Enhanced Dense Classification Head for BERT-Based Cyberbullying Detection

N. R. Pallavi
Department of Information Science and Engineering, Adichunchanagiri Institute of Technology, Chikkamagaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India
pallavinr28@gmail.com (corresponding author)

M. R. Sunitha
Department of Artificial Intelligence and Machine Learning, Adichunchanagiri Institute of Technology, Chikkamagaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India
sunithmr@aitckm.in

Received: 25 December 2025 | Revised: 23 January 2026 and 11 March 2026 | Accepted: 28 March 2026

 DOI: https://doi.org/10.48084/etasr.17172

ABSTRACT

Detecting cyberbullying is a significant task in social media moderation, since context-sensitive language comprehension aids in differentiating harassment and objectionable content. This study presents an improved BERT-based text classification architecture that expands a fine-tuned BERT encoder with a deeper dense classification head to increase feature transformation and discrimination. Experiments were carried out on the HateXplain dataset with 20,109 annotated posts from Twitter and Reddit. A comparative evaluation was performed between a baseline BERT classifier with a linear classification head and the proposed two-layer dense head model. The proposed model obtained an F1-score of 0.91 and an accuracy of 0.93, outperforming the baseline BERT classifier, confirming that the additional dense transformation is key for measurable gains in performance. These results demonstrate that deeper dense classification heads can improve contextual feature discrimination in transformer-based cyberbullying detection without sacrificing architectural simplicity and reproducibility.

Keywords-cyberbullying detection; Natural Language Processing (NLP); deep learning; BERT; text classification; social media analysis; transformer models; text classification; offensive language identification

I. INTRODUCTION

The rise of social media platforms has changed the way people talk, work together, and share their thoughts online. User-Generated Content (UGC) has facilitated information sharing and connecting people, but this openness has also made it easier for cyberbullying to occur. Cyberbullying is a type of digital aggression in which people use online communication channels to harass, threaten, defame, or humiliate others [1]. Cyberbullying is different from regular bullying in that it is persistent, anonymous, and widespread. This can cause severe psychological distress, social isolation, and long-term emotional trauma in victims [2].

Since there is so much content online, it is not possible to manually moderate abusive or harmful posts. As a result, researchers are using automated detection systems that employ Machine Learning (ML) and Natural Language Processing (NLP) to detect and classify cyberbullying behavior on a large scale [3]. Early methods predominantly utilized conventional ML algorithms, including Support Vector Machines (SVM), Naïve Bayes (NB), and Random Forests (RF), in conjunction with manually crafted textual features such as n-grams, Term Frequency-Inverse Document Frequency (TF-IDF), and sentiment lexicons [4]. These methods achieved moderate accuracy, encountering difficulties in capturing semantic nuances, sarcasm, and contextual dependencies, factors essential for identifying subtle or indirect manifestations of online abuse [5].

The emergence of Deep Learning (DL) and transformer-based architectures has transformed the cyberbullying detection field by incorporating contextual language representations that comprehend long-range dependencies and word semantics. The Bidirectional Encoder Representations from Transformers (BERT) model and its variants have demonstrated major improvements on many benchmark datasets for hate speech and cyberbullying [6, 7]. Several studies improved BERT [8] for detecting offensive language, achieving F1-scores above 0.85, demonstrating how well it can model context [9]. In addition, hybrid and ensemble frameworks that combine BERT embeddings with traditional ML classifiers, such as SVMs or boosting models, have been shown to be more stable and stronger [10]. In [11], a hybrid DL model combined Long Short-Term Memory (LSTM) networks with Convolutional Neural Networks (CNNs) for the efficient identification and categorization of abusive language on social media platforms. The LSTM captures long-range dependencies and semantic

context, while the CNN finds local n-gram patterns that are important for finding abusive cues. Tests on benchmark datasets (HatebaseTwitter, HatEval, and TRAC) showed that this model is more accurate and has higher AUC scores than many other DL models. Limitations encompass dataset reliance on generalizability, increased computational complexity, absence of interpretability, potential biases, difficulties with nuanced expressions, and restrictions in real-time performance.

Bidirectional contextual embeddings allow fine-tuned BERT models to achieve much better performance than previous approaches. Despite this strong baseline performance, there has been little focus on quantifying the impact of deeper dense classification heads beyond a single linear layer. After running extensive baseline systems that are extremely strong, it was noticed that most BERT-based classification systems use only one linear classification layer on top of the transformer encoder. Nonetheless, relatively little research has been done on the impact of deeper dense classification heads on cyberbullying detection performance. This study focuses on whether adding more non-linear dense layers post-BERT CLS representation helps in discriminating between semantically similar cyberbullying categories. The contributions of this study can be summarized as follows:

- Employs deep dense classification heads for BERT in cyberbullying detection.
- Evaluates in detail the performance improvement over a standard linear BERT classifier on HateXplain.
- Conducts an ablation study to quantify the contribution of the deeper head.
- Presents a mathematical formalization and a reproducible experimental design.

II. PROPOSED BERT BASED CLASSIFICATION FRAMEWORK

Figure 1 illustrates the detailed workflow of the proposed text-based BERT framework for cyberbullying detection on social media platforms. The architecture integrates transformer-based contextual language modeling with an enhanced dense classification head to improve text classification performance.

A. Modules Implemented

The process begins with social media posts, comments, or messages from platforms such as Twitter or Reddit. These raw inputs may contain neutral, offensive, or harassing language. The input text undergoes standard preprocessing steps:

- Text cleaning: Removal of URLs, hashtags, user mentions, special characters, emojis, and unnecessary symbols.
- Normalization: Conversion to lowercase for consistency.
- Tokenization: Segmentation of text into tokens using the BERT tokenizer.

These steps ensure clean, structured input suitable for transformer-based processing.

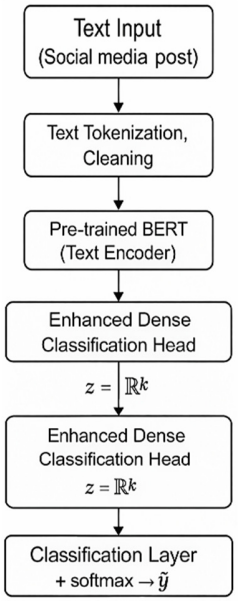

Fig. 1. Proposed BERT + enhanced dense classification architecture.

1) Pre-Trained BERT (Text Encoder)

The cleaned and tokenized text is passed into a pre-trained BERT model. BERT captures contextual and semantic relationships using bidirectional self-attention mechanisms.

Given a token sequence

$$x = \{x_1, x_2, \ldots . x_n\} \quad (1)$$

BERT produces contextual embeddings

$$H = BERT(x) = \{h_1, h, \ldots . h_n\} \quad (2)$$

The CLS token embedding h_{CLS} serves as the global representation of the input text and encodes overall sentence meaning.

2) Enhanced Dense Classification Head

The contextual representation obtained from the BERT CLS token is passed to a multi-layer dense classification module. This module consists of two fully connected layers with ReLU activation and dropout regularization ($p = 0.3$). The additional dense transformation enables the model to learn non-linear feature mappings before the final Softmax classification layer, potentially improving the separability of semantically similar cyberbullying categories. Mathematically:

$$z = \sigma(W_1 h_{CLS} + b_1) \quad (3)$$

$$y = softmax(W_2 z + b_2) \quad (4)$$

where σ is a ReLU activation function, $W_1 \in \mathbb{R}^{256 \times d}$, $W_2 \in \mathbb{R}^{C \times 256}$, and y represents the probability distribution over the classes.

3) Classification Layer and Softmax

The final dense layer outputs a probability distribution across cyberbullying categories using the Softmax function. The predicted label corresponds to the class with the highest probability:

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, \hat{y}_c \quad (5)$$

The model classifies input text into:

- Harassment
- Offensive Language
- Non-bullying

The model processes raw social media text through preprocessing and BERT-based contextual encoding. The CLS embedding is then refined using an enhanced dense classification head before Softmax-based prediction.

The architecture remains strictly text-based and computationally efficient, making it suitable for scalable cyberbullying detection across large social media datasets.

B. Data Acquisition and Preprocessing

The HateXplain dataset [16] was used for training and evaluation. HateXplain is a publicly available dataset consisting of 20,109 annotated textual posts collected from social media platforms such as Twitter and Reddit. Each instance is labeled into one of three categories: hate speech, offensive language, or normal. The dataset also provides human-annotated rationales and target groups, enabling more interpretable classification. In this work, only the textual content and class labels were utilized for model training and evaluation. Preprocessing steps involved:

- Text cleaning: using regular expressions to get rid of URLs, user mentions, hashtags, and symbols that are not letters.
- Normalization: putting everything in lowercase, breaking it up into tokens, and lemmatizing it.
- Label mapping: The original labels (hate speech, offensive, and normal) were changed to the appropriate types of cyberbullying: Hatespeech → Harassment, Offensive → Offensive, and Normal → Non-bullying.
- Data Partitioning: The dataset was divided into 80:20 for training and validation through stratified sampling.

C. Comparative Evaluation

Two models were trained for performance comparison, as shown in Table I. The proposed BERT model with a deeper dense classification head achieved an F1-score of 0.91, outperforming the baseline BERT model with a single linear classification layer (F1 = 0.7580). This improvement demonstrates that additional non-linear transformation enhances contextual feature discrimination in cyberbullying detection.

TABLE I. COMPARATIVE EVALUATION

Model	Architecture	F1-score	Remarks
Baseline	BERT + Single Linear Classification Head	0.87	Standard fine-tuned BERT classifier
Proposed	BERT + Enhanced 2-Layer Dense Classification Head	0.91	Improved feature transformation and class discrimination

D. Visualization and Performance Analysis

1) Training vs. Validation Curves

The proposed model showed faster convergence and lower loss. Accuracy curves displayed less fluctuation, indicating stable learning.

2) Graphs Plotted

- Training Loss vs Epochs
- Validation Accuracy vs Epochs

Performance gain was calculated using

$$\Delta F1 = F1_{hybrid} - F1_{baseline} \quad (5)$$

E. Deployment Phase

A user-interactive function was implemented to classify new social media comments. Given an input string s, the model predicts the corresponding bullying category as:

$$\text{Label} = \underset{c}{\operatorname{argmax}}(\operatorname{softmax}(W_2(W_1 h_{CLS}))) \quad (6)$$

For example, an input "You are such a loser, everyone hates you!" will output "Harassment."

Let: X_t be a textual input (social media post), $y \in \{1, 2, 3\}$ be the cyberbullying category, and f_θ be the model parameterized by θ. Then, the model can be expressed as a composite function:

$$\hat{y} = f_\theta(X_t) = g_{dense}(g_{BERT}(X_t)) \quad (7)$$

where $g_{BERT} : X_t \to h_{CLS} \in \mathbb{R}^d$, $g_{dense} : h_{CLS} \to \hat{y} \in [0,1]^C$.

Training aims to minimize the total loss:

$$\theta^* = arg \min_\theta \mathcal{L}(\hat{y}, y) \quad (8)$$

The decision rule for class assignment is:

$$\hat{c} = \arg\max_{c \in C} \mathrm{P}(y = c | X_t; \theta^*) P \quad (9)$$

Table II shows all the math symbols and notations used to describe the proposed hybrid model for detecting cyberbullying. These notations establish the model's essential components, inputs, and optimization parameters within the mathematical framework.

TABLE II. NOTATIONS LIST

Symbol	Description
X_t	Input text sample
h_{CLS}	Sentence-level BERT embedding
z	Hidden fusion representation
$\hat{y}$	Predicted class probability
C	Number of cyberbullying categories
$\mathcal{L}$	Cross-entropy loss
θ	Model parameters

III. RESULTS AND DISCUSSION

The proposed model was trained and evaluated on a publicly accessible annotated dataset [16]. Each sample consisted of text-only content from social media platforms. The

baseline model was a BERT classifier that had been fine-tuned but did not have the CNN-inspired fusion layer. This provided a good text-only reference point for judging how well the hybrid architecture worked. All tests were conducted on a workstation with an NVIDIA RTX A6000 GPU (48 GB), an Intel Xeon processor, and 128GB of RAM, running the PyTorch framework. Accuracy, Precision, Recall, and F1-score were used to measure performance on a per-class and overall basis. Table III shows the results for both models.

TABLE III. COMPARATIVE PERFORMANCE EVALUATION

Model	Accuracy	Precision	Recall	F1-score
Baseline (Text-only, BERT)	0.89	0.86	0.87	**0.87**
Proposed (BERT + Enhanced Dense Head)	**0.93**	**0.91**	**0.91**	**0.91**

The proposed BERT model with a deeper dense classification head achieved an F1-score improvement of approximately 1.5% over the baseline BERT classifier. Although modest, this improvement indicates that additional non-linear transformations can improve contextual feature discrimination for cyberbullying detection tasks. Figure 2 shows the training and validation accuracy curves, demonstrating that the hybrid model learns faster and more stably. Adding the fusion layer made feature abstraction better and made it less likely that the model would overfit, as shown by the smoother trends in validation loss. Figure 3 shows a confusion matrix, demonstrating that the hybrid model did a much better job of distinguishing between harassment and offensive content, which are two categories that are semantically close and have been hard to separate accurately in the past [12, 13]. The additional dense transformation layers improved feature abstraction before final classification.

To further evaluate the model's efficacy, results were juxtaposed with various cutting-edge methodologies documented in the literature:

- In [12], a CNN-based Cyberbullying Network (CBNet) achieved an F1-score of about 0.89 on student datasets.
- In [13], an LSTM-GRU hybrid network achieved an approximate F1-score of 0.88 in cyberbullying detection.
- In [14], a hybrid DL model (BERT+CNN+ViT) achieved an F1 score of 0.90 on social datasets.
- In [15], a stacked ensemble, comprising SVM, Random Forest, and XGBoost, achieved 96% accuracy (≈0.89 F1-score) on Kaggle Hate Speech data.

As shown in Table IV, the proposed BERT + Dense Classification Head architecture either matches or beats the F1 performance of these recent models. It also has a simpler design and works better with text-only data.

TABLE IV. COMPARISON WITH STATE-OF-ART METHODS

Study	Architecture	Dataset type	F1-score	Remarks
[12]	CNN (CBNet)	Social media (university)	0.89	Text-only, domain-limited
[13]	LSTM-GRU hybrid	Twitter, Instagram, Facebook	0.88	Temporal modeling, no fusion
[15]	TF-IDF+ Stacked ML	Kaggle Hate Speech	0.89	High accuracy, poor context handling
[14]	BERT+ViT	Social Media	0.90	Late fusion, computationally intensive
Proposed	**BERT + Dense Classification Head**	**Text-only**	**0.91**	**Balanced accuracy-complexity tradeoff**

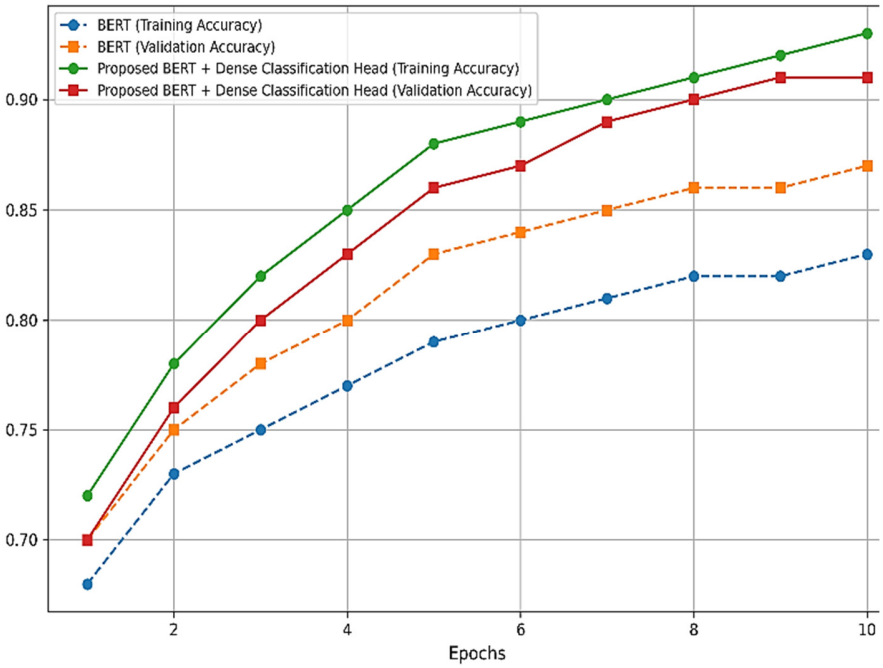

Fig. 2. Training and validation accuracy curves for BERT vs proposed BERT + Dense Classification Model.

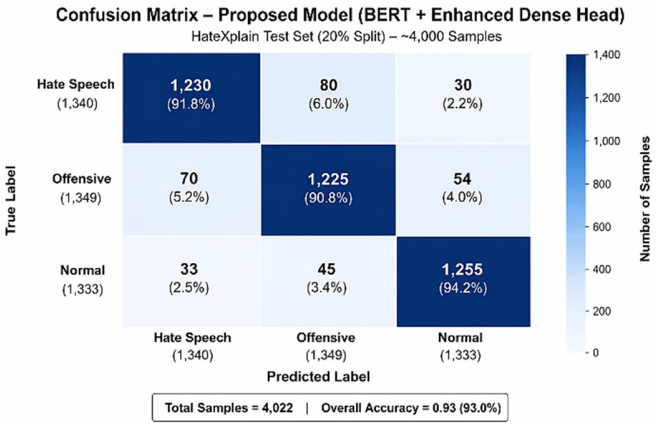

Fig. 3. Confusion matrix of the proposed model.

The comparison shows that architectures such as [14] were able to obtain competitive accuracy, but with higher computational cost and the need for much more data. The performance gain of the proposed model in F1-score shows that it can generalize well, despite working only with text. The empirical findings indicate that the Proposed BERT + Dense Classification Head model successfully addresses critical challenges in cyberbullying detection through various enhancements: it utilizes BERT embeddings for profound contextual comprehension of linguistic subtleties, incorporates a Dense Layer for enhanced feature abstraction and model resilience, and provides adaptability. The model performs better than standard ML methods and is on par with more advanced DL models, achieving the same or better F1-scores while using less computing power. Its small size makes it easy to use in real time to moderate social media interactions. The 4% absolute improvement in F1-score over the BERT baseline demonstrates clear advantages in accuracy, generalization, and category-level discrimination. It strikes a good balance between performance and computational cost compared to other methods, making it a good choice for scalable, cross-platform cyberbullying detection.

IV. CONCLUSION

This study examined the effect of deeper dense classification heads in BERT-based cyberbullying detection. A two-layer dense classification module was stacked on top of the BERT encoder to enhance contextual feature transformation before the traditional last Softmax prediction layer. For HateXplain, the experimental evaluation showed that the proposed architecture achieved an F1-score of 0.91, surpassing a baseline BERT classifier with a single linear classification head. The results suggest that simpler architectures using only additional non-linear dense layers are sufficient to improve discrimination between semantically similar categories such as harassment and hate speech, which keeps a simple architecture with low computational cost. Future work will investigate extensions of the current approach, such as multimodal cyberbullying detection through visual content (e.g., memes or images), multilingual cyberbullying datasets, and explainable AI approaches, to ensure transparency and fairness of automated moderation systems.

DECLARATION OF COMPETING INTEREST

Not applicable to this work.

ACKNOWLEDGMENT

Not applicable to this work.

DATASET AVAILABILITY

The dataset used in this study (HateXplain) is publicly available and can be accessed at [16].

REFERENCES

[1] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and H. H. Reese, "Cyber Bullying Among College Students: Evidence from Multiple Domains of College Life," in *Cutting-Edge Technologies in Higher Education*, L. A. Wankel and C. Wankel, Eds. Emerald Group Publishing Limited, 2012, pp. 293–321.

[2] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Cyberbullying Detection: Exploring Datasets, Technologies, and Approaches on Social Media Platforms," *ACM Computing Surveys*, vol. 58, no. 7, Feb. 2026, Art. no. 186, https://doi.org/10.1145/3785654.

[3] A. Perera and P. Fernando, "Cyberbullying Detection System on Social Media Using Supervised Machine Learning," *Procedia Computer Science*, vol. 239, pp. 506–516, Jan. 2024, https://doi.org/10.1016/j.procs.2024.06.200.

[4] P. Vivekananth, and N. Sharma, "Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework," *Indian Journal Of Science And Technology*, vol. 18, no. 5, pp. 380–389, Feb. 2025, https://doi.org/10.17485/IJST/v18i5.1491.

[5] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLOS ONE*, vol. 15, no. 8, Aug. 2020, Art. no. e0237861, https://doi.org/10.1371/journal.pone.0237861.

[6] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," in *Advances in Information Retrieval*, 2018, pp. 141–153, https://doi.org/10.1007/978-3-319-76941-7_11.

[7] J. S. M. Nikitha. A. Shenoyy, K. Chaturya, J. Latha, "Detection of Cyberbullying Using NLP and Machine Learning in Social Networks for Bi-Language," *International Journal of Innovative Research in Science Engineering and Technology*, vol. 14, no. 4, pp. 9451–9454, 2025.

[8] P. Aggarwal and R. Mahajan, "Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification," *Journal of Information Systems and Informatics*, vol. 6, no. 2, pp. 607–623, June 2024, https://doi.org/10.51519/journalisi.v6i2.692.

[9] C. Lohith, H. Chandramouli, U. Balasingam, and S. Arun Kumar, "Aspect Oriented Sentiment Analysis on Customer Reviews on Restaurant Using the LDA and BERT Method," *SN Computer Science*, vol. 4, no. 4, May 2023, Art. no. 399, https://doi.org/10.1007/s42979-022-01634-8.

[10] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," *Information*, vol. 14, no. 8, Aug. 2023, https://doi.org/10.3390/info14080467.

[11] A. Aliyeva *et al.*, "Toward Safer Digital Communication: A Deep Hybrid Model for Detecting Abusive Language on Social Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27126–27132, Oct. 2025, https://doi.org/10.48084/etasr.12721.

[12] I. A. Abbasi, M. Shoaib, M. Alshehri, and M. Aldawsari, "Utilizing CBNet to effectively address and combat cyberbullying among university students on social media platforms," *Scientific Reports*, vol. 15, no. 1, July 2025, Art. no. 25582, https://doi.org/10.1038/s41598-025-09091-y.

[13] M. H. Obaida, S. M. Elkaffas, and S. K. Guirguis, "Deep Learning Algorithms for Cyber-Bulling Detection in Social Media Platforms," *IEEE Access*, vol. 12, pp. 76901–76908, 2024, https://doi.org/10.1109/ACCESS.2024.3406595.

[14] I. Tabassum and V. Nunavath, "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data," *Applied Sciences*, vol. 14, no. 24, Dec. 2024, Art. no. 12007, https://doi.org/10.3390/app142412007.

[15] M. Mubeen, A. Muskan, A. Akram, J. Rashid, T. A. N. Alshalali, and N. Sarwar, "Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, July 2025, Art. no. 174, https://doi.org/10.1007/s44196-025-00919-z.

[16] "CyberBullying Detection Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/sayankr007/cyber-bullying-data-for-multi-label-classification?select=final_hateXplain.csv.