evaluation metrics, demonstrating its superior capability in capturing contextual semantics and improving multiclass cyberbullying detection performance.

5. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed Explainable Transformer-Based Cyberbullying Detection framework offers several significant advantages over traditional machine learning and conventional deep learning approaches. These advantages enhance the system's accuracy, interpretability, scalability, and real-world applicability.

1. **High Classification Accuracy**
   The proposed model utilizes a fine-tuned Transformer architecture (BERT), which captures deep contextual relationships within textual data. This enables the system to achieve higher prediction accuracy compared to traditional models such as CNN and LSTM. Improved accuracy ensures reliable detection of cyberbullying content, reducing both false positives and false negatives, which is critical for effective moderation and safe online communication.

2. **Context-Aware Detection**
   Unlike conventional approaches that rely on surface-level textual features, the Transformer model understands contextual semantics through bidirectional attention mechanisms. This allows the system to correctly interpret sarcasm, implicit aggression, and context-dependent abusive language. As a result, the model can detect subtle and complex forms of cyberbullying that may be overlooked by traditional text classification methods.

3. **Multiclass Toxicity Detection**
   The proposed system performs multiclass classification rather than simple binary toxic/non-toxic detection. It identifies specific categories such as insult, threat, hate speech, obscene language, and identity-based attacks. This fine-grained classification improves moderation capability by enabling targeted responses, better policy enforcement, and more detailed analysis of harmful online behavior.

4. **Explainable and Interpretable Predictions**
   A major limitation of deep learning systems is their black-box nature. To address this, the proposed framework integrates Explainable Artificial Intelligence using SHAP (SHapley Additive exPlanations). SHAP provides word- level importance visualization, highlighting which terms contributed most to the classification decision. This improves transparency, builds trust in automated moderation systems, and supports ethical and accountable AI deployment.

5. **Scalable for Real-Time Applications**
   The architecture is designed to be scalable and deployable in real-time environments such as social media monitoring systems, online forums, and content moderation platforms. With optimized inference and efficient Transformer implementation, the system can process large volumes of textual data and detect cyberbullying in near real-time, making it suitable for practical, large-scale deployment.

6. **Robustness and Generalization**
   The use of pre-trained language models enables the system to generalize well across diverse textual styles, slang, and informal language commonly found in social media. Fine-tuning further adapts the model to domain-specific cyberbullying patterns, improving robustness across different datasets and online platforms.

7. **Reduced Manual Feature Engineering**
   Traditional machine learning models require extensive manual feature extraction such as TF-IDF, n-grams, and linguistic rules. The proposed Transformer-based approach automatically learns feature representations from raw text, reducing human effort while improving model performance and adaptability.
   Overall, the proposed system provides a powerful, accurate, interpretable, and scalable solution for automated cyberbullying detection, making it highly suitable for real-world deployment in modern digital communication environments.

6. CONCLUSION

This research presented an Explainable Transformer-Based framework for multiclass cyberbullying detection. The integration of BERT with Explainable AI techniques significantly improved classification performance and interpretability. The proposed system effectively identifies various forms of online toxicity and provides transparency in decision-making. This framework can assist social media platforms in automated moderation and early intervention, contributing to safer digital environments.
Future work may include: Multilingual cyberbullying detection, Real-time API deployment, Integration with image and audio toxicity detection

7. ACKNOWLEDGEMENT

The authors thank Prof. V. L. Tohare and the Department of Computer Science & Engineering, Siddhivinayak Technical Campus Shegaon, for their valuable support and guidance

8. REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for
[2] Language Understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.
[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Proc. ICLR*, 2013.
[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735– 1780, 1997.
[6] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. EMNLP*, pp. 1746–1751, 2014.
[7] S. Ji, C. Pan, X. Cambria, P. Marttinen, and S. Yu, "A Survey on Mental Health Detection Using Social Media Text," *IEEE Trans. Computational Social Systems*, vol. 8, no. 4, pp. 929–946, 2021.
[8] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. NeurIPS*, pp. 4765– 4774, 2017.
[9] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," *Proc. ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.
[10] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," *Proc. ICWSM*, 2013.
[11] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.
[12] A. Vaswani et al., "Attention Is All You Need," *Proc. NeurIPS*, pp. 5998–6008, 2017.
[13] Kaggle, "Toxic Comment Classification Challenge Dataset,"
[14] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[15] World Health Organization, "Depression and Other Common Mental Disorders: Global Health Estimates," WHO Press, 2021.
[16] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," *Advances in Neural Information Processing Systems*, 2015.