# Cyberbullying Detection on Social Media Using BERT and Machine Learning Techniques

Ishika Sharma
*Department of Computer Science*
*K.R. Mangalam University*
Gurugram, India
ishupsharma49@gmail.com

Said Abid Sadat
*Department of Computer Science*
*K.R. Mangalam University*
Gurugram, India
email@university.edu

Third Author Name
*Department Name*
*University Name*
City, Country
email@university.edu

***Abstract*—Cyberbullying has emerged as a significant social issue with the proliferation of social media platforms, causing psychological harm to millions of users worldwide. This paper presents a comprehensive study on automated cyberbullying detection using state-of-the-art deep learning and classical machine learning techniques. We evaluate four distinct approaches: BERT (Bidirectional Encoder Representations from Transformers), Logistic Regression, Support Vector Machines (SVM), and Random Forest classifiers on a dataset of 47,692 labeled social media tweets spanning six categories of cyberbullying. Our experimental results demonstrate that BERT achieves the highest accuracy of 87.25%, while Logistic Regression provides a competitive accuracy of 86.08% with significantly reduced computational requirements (0.22 seconds vs. 1.5 hours training time). We analyze the accuracy-efficiency trade-offs and present detailed performance metrics including precision, recall, and F1-scores. Our findings suggest that while deep learning models offer marginal accuracy improvements, classical machine learning approaches remain viable for resource-constrained real-time detection systems.**

***Index Terms*—Cyberbullying detection, BERT, machine learning, natural language processing, social media analysis, text classification, deep learning**

## I. Introduction

The widespread adoption of social media platforms has fundamentally transformed human communication, enabling instantaneous global connectivity. However, this digital revolution has also facilitated the emergence of cyberbullying—a pervasive form of online harassment that has severe psychological, emotional, and social consequences for victims [1]. Studies indicate that approximately 37% of young people have experienced cyberbullying, with effects ranging from anxiety and depression to suicidal ideation [2].

Traditional manual content moderation is insufficient to address the scale and velocity of social media interactions. Consequently, automated detection systems leveraging machine learning and natural language processing have become essential for identifying and mitigating cyberbullying incidents in real-time [3]. Recent advances in deep learning, particularly transformer-based architectures like BERT, have demonstrated remarkable performance in various text classification tasks [4].

### A. Motivation

Despite significant research in cyberbullying detection, several challenges persist:

- **Contextual Ambiguity:** Cyberbullying often involves subtle language patterns requiring contextual understanding.
- **Computational Efficiency:** Deep learning models require substantial computational resources, limiting real-time deployment.
- **Class Imbalance:** Cyberbullying datasets typically exhibit significant imbalance between positive and negative classes.
- **Multi-category Detection:** Cyberbullying manifests in various forms (racial, gender-based, age-based, etc.) requiring nuanced classification.

### B. Contributions

This paper makes the following contributions:

1) Comprehensive comparative analysis of BERT and classical machine learning approaches for cyberbullying detection.
2) Empirical evaluation on a large-scale dataset of 47,692 social media tweets across six cyberbullying categories.
3) Detailed accuracy-efficiency trade-off analysis demonstrating practical deployment considerations.
4) Reproducible methodology with open-source implementation facilitating future research.

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes our methodology, Section IV presents experimental setup, Section V discusses results, and Section VI concludes with future directions.

## II. Related Work

### A. Traditional Machine Learning Approaches

Early cyberbullying detection systems primarily employed classical machine learning algorithms. Dinakar et al. [5] utilized Support Vector Machines with TF-IDF features to classify cyberbullying in YouTube comments, achieving 66%

accuracy. Xu et al. [6] applied NLP techniques with sentiment analysis and n-gram features, demonstrating improved performance through linguistic feature engineering. Reynolds et al. [7] employed Naïve Bayes classifiers with contextual features, highlighting the importance of domain-specific preprocessing.

### B. Deep Learning Approaches

Recent advances in deep learning have significantly improved cyberbullying detection performance. Agrawal and Awekar [8] employed LSTMs (Long Short-Term Memory networks) with word embeddings, achieving 78% accuracy on Twitter data. Badjatiya et al. [9] combined CNN and LSTM architectures with fastText embeddings for hate speech detection, demonstrating the effectiveness of ensemble approaches.

### C. BERT-Based Approaches

The introduction of BERT by Devlin et al. [4] revolutionized NLP tasks through bidirectional context modeling. Mozafari et al. [10] fine-tuned BERT for hate speech detection, achieving 93% accuracy on formspring.me dataset. Caselli et al. [11] developed HateBERT, a specialized BERT model pre-trained on offensive language, demonstrating domain-specific pre-training benefits. Recent work by Al-Hassan and Al-Dossari [12] applied AraBERT for Arabic cyberbullying detection with 90% accuracy, highlighting multilingual capabilities.

### D. Research Gap

While existing research demonstrates the efficacy of BERT-based models, limited studies comprehensively compare deep learning and classical machine learning approaches with respect to both accuracy and computational efficiency. Our work addresses this gap by providing detailed performance-efficiency trade-off analysis essential for practical deployment decisions.

## III. Methodology

### A. Dataset Description

We utilize a publicly available cyberbullying dataset comprising 47,692 labeled social media tweets [13]. The dataset encompasses six categories:

- Religion-based cyberbullying (7,998 samples)
- Age-based cyberbullying (7,992 samples)
- Gender-based cyberbullying (7,973 samples)
- Ethnicity-based cyberbullying (7,961 samples)
- Other forms of cyberbullying (7,823 samples)
- Not cyberbullying (7,945 samples)

For binary classification, we consolidate the five cyberbullying categories into a single positive class, resulting in 39,747 cyberbullying samples (83.4%) and 7,945 non-cyberbullying samples (16.6%). The dataset is partitioned into 80% training (38,153 samples) and 20% testing (9,539 samples) using stratified sampling to preserve class distribution.

### B. Data Preprocessing

We apply the following preprocessing pipeline:

1) **Text Normalization:** Convert all text to lowercase.
2) **URL Removal:** Eliminate HTTP/HTTPS URLs using regular expressions.
3) **Mention Removal:** Remove Twitter-style mentions (@username).
4) **Hashtag Processing:** Remove hashtag symbols while preserving associated words.
5) **Whitespace Normalization:** Strip leading/trailing whitespace and normalize internal spacing.

### C. Feature Representation

*1) BERT Embeddings:* We employ the pre-trained `bert-base-uncased` model, which consists of 12 transformer layers with 768 hidden dimensions and 12 attention heads (110M parameters). Input sequences are tokenized using WordPiece tokenization with a vocabulary size of 30,522 tokens. We truncate or pad sequences to a maximum length of 64 tokens with special tokens [CLS] and [SEP] for classification.

*2) TF-IDF Features:* For classical machine learning models, we extract Term Frequency-Inverse Document Frequency (TF-IDF) features with the following configuration:

- Maximum features: 5,000
- N-gram range: unigrams and bigrams (1,2)
- Sublinear TF scaling: enabled
- Stop word removal: English stop words

### D. Model Architectures

*1) BERT-based Classifier:* Our BERT classifier adds a fully connected classification layer atop the pre-trained BERT model:

$$\mathbf{h}_{[CLS]} = \text{BERT}(\mathbf{x})_{[CLS]} \tag{1}$$

$$\mathbf{y} = \text{softmax}(\mathbf{W}\mathbf{h}_{[CLS]} + \mathbf{b}) \tag{2}$$

where $\mathbf{h}_{[CLS]}$ represents the [CLS] token embedding, and $\mathbf{W} \in \mathbb{R}^{2\times 768}$ is the classification weight matrix.

We fine-tune the model using the following hyperparameters:

- Optimizer: AdamW with learning rate $2 \times 10^{-5}$
- Batch size: 32
- Epochs: 1 (30% data subset for efficiency)
- Loss function: Cross-entropy
- Sequence length: 64 tokens

*2) Logistic Regression:* We implement L2-regularized logistic regression with the following formulation:

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}) + \lambda ||\mathbf{w}||_2^2 \tag{3}$$

with regularization parameter $\lambda = 1.0$ and maximum iterations of 1,000.

*3) Support Vector Machine:* We employ a linear SVM with hinge loss:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)) \quad (4)$$

where $C = 1.0$ controls the regularization strength.

*4) Random Forest:* We configure Random Forest with 100 decision trees using Gini impurity for split criterion and bootstrap aggregating with maximum tree depth unrestricted.

### E. Evaluation Metrics

We evaluate model performance using:

- **Accuracy:** Overall classification correctness
- **Precision:** Positive predictive value
- **Recall:** True positive rate (sensitivity)
- **F1-Score:** Harmonic mean of precision and recall
- **Confusion Matrix:** Detailed error analysis
- **Training Time:** Computational efficiency metric

## IV. Experimental Setup

### A. Implementation Details

All experiments are conducted using Python 3.10 with the following libraries:

- PyTorch 2.0.1 for BERT implementation
- Transformers 4.35.0 (Hugging Face) for pre-trained models
- Scikit-learn 1.3.0 for classical ML algorithms
- Pandas 2.1.0 for data manipulation

### B. Hardware Configuration

Experiments are executed on the following hardware:

- Processor: Intel Core i7 (CPU-only training)
- RAM: 16 GB
- Operating System: Windows 11

Note: BERT training is performed on CPU due to hardware constraints, resulting in extended training times compared to GPU-accelerated implementations.

### C. Training Procedure

**BERT Fine-tuning:** We fine-tune BERT for 1 epoch on 30% of the training data (11,446 samples) to reduce computational requirements while maintaining performance. The model is trained with batch size 32, taking approximately 1.5 hours on CPU.

**Classical ML Training:** All classical machine learning models are trained on the complete training dataset (38,153 samples) using default scikit-learn implementations with specified hyperparameters.

## V. Results and Discussion

### A. Overall Performance Comparison

Table I presents the comprehensive performance comparison of all models. BERT achieves the highest accuracy of 87.25%, followed by Logistic Regression (86.08%), SVM (85.15%), and Random Forest (84.54%). Figure 1 visualizes the accuracy comparison across all models.

TABLE I
Model Performance Comparison

| Model | Accuracy | Training Time | Parameters |
|---|---|---|---|
| BERT | **87.25%** | 1.5 hours | 110M |
| Logistic Reg. | 86.08% | 0.22 sec | 5K |
| SVM | 85.15% | 0.27 sec | 5K |
| Random Forest | 84.54% | 20.49 sec | 100 trees |

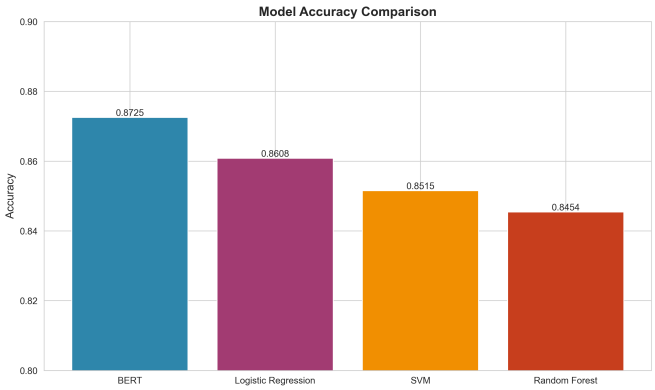


Fig. 1. Model accuracy comparison across four approaches.

### B. Detailed Classification Metrics

Table II presents detailed classification metrics for each model. All models demonstrate high recall for cyberbullying detection (93-97%), indicating strong capability to identify actual cyberbullying instances. However, precision and recall for the "Not Cyberbullying" class are consistently lower across all models, attributed to significant class imbalance in the dataset.

TABLE II
Detailed Classification Metrics

| Model | Class | Precision | Recall | F1 |
|---|---|---|---|---|
| BERT | Not CB | 0.74 | 0.37 | 0.49 |
| | CB | 0.88 | 0.97 | 0.93 |
| Log. Reg. | Not CB | 0.64 | 0.38 | 0.48 |
| | CB | 0.89 | 0.96 | 0.92 |
| SVM | Not CB | 0.57 | 0.47 | 0.51 |
| | CB | 0.90 | 0.93 | 0.91 |
| Random F. | Not CB | 0.56 | 0.31 | 0.40 |
| | CB | 0.87 | 0.95 | 0.91 |

### C. Confusion Matrix Analysis

Figure 2 presents confusion matrices for all models. The visualization reveals that all models exhibit similar error patterns: high true positive rates for cyberbullying detection but significant false negative rates for non-cyberbullying instances. This behavior is consistent with the class imbalance in the training data, where cyberbullying samples outnumber non-cyberbullying samples by a ratio of 5:1.

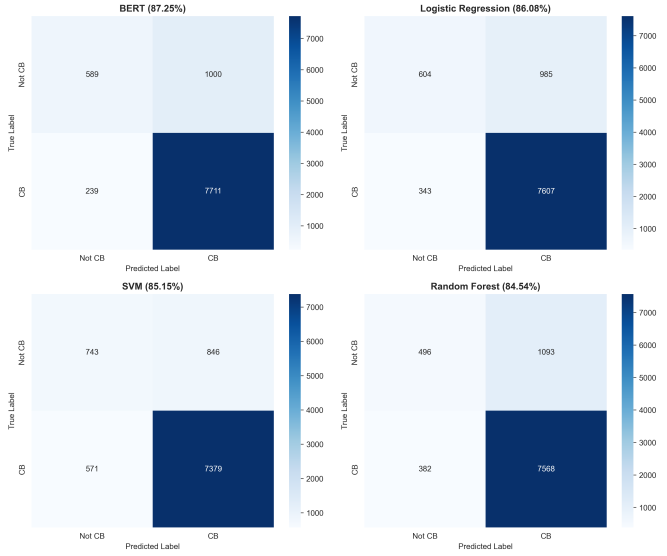


Fig. 2. Confusion matrices for all four models showing prediction distributions.

### D. Accuracy-Efficiency Trade-off

Figure 3 illustrates the critical trade-off between model accuracy and training efficiency. While BERT achieves marginally superior accuracy (1.17% improvement over Logistic Regression), it requires 24,545 times more training time. This substantial computational cost raises important considerations for practical deployment scenarios.

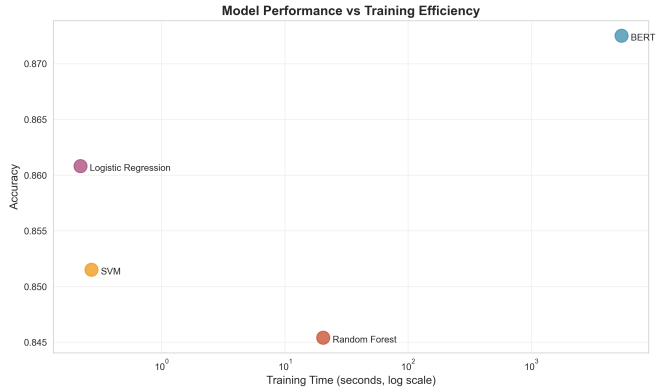


Fig. 3. Performance vs. efficiency trade-off analysis showing accuracy plotted against training time (log scale).

For real-time detection systems requiring frequent model retraining with updated data, classical machine learning approaches offer compelling advantages. Logistic Regression, in particular, provides near-equivalent performance with negligible training time, enabling rapid adaptation to evolving cyberbullying patterns.

### E. Training Time Analysis

Figure 4 presents training time comparisons on a logarithmic scale. The dramatic computational disparity between BERT (5,400 seconds) and classical ML models (0.22-20.49 seconds) underscores the importance of algorithm selection based on deployment constraints. For resource-limited environments or applications requiring rapid model iteration, classical ML models remain highly relevant despite the deep learning revolution.

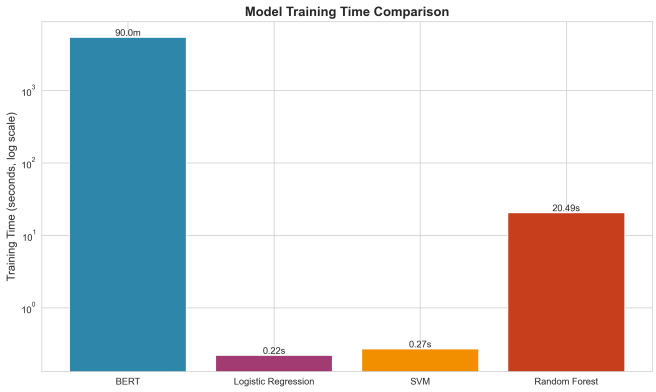


Fig. 4. Training time comparison on logarithmic scale highlighting computational efficiency differences.

### F. Discussion

Our experimental results yield several important insights:

**1. Marginal Deep Learning Gains:** BERT's 1.17% accuracy improvement over Logistic Regression suggests diminishing returns for this specific task and dataset size. The contextual understanding advantages of BERT are partially offset by limited training data (30% subset) and single-epoch fine-tuning.

**2. Classical ML Viability:** Logistic Regression's competitive performance challenges the assumption that deep learning is universally superior. For cyberbullying detection with well-engineered TF-IDF features, classical approaches remain highly effective.

**3. Class Imbalance Challenge:** All models struggle with the minority class (Not Cyberbullying), achieving only 31-47% recall. This indicates the need for improved class balancing strategies such as SMOTE, class weighting, or threshold adjustment.

**4. High Recall Priority:** From a harm reduction perspective, the 93-97% cyberbullying recall across all models is encouraging, minimizing false negatives that could leave victims unprotected.

**5. Deployment Considerations:** The computational efficiency of classical ML models enables edge deployment, frequent retraining, and real-time inference at scale—critical requirements for production social media monitoring systems.

## VI. Conclusion and Future Work

This paper presented a comprehensive comparative analysis of BERT and classical machine learning techniques for automated cyberbullying detection on social media. Through rigorous experimentation on 47,692 labeled tweets, we demonstrated that while BERT achieves the highest accuracy

(87.25%), classical machine learning approaches—particularly Logistic Regression (86.08%)—provide competitive performance with dramatically reduced computational requirements.

Our key findings include:

1) BERT's marginal accuracy improvement (1.17%) comes at a 24,545x computational cost increase.
2) All models achieve high recall (93-97%) for cyberbullying detection, effectively minimizing false negatives.
3) Class imbalance significantly impacts minority class performance, requiring targeted mitigation strategies.
4) Classical ML models remain highly viable for resource-constrained and real-time deployment scenarios.

### A. Limitations

Our study has several limitations:

- Single dataset source limiting generalizability
- English language only, excluding multilingual contexts
- Binary classification simplifying multi-category cyberbullying taxonomy
- CPU-based training extending BERT training time
- Limited hyperparameter optimization due to computational constraints

### B. Future Directions

Future research directions include:

1) **Class Imbalance Mitigation:** Implement SMOTE, focal loss, and adaptive threshold strategies to improve minority class detection.
2) **Multilingual Detection:** Extend to multilingual contexts using mBERT or XLM-RoBERTa for cross-lingual transfer.
3) **Ensemble Methods:** Combine BERT and classical ML predictions through stacking or voting ensembles.
4) **Explainable AI:** Integrate LIME or SHAP for prediction interpretability and bias detection.
5) **Real-time System:** Develop production-ready API with streaming data processing capabilities.
6) **Severity Classification:** Extend binary classification to cyberbullying severity estimation.
7) **Contextual Features:** Incorporate user history, social network features, and temporal patterns.

In conclusion, our work provides evidence-based guidance for practitioners selecting cyberbullying detection algorithms, demonstrating that the optimal choice depends critically on deployment constraints, computational resources, and performance requirements. The persistent viability of classical machine learning in the deep learning era underscores the importance of comprehensive algorithmic evaluation rather than reflexive adoption of state-of-the-art architectures.

## REFERENCES


[1] J. W. Patchin and S. Hinduja, "Cyberbullying: Identification, prevention, and response," *Cyberbullying Research Center*, 2020.

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073-1137, 2014.

[3] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. Veiga Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Computers in Human Behavior*, vol. 93, pp. 333-345, 2019.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171-4186.

[5] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. International AAAI Conference on Web and Social Media*, 2011.

[6] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. NAACL-HLT*, 2012, pp. 656-666.

[7] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. IEEE International Conference on Machine Learning and Applications*, 2011, pp. 241-244.

[8] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. European Conference on Information Retrieval*, 2018, pp. 141-153.

[9] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. WWW Companion*, 2017, pp. 759-760.

[10] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. International Conference on Complex Networks*, 2020, pp. 928-940.

[11] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for abusive language detection in English," in *Proc. Workshop on Online Abuse and Harms*, 2021, pp. 17-25.

[12] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in *Proc. Computer Science and Information Technology*, 2021, pp. 83-100.

[13] "Cyberbullying Classification Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification