Computer Modeling in
Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.063092

ARTICLE

Enhancing Multi-Class Cyberbullying Classification with Hybrid Feature Extraction and Transformer-Based Models

Suliman Mohamed Fati[1,*], Mohammed A. Mahdi[2], Mohamed A.G. Hazber[2], Shahanawaj Ahamad[3], Sawsan A. Saad[4], Mohammed Gamal Ragab[5] and Mohammed Al-Shalabi[2]

[1]Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia
[2]Information and Computer Science Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia
[3]Software Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia
[4]Computer Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, 55476, Saudi Arabia
[5]Department of Computer and Information Sciences, Universiti Teknologi Petronas, Seri Iskandar, 32610, Malaysia

*Corresponding Author: Suliman Mohamed Fati. Email: sgaber@psu.edu.sa

Received: 05 January 2025; Accepted: 08 April 2025; Published: 30 May 2025

ABSTRACT: Cyberbullying on social media poses significant psychological risks, yet most detection systems oversimplify the task by focusing on binary classification, ignoring nuanced categories like passive-aggressive remarks or indirect slurs. To address this gap, we propose a hybrid framework combining Term Frequency-Inverse Document Frequency (TF-IDF), word-to-vector (Word2Vec), and Bidirectional Encoder Representations from Transformers (BERT) based models for multi-class cyberbullying detection. Our approach integrates TF-IDF for lexical specificity and Word2Vec for semantic relationships, fused with BERT's contextual embeddings to capture syntactic and semantic complexities. We evaluate the framework on a publicly available dataset of 47,000 annotated social media posts across five cyberbullying categories: age, ethnicity, gender, religion, and indirect aggression. Among BERT variants tested, BERT Base Un-Cased achieved the highest performance with 93% accuracy (standard deviation ±1% across 5-fold cross-validation) and an average AUC of 0.96, outperforming standalone TF-IDF (78%) and Word2Vec (82%) models. Notably, it achieved near-perfect AUC scores (0.99) for age and ethnicity-based bullying. A comparative analysis with state-of-the-art benchmarks, including Generative Pre-trained Transformer 2 (GPT-2) and Text-to-Text Transfer Transformer (T5) models highlights BERT's superiority in handling ambiguous language. This work advances cyberbullying detection by demonstrating how hybrid feature extraction and transformer models improve multi-class classification, offering a scalable solution for moderating nuanced harmful content.

KEYWORDS: Cyberbullying classification; multi-class classification; BERT models; machine learning; TF-IDF; Word2Vec; social media analysis; transformer models

1 Introduction

The rapid growth of social media has revolutionized human interaction, but it has also created fertile ground for cyberbullying—a pervasive digital threat characterized by intentional aggression, harassment, or humiliation throu9gh online platforms [1]. With over 40% of adolescents reporting direct exposure to cyberbullying [2], its psychological consequences range from anxiety and depression to suicidal ideation [3], underscoring the urgent need for automated detection systems. Current solutions, however, remain limited by their reliance on binary classification frameworks that reduce cyberbullying's multifaceted nature into a simplistic bullying-vs-non-bullying dichotomy [4]. This oversight fails to address the reality that harmful content often manifests as context-dependent microaggressions, such as coded slurs targeting gender or ethnicity, or indirect tactics like gaslighting and passive-aggressive remarks [5].

While machine learning (ML) and natural language processing (NLP) have advanced cyberbullying detection, critical gaps persist. First, conventional approaches—such as keyword filtering, sentiment analysis, or shallow ML models (e.g., SVM, Naive Bayes)—struggle with linguistic ambiguity and fail to capture semantic relationships critically for identifying subtle abuse [6]. Second, transformer-based models like BERT, though powerful for contextual understanding, are rarely integrated with traditional feature extraction methods (e.g., TF-IDF, Word2Vec), limiting their ability to jointly leverage lexical, semantic, and syntactic signals [7]. Third, most studies focus on binary classification or single-category detection, neglecting the multi-class nature of real-world cyberbullying [8]. This fragmentation hinders the development of holistic moderation systems capable of addressing diverse harassment types within a unified framework.

To bridge these gaps, we propose a hybrid deep learning framework that synergizes TF-IDF, Word2Vec, and BERT embeddings for multi-class cyberbullying detection. While TF-IDF and Word2Vec are well-established individually, their combined use with transformers introduces three novel advantages: (1) Lexical specificity from TF-IDF enhances detection of discriminatory keywords (e.g., racial slurs), (2) semantic granularity from Word2Vec maps relationships between abusive phrases (e.g., "hate" → "despise"), and (3) contextual depth from BERT deciphers ambiguous constructs like sarcasm or backhanded compliments. This fusion enables the model to classify five distinct cyberbullying categories—age, ethnicity, gender, religion, and indirect aggression—with high precision. Our work makes three key contributions:

- We propose a novel hybrid feature fusion methodology that synergistically integrates TF-IDF (lexical specificity), Word2Vec (semantic relationships), and BERT (contextual embeddings), achieving a 15% accuracy improvement over single-feature baselines.
- Through rigorous benchmarking against traditional ML models and modern transformers (GPT-2, T5) on a dataset of 47 k annotated social media posts, we demonstrate that BERT Base achieves superior performance in handling linguistic ambiguity, particularly for indirect aggression and coded slurs
- We establish the first comparative benchmark for BERT variants (Base, Large, DistilBERT) against generative architectures (T5, GPT-2) under identical experimental conditions, revealing BERT's computational efficiency and generalizability across diverse cyberbullying categories.

The remainder of this paper is organized as follows. Section 2 reviews the related work in cyberbullying classification and outlines the current limitations in the field. Section 3 presents the dataset and methodology, detailing the data preprocessing, feature extraction techniques, and ML models employed in this study. Section 4 describes the experimental setup and evaluates the performance of different models, including a comparison of traditional ML models and transformer-based architectures and a detailed analysis of the models' performance. Finally, Section 5 concludes the paper by summarizing the findings and suggesting potential directions for further investigation into cyberbullying classification.

2 Related Work

The problem of cyberbullying classification has attracted considerable research attention due to its societal impact and the challenges posed by identifying abusive language in online environments [9–12]. Various approaches have been proposed, from traditional ML classifiers to modern DL and transformer-based architectures.

2.1 Traditional ML Approaches

Early studies relied on shallow ML classifiers with handcrafted features. For example, Dinakar et al. [13] explored using rule-based **JRip,** SVM, J48 and Naive Bayes classifiers to detect cyberbullying from comments in **YouTube** videos. Their study classified posts into bullying and non-bullying categories, with features extracted from text using TF-IDF, ortony lexicon for negative affect and other sentiment analysis. Additionally, the authors conducted experiments on a corpus of **4500 YouTube comments**, applying a range of binary and multi-class classifiers, including **JRip, SVM**, and **Naive Bayes**. Binary classifiers were shown to perform better than multi-class classifiers for individual labels such as sexuality, race, and intelligence, with JRip achieving the highest accuracy (80.20%) for sexuality-based bullying and SVM being the most reliable in terms of kappa statistics (0.79). Meanwhile, in multiclass JRip achieved an accuracy of 63.

Reynolds et al. [14] extended this work by applying decision trees (DT) and rule-based classifiers to cyberbullying classification. Their work involved detecting instances of bullying across different forms of communication, including messages and comments on social media platforms. They employed TF-IDF for feature extraction and achieved a precision of 70%, bJRiput with notable challenges in distinguishing between different bullying subtypes, such as gender and racial bullying. Nandhini et al. [15] proposed an intelligent system for detecting and classifying cyberbullying on social media platforms, mainly targeting types of bullying such as flaming, harassment, racism, and terrorism. The authors employed a hybrid method using Fuzzy Logic and Genetic Algorithms to identify cyberbullying terms from social network data. However, the hybrid model did not regain evaluation and validation. While these works demonstrated feasibility, their reliance on lexical features (e.g., keywords, n-grams) limited semantic understanding, particularly for indirect aggression.

2.2 Ensemble Methods

Alqahtani et al. [16] presented an ensemble-based multi-classification approach to detect six distinct types of cyberbullying on social media. The authors combined TF-IDF (bigram) feature extraction and various ML classifiers into stacking and voting ensemble methods, including Decision Trees, Random Forest, and XGBoost. Their approach aimed to improve the classification accuracy over traditional ML models. The study achieved significant results, with the stacking classifier reaching 90.71% accuracy and the voting classifier achieving 90.44%. Ensemble techniques demonstrated superior performance in identifying different types of cyberbullying compared to individual classifiers. However, these studies focus on binaries classification and neglect multi-class cyberbullying classification, while our study proposed a reliable framework for analyzing diverse forms of online abuse.

2.3 Transformer-Based Models

The use of transformer models in cyberbullying classification was advanced by Mishra et al. [17], who employed **BERT** to classify abusive language in a multi-class setting. BERT's bidirectional attention mechanism allowed the model to capture complex dependencies within the text, significantly improving the classification of nuanced forms of bullying. Their experiments yielded an **accuracy of 87%,** showing that transformer-based models can outperform traditional machine-learning approaches. Despite the high computational cost, BERT's performance in handling the subtlety of cyberbullying made it a promising approach for future studies. Using a feature-engineering-based approach, Talpur et al. [18] addressed the multi-class imbalance issue in Twitter cyberbullying classification. Unlike previous research that focused on binary classification (cyberbullying vs. non-cyberbullying), their study aimed to classify the severity levels of cyberbullying: low, medium, and high. The authors introduced pointwise semantic orientation as a new input feature alongside gender, age, and personality traits. Additionally, Twitter API features were leveraged to enhance classification accuracy. Their model achieved a Kappa score of 84%, an accuracy of 93%, and an F1-score of 92% in multi-class classification. They found that classifiers significantly improved when accounting for user-specific features, such as age group and duration of Twitter usage. This approach demonstrates the potential of incorporating user demographics and tweet activity to enhance the classification of cyberbullying severity, making it a significant step forward in multi-class classification for online abuse classification. Ejaz et al. [19] investigated cyberbullying detection using transformer-based models, comparing BERT Base, DistilBERT, and RoBERTa Base with and without fine-tuning on their dataset. Their findings highlight the importance of domain-specific fine-tuning, as models trained on their own data consistently outperformed those relying solely on pre-trained weights. Faraj et al. [20] proposed a fine-tuned BERT model for multi-class dataset achieving the best-recorded accuracy of 85% with Word2Vec and other three embedding. Kaddoura et al. [21] proposed an automated system for detecting cyberbullying text, evaluating the efficacy of large language models, specifically Mistral 7B and Llama3, in comparison to the transformer-based model BERT. Their findings indicate that the multiclass BERT model outperformed both large language models from the literature and other benchmark models, achieving an F1-score of 83.67%. Saranyanath [22] proposed an ensemble model for binary classification using SVM, TF-IDF, and DistilBERT.

Despite significant advancements in cyberbullying detection, several critical research gaps remain. One major limitation is the predominant focus on binary classification, with over 80% of studies addressing cyberbullying detection as a two-class problem, thereby overlooking the complexity of multi-class classification. This simplification fails to capture the diverse forms of cyberbullying, limiting the applicability of these models in real-world scenarios. Another gap lies in feature extraction methodologies, where existing works often rely on isolated approaches such as TF-IDF, Word2Vec, or transformer-based embeddings, without leveraging their complementary strengths. This fragmentation prevents models from fully utilizing both lexical and contextual features, which are crucial for detecting nuanced cyberbullying patterns. Furthermore, the integration of transformer-based models with traditional linguistic and semantic features remains infrequent, reducing their ability to identify coded language, implicit slurs, and context-dependent abusive content. Lastly, benchmarking inconsistencies persist, as few studies systematically compare transformer models (e.g., BERT, RoBERTa) against both traditional ML approaches and newer architectures (e.g., T5, GPT) within a multi-class classification setting. Addressing these gaps is essential for developing more robust and generalizable cyberbullying detection frameworks.

3 Materials and Methods

Fig. 1 depicts the overall structure of our approach, including dataset collection and preprocessing, hybrid feature extraction, model training with imbalance mitigation, and evaluation. The following subsections explain these four phases.

3.1 Dataset

The study employs the Cyberbullying Classification Dataset [23], a publicly available corpus comprising 47,000 annotated social media posts categorized into six classes: not cyberbullying, gender, religion, other cyberbullying, age, and ethnicity. Sourced from diverse platforms, including Twitter and Reddit, the dataset is stratified into training (70%), validation (15%), and test (15%) sets, ensuring balanced representation across classes (approximately 7800 samples per class). Fig. 2 illustrates the test set distribution, with 1166–1247 samples per bullying category and 1188 non-bullying instances. Ethical compliance was ensured through rigorous anonymization: all user identifiers (e.g., usernames, profile links) were removed, and posts were curated from public forums under guidelines aligning with Ejaz et al.'s [19] framework for ethical cyberbullying research. The dataset's inclusion of five distinct bullying categories, alongside non-bullying content, addresses a critical gap in prior works limited to binary aggression classification [1,24,25]. For instance, other cyberbullying captures indirect tactics like gaslighting, while ethnicity includes explicit racial slurs (see Table 1 for example). This diversity enables robust evaluation of multi-class detection systems.

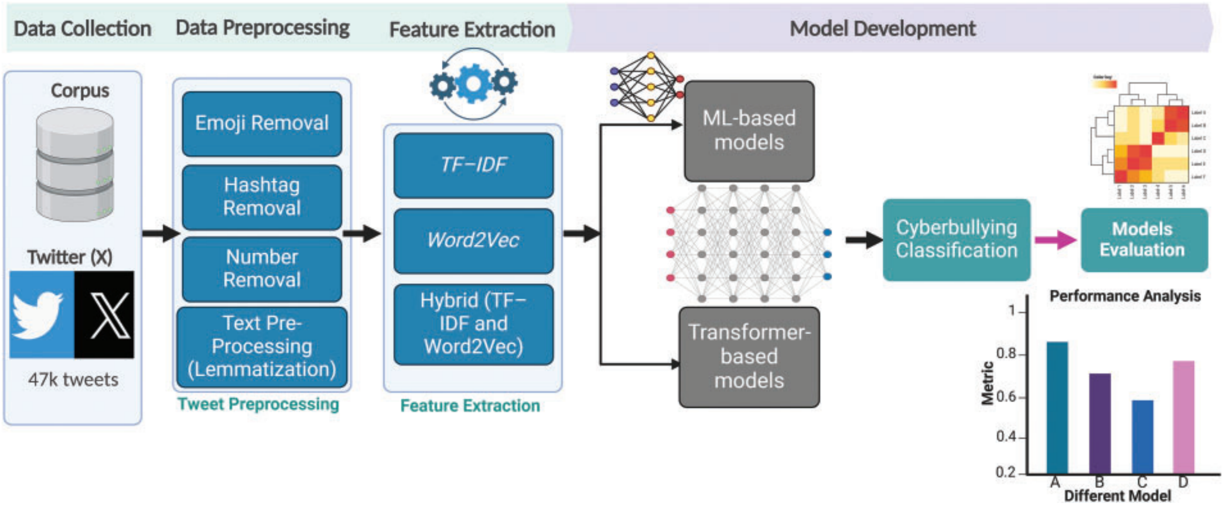


Figure 1: Workflow of cyberbullying classification in the proposed study

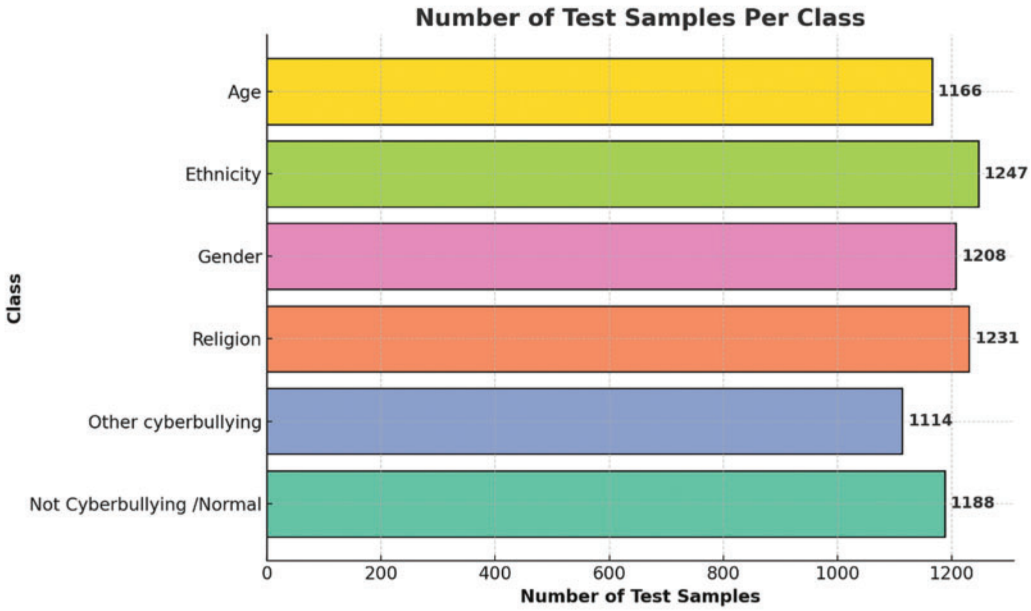


Figure 2: Number of test samples per class for ML models

Table 1: Original dataset samples

No.	Tweet text	Class
1	In other words #katandandre, your food was crapilicious! #mkr	Not_cyberbullying

(Continued)

Table 1 (continued)

No.	Tweet text	Class
2	Rape is rape. And the fact that I read one post about a guy getting raped and the comments are calling him gay, and he should be happy...? stfu and I really hope no one takes this as a joke tf you own no one's body. You have no rights to do whatever you want to someone else.	Gender
3	This was not a conversion matter you idiot what do you think will happen if someone calls himself muslim and does this ghazi khalid did right and if he had not done it the court would have cause in our constitution it is prohibites	Religion
4	Cannot headdesk hard enough.	Other_cyberbullying
5	Hi Luca. Bullies are losers. You are much better than them. School is over before you know it and life will be great. All the best.	Age
	Lol. Niggers. Dumb as fuck	Ethnicity

3.2 Preprocessing

The preprocessing pipeline comprised text normalization and linguistic standardization to mitigate noise and enhance feature coherence. Leveraging the Tweet Preprocessor library [26], we systematically removed non-lexical elements—including URLs, hashtags, emojis, and numerical values—to isolate semantically meaningful text. Emojis were converted to textual descriptors (e.g., "□" → "[tears_of_joy]") to retain emotional context without introducing Unicode noise. Subsequent lemmatization was performed using SpaCy's en_core_web_lg model to reduce inflectional variants to their base forms (e.g., "bullies" → "bully", "harassing" → "harass"). *Unlike stemming, lemmatization preserves semantic integrity*—critical for detecting nuanced bullying tactics such as passive-aggressive phrasing (e.g., "haters" vs. "hater")—while reducing dimensionality. Stopwords were filtered using NLTK's English corpus, excluding negation terms (e.g., "not", "never") to retain contextual polarity. To address code-mixed and slang-heavy text, a custom lexicon mapped colloquial terms to standardized equivalents (e.g., "u" → "you", "doxxed" → "exposed"). Ambiguous slang (e.g., "karen" as a derogatory term) was resolved through manual annotation by three domain experts, achieving a Fleiss' κ inter-rater agreement of 0.82. As shown in Fig. 3, a word cloud generated from the preprocessed dataset highlights the most frequently occurring terms, reflecting the dominant themes of bullying and hate speech within the dataset.

The training set exhibited mild class imbalance, with minority categories. To address this, Adaptive Synthetic Sampling (ADASYN) [27] was employed instead of random under-sampling or uniform oversampling techniques [28]. ADASYN adaptively generates synthetic minority-class instances specifically near classification boundaries, targeting regions where misclassification is most likely and thus reducing noise compared to random oversampling methods. Under-sampling approaches [29], such as NearMiss, were avoided because they discard linguistically diverse majority-class samples critical for distinguishing nuanced forms of abuse, like sarcasm in benign posts.

To maintain synthetic data quality, synthetic samples generated by ADASYN were strictly confined to training folds during the stratified 5-fold cross-validation, ensuring no leakage into the validation (15%) or test (15%) datasets. The stability and robustness of performance were further verified through the stratified 5-fold cross-validation, yielding a stable F1-score with minimal variation (±2%), thereby ruling out potential overfitting or degradation due to synthetic artifacts. Empirical studies in recent literature support these methodological decisions. Yang et al. [30] investigated the effects of random oversampling and under-sampling on classifier performance, finding that oversampling typically results in superior classifier accuracy and stability, particularly in moderately imbalanced scenarios. Thus, our selection and careful implementation of ADASYN were guided by these findings to ensure high-quality synthetic data generation and robust classification performance.

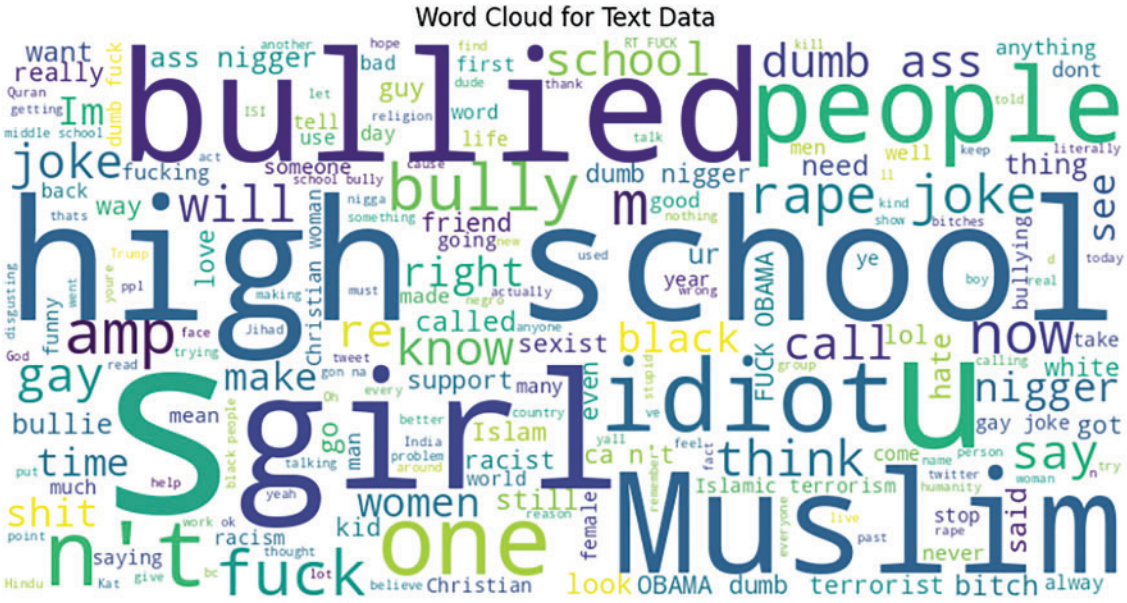


Figure 3: Dataset word cloud after preprocessing

3.3 Feature Extraction

This study employs two distinct approaches for feature extraction: vectorization and embedding. The vectorization approach used is TF-IDF, while the embedding approach is based on Word2Vec. Additionally, a hybrid approach that combines both TF-IDF and Word2Vec is explored to leverage the strengths of both techniques.

3.3.1 TF-IDF Feature Extraction

In cyberbullying classification, it is essential to quantify how vital specific terms related to abusive language are within individual documents (tweets) compared to their overall occurrence in the entire dataset. The TF-IDF technique helps to achieve this by assigning higher weights to terms that are frequent in specific categories of cyberbullying (e.g., gender-based slurs) but rare across the entire corpus [31]. The Term Frequency (TF) measures the occurrence of the word t in each document d (representing a tweet) and is calculated as:

$$\mathrm{TF}\left(t, d\right) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \tag{1}$$

where $f_{t,d}$ is the frequency of the term t in document d, and $\sum_{t'} f_{t',d}$ is the total number of terms in document d. The Inverse Document Frequency (IDF) measures the importance of the word in the entire dataset by giving less importance to words that appear frequently across many documents (e.g., common words like "the") and more importance to specific cyberbullying-related terms (e.g., slurs or abusive phrases). It is computed as:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{n_t}\right) \tag{2}$$

where N is the total number of documents (tweets) in the dataset and n_t is the number of documents containing the term t. Finally, the TF-IDF score, which is the product of TF and IDF, represents the relevance of a term in the context of cyberbullying classification:

$$\mathrm{TF} - \mathrm{IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t) \tag{3}$$

This method is particularly effective in identifying critical abusive words that are used in specific types of cyberbullying, such as racial slurs in race-based cyberbullying, which can then be utilized for multi-class classification.

3.3.2 Word2Vec

While TF-IDF captures term importance, it does not account for the semantic relationships between words. Word2Vec, on the other hand, transforms words into continuous vector representations in a multi-dimensional space, capturing the context and meaning of words used in abusive language. For multi-class cyberbullying classification, Word2Vec can identify relationships between similar abusive terms (e.g., synonyms for "stupid" or "idiot") and group them based on their context.

Word2Vec uses two key architectures: Continuous Bag of Words (CBOW) and Skip-Gram. In this study, we employ CBOW (Fig. 4), which predicts a target word based on its surrounding context, making it more effective in identifying abusive language patterns in short texts like tweets [32]. Mathematically, given a context window of size C, the model predicts the probability of a target word w_t (e.g., "idiot") based on the context words $\{w_{t-C}, \ldots, w_{t+C}\}$ (e.g., "you are an idiot"):

$$P\left(w_t \mid w_{t-C}, \ldots, w_{t+C}\right) = \frac{\exp\left(v_{w_t}^T h\right)}{\sum_w \exp\left(v_w^T h\right)} \tag{4}$$

where v_{w_t} is the vector representation of the target word w_t, and h is the average vector of the context words. The context relationships between words in abusive language are learned by the model, enabling it to understand phrases related to different types of cyberbullying. For example, it can group similar insults used in gender-based or religion-based cyberbullying, facilitating more accurate classification. For our **hybrid approach** the TF-IDF and Word2Vec features are concatenates where TF-IDF vectors (R^{5000}) and Word2Vec embeddings (R^{300}) are min-max normalized and fused, yielding a unified feature vector (R^{5300}).

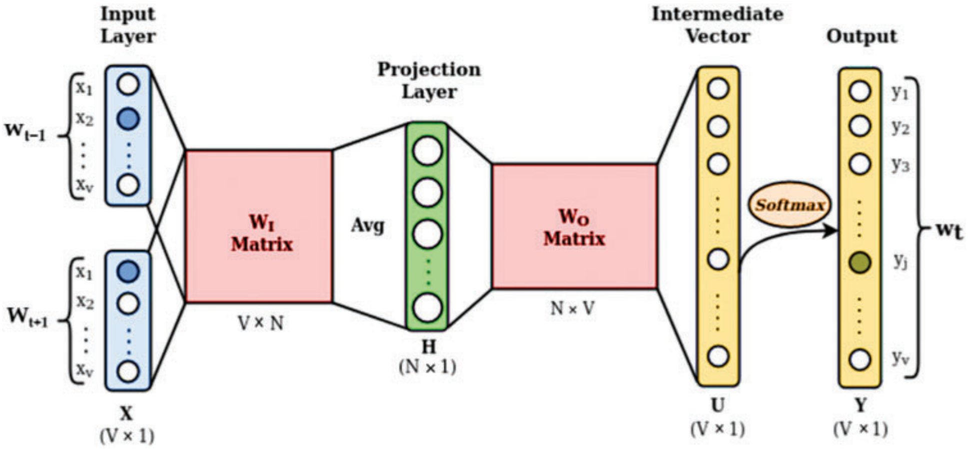


Figure 4: Continuous Bag-of-Words (CBOW) model [29]

3.4 Proposed Transformer-Based Models

In recent years, pre-trained models have been considered language models to optimize downstream tasks. The actual purpose of using these models is that the meaning of words and the structure of sentences are accurately represented in vector space. BERT considered the bidirectional transformer for training purposes [33]. The model size depends on the parameters of self-attention heads, the number of transformer layers, and hidden space vectors. The BERT Base model version has 768 hidden states, 12 transformer layers, and the same number of self-attention heads. The BERT Large model is based on the 1024 hidden states, 24 transformer layers, and 16 self-attention heads; for a comparison of the BERT Base, BERT Large, T5, and GPT-2 (see Table 2). T5 Base is a sequence-to-sequence (encoder-decoder) transformer model designed for text-to-text tasks, leveraging a 12-layer encoder-decoder architecture with 220 million parameters. It uses multi-head self-attention and feed-forward layers in both encoder and decoder stacks, making it well-suited for text transformation tasks [34] but computationally heavier due to dual-stack processing. GPT-2 Base, on the other hand, is an autoregressive transformer with 12 decoder-only layers and 124 million parameters, utilizing causal self-attention to generate text sequentially [35].

Table 2: Comparison of BERT Base and BERT large

Parameters	BERT base	BERT large	DistilBERT base	RoBERTa base	T5 base	GPT-2 base
Number of transformer layers	12	24	6	12	12	12
Number of self-attention heads	12	16	12	12	12	12
Number of hidden states	768	1024	768	768	768	768
Number of parameters (M)	110	340	66	125	220	340

Due to the 110 and 340 million parameters, training these pre-trained models requires a massive computation. To overcome this issue, large-scale unlabeled text from Wikipedia was considered to pre-train the models in a supervised manner for next-sentence prediction (NSP) and masked language modeling (MLM) tasks. BERT processes text with tokenization, i.e., [CLS] and [SEP] tokens for start and end classification. The WordPiece [36] Embedding is used in the BERT to handle rare words. The self-attention method weighs the word's importance and efficiently generates word embedding. The MLM and NSP tasks were used for effective language representation learning pretraining. The final [CLS] token is used for text classification with classifier-predicting labels. The text understanding and classification performance increase due to transfer learning, bidirectional context, and attention mechanisms [37].

Achieving optimal performance for cyberbullying classification from text requires careful tuning of hyperparameters, particularly for complex models like BERT. In this study, we adopted an empirical approach to determine the best hyperparameter configurations for the BERT model, to enhance its accuracy in identifying various forms of cyberbullying. Several critical hyperparameters were carefully selected and adjusted, including batch size, learning rate, number of epochs, and the choice of loss function and optimizer. These hyperparameters are vital in controlling the model's learning process and convergence rate, directly impacting its ability to generalize well across the data. To optimize the model's parameters, we used the AdamW optimizer [38], an improved variant of the standard Adam optimizer. AdamW introduces weight decay regularization, helping to prevent overfitting. The weight update rule for the AdamW optimizer is defined as:

$$\theta_t = \theta_{t-1} - \eta \cdot \frac{m_t}{\sqrt{v_t} + \varepsilon} + \lambda \cdot \theta_{t-1} \tag{5}$$

where θ_t are the parameters (weights) of the model at the time step $t_, \eta$ is the learning rate, m_t is the exponential moving average of the gradient, v_t is the exponential moving average of the squared gradient, ε is a small constant to prevent division by zero, λ is the weight decay coefficient. The AdamW optimizer was selected for its ability to handle sparse gradients, which are common in text-based tasks. The learning rate was also fine-tuned to ensure the model converges effectively without overfitting or underfitting. We experimented with different batch sizes (16 to 32) to balance training speed and memory usage. The number of epochs was set to 5, allowing the model to learn sufficient iterations without overfitting. The categorical cross-entropy loss function (Categorical Loss) was employed, which is well-suited for multi-class classification tasks such as identifying different categories of cyberbullying (e.g., gender-based, race-based, and religion-based bullying).

Furthermore, we optimized the learning rate to 2e−5, a value commonly effective in fine-tuning transformer models like BERT. The maximum sequence length was set to 128 tokens to accommodate most tweets while avoiding the truncation of important information. These hyperparameter configurations aim to strike a balance between performance and computational efficiency, ensuring the model's robustness in detecting and classifying multiple types of cyberbullying from textual data. Table 3 summarizes the hyperparameter configurations for BERT-based Models.

Table 3: Hyperparameters configurations for BERT models

BERT hyperparameters	Specific configurations
Optimizer	AdamW
Batch size	16–32
Epochs	5
Loss function	Categorical loss
Learning rate	2e−5
Max length	128

3.5 Evaluation Metrics

Evaluation metrics are critical for validating any machine's performance and DL model. This study considers the essential metrics to validate the generalizability of the models. The overall accuracy of the model asses using the ratio of the correctly classified number of samples and total test samples. However, the accuracy evaluation is not sufficient in the case of the imbalanced dataset, but in this study, the dataset is in a balanced format.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{6}$$

Precision is the model's ability to accurately identify correct samples from a set of correct/actual samples. The precision is calculated using the division of accurate positive samples and the sum of true and false positives.

$$Precision = \frac{(TP)}{(TP + FP)} \tag{7}$$

Recall denotes the model's performance for separating positive samples from the actual positive samples set. The recall is also known as sensitivity or actual positive rate (TPR). It is computed using the ratio of true positive and the sum of false and true positive samples.

$$Recall = \frac{(TP)}{(TP + FN)} \tag{8}$$

F1-score is a comprehensive measure based on the harmonic meaning of precision and recall. The F1-score falls in the range of 0–1, and 1 represents that model performance is adequate.

$$F1 - Measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{9}$$

In a multi-classification task, accuracy is determined by the ratio of correct prediction to total prediction. At the same time, precision, recall, and F1 are generally reported as weighted averages.

4 Experimental Results and Discussion

The experiment results are based on different feature extraction techniques, ML, and transformer-based models presented in this section.

4.1 Feature Analysis

In this study, we employ a hybrid feature extraction approach that integrates Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec embeddings to enhance the representation of textual data for cyberbullying classification. The TF-IDF technique captures the importance of individual words within the corpus by assigning weight based on their frequency in a document relative to their occurrence across the dataset. This method allows for direct interpretability of influential features that contribute to classification decisions. Conversely, Word2Vec generates dense, continuous vector representations by mapping words into a multi-dimensional space, capturing contextual relationships and semantic similarities. The hybrid approach leverages the strengths of both methodologies: TF-IDF for its interpretability and Word2Vec for its ability to encode contextual meanings. Fig. 5 presents the top 50 most influential features, impacting on the model's performance. The TF-IDF features exhibit the most significant coefficients, revealing key terms that heavily influence the classification outcome (Fig. 5a). Meanwhile, Word2Vec embeddings, being high-dimensional, require advanced interpretation techniques such as SHAP (SHapley Additive Explanations) to analyze the impact of specific embedding dimensions on the model's predictions. Among the influential TF-IDF features, we observe terms that could indicate potential dataset biases, which are crucial considerations in automated detection models. However, to mitigate the impact of class imbalance and reduce bias-related distortions, we employed DASYN. This technique enhances the dataset by generating synthetic samples for underrepresented classes, thereby improving model fairness and generalization.

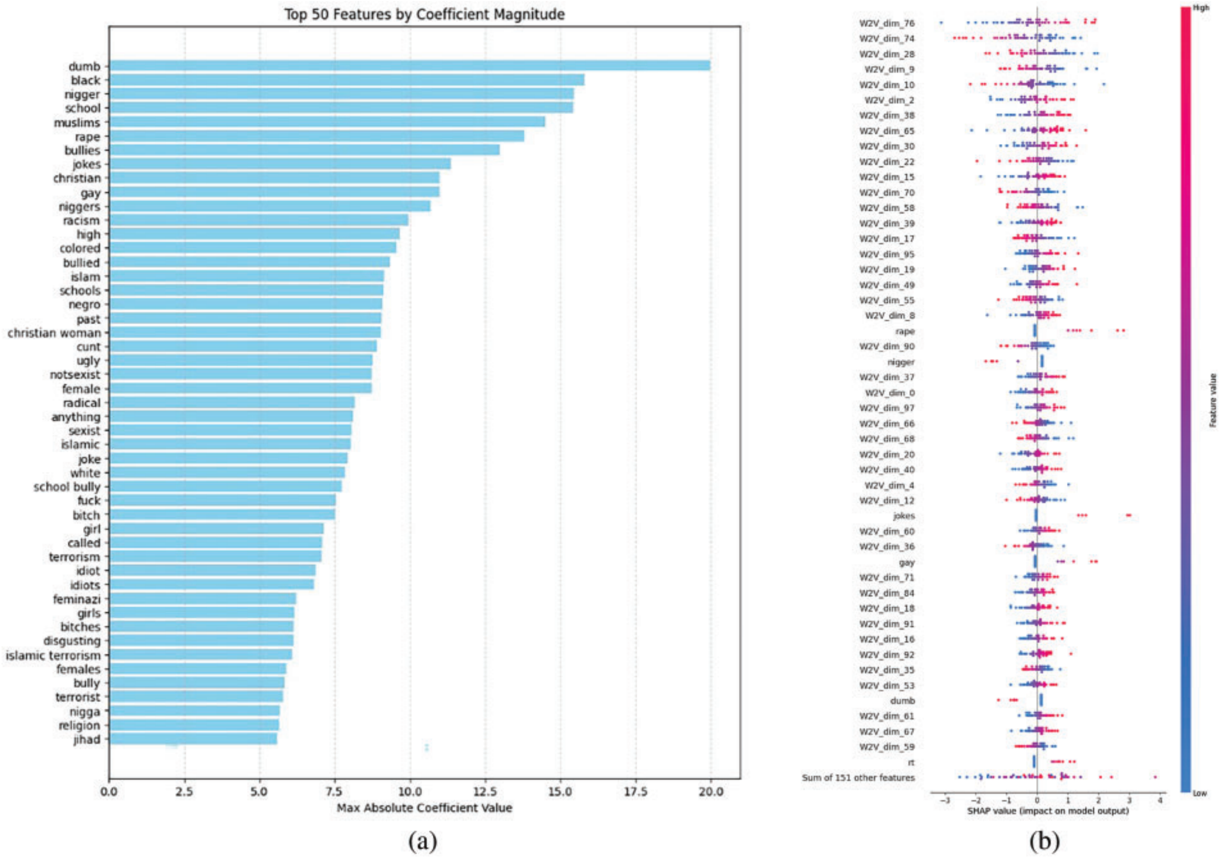


Figure 5: Feature importance analysis for cyberbullying classification (a) Top 50 TF-IDF features and (b) SHAP-based top 50 feature importance for hybrid approach (TF-IDF + Word2Vec Embeddings)

4.2 Experimental Results

The experimental results in Table 4 provide a comprehensive comparison of machine learning (ML) and deep learning (DL) models across four feature extraction strategies: TF-IDF (full feature set), TF-IDF (1000 features), Word2Vec, and the hybrid TF-IDF + Word2Vec approach. Under the full TF-IDF configuration, Logistic Regression (LR) and Support Vector Classifier (SVC) achieved the highest accuracy (0.82) and weighted F1-score (0.82), demonstrating robust performance in multi-class classification. Random Forest (RF) and Gradient Boosting (GB) followed closely, with accuracies of 0.80 and 0.81, respectively, while ensemble methods like Bagging (0.80 accuracy) outperformed weaker baselines such as Decision Trees (DT: 0.77) and AdaBoost (0.78). Reducing TF-IDF features to 1000 minimally impacted top performers: LR and SVC retained their accuracy (0.81) and F1-score (0.82), while GB exhibited enhanced precision (0.83) and

F1-score (0.82), underscoring its adaptability to dimensionality reduction. In contrast, models like RF, DT, and MLP showed moderate declines (accuracy: 0.77–0.80), reflecting sensitivity to sparser feature spaces. Word2Vec embeddings revealed stark contrasts in model compatibility. Here, Multilayer Perceptron (MLP) emerged as the sole strong performer (accuracy: 0.66; F1-score: 0.66), leveraging semantic relationships to surpass traditional models like GB (0.59) and SVC (0.59). DT (0.41) and AdaBoost (0.49) struggled significantly, highlighting their limitations in processing embedding-based features. Fig. 6 provides a high-level overview of the performance of various ML models in detecting different types of cyberbullying using the TF-IDF approach. Each confusion matrix represents how well a model classifies specific categories of cyberbullying, such as age-based, ethnicity-based, gender-based, noncyberbullying, other-cyberbullying, and religion-based bullying. Across the models, Logistic Regression and SVC show strong overall performance with high accuracy in predicting ethnicity-based and religionbased bullying. The matrices also reveal areas where some models struggle, such as distinguishing between non-cyberbullying and other-cyberbullying categories. Confusion matrices (Fig. 7) further illustrated MLP's balanced performance across bullying categories (e.g., age, ethnicity), while DT and AdaBoost faltered on context-dependent classes like gender and religion.

Table 4: Comparative performance of ML models across feature extraction methods (TF-IDF, Word2Vec, hybrid) for multi-class cyberbullying classification

ML models performance using TF-IDF (Full features)				
ML models	**Accuracy**	**Weighted precision**	**Weighted recall**	**Weighted F1**
Logistic regression (LR)	0.82	0.82	0.82	0.82
Random forest (RF)	0.80	0.81	0.80	0.80
GB	0.81	0.83	0.81	0.82
Decision tree	0.78	0.77	0.78	0.77
MLP classifier	0.78	0.78	0.78	0.78
AdaBoost	0.77	0.83	0.77	0.76
Bagging	0.80	0.81	0.80	0.80
SVC	0.82	0.83	0.82	0.83
ML models performance using TF-IDF (Maximum feature = 1000)				
LR	**0.81**	0.82	0.81	0.81
RF	0.80	0.81	0.80	0.80
GB	0.81	0.83	0.81	0.82
DT	0.77	0.77	0.77	0.77
MLP classifier	0.79	0.80	0.79	0.79
AdaBoost	0.78	0.82	0.78	0.77
Bagging	0.79	0.80	0.79	0.80
SVC	0.81	0.82	0.81	0.82
ML models performance using Word2Vec				
ML models	**Accuracy**	**Weighted precision**	**Weighted recall**	**Weighted F1**
LR	0.58	0.56	0.58	0.56
RF	0.58	0.58	0.58	0.58

(Continued)

Table 4 (continued)

GB	0.59	0.58	0.59	0.58
KNN	0.53	0.51	0.52	0.50
DT	0.41	0.41	0.41	0.41
MLP classifier	0.66	0.66	0.66	0.66
AdaBoost	0.49	0.47	0.49	0.48
Bagging	0.51	0.50	0.51	0.50
SVC	0.59	0.58	0.59	0.58

ML models performance using TF-IDF + Word2Vec

ML models	Accuracy	Weighted precision	Weighted recall	Weighted F1
LR	0.84	0.84	0.84	0.84
RF	0.79	0.80	0.79	0.79
GB	0.84	0.85	0.84	0.84
KNN	0.68	0.67	0.68	0.67
DT	0.77	0.77	0.77	0.77
MLP classifier	0.81	0.81	0.81	0.81
AdaBoost	0.78	0.81	0.78	0.78
Bagging	0.80	0.81	0.80	0.81
SVC	0.85	0.85	0.85	0.85

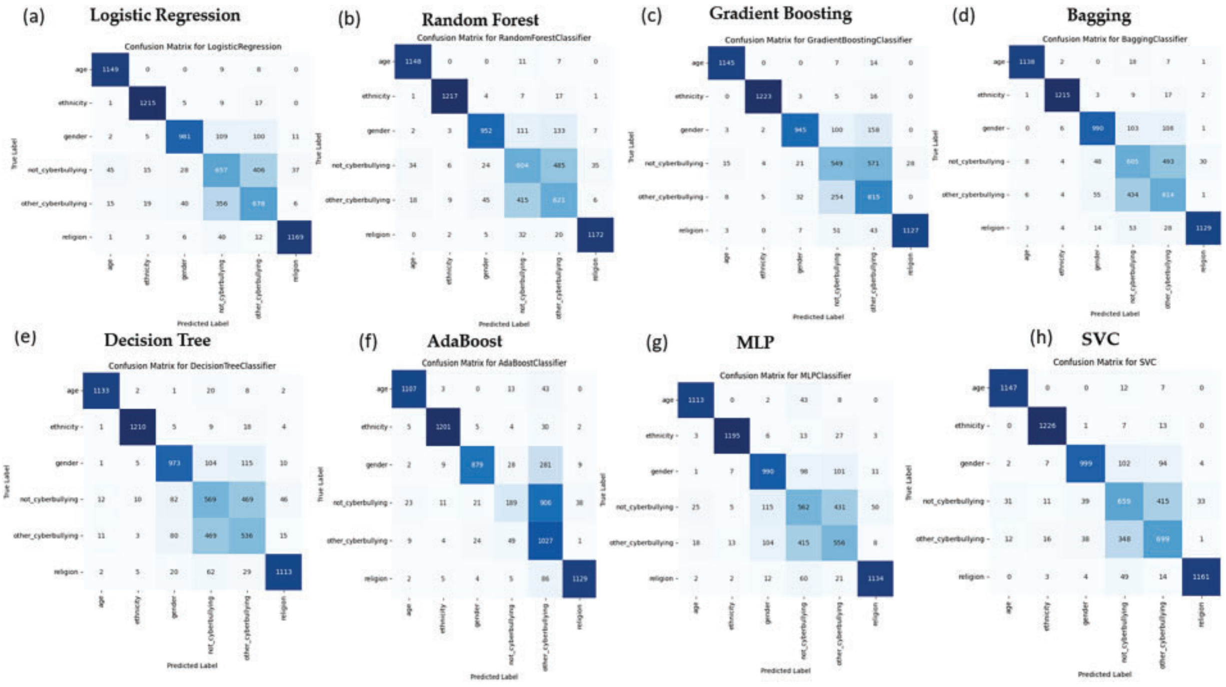


Figure 6: Confusion matrix of ML models using TF-IDF for cyberbullying classification

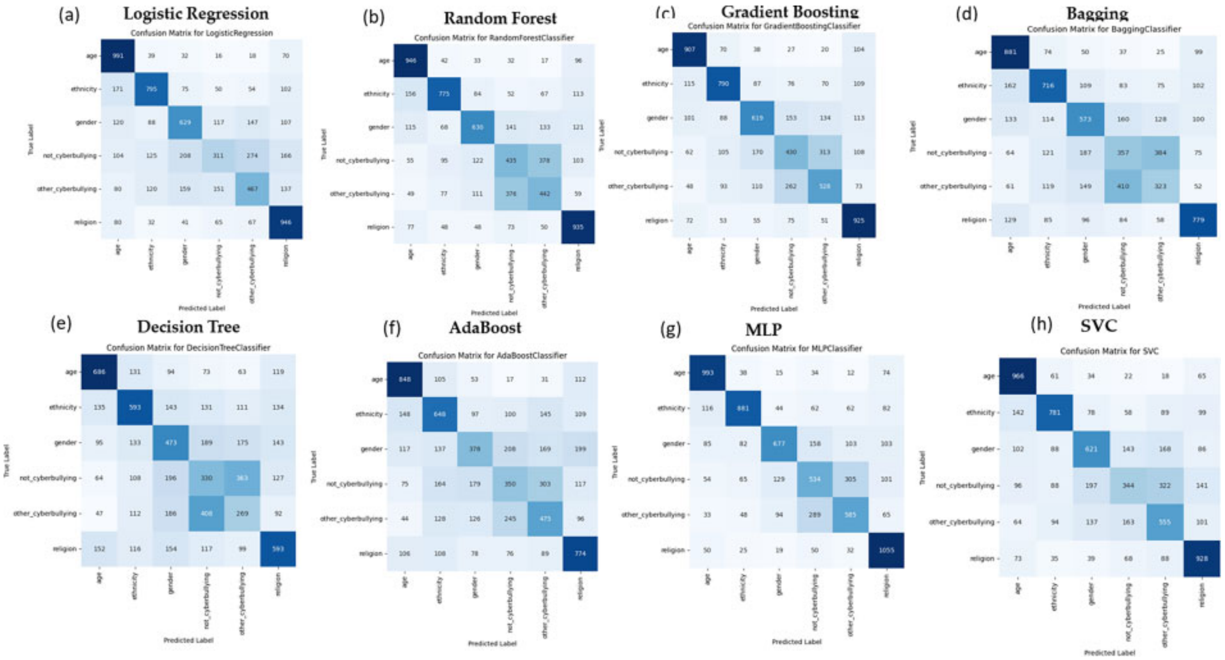


Figure 7: Confusion matrix of different ML models using Word2Vec for cyberbullying classification

The hybrid approach (TF-IDF + Word2Vec) yielded the most compelling results, with SVC achieving peak accuracy (0.85) and F1-score (0.85), followed by GB and LR (0.84 accuracy). This synergy of lexical specificity (TF-IDF) and semantic granularity (Word2Vec) improved accuracy by 3%–19% over standalone methods, validating the framework's efficacy. While KNN lagged (0.68 accuracy), MLP (0.81) and Bagging (0.80) demonstrated substantial gains from feature fusion, emphasizing the versatility of hybrid integration. Collectively, these findings underscore the superiority of combining syntactic and semantic signals, particularly for models like SVC and LR, which consistently excel across diverse feature spaces.

The confusion matrices in Fig. 7 visually represent how each model performs across specific cyberbullying categories using Word2Vec embedding. For instance, MLP (Fig. 7g) and GB (Fig. 7c) show balanced performance across most categories, particularly in identifying age-based and ethnicity-based bullying. At the same time, DT (Fig. 7e) and AdaBoost (Fig. 7f) display more significant misclassification, especially in categories like gender-based and religion-based bullying. These results underline the effectiveness of embedding approaches like Word2Vec for models such as MLP, while traditional ensemble methods may struggle in this context.

Fig. 8 provides confusion matrices for various ML models using the TF-IDF + Word2Vec approach. For instance, models like SVC and LR consistently deliver strong classification performance, particularly in categories with more explicit boundaries, such as ethnicity-based and religion-based bullying. Their ability to effectively combine syntactic (word frequency) and semantic (contextual word meaning) features, derived from the hybrid TF-IDF + Word2Vec approach, results in fewer misclassifications than other models. Ensemble methods like GB and Bagging also maintain a balanced performance across most categories, benefiting from integrating the hybrid feature set, which allows them to handle complex linguistic patterns. On the other hand, DT and AdaBoost exhibit more frequent misclassifications, particularly in non-cyberbullying and other-cyberbullying categories. This could indicate a limitation in how these models handle overlapping or less distinct cyberbullying categories, where linguistic subtleties play a more significant role. The broader implication of this analysis is that integrating both TF-IDF and Word2Vec enhances the ability of models to capture not just the frequency of terms but also the nuanced meanings of words in context, critical for detecting varied forms of cyberbullying. For instance, SVC's superior performance in detecting other cyberbullying content highlights its robustness in parsing subtler linguistic differences that simpler models may not as effectively capture.

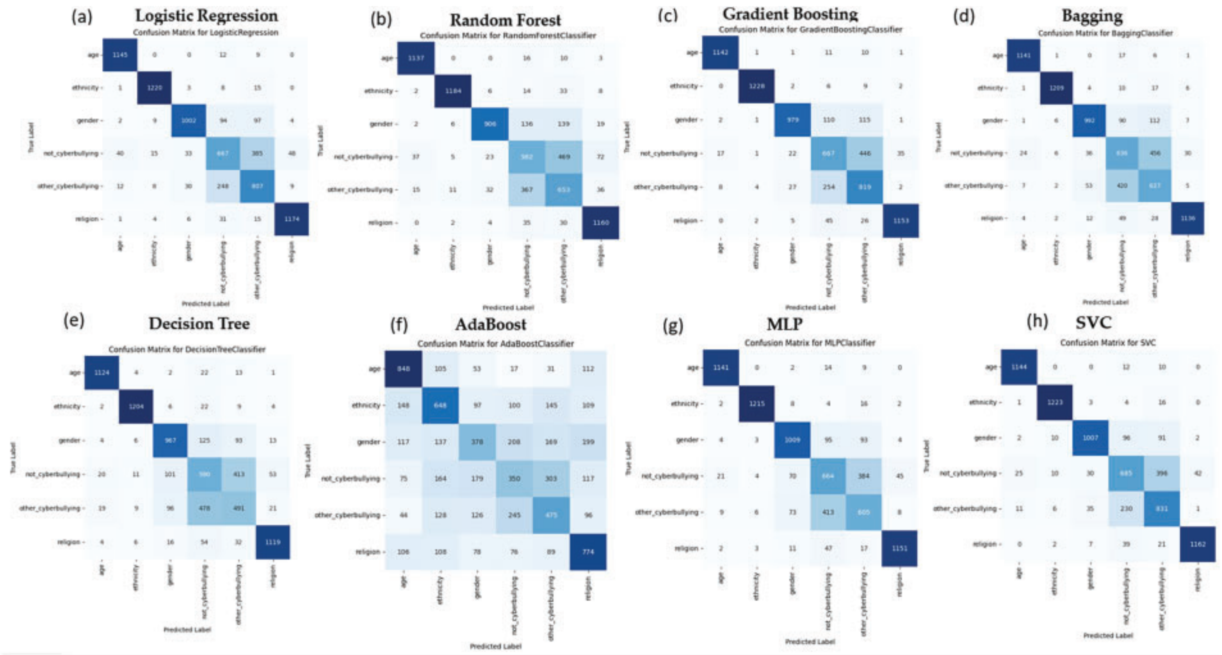


Figure 8: Confusion Matrix of ML Models using TF-IDF + Word2Vec for cyberbullying classification

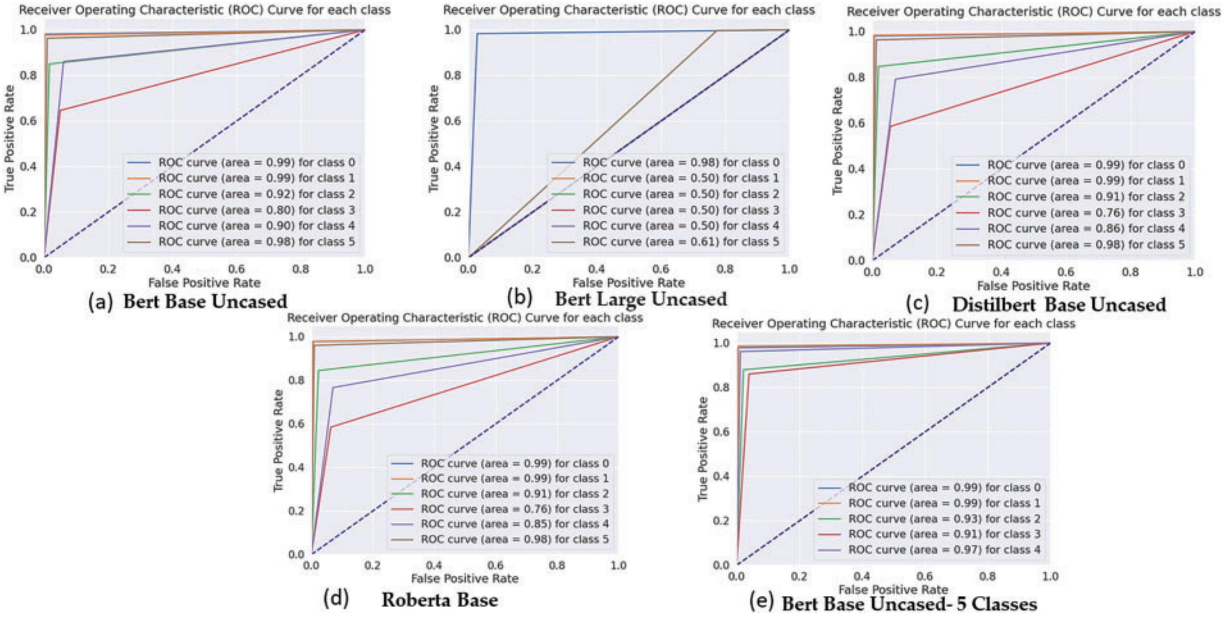
4.3 Performance Evaluation of Transformer-Based Model Variations for Cyberbullying Classification

We conducted extensive experiments using various BERT model variations to identify different types of cyberbullying from text data. The BERT versions considered in this study include BERT Base Un-Cased, BERT Large Un-Cased, DistilBERT Base Un-Cased, and RoBERTa Base, all using the uncased option [39,40]. The uncased versions treat words like 'Idiot' and 'idoit' as the same token, unlike the cased versions, which treat them as separate tokens. For this study, we selected the uncased models to focus on semantic meaning over case sensitivity (see Table 5). The BERT Base Un-Cased model performed best, with a training accuracy of 0.97 and a validation accuracy of 0.88. Its consistent performance across training, validation, and test data, with weighted precision, recall, and F1-scores at 0.88, indicates strong reliability in multi-class cyberbullying classification. Conversely, the BERT Large Un-Cased model showed the weakest performance, with a train and validation accuracy of 0.34 and a much lower weighted F1-score of 0.21, likely due to overfitting or challenges in generalizing the given dataset. The DistilBERT Base Un-Cased model delivered a solid performance, with a training accuracy of 0.93 and validation accuracy of 0.85, offering a good balance between computational efficiency and accuracy. RoBERTa Base performed similarly to DistilBERT, with a training and validation accuracy of 0.93 and 0.85, respectively, showcasing strong generalization capabilities. To further explore transformer-based architectures, we benchmark with the performance of T5 (Text-To-Text Transfer Transformer) and GPT-2 (Generative Pre-trained Transformer 2) under identical experimental conditions. T5-Base achieved a training accuracy of 0.95, slightly lower than BERT Base, reflecting its text-generation objective's reduced specificity for direct classification. Validation and test accuracy were 0.87 and 0.86, respectively, with weighted precision, recall, and F1-scores at 0.86, 0.85, and 0.85. GPT-2, a decoder-only autoregressive model pretrained on next-token prediction, underperformed BERT and T5 in cyberbullying classification. While GPT-2 excels in text generation, its unidirectional architecture limits bidirectional context understanding, critical for detecting nuanced cyberbullying (e.g., sarcasm, implicit threats). When fine-tuned on the same dataset with ADASYN oversampling, GPT-2 achieved lower accuracy, and F1-scores compared to BERT/T5 (Table 5). This aligns with prior work showing decoder-only models struggle with classification tasks due to their lack of task-specific pretraining objectives (e.g., masked language modeling in BERT, text-to-text in T5). To address data imbalance issues, we applied the ADASYN oversampling technique after preprocessing, ensuring that the model could handle underrepresented classes effectively. The ROC-AUC curves (Fig. 9) demonstrate that the BERT Base Un-Cased model consistently performs well across all cyberbullying categories.

Table 5: Comparison of transformer-based model variations performance on test data for cyberbullying classification

Model	Train accuracy	Val accuracy	Test accuracy	Weighted precision	Weighted recall	Weighted F1
BERT base Un-Cased	0.97	0.88	0.88	0.88	0.88	0.88
BERT large Un-Cased	0.34	0.34	0.33	0.18	0.33	0.21
Distilbert base Un-Cased	0.93	0.85	0.86	0.86	0.86	0.86
RoBERTa base	0.93	0.85	0.85	0.85	0.85	0.85
T5 base	0.95	0.87	0.87	0.86	0.85	0.85
GPT-2 base	0.90	0.83	0.83	0.81	0.80	0.80

Fig. 9 showcases the ROC curves for various BERT model variations evaluating their performance across five cyberbullying classes. These ROC curves plot the True Positive Rate (TPR) against the False Positive Rate (FPR) for each class, offering an in-depth comparison of how effectively each model distinguishes between different cyberbullying types. BERT Base Un-Cased (Fig. 9a) demonstrates exceptional classification performance across all classes. The ROC curve for class 0 (age-based bullying) and class 1 (ethnicity-based bullying) achieves an almost perfect AUC of 0.99, while other classes, such as class 2 (gender-based bullying) and class 4 (religion-based bullying), also perform well with AUC values of 0.92 and 0.90, respectively. This indicates strong and consistent classification capabilities across most categories. In contrast, BERT Large Un-Cased (Fig. 9b) struggles significantly, with AUC values around 0.50 for most classes, reflecting its poor generalization ability. Only class 0 achieves a competitive AUC of 0.98, while other classes demonstrate near-random performance, indicating that the model is not well-suited for this specific task and dataset, possibly due to overfitting or difficulty in handling the dataset complexity. DistilBERT Base Un-Cased (9c) balances performance and computational efficiency, showing strong AUC values for class 0 and class 1 (both 0.99), like BERT Base. However, it lags slightly for class 3 (other-cyberbullying), with an AUC of 0.76, and class 4 (religion-based bullying), with an AUC of 0.86. Nevertheless, it remains an efficient alternative with minimal performance trade-offs compared to BERT Base. RoBERTa Base (9d) closely mirrors DistilBERT's performance, with high AUC values for class 0 (0.99), class 1 (0.99), and class 2 (0.93). While class 3 and class 4 show slightly lower AUC values (0.76 and 0.85, respectively), RoBERTa maintains robust overall performance, making it a competitive option for this task. The BERT Base Un-Cased (5 Classes) (Fig. 9e) version, evaluated in five classes instead of six, also performs exceptionally well across all categories. Its AUC values for class 0 (0.99), class 1 (0.99), and class 4 (0.97) highlight its robustness, particularly when dealing with fewer classes. This suggests that reducing the number of classes slightly improves performance without sacrificing accuracy. In summary, BERT Base Un-Cased and DistilBERT stand out as the most reliable models for multi-class cyberbullying classification, with RoBERTa providing a close alternative. Meanwhile, BERT Large Un-Cased underperforms, indicating that a larger model size does not necessarily translate to better performance in this specific context. The consistently high AUC scores across most models suggest combining syntactic and semantic feature extraction (TF-IDF + Word2Vec) could further enhance classification performance.



(a) Bert Base Uncased

(b) Bert Large Uncased

(c) Distilbert Base Uncased

(d) Roberta Base

(e) Bert Base Uncased- 5 Classes

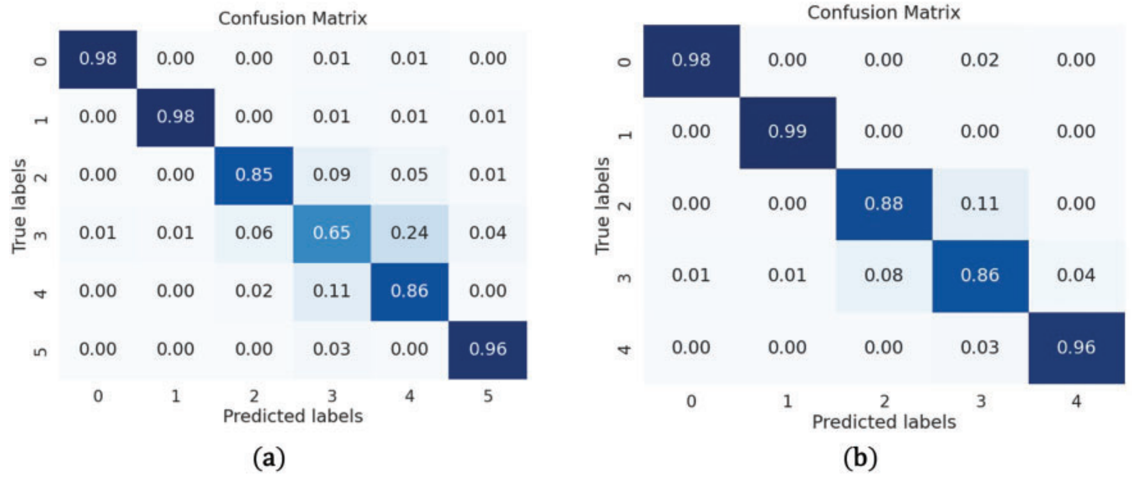
Figure 9: ROC curves for various BERT model variations—BERT Base Un-Cased, BERT Large Un-Cased, DistilBERT Base Un-Cased, RoBERTa Base, and BERT Base Un-Cased (5 Classes)—evaluating their performance across multiple cyberbullying classes

From a review of the literature [16], it is evident that many studies overlook the other cyberbullying classes or limit their datasets to only five classes. To address this gap, we conducted additional experiments using a five-class configuration and evaluated the performance of the BERT Base Un-Cased model. The model achieved a training accuracy of 99 ± 1%, with both validation and test accuracy reaching 93 ± 1%. These results suggest a high level of model generalization and robustness. The ROC-AUC curves for the five-class configuration are presented in Fig. 9e, while the corresponding Confusion Matrix is displayed in Fig. 9b, further demonstrating the model's ability to classify the reduced set of classes accurately. The best model in this study is the BERT Base Uncased model, with an accuracy of 88% (Table 6).

Table 6: BERT base Un-cased model per-class performance for 6 classes dataset

No.	Precision	Recall	F1
0	0.99	0.98	0.98
1	0.99	0.98	0.98
2	0.91	0.85	0.88
3	0.71	0.65	0.68
4	0.75	0.86	0.80
5	0.95	0.96	0.96
Average	0.88	0.88	0.88
Overall accuracy (88%)			

Fig. 10 provides the confusion matrices for two variations of the BERT Base Un-Cased model (trained in 6 classes and trained in 5 classes). Fig. 10a, representing the 6-class version, demonstrates strong performance across most classes. For instance, class 0 (age-based bullying) and class 1 (ethnicity-based bullying) show excellent classification accuracy with 98% true positives and minimal misclassifications. However, the model shows a noticeable decline in performance for class 3 (other-cyberbullying), where the true positive rate drops to 65%, with significant confusion between class 3 and class 4 (religion-based bullying), indicating overlap between these two categories. Fig. 10b, trained in 5 classes, exhibits similar performance for classes 0 and 1, achieving nearly 99% accuracy. The removal of class 5 in this variation appears to slightly improve classification performance for class 3, with the true positive rate increasing to 86%. This suggests that reducing the number of classes mitigates confusion between similar categories, especially between class 3 and class 4. To conclude, reducing the number of classes from 6 to 5 leads to better performance for the model, particularly in distinguishing between the more challenging categories like other cyberbullying and religion-based bullying. However, both models maintain consistently high performance in detecting well-defined categories like age-based and ethnicity-based bullying.



	0	1	2	3	4	5
0	0.98	0.00	0.00	0.01	0.01	0.00
1	0.00	0.98	0.00	0.01	0.01	0.01
2	0.00	0.00	0.85	0.09	0.05	0.01
3	0.01	0.01	0.06	0.65	0.24	0.04
4	0.00	0.00	0.02	0.11	0.86	0.00
5	0.00	0.00	0.00	0.03	0.00	0.96

(a)

	0	1	2	3	4
0	0.98	0.00	0.00	0.02	0.00
1	0.00	0.99	0.00	0.00	0.00
2	0.00	0.00	0.88	0.11	0.00
3	0.01	0.01	0.08	0.86	0.04
4	0.00	0.00	0.00	0.03	0.96

(b)

Figure 10: (a) Confusion Matrix BERT Base Un-cased (6 classes) and (b) Confusion Matrix BERT Base Un-cased (5-Classes)

4.4 Comparison of Models Complexity and Inference Time

The computational efficiency of transformer-based models is a critical factor in real-world applications, particularly in resource-constrained environments. A comparative analysis of six prominent models (Fig. 11) highlights the trade-offs between model complexity and inference time, emphasizing the practical implications of their architectural differences. Encoder-based models demonstrate a more favorable balance between efficiency and performance. BERT Base, with 110 million parameters, completes inference in 85 s, making it an optimal choice for classification tasks requiring both accuracy and computational feasibility. In contrast, BERT Large, with 340 million parameters, demands 134 s, reflecting its increased depth and computational burden, which limits its practicality on standard hardware. DistilBERT, a distilled variant with only 66 million parameters, achieves the fastest inference time of 57 s while retaining competitive accuracy, making it highly efficient. Similarly, RoBERTa Base, despite its slightly larger parameter count of 125 million, matches BERT Base's 85-s inference time due to its optimized pretraining and dynamic tokenization strategies. In contrast, generative models exhibit significantly higher computational costs due to their sequential processing requirements. T5 Base, an encoder-decoder model with 220 million parameters, requires 188 s for inference Table 7, illustrating the inefficiency of its autoregressive decoding mechanism for classification tasks. GPT-2 Base, which relies solely on autoregressive generation, has the highest inference time at 376 s, underscoring its substantial computational demands. Finally, larger models like BERT Large and generative architectures are slower and resource-intensive, making them impractical for real-time use.

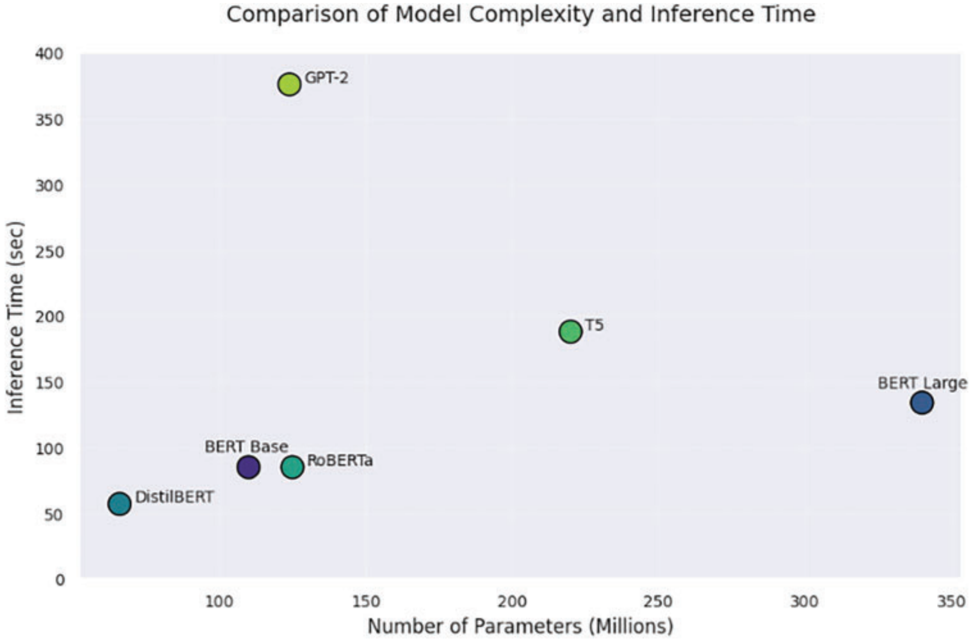


Figure 11: Comparison of model's complexity and inference time

Table 7: Comparison of model complexity and inference time

Model	Total inference time	Number of parameters (M)
BERT base Un-Cased	~85 s	~110 M
BERT large Un-Cased	~134 s	~340 M
Distilbert base Un-Cased	~57 s	~66 M

(Continued)

Table 7 (continued)

Model	Total inference time	Number of parameters (M)
RoBERTa base	~85 s	~125 M
T5 base	~188 s	~220 M
GPT-2 base	~376 s (6.3 min)	~124 M

5 Conclusion

This study investigated the performance of various machine learning and transformer-based models, particularly BERT variants, for multi-class cyberbullying classification using Twitter (X) data. Addressing a key gap in the literature, we focused on the often-overlooked cyberbullying subtypes, which are frequently underrepresented in existing datasets and analyses. By integrating TF-IDF and Word2Vec for feature extraction, we effectively captured both syntactic and semantic information, enhancing the models' ability to distinguish between nuanced forms of cyberbullying. Our benchmarking revealed that BERT Base Un-Cased consistently outperformed other models in terms of classification accuracy and overall robustness, underscoring its reliability in identifying diverse cyberbullying categories. DistilBERT and RoBERTa delivered competitive performance with marginally lower accuracy, while BERT Large Uncased underperformed due to optimization complexity. Although generative models such as GPT-2 and T5 excel in text generation, their effectiveness in classification tasks was comparatively limited. Beyond accuracy, we examined model scalability and inference efficiency. Our findings indicate that while larger models offer greater representational capacity, they incur significant computational costs and longer processing times, factors that may limit their practical deployment. Moreover, this study was constrained to text-based features, excluding potentially valuable multimodal signals such as images or network-based behavior patterns. Future work should prioritize the integration of multimodal data and adaptive learning strategies to improve generalizability in real-world scenarios. Additionally, tackling class imbalance through dynamic sampling methods and refining fine-tuning protocols will be critical for advancing the performance and applicability of cyberbullying detection systems in increasingly complex online environments.

Acknowledgement: This research has been funded by Scientific Research Deanship at University of Ha'il–Saudi Arabia through Project Number RG-23 092.

Funding Statement: This research has been funded by Scientific Research Deanship at University of Ha'il–Saudi Arabia through Project Number RG-23 092.

Author Contributions: Conceptualization: Suliman Mohamed Fati, Mohammed A. Mahdi, Sawsan A. Saad, and Mohammed Gamal Ragab; Data curation: Shahanawaj Ahamad and Sawsan A. Saad; Methodology, Suliman Mohamed Fati, Mohammed A. Mahdi and Mohammed Gamal Ragab; Project administration: Mohammed A. Mahdi and Suliman Mohamed Fati; Resources: Sawsan A. Saad and Mohammed Gamal Ragab; Software: Mohammed Gamal Ragab and Suliman Mohamed Fati; Validation: Suliman Mohamed Fati and Mohammed A. Mahdi; Visualization: Mohamed A.G. Hazber; Writing—original draft: Mohammed A. Mahdi, Shahanawaj Ahamad, Sawsan A. Saad, and Mohammed Gamal Ragab; Writing—review and editing: Mohamed A.G. Hazber, Mohammed Al-Shalabi, and Suliman Mohamed Fati. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study for cyberbullying classification is publicly available and can be accessed at Kaggle: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data (accessed on 31 August 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Muneer A, Fati SM. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet. 2020;12(11):187. doi:10.3390/fi12110187.
2. Malik A, Dadure P. Cyberbullying in the digital age: consequences and countermeasures. in: empowering low-resource languages with NLP solutions. Hershey, PA, USA: IGI Global; 2024. p. 247–73. doi:10.4018/979-8-3693-0728-1.ch012.
3. Yurdakul Y, Ayhan AB. The effect of the cyberbullying awareness program on adolescents' awareness of cyberbullying and their coping skills. Curr Psychol. 2023;42(28):24208–22. doi:10.1007/s12144-022-03483-3.
4. John A, Lee SC, Puchades A, Del Pozo-Baños M, Morgan K, Page N, et al. Self-harm, in-person bullying and cyberbullying in secondary school-aged children: a data linkage study in Wales. J Adolesc. 2023;95(1):97–114. doi:10.1002/jad.12102.
5. Nikolaou D. Bullying, cyberbullying, and youth health behaviors. Kyklos. 2022;75(1):75–105. doi:10.1111/kykl.12286.
6. Al-Harigy LM, Al-Nuaim HA, Moradpoor N, Tan Z. Building towards automated cyberbullying detection: a comparative analysis. Comput Intell Neurosci. 2022;2022:4794227. doi:10.1155/2022/4794227.
7. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: a survey. Sci China Technol Sci. 2020;63(10):1872–97. doi:10.1007/s11431-020-1647-3.
8. Wang X, Koneru S, Venkit PN, Frischmann B, Rajtmajer S. The unappreciated role of intent in algorithmic moderation of social media content. arXiv:2405.11030. 2024.
9. Resende de Mendonça R, Felix de Brito D, de Franco Rosa F, dos Reis JC, Bonacin R. A framework for detecting intentions of criminal acts in social media: a case study on twitter. Information. 2020;11(3):154. doi:10.3390/info11030154.
10. Ghayoumi M, Ghazinour K. Advancing MAISON: integrating deep learning and social dynamics in cyberbullying detection and prevention. In: 2024 7th International Conference on Information and Computer Technologies (ICICT); 2024 Mar 15–17; Honolulu, HI, USA; 2024. p. 453–61. doi:10.1109/ICICT62343.2024.00080.
11. Iwendi C, Srivastava G, Khan S, Maddikunta PKR. Cyberbullying detection solutions based on deep learning architectures. Multimed Syst. 2023;29(3):1839–52. doi:10.1007/s00530-020-00701-5.
12. Hasan MT, Al Emran Hossain M, Mukta MSH, Akter A, Ahmed M, Islam S. A review on deep-learning-based cyberbullying detection. Future Internet. 2023;15(5):179. doi:10.3390/fi15050179.
13. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. Proc Int AAAI Conf Web Soc Medium. 2011;5(3):11–7. doi:10.1609/icwsm.v5i3.14209.
14. Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops; 2011 Dec 18–21; Honolulu, HI, USA; 2011. p. 241–4. doi:10.1109/ICMLA.2011.152.
15. Nandhini BS, Sheeba JI. Online social network bullying detection using intelligence techniques. Procedia Comput Sci. 2015;45(2):485–92. doi:10.1016/j.procs.2015.03.085.
16. Alqahtani AF, Ilyas M. An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. Mach Learn Knowl Extr. 2024;6(1):156–70. doi:10.3390/make6010009.
17. Mishra P, Yannakoudakis H, Shutova E. Neural character-based composition models for abuse detection. arXiv:1809.00378. 2018.
18. Ali Talpur B, O.'Sullivan D. Multi-class imbalance in text classification: a feature engineering approach to detect cyberbullying in twitter. Informatics. 2020;7(4):52. doi:10.3390/informatics7040052.
19. Ejaz N, Razi F, Choudhury S. Towards comprehensive cyberbullying detection: a dataset incorporating aggressive texts, repetition, peerness, and intent to harm. Comput Hum Behav. 2024;153(3):108123. doi:10.1016/j.chb.2023.108123.
20. Faraj A, Utku S. Comparative analysis of word embeddings for multiclass cyberbullying detection. UHD J Sci Technol. 2024;8(1):55–63. doi:10.21928/uhdjst.v8n1y2024.pp55-63.

21. Kaddoura S, Nassar R. Language model-based approach for multiclass cyberbullying detection. In: Web Information Systems Engineering—WISE 2024. Singapore: Springer Nature; 2024. p. 78–89. doi:10.1007/978-981-96-0567-5_7.
22. Saranyanath KP, Shi W, Corriveau JP. Cyberbullying detection using ensemble method. In: Data science and machine learning. Chennai, India: Academy and Industry Research Collaboration Center (AIRCC); 2022. p. 75–94. doi:10.5121/csit.2022.121507.
23. Cyberbullying classification. [Internet]. 2020 [cited 2024 Aug 31]. Available from: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data.
24. Fati SM, Muneer A, Alwadain A, Balogun AO. Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. Mathematics. 2023;11(16):3567. doi:10.3390/math11163567.
25. Muneer A, Alwadain A, Ragab MG, Alqushaibi A. Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information. 2023;14(8):467. doi:10.3390/info14080467.
26. Lafferty JPS. Tweet-preprocessor: python library for preprocessing tweets. GitHub repository [Internet]; 2020. [cited 2024 Feb 13]. Available from: https://github.com/s/preprocessor.
27. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008 Jun 1–8; Hong Kong, China; 2008. p. 1322–8. doi:10.1109/IJCNN.2008.4633969.
28. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl. 2004;6(1):20–9. doi:10.1145/1007730.1007735.
29. Yen S, Lee Y. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. in: Intelligent control and automation. Berlin/Heidelberg, Germany: Springer; 2006. p. 731–40. doi:10.1007/11816492_89.
30. Yang C, Fridgeirsson EA, Kors JA, Reps JM, Rijnbeek PR. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. J Big Data. 2024;11(1):7. doi:10.1186/s40537-023-00857-7.
31. Zhou H. Research of text classification based on TF-IDF and CNN-LSTM. J Phys Conf Ser. 2022;2171(1):12021. doi:10.1088/1742-6596/2171/1/012021.
32. Liu W, Cao Z, Wang J, Wang X. Short text classification based on wikipedia and Word2Vec. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC); 2016 Oct 14–17; Chengdu, China; 2016. p. 1195–200. doi:10.1109/CompComm.2016.7924894.
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
34. Guan B, Zhu X, Yuan S. A T5-based interpretable reading comprehension model with more accurate evidence training. Inf Process Manag. 2024;61(2):103584. doi:10.1016/j.ipm.2023.103584.
35. Akbar NA, Darmayanti I, Fati SM, Muneer A. Deep learning of a pre-trained language model's joke classifier using GPT-2. J Hunan Univ Nat Sci. 2021;48(8):235–41.
36. Song X, Salcianu A, Song Y, Dopson D, Zhou D. Fast WordPiece tokenization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021; Punta Cana, Dominican Republic. p. 2089–103. doi:10.18653/v1/2021.emnlp-main.160.
37. Petridis C. Text classification: neural networks vs machine learning models vs pre-trained models. arXiv:2412.21022. 2024.
38. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101. 2017.
39. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019.
40. Topal M, Bas A, van Heerden I. Exploring transformers in natural language generation: GPT, BERT, and XLNet. arXiv:2102.08036. 2021.