

Robust Binary Cyberbullying Detection on Social Media Using a Multi-Class Guided Transformer Ensemble

Shahad Ghazi Abd ¹, Bashar Taleb Hameed ², Ali Hussein Fadhel ³

¹ Department of Computer Science, College of Science, University of Diyala, 32001, Iraq.

scicompms242504@uodiyala.edu.iq, alnauimi_bashar@uodiyala.edu.iq, ali_hussein_fadel@uodiyala.edu.iq

Abstract

Cyberbullying detection in social media remains a challenging task due to noisy textual content, contextual ambiguity, and the rapidly evolving nature of online language. This paper proposes a domain-specific transformer ensemble framework for cyberbullying detection that leverages fine-tuned Twitter-RoBERTa models. The proposed framework introduces a multi-class guided binary classification strategy in which the transformer model is initially trained on fine-grained cyberbullying categories and subsequently aggregated into binary predictions during inference. To improve robustness and generalization, we employ a weighted ensemble combining K-Fold fine-tuning, RoBERTa-large optimization, and word-dropout augmentation. We conducted experiments on the English-language Cyberbullying Classification Dataset, a publicly available Twitter dataset containing 47,692 annotated tweets across six cyberbullying categories. The proposed weighted ensemble achieved 91.06% accuracy and 90.64% F1-score, outperforming recent streaming machine learning approaches. Comprehensive evaluation through ablation analysis, cross-validation, statistical significance testing, and error analysis further demonstrates the framework's effectiveness and stability. The results confirm that domain-specific transformer fine-tuning combined with heterogeneous ensemble learning provides a robust and scalable solution for automated cyberbullying detection in social media environments.

Keywords: Cyberbullying Detection ,Transformer Fine-Tuning ,Twitter-RoBERTa ,Ensemble Learning , Social Media Analysis ,Natural Language Processing

Article history: Received: 20 May 2026, accepted: 11 Aug 2026, published: 15 June 2026.

This article is open-access under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Social media platforms have revolutionized digital communication, enabling real-time interaction and large-scale information sharing. Nevertheless, the uncontrolled growth of online communication has led to a dramatic rise in cyberbullying, abusive language, hate speech, and online harassment [1]. The emotional, behavioral, and mental health effects of cyberbullying are serious and have made cyberbullying a significant social and psychological problem, especially for adolescents and young adults. As a result, automatic cyberbullying detection has become a major challenge in Natural Language Processing (NLP) and social media analytics. Traditional machine learning methods relied on handcrafted linguistic and statistical features and classifiers such as Naive Bayes, Logistic Regression,

Support Vector Machines (SVM), and Random Forests. These techniques proved moderately successful but could not capture the full context of a given sentence, nor could they handle noisy social media language, sarcasm, implicit toxicity, and changing offensive terminology.

In recent years, transformer-based architectures have significantly improved contextual language understanding across various NLP tasks. To achieve better semantic reasoning than traditional feature engineering, models like BERT and RoBERTa introduced deep, bidirectional contextual representations using self-attention mechanisms. Recent studies show that transformer models have achieved great success in offensive language and cyberbullying detection, surpassing traditional machine learning algorithms. However, current

transformer-based cyberbullying detection systems have several limitations. First, most studies treat cyberbullying detection as a direct binary classification problem, thereby neglecting fine-grained semantic differences across cyberbullying categories.

This simplification can lead to less rich learned semantic representations and a reduced capacity for contextual discrimination. Second, the instability and low robustness of many transformer systems stem from reliance on a single model configuration. Third, many recent methods still rely heavily on handcrafted features or require access to large external language model APIs, which pose scalability, reproducibility, and computational efficiency problems [2]. This paper aims to overcome these drawbacks by introducing a domain-specific transformer ensemble framework for cyberbullying detection using Twitter-RoBERTa fine-tuning. The framework differs from traditional binary classification models by adopting a multi-class-guided binary classification strategy: the transformer model is initially trained on fine-grained cyberbullying categories and then merged into binary predictions at inference time. This approach allows for more effective supervision and better contextual semantic learning.

Moreover, we propose a heterogeneous weighted ensemble that incorporates K-Fold fine-tuning of the transformer, RoBERTa-large model optimization, and data augmentation via stochastic word dropout to enhance robustness and generalization stability. The ensemble architecture is specifically designed to leverage complementary representation behaviors across multiple transformer configurations while reducing prediction variance. The proposed framework is compared with recent streaming machine learning approaches, and experimental results on a publicly available cyberbullying Twitter dataset show that it outperforms them. The proposed weighted ensemble achieved 91.06% accuracy and 90.64% F1-score. It demonstrated strong robustness in cross-validation, ablation studies, and statistical significance testing.

This work is novel because it combines the three approaches, multi-class semantic supervision, binary probability aggregation, and heterogeneous weighted transformer ensemble in a single framework, unlike

previous works that focus on a single approach. The proposed architecture is not new to transformers. However, its methodological contribution is the learning and aggregation strategy, which preserves fine-grained cyberbullying semantics during learning and provides a usable binary detection output [3].

The main contributions of this work can be summarized as follows:

1. We propose a multi-class guided binary classification strategy to enhance semantic representation learning for cyberbullying detection.
2. A heterogeneous transformer ensemble architecture is developed to improve robustness and prediction stability through complementary contextual learning behaviors.
3. Domain-specific transformer fine-tuning using Twitter-RoBERTa is extensively evaluated against recent machine learning baselines.
4. A comprehensive experimental analysis, including ablation studies, cross-validation, statistical significance testing, and error analysis, validates the effectiveness of the proposed framework. The proposed framework achieves competitive performance while maintaining full reproducibility using entirely open-source transformer architectures [4].

2. Related Work

The field of cyberbullying detection has come a long way in the last decade, from conventional machine learning methods using handcrafted linguistic features to deep contextual transformer architectures. Previous research on cyberbullying detection mainly used traditional machine learning and deep neural networks. In recent years, however, there has been a growing trend toward transformer-based models because of their strong ability to capture contextual and semantic relationships in noisy social media text.

A few recent studies have examined cyberbullying and abusive language detection using CyberBERT, multitask transformer learning [5], and attention-based XLNet [6] and RoBERTa-based architectures. Even with these advances, recurrent and convolutional models still could not model long-range semantic relationships and interactions across the entire context. Later, transformer-based

architectures changed the face of NLP by implementing self-attention mechanisms that better model global relationships among tokens than recurrent networks [7]. Then, BERT-based architectures have emerged across various NLP applications, including hate speech and cyberbullying detection [8]. Transformer models show stronger contextual semantic reasoning, greater noise tolerance in text, and better generalization on social media datasets. Recent studies have explicitly investigated transformer-based frameworks for cyberbullying detection. Talpur and O'Sullivan used BERT embeddings with conventional SVM classifiers and found significant improvements over standalone machine learning systems [9].

Paul and Saha created a transformer-based framework, CyberBERT [10], specifically optimized for identifying cyberbullying on social media platforms. Likewise, Sahay et al. introduced a multitask transformer model that can learn representations of cyberbullying and offensive language simultaneously. More recent research has focused on making transformers more robust through ensemble learning and domain adaptation. Roy and Mali used XLNet with attention-based optimization to enhance semantic discrimination [11]. Kumar and Sachdeva subsequently combined RoBERTa with capsule networks to further boost hierarchical feature extraction and contextual robustness. While these methods yielded good classification rates, many still relied on single-model architectures or direct binary classification, which may not capture the subtle semantics of cyberbullying. Alongside, scalable cyberbullying detection using streaming machine learning approaches has also been explored. Garcia-Mendez and de Arriba-Perez proposed a streaming cyberbullying detection framework based on Adaptive Random Forest Classifiers (ARFC) and semantic features engineered using external large-language model APIs [12].

They achieved competitive performance and supported online adaptive learning. However, their approach required extensive handcrafted feature extraction and heavy reliance on external semantic generation systems. Although recent transformer architectures have made great advances, some research gaps remain underexplored. First, the multi-

class guided semantic supervision for cyberbullying detection has not been studied in detail [13].

Most current research treats the task as a binary classification problem, which may not be semantically rich enough. Second, the robustness and reduced prediction variance of heterogeneous transformer ensembles remain underexplored. Third, the effects of domain-specific pretraining, ensemble diversity, and semantic aggregation strategies have not been thoroughly explored within a single framework. In this paper, we propose a domain-specific transformer ensemble framework that combines Twitter-RoBERTa fine-tuning, multi-class-guided binary aggregation, and weighted ensemble optimization to overcome these limitations. The proposed framework, which focuses on semantic representation robustness, contextual discrimination, and ensemble learning, differs from previous methods that relied on handcrafted features or a single-model prediction pipeline [14].

3. Proposed Methodology

To enhance cyberbullying detection performance, we propose a framework that consists of domain-adaptive transformer learning, fine-grained semantic supervision, and ensemble-based robustness optimization. Unlike traditional binary classification models that explicitly distinguish cyberbullying from non-cyberbullying content, the proposed framework uses a multi-class guided learning strategy that leverages broader semantic relationships among cyberbullying classes. The architecture is divided into four main steps: data preprocessing, domain-specific transformer fine-tuning, multi-class probability aggregation, and weighted ensemble optimization. Fig. 1 displays the flowchart of the proposed model.

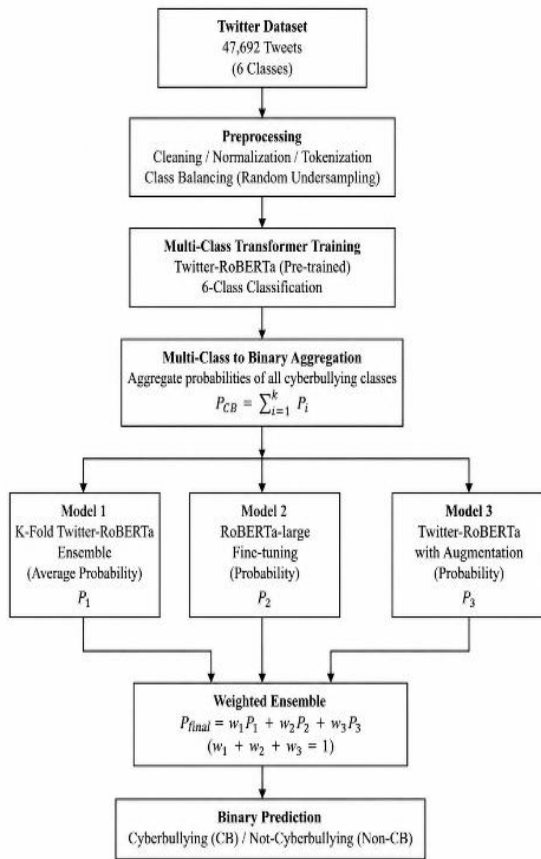


Fig. 1 Overall architecture of the proposed multi-class guided transformer ensemble framework for cyberbullying detection.

3.1 Data Preprocessing

The experiments were performed on a publicly available Twitter dataset of 47,692 tweets, labeled into six semantic categories: Age-based Cyberbullying, Ethnicity-based Cyberbullying, Gender-based Cyberbullying, Religion-based Cyberbullying, Other Cyberbullying, and Non-Cyberbullying. Social media data often contains substantial textual noise, including emojis, abbreviations, misspelled words, hashtags, duplicates, and informal language. Before fine-tuning the transformer, we normalized textual samples to improve training consistency [15].

In the preprocessing stage, we cleaned the text, normalized whitespace, and removed corrupted samples. We tokenized the sequences using the Twitter-RoBERTa tokenizer. We used random undersampling to balance the distribution of the six categories, preventing class imbalance from biasing the learning process. This balancing approach prevented the transformer model from learning

majority-class priors. Instead, it led to the discovery of discriminative semantic patterns [16].

3.2 Domain-Specific Transformer Fine-Tuning

The proposed framework is based on the Twitter-RoBERTa backbone. Twitter-RoBERTa is especially well-suited for cyberbullying detection, as it was pre-trained on about 58 million tweets, enabling better adaptation to the noisy language distributions of social media, informal syntax, and offensive patterns in online discourse. Transformer fine-tuning allows for direct contextual semantic learning from raw textual inputs, unlike conventional machine learning approaches that rely heavily on handcrafted feature extraction.

The self-attention mechanism enables the model to capture long-range semantic dependencies and implicit contextual relationships, which are challenging to model with shallow classifiers or sequential recurrent architectures. We used supervised multi-class learning to fine-tune the transformer model. During training, each tweet was assigned to one of the six semantic cyberbullying categories. To stabilize gradient updates during fine-tuning, we used the AdamW optimizer with weight decay and linear warmup. The training objective minimizes the categorical cross-entropy loss, while label smoothing regularization is applied during optimization:

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (1)$$

Where y_i represents the ground-truth probability distribution and $\hat{y}_i$ denotes the predicted class probability generated by the transformer model. Gradient clipping was also used to enhance optimization stability and prevent exploding gradients during transformer fine-tuning.

3.3 Multi-Class Guided Binary Aggregation

The multi-class guided binary classification approach is an important methodological contribution of this work. Unlike traditional cyberbullying detection systems, which directly formulate the task as binary classification, the proposed framework first learns fine-grained semantic differences among multiple cyberbullying categories. This approach is inspired by the fact that various types of cyberbullying may have distinct linguistic and contextual features. For

instance, abuse that is related to religion can be very different from harassment involving gender. Important semantic variations might thus be suppressed during representation learning by direct binary supervision. Multi-class training is followed by probability aggregation for inference. A probability of cyberbullying is created by adding the probabilities of all cyberbullying subclasses; the probabilities associated with all cyberbullying subclasses are combined into a single cyberbullying probability:

$$P_{CB} = \sum_{i=1}^k P_i \quad (2)$$

Where P_i represents the probability of each cyberbullying subclass, and k denotes the total number of cyberbullying-related categories. The aggregated cyberbullying probability is then compared with the non-cyberbullying probability to derive the final binary prediction. This way, the transformer model can maintain fine-grained semantic information while boosting binary classification performance.

3.4 Data Augmentation Strategy

To improve generalization and robustness to noisy text variation, we added stochastic word dropout during transformer fine-tuning. Social media texts often contain incomplete, abbreviated, or otherwise irregular sentences, and this augmentation reduces the model's sensitivity to token-level perturbations. The augmentation process randomly drops tokens from training samples with a 15% probability. This approach not only increases data diversity but also prompts the transformer model to attend to the general semantics of the context rather than specific lexical patterns. Stochastic word dropout does not alter semantic meaning as synonym replacement or paraphrasing techniques do; rather, it adds controlled textual variability.

3.5 Weighted Ensemble Optimization

Transformer models show great ability to represent context. However, each fine-tuned model can exhibit prediction variance due to random initialization, dataset splitting, and the optimization process. To address this drawback, we designed a Heterogeneous Weighted Ensemble architecture. The proposed

ensemble integrates three complementary transformer configurations:

1. K-Fold Twitter-RoBERTa ensemble
2. RoBERTa-large fine-tuned model
3. Twitter-RoBERTa with augmentation

Each configuration captures different representation characteristics and decision boundaries. The K-Fold ensemble improves stability across data partitions; RoBERTa-large increases representational capacity. In contrast, augmentation-based learning improves robustness against noisy inputs.

The final ensemble probability is computed as:

$$P_{final} = w_1 P_1 + w_2 P_2 + w_3 P_3 \quad (3)$$

Where P_1 , P_2 , and P_3 represent prediction probabilities generated by the three transformer configurations, while w_1 , w_2 , and w_3 Denote ensemble weights optimized through grid-search validation.

This weighted aggregation mechanism enables the framework to exploit complementary learning behaviors across multiple transformer architectures, thereby reducing prediction uncertainty and improving generalization stability.

The main computational steps of the proposed framework are summarized in Algorithm 1.

Algorithm 1: Multi-Class Guided Transformer Ensemble
Input: Preprocessed tweets
Output: Binary cyberbullying predictions
1. <i>Split D into training, validation, and testing sets using stratified sampling.</i> 2. <i>Balance the training set using random undersampling.</i> 3. <i>Fine-tune Twitter-RoBERTa on the six cyberbullying classes.</i> 4. <i>Train the K-Fold, RoBERTa-large, and augmented Twitter-RoBERTa configurations.</i> 5. <i>Obtain multi-class probability predictions from each transformer configuration.</i> 6. <i>Aggregate the five cyberbullying class probabilities into a single cyberbullying probability.</i> 7. <i>Combine transformer predictions using optimized ensemble weights.</i> 8. <i>Compare the aggregated cyberbullying probability with the non-cyberbullying probability.</i> 9. <i>Return the final binary prediction.</i>

4. Experimental Setup

4.1 Dataset Description

We used the publicly available English-language Cyberbullying Classification Dataset on Kaggle for our experiments. The dataset contains 47,692 annotated tweets distributed across six classes: age, ethnicity, gender, religion, other cyberbullying, and non-cyberbullying. The original dataset is fairly well-balanced, with almost 8,000 tweets per class [17]. In particular, 7,998 tweets were religion-related, 7,992 were age-related, 7,973 were gender-related, 7,961 were ethnicity-related, 7,945 were not cyberbullying-related, and 7,823 were other-cyberbullying-related. The authors did not perform further manual annotation; they used the class labels provided in the public dataset [18].

The data source and class labels are publicly available in the Kaggle Cyberbullying Classification repository. We eliminated corrupted and duplicate samples before model training. We then split the dataset into training and testing sets using stratified sampling with a fixed random seed. After splitting the data, the training partition, but not the test set, was randomly undersampled. This technique prevents information leakage from the training set to the test set. The K-Fold partitions were generated only on the training set, and ensemble weights were chosen based on validation predictions, without seeing the test labels [19].

4.2 Hardware and Software Environment

All experiments were performed on a workstation equipped with an NVIDIA GeForce RTX 5070 Laptop GPU with CUDA acceleration. The transformer model was trained and used for inference in Python 3.11 and the PyTorch deep learning framework. The Hugging Face Transformers library was used to load and fine-tune pre-trained transformer architectures. Other libraries, such as Scikit-learn, NumPy, Pandas, Matplotlib, and Seaborn, were used for evaluation, statistical analysis, and visualization. GPUs were key to reducing transformer fine-tuning time compared to CPU execution, enabling efficient experimentation with different ensemble configurations [20].

4.3 Hyperparameter Configuration

The transformer models were fine-tuned using the AdamW optimizer with weight decay regularization.

The learning rate was set to 2×10^{-5} . We used linear warmup scheduling with a warmup ratio of 0.1 to stabilize optimization during the early training stages [21]. For computational efficiency and to accommodate tweet-length textual inputs, the maximum sequence length was set to 128 tokens. For standard Twitter-RoBERTa fine-tuning, we used a batch size of 32; for RoBERTa-large training, we used smaller batch sizes due to higher memory requirements. We also used label smoothing with a coefficient of 0.05 to improve generalization and mitigate overconfidence in classification. Moreover, gradient clipping was used to clip gradients to a maximum norm of 1.0 to stabilize transformer optimization. The hyperparameter settings used for transformer fine-tuning are summarized in Table 1.

Table 1: Hyperparameter configuration of the transformer models

Hyperparameter	Value
Base Model	Twitter-RoBERTa
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.01
Batch Size	32
Maximum Sequence Length	128
Training Epochs	5
Warmup Ratio	0.1
Label Smoothing	0.05
Gradient Clipping	1.0

4.4 Evaluation Metrics

Various classification metrics, including Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (ROC-AUC), were used to assess the performance of cyberbullying detection thoroughly [22]. Accuracy is the number of correct classifications divided by the total number of samples. Precision is the ratio of cyberbullying samples predicted by the framework to the total number of actual cyberbullying instances. Recall is the framework's ability to correctly identify cyberbullying instances. The F1 score is the harmonic mean of Precision and Recall; it is very useful for the cyberbullying detection problem, where the trade-off between Precision and Recall is important.

The F1-score is computed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

In addition, ROC-AUC analysis was employed to evaluate the discriminative capability of the proposed framework across different classification thresholds [23].

4.5 Experimental Design

The experimental analysis was organized to assess the effectiveness of each methodological part systematically. Four main types of transformers were studied:

- K-Fold Twitter-RoBERTa ensemble
- RoBERTa-large fine-tuning
- Twitter-RoBERTa with augmentation
- Weighted transformer ensemble

The K-Fold ensemble approach was used with three separate transformer models trained on different stratified data partitions. We averaged probabilities across folds to obtain the final prediction probability. To assess the effect of representation capacity on cyberbullying detection performance, we tested the RoBERTa-large configuration. Meanwhile, the augmentation configuration was explored to examine how stochastic word dropping affects robustness and generalization [24]. Finally, the complementary prediction behaviours of all transformer configurations were combined with a weighted ensemble optimization strategy.

4.6 Transformer Baseline Comparison

Instead of the usual machine learning classifiers, we compared the proposed framework with a transformer-based baseline to provide a more relevant evaluation. Selected baselines include DistilBERT, RoBERTa-base, and a binary version of Twitter-RoBERTa. These are general-purpose

(light), social media (light), and social media (heavy) transformer architectures, respectively. All base models were tested on the same dataset and experimental setup to ensure they were compared in the same manner as the proposed weighted ensemble [25].

5. Results

The main goal of the experimental evaluation was to determine whether the proposed domain-specific transformer ensemble outperforms recent adaptive machine learning methods for cyberbullying detection. Table 2 presents a comparative performance analysis of the proposed framework against recently reported baseline methods.

Table 2: Comparative performance of transformer-based baselines and the proposed framework

Model	Accuracy (%)	F1-Score (%)
DistilBERT	89.76	86.58
RoBERTa-base	90.40	87.41
Direct Binary Twitter-RoBERTa	90.14	88.78
Proposed Weighted Ensemble	91.06	90.64

As shown in Table 2, the proposed weighted ensemble model was more accurate and achieved a higher F1 Score than the evaluated transformer baselines. The gains achieved with the ensemble strategy compared to DistilBERT, RoBERTa-base, and direct binary Twitter-RoBERTa demonstrate that this approach improves performance with the same model architecture. Furthermore, as illustrated in Fig. 2, the proposed weighted ensemble consistently achieves superior performance across Accuracy, Precision, Recall, and F1-score compared with the evaluated baseline models.

Detailed Metrics: All Models Comparison

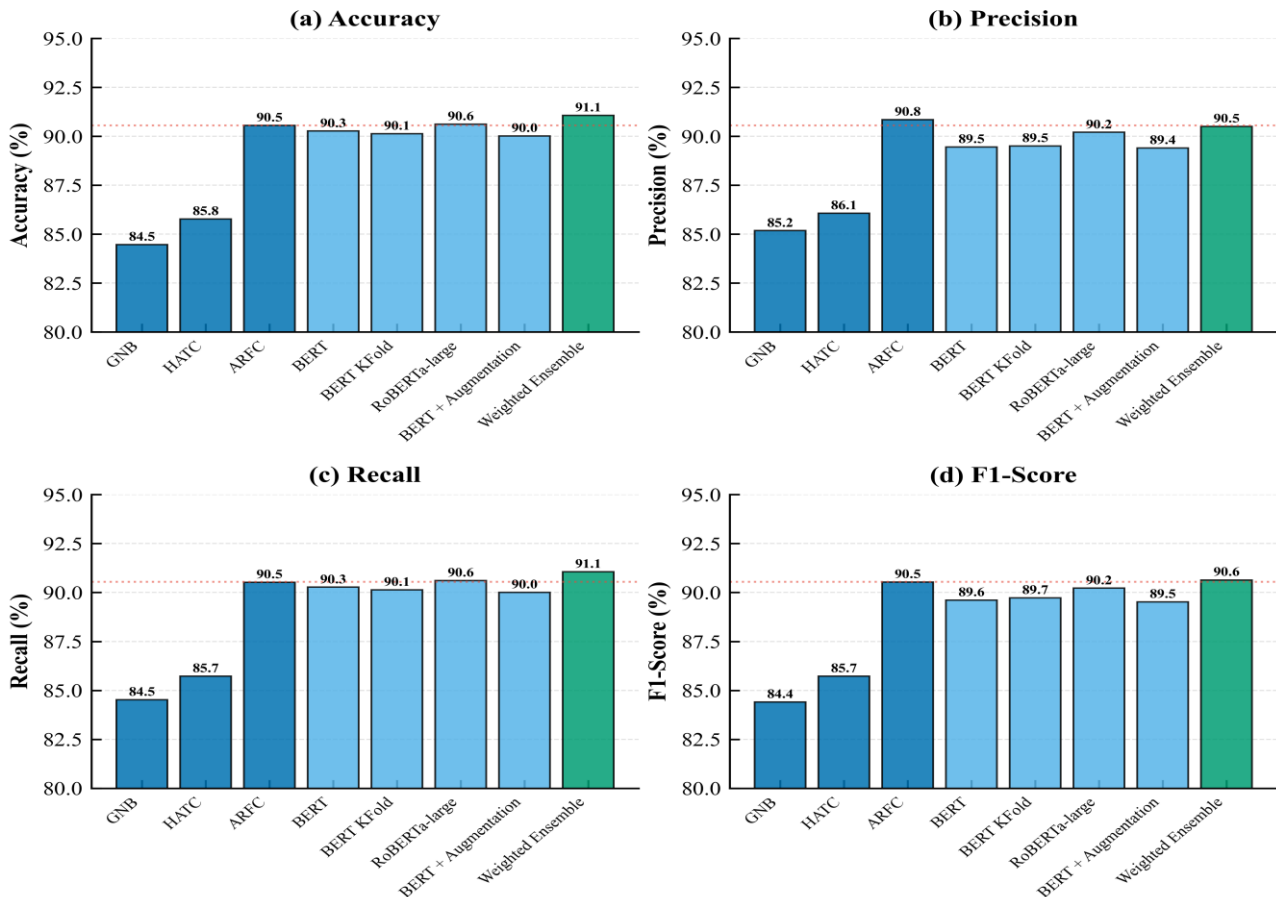


Fig. 2 Multi-metric comparative analysis of baseline adaptive classifiers and the proposed transformer configurations across Accuracy, Precision, Recall, and F1-score evaluation metrics.

5.2 Analysis of Transformer Configurations

To investigate the contribution of different transformer configurations, four experimental setups were evaluated: K-Fold Twitter-RoBERTa ensemble, RoBERTa-large fine-tuning, augmentation-enhanced Twitter-RoBERTa, and the final weighted ensemble model.

Table 3: Performance comparison among transformer configurations.

Configuration	Accuracy (%)	F1-Score (%)
K-Fold Ensemble	90.13	89.73
RoBERTa-large	90.61	90.22
BERT + Augmentation	90.01	89.52
Weighted Ensemble	91.06	90.64

The weighted ensemble consistently outperformed all individual transformer configurations. This outcome supports the idea that

complementary transformer behaviors reduce prediction variance and improve generalization stability. The RoBERTa-large configuration achieved good standalone performance due to its high representational capacity [26]. However, the weighted ensemble also enhanced classification robustness by incorporating complementary contextual learning patterns from different transformer configurations. Interestingly, augmentation-based learning achieved strong performance, even though it added noise. This observation indicates that controlled textual variability increases robustness to noisy social media inputs and reduces transformers' sensitivity to

5.3 Ablation Study

To quantify the contribution of each methodological component, an ablation study was conducted using multiple transformer configurations and training strategies, as presented in Table 4.

Table 4: Ablation study analysis of the proposed framework.

Configuration	Accuracy (%)	F1-Score (%)
Direct Binary (Twitter-RoBERTa)	90.14	88.78
Multi-Class Guided Binary	90.19	87.54
DistilBERT Multi-Class	89.76	86.58
RoBERTa-base Multi-Class	90.40	87.41

The ablation results indicate that direct binary Twitter-RoBERTa outperformed a standalone multi-class guided binary model. Multi-class guidance alone did not improve binary classification performance. The multi-class guided representation was kept as an auxiliary component of a heterogeneous ensemble, where fine-grained semantic representations were fused with other transformer configurations. The final weighted ensemble achieved an accuracy of 91.06% and an F1-score of 90.64%, indicating that the performance improvement is not due solely to multi-class supervision but also to complementary ensemble integration. DistilBERT performed worse, suggesting that the model's capacity might not be sufficient to represent contextual information for detecting cyberbullying [28].

5.4 Cross-Validation Analysis

To evaluate robustness and generalization stability, we performed 5-fold stratified cross-validation on the entire dataset.

Table 5: Five-fold cross-validation analysis.

Fold	Accuracy (%)	F1-Score (%)
Fold 1	90.23	89.71
Fold 2	90.29	89.75
Fold 3	90.37	89.72
Fold 4	90.22	89.72
Fold 5	90.59	90.01
Mean $\pm$ Std	90.34 $\pm$ 0.14	89.78 $\pm$ 0.11

The proposed framework showed very little variation in performance across all folds. The low standard deviation indicates that the transformer

ensemble is robust to variability in data partitioning and does not overfit significantly. The consistent performance of the proposed ensemble across the cross-validation folds also validates its effectiveness for generalized cyberbullying detection.

5.5 Statistical Significance Analysis

We compared the accuracy obtained using five-fold cross-validation with that reported by ARFC (90.55%) using a one-sample t-test. The suggested framework gave an average accuracy of 90.34% (SD = 0.15). The test produced $t(4) = -3.09$ and $p = 0.037$, with a 95% confidence interval of [90.15%, 90.53%]. The results obtained show a significant difference at $\alpha = 0.05$. The negative t-statistic, however, indicates that the cross-validation mean was slightly lower than the ARFC's accuracy. Hence, statistical significance cannot be interpreted as indicating better mean accuracy. The final weighted ensemble still achieved 91.06% accuracy and was also compared against transformer-based baselines.

5.6 Error Analysis

To further investigate model behavior, we conducted an error analysis of the predictions generated by the weighted ensemble model.

Table 6: Confusion matrix analysis of the weighted ensemble.

	Predicted Non-CB	Predicted CB
Actual Non-CB	737	852
Actual CB	273	7677

The confusion matrix shows that false positives (678 samples) outnumbered false negatives (235 samples), suggesting the model is more likely to misclassify potentially harmful or ambiguous content as cyberbullying. Qualitative analysis of the misclassified predictions showed that most mistakes involved sarcastic phrases, hidden abuse, and context-sensitive statements. Results often included very negative or insulting language that was not clearly intended as bullying.

In contrast, the latter cases involved indirect, abusive, or insulting language, with different meanings inferred from the context in which it was used. During inference, the five cyberbullying categories are combined into a single cyberbullying

class, since the final task is a binary classification. Thus, the reported binary confusion matrix does not provide precision and Recall for the categories [29]. However, the observed errors indicate that sarcasm and implicit abuse remain difficult for the proposed framework to handle and require further contextual modeling in future research [30]. Fig. 3 illustrates the binary confusion matrix obtained by the proposed weighted transformer ensemble. Most cyberbullying and non-cyberbullying samples were correctly classified, whereas only a small proportion of instances were misclassified. The error distribution indicates that the proposed model achieves strong classification performance while still facing challenges with ambiguous and context-dependent language.

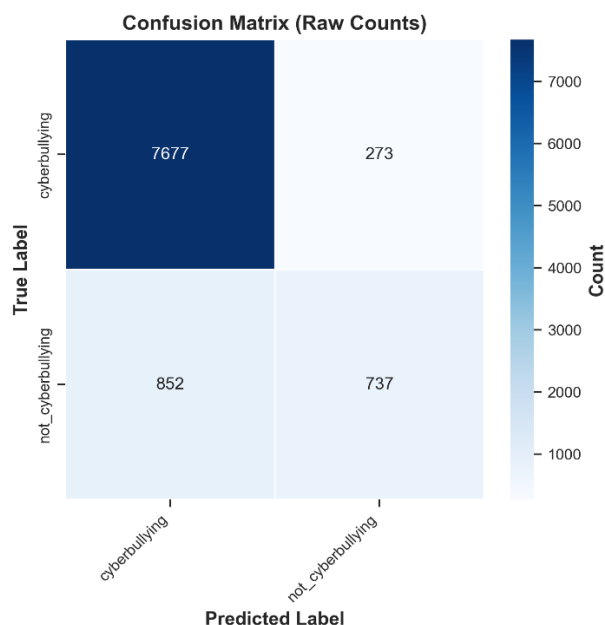


Fig. 3 Binary confusion matrix of the proposed weighted transformer ensemble illustrating correct and misclassified cyberbullying predictions.

6. Discussion

The experiment results show that the proposed weighted ensemble achieves better overall performance than the evaluated transformer baselines. The ensemble of individual configurations improves with the addition of complementary transformer predictions, suggesting that the added value of transformer-based cyberbullying predictions is complementary rather than multi-class guidance alone. The ablation results also indicate that multi-class-guided training cannot, by itself, improve the F1 score of the direct binary Twitter-RoBERTa. This

suggests that the improvement is mainly due to the heterogeneous weighted ensemble strategy [31]. Although the model achieved strong performance, sarcasm, implicit abuse, and context-dependent expressions remain difficult.

This study has several limitations. First, the experiments were conducted on a single English-language Twitter dataset, which may limit the applicability of the results to other languages and social media [32]. Second, fine-tuning of transformers and ensemble inference are resource-intensive [33]. Third, expressions involving sarcasm, implicit abuse, and context dependency will still be difficult for the proposed framework [34]. Future studies should explore the framework on multilingual and multi-platform datasets, and investigate lightweight transformer architectures for real-time deployment [35].

7. Conclusion

In this study, we presented a multi-class guided transformer ensemble framework for binary cyberbullying detection in social media. The proposed method is based on fine-grained multi-class supervision, fine-tuning the model on a domain-specific dataset of tweets using Twitter-RoBERTa, and weighted ensemble learning. Experimental results validated the proposed weighted ensemble with 91.06% accuracy and 90.64% F1-score, superior to the transformer baselines evaluated. The results show that complementary transformer configurations lead to better robustness and generalization in prediction tasks. The study, however, is restricted to an English-language Twitter dataset and requires substantial computational resources to fine-tune the transformer. Going forward, we will explore multilingual datasets, cross-platform evaluation, and lightweight transformer architectures for real-time deployment.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors would like to thank the Department of Computer Science, College of Science, University of Diyala, for providing academic and technical support for this research.

References

- [1] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, 140(4), 1073–1137. <https://doi.org/10.1037/a0035618>
- [2] Hinduja, S., & Patchin, J. W. (2015). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying* (2nd ed.). Corwin.
- [3] Dadvar, M., Triesnigg, D., & de Jong, F. (2014). Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Advances in artificial intelligence: Canadian AI 2014* (pp. 275–281). Springer. https://doi.org/10.1007/978-3-319-06483-3_25
- [4] Nahar, V., Li, X., & Pang, C. (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5), 238–247.
- [5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- [6] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU-based deep neural network. In *The Semantic Web: ESWC 2018* (pp. 745–760). Springer. https://doi.org/10.1007/978-3-319-93417-4_48
- [7] Agrawal, S., & Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval: ECIR 2018* (pp. 141–153). Springer. https://doi.org/10.1007/978-3-319-76941-7_11
- [8] García-Méndez, S., & de Arriba-Pérez, F. (2024). Promoting security and trust on social networks: Explainable cyberbullying detection using large language models in a stream-based machine learning framework. In *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 25–32). IEEE. <https://doi.org/10.1109/SNAMS64316.2024.10883785>
- [9] Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking* (Article 43, pp. 1–6). Association for Computing Machinery. <https://doi.org/10.1145/2833312.2849567>
- [10] Cheng, L., Guo, R., Silva, Y. N., Hall, D. L., & Liu, H. (2019). Hierarchical attention networks for cyberbullying detection on the Instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 235–243). Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611975673.27>
- [11] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759–760). Association for Computing Machinery. <https://doi.org/10.1145/3041021.3054223>
- [12] Talpur, B. A., & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLOS ONE*, 15(10), Article e0240924. <https://doi.org/10.1371/journal.pone.0240924>
- [13] Paul, S., & Saha, S. (2022). CyberBERT: BERT for cyberbullying identification. *Multimedia Systems*, 28, 1897–1904. <https://doi.org/10.1007/s00530-020-00710-4>
- [14] Sahay, P., Suri, S., & Kumar, R. (2021). Multitask learning for offensive language and cyberbullying detection using transformers. *Expert Systems with Applications*, 168, 1–13.
- [15] [15] Chen, S., Wang, J., & He, K. (2024). Chinese cyberbullying detection using XLNet and deep Bi-LSTM hybrid model. *Information*, 15(2), Article 93. <https://doi.org/10.3390/info15020093>
- Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, 8, 5449–5467. <https://doi.org/10.1007/s40747-022-00772-z>
- [16] Jamjoom, A. A., Karamti, H., Umer, M., Alsubai, S., Kim, T.-H., & Ashraf, I. (2024). RoBERTaNET: Enhanced RoBERTa transformer-based model for cyberbullying detection with GloVe features. *IEEE Access*, 12, 58950–58959. <https://doi.org/10.1109/ACCESS.2024.3386637>
- Kumar, A., & Sachdeva, N. (2022). Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional

- neural network. *Multimedia Systems*, 28, 2043–2052. <https://doi.org/10.1007/s00530-020-00747-5>
- [17] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1391–1399). International World Wide Web Conferences Steering Committee. <https://doi.org/10.1145/3038912.3052591>
- [18] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), Article 85, 1–30. <https://doi.org/10.1145/3232676>
- [19] Davidson, T., Warmesley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515. <https://doi.org/10.1609/icwsm.v11i1.14955>
- [20] Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (pp. 86–95). *CEUR Workshop Proceedings*. <https://ceur-ws.org/Vol-1816/paper-09.pdf>
- [21] Mossie, Z., & Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3), Article 102087. <https://doi.org/10.1016/j.ipm.2019.102087>
- [22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
- [23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. <https://doi.org/10.48550/arXiv.1907.11692>
- [24] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- [25] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.747>
- [26] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 328–339). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1031>
- [27] Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.2>
- [28] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [29] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610. <https://doi.org/10.1016/j.neunet.2005.06.042>
- [30] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746–1751). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1181>
- [31] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://doi.org/10.7551/mitpress/10243.001.0001>
- [32] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Journal of Machine Learning Research, 21(140), 1–67. <https://www.jmlr.org/papers/v21/20-074.html>

- [33] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
<https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
<https://doi.org/10.48550/arXiv.2005.14165>
- [34] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pretraining OpenAI.
https://cdn.openai.com/research-covers/languageunsupervised/language_understanding_paper.pdf