Cyberbullying Detection using Transformer Architectures: A Comparative Experimental Study

Drashti Bhikadiya
P P Savani University
Kosamba, Surat, India

Hemangi Kacha
P P Savani University
Kosamba, Surat, India

Abhijeetsinh Jadeja
Sankalchand Patel University,
Visnagar, India

Jayashri Patil
P P Savani University,
Kosamba, Surat, India

ABSTRACT

The rapid growth of social media platforms has made the automatic detection of online harassment a pressing requirement for safe digital communication. Recent advances in deep learning, including Bi-LSTM and CNN based models, have shown strong results in identifying online hate speech, but most prior studies restrict their evaluation to a small number of explicit, attribute-specific categories. In this work, two Transformer-based architectures, RoBERTa and DistilBERT, are fine-tuned and evaluated on a challenging six-class cyberbullying classification dataset comprising the categories Age, Ethnicity, Gender, Religion, Other_Cyberbullying, and Not_Cyberbullying. RoBERTa achieved the best overall performance, with a test accuracy of 87.79% and a weighted F1-score of 0.88. DistilBERT achieved a comparable test accuracy of 87.19% (weighted F1 = 0.87) while using approximately 47% fewer parameters. An ablation study and a scenario-based evaluation further show that the difficulty is concentrated almost entirely in distinguishing generalised harassment from non-harassment content.

General Terms

Natural Language Processing, Deep Learning, Text Classification, Hate Speech Detection

Keywords

Cyberbullying Detection, Transformer Models, RoBERTa, DistilBERT, Multi-class Classification, Social Media Analysis, Fine-tuning

1. INTRODUCTION

Social media platforms have seen a significant increase in user-generated content, which has been a major contributor to the growth of cyberbullying. Cyberbullying is a form of bullying that is intentional and happens repeatedly through the use of technology. Unlike traditional forms of bullying, cyberbullying is digital, leaves a lasting digital footprint, and allows the perpetrator to hide behind the anonymity of the internet. This has made automated cyberbullying detection a notable and growing research area in Natural Language Processing (NLP).

The first attempts at cyberbullying detection relied heavily on feature engineering. Reynolds, Kontostathis and Edwards [1] demonstrated that a J48 decision tree could classify bullying with 78.5% accuracy. User context was shown to matter by Dadvar et al. [2], where incorporating a user's gender improved precision by 39%.

A fundamental limitation persists in state-of-the-art cyberbullying studies: experiments are typically restricted to five explicit, attribute-specific categories, leaving out the more complex, generalised harassment. This research addresses that gap by studying the performance of two Transformer models, RoBERTa and DistilBERT, on a challenging six-class classification problem that includes the ambiguous Other_Cyberbullying class, using the dataset described in [3].

2. DATA COLLECTION AND PREPARATION

The dataset was gathered from Kaggle's Cyberbullying Classification Dataset [3] for social media posts, comprising posts divided into six categories: age, ethnicity, gender, religion, other_cyberbullying, and not_cyberbullying. The preprocessing workflow was as follows:

(1) The dataset was downloaded from Kaggle [3] and inspected for column data types and distribution of classes.

(2) The target column cyberbullying_type was converted from string data to integers to prepare for supervised learning.

(3) The dataset was divided into 80% training and 20% testing using stratified sampling to maintain class balance.

(4) Each text was tokenised using the model-specific pre-trained tokenizer, converting text into subword tokens.

(5) Tokenised sequences were padded and truncated to a maximum of 128 tokens.

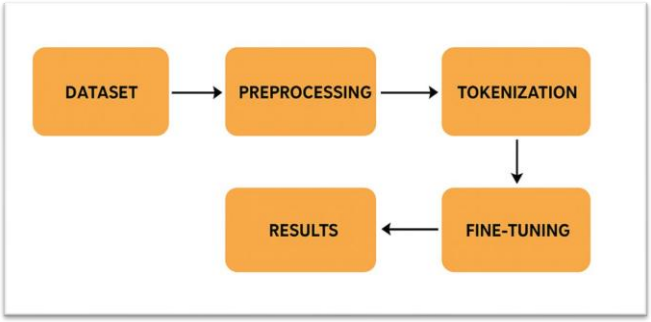

Fig. 1. Working flow of the proposed cyberbullying detection system.

3. LITERATURE REVIEW

Reynolds, Kontostathis, and Edwards [1] trained classifiers on datasets from Formspring.me. Their 296-word insult lexicon with normalised count features and J48 decision tree achieved 78.5% true positive accuracy, highlighting the importance of positive instance weighting for imbalanced datasets.

Dadvar et al. [2] analysed 4,500 YouTube comments using an SVM classifier. Incorporating gender information improved precision by 39% (0.31 to 0.43) and F-measure by 15%, demonstrating the value of author attributes in harassment detection.

Wang et al. [4] proposed a hybrid framework using Word2Vec embeddings, Graph Convolutional Networks (GCN), and XGBoost, achieving accuracy of 0.86.

Aldhyani et al. [5] repurposed the Sentiment140 dataset for attribute-specific cyberbullying detection using 47,000 tweets. A

Naive Bayes classifier achieved 85% accuracy, while a Bi-LSTM model improved this to 93%.

Orelaja and Akinola [6] built binary and multiclass datasets from Wikipedia Talk pages and Twitter. A standalone Bi-LSTM achieved 94.1% on binary and 99% on the five-class multiclass dataset.

Patil et al. [7] developed an enhanced depression detection system on social media using advanced machine learning and linguistic analysis, demonstrating that combining text-derived linguistic features with supervised classifiers substantially improves detection of at-risk posts. Kacha et al. [8] evaluated visual motion and multimodal strategies for depression recognition. Patil et al. [9] surveyed Transformer model dominance and multimodal approaches for depression detection, noting that pre-trained language models substantially outperform classical machine learning on nuanced affect classification.

Patil and Sheth [10] applied deep learning and machine learning to classify Big Five personality traits from text, while their survey [11] compared feature engineering and representation learning approaches, concluding that contextual embeddings consistently outperform hand-crafted lexical features. Patil and Sheth [12] addressed data preparation and quality challenges for personality recognition in Indian languages. Early work by Patil and Godhwani [13] on Named Entity Recognition in Marathi demonstrated that language-specific adaptation is needed when extending NLP pipelines to low-resource settings.

Verma et al. [14] integrated NLP with Conditional GANs to generate customised synthetic training scenarios, illustrating how generative techniques can address data scarcity. Pandya et al. [15] evaluated text-based mining and sentiment analysis for social media. Pandya et al. [16] demonstrated the effectiveness of n-gram features for text classification. Pandya et al. [17] investigated diagnostic criteria for depression using static and dynamic visual features.

4. METHODOLOGY

4.1 Dataset

This research uses the Cyberbullying Classification Dataset from Kaggle [3], containing social media posts categorised into six classes: age, ethnicity, gender, religion, other_cyberbullying, and not_cyberbullying. Each class comprised approximately 7,000–8,000 instances, yielding a balanced multi-class dataset split 80% training / 20% testing via stratified sampling.

4.2 Text Preprocessing

- Label Encoding: The cyberbullying_type column was converted from categorical labels to integer values for supervised learning.
- Tokenisation: Texts were tokenised using model-tied pre-trained tokenizers, the DistilBERT tokenizer for DistilBERT and the RoBERTa tokenizer for RoBERTa.
- Padding and Truncation: Sequences were padded or truncated to a uniform length of 128 tokens.
- Attention Masks: Masks were generated to distinguish actual tokens from padding tokens.

4.3 Model Architectures

4.3.1 RoBERTa (roberta-base)

RoBERTa (Robustly Optimised BERT Pre-training Approach) improves upon BERT by removing the next-sentence prediction objective and training on larger datasets. Specifications: 12 Transformer Layers, 768 Hidden Dimensions, 12 Attention Heads, approximately 125 Million Parameters. A six-class classification head (fully connected layer with softmax) was appended to the [CLS] token representation.

4.3.2 DistilBERT (distilbert-base-uncased)

DistilBERT is a distilled version of BERT retaining approximately 97% performance with significantly fewer parameters: 6 Transformer Layers, approximately 66 Million Parameters. An analogous six-class classification head was applied.

4.4 Fine-Tuning Configuration

Training used the Hugging Face Transformers Trainer API. Evaluation was performed at the end of each epoch to monitor convergence and avoid overfitting. The fine-tuning hyperparameters common to both models are summarised in Table 1.

Table 1. Fine-tuning hyperparameters for RoBERTa and DistilBERT

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	16
Epochs	4
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	500
Max Sequence Length	128
Loss Function	Cross-Entropy

4.5 Evaluation Metrics

- Accuracy: Overall correctness of predictions across all classes.
- Precision: Proportion of correctly predicted positive instances per class.
- Recall: Ability to identify all actual positive instances.
- Weighted F1-Score: Harmonic mean of precision and recall, weighted by class support (primary metric).
- Confusion Matrix: Visualises classification performance and misclassification patterns.

4.6 Ablation Study

To isolate the impact of the other_cyberbullying class, an ablation study was conducted by removing this class, reducing the problem to five-class classification. Both models were independently retrained and evaluated on this reduced dataset.

4.7 Experimental Workflow

The end-to-end experimental workflow pipeline followed for both models is summarised in Table 2, and the corresponding architecture overview is shown in Figure 2.

Table 2. End-to-end experimental workflow pipeline

Step	Description
1. Data Ingestion	Load Kaggle dataset; inspect distribution
2. Label Encoding	Convert string labels to integer indices (0-5)
3. Tokenisation	Apply tokenizer; pad & truncate to 128 tokens
4. Dataset Split	Stratified 80/20 train-test split

Step	Description	Step	Description
5. Model Init	Load pre-trained weights; attach 6-class head	**7. Evaluation**	Compute accuracy, F1, precision, recall
6. Fine-Tuning	Train 4 epochs using AdamW with warmup	**8. Ablation**	Repeat on 5-class (no other_cyberbullying)

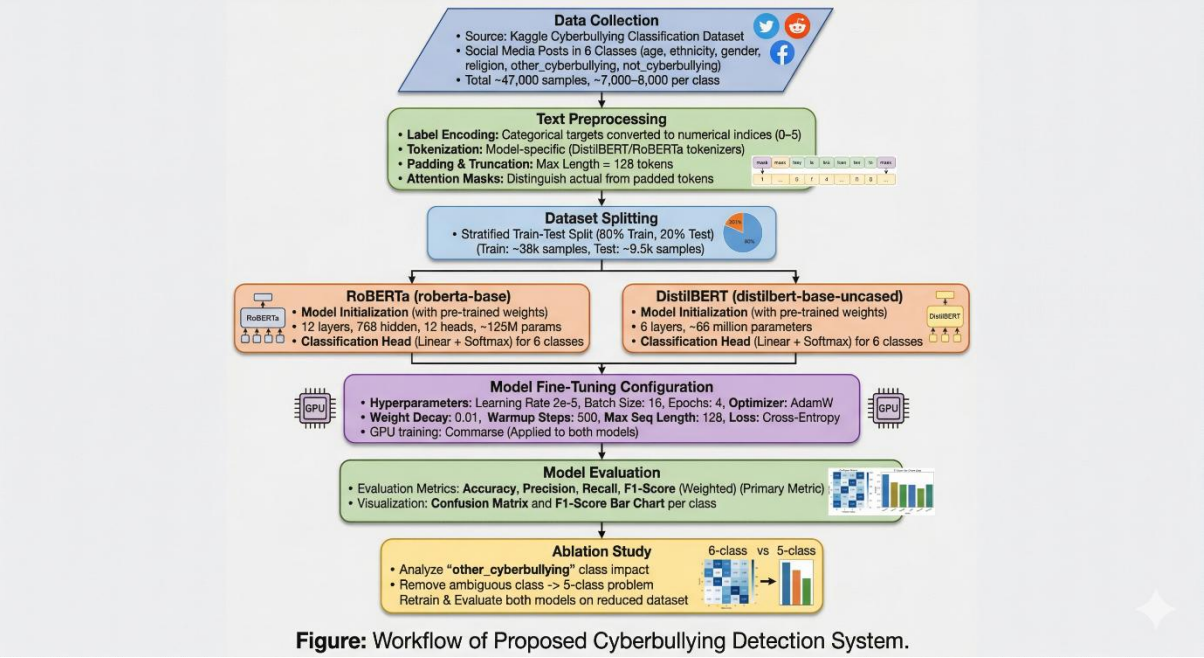

Fig. 2. Architecture overview and experimental results summary.

5. RESULTS

The complexity of the dataset significantly influenced performance compared to prior literature. Aldhyani et al. [5] reported 99% accuracy and Orelaja and Akinola [6] achieved 93% using Bi-LSTM; however, both were restricted to five explicit, attribute-specific classes. By introducing other_cyberbullying, representing generalised harassment without specific attribute keywords, this research creates a substantially harder problem.

Six-class Results: RoBERTa achieved a test accuracy of 87.79% and a weighted F1-score of 0.88, with near-perfect F1-scores for Age (0.99), Ethnicity (0.98), and Religion (0.97). DistilBERT achieved 87.19% accuracy and a weighted F1-score of 0.87, matching RoBERTa on Age (F1 = 0.99). Both models struggled with not_cyberbullying and other_cyberbullying due to their ambiguous, keyword-sparse nature.

5.1 Overall Performance

Ablation Study (Five-class): Removing other_cyberbullying substantially improved performance: both RoBERTa and DistilBERT achieved 95% weighted F1, confirming that generalised harassment is the primary driver of difficulty. The classification reports are shown in Figures 3–6.

Table 3. Classification Report – DistilBERT (with other_cyberbullying)

CATEGORY	PRECISION	RECALL	F1 SCORE	SUPPORT
age	0.99	0.98	0.99	1598
ethnicity	0.98	0.98	0.98	1592
gender	0.89	0.91	0.9	1595
not_cyberbullying	0.72	0.61	0.66	1589
other_cyberbullying	0.68	0.78	0.73	1565
religion	0.96	0.97	0.97	1600
accuracy	N/A	N/A	0.87	9539
macro avg	0.87	0.87	0.87	9539
weighted avg	0.87	0.87	0.87	9539

Table 4. Classification Report – RoBERTa (with other_cyberbullying)

CATEGORY	PRECISION	RECALL	F1-SCORE	SUPPORT
age	0.99	0.98	0.99	1598
ethnicity	0.98	0.97	0.98	1592
gender	0.89	0.92	0.9	1595
not_cyberbullying	0.76	0.6	0.67	1589
other_cyberbullying	0.69	0.82	0.75	1565
religion	0.96	0.97	0.97	1600
accuracy	N/A	N/A	0.88	9539
macro avg	0.88	0.88	0.88	9539
weighted avg	0.88	0.88	0.88	9539

Table 5. Classification Report – DistilBERT (without other_cyberbullying)

METRIC	PRECISION	RECALL	F1-SCORE	SUPPORT
age	0.99	0.98	0.99	1598
ethnicity	0.99	0.98	0.98	1592
gender	0.95	0.92	0.93	1595
not_cyberbullying	0.89	0.9	0.89	1589
religion	0.95	0.98	0.97	1600
accuracy	N/A	N/A	N/A	7974
macro avg	0.95	0.95	0.95	7974
weighted avg	0.95	0.95	0.95	7974

Table 6. Classification Report – RoBERTa (without other_cyberbullying)

METRIC	PRECISION	RECALL	F1-SCORE	SUPPORT
age	0.99	0.98	0.99	1598
ethnicity	0.99	0.98	0.98	1592
gender	0.95	0.92	0.93	1595
not_cyberbullying	0.89	0.9	0.89	1589
religion	0.95	0.98	0.97	1600
accuracy	N/A	N/A	N/A	7974
macro avg	0.95	0.95	0.95	7974
weighted avg	0.95	0.95	0.95	7974

5.3 Computational Comparison

RoBERTa (approximately 125 million parameters, 12 transformer layers) and DistilBERT (approximately 66 million parameters, 6 transformer layers) were fine-tuned under an identical configuration (Table 1). DistilBERT has roughly 47% fewer parameters and half the transformer layers of RoBERTa, yet on the six-class task it reached 87.19% accuracy compared with RoBERTa's 87.79%, a gap of only 0.6 percentage points. On the Age class specifically, DistilBERT matched RoBERTa's F1-score of 0.99. The training loss curves in Figures 9 and 10 show both models converging to a similar loss range (approximately 0.15–0.2) within the first training epoch.

5.4 Error and Misclassification Analysis

The confusion matrices in Figures 7 and 8 reveal that errors are heavily concentrated in a single off-diagonal block: confusion between not_cyberbullying and other_cyberbullying. For DistilBERT, 466 of 1589 not_cyberbullying instances were misclassified as other_cyberbullying, and 255 of 1565 other_cyberbullying instances were misclassified as not_cyberbullying. For RoBERTa, the corresponding counts were

512 of 1589 and 148 of 1565 respectively. The identity-attribute classes (age, ethnicity, religion) each have fewer than 20 misclassified instances out of approximately 1,600.

This pattern explains why not_cyberbullying (F1 = 0.66–0.67) and other_cyberbullying (F1 = 0.73–0.75) have substantially lower F1-scores than the identity-attribute classes (0.97–0.99). The two classes are defined by the absence of an explicit, attribute-linked slur or identity term rather than by its presence. The gender class occupies an intermediate position (F1 of 0.90 for both models), with 119–135 instances confused mainly with not_cyberbullying and other_cyberbullying.

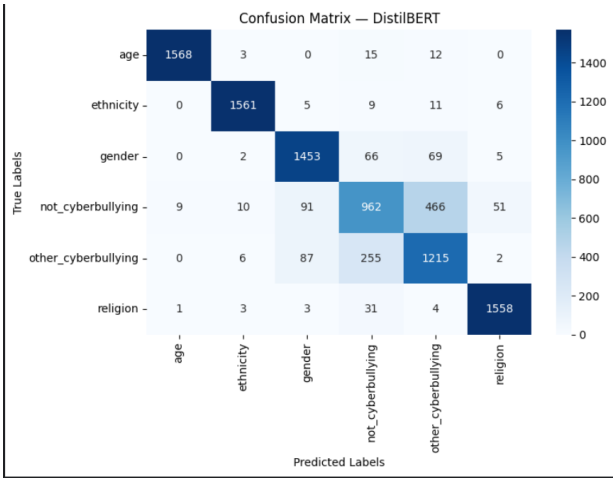

Fig. 3. Confusion Matrix – DistilBERT.

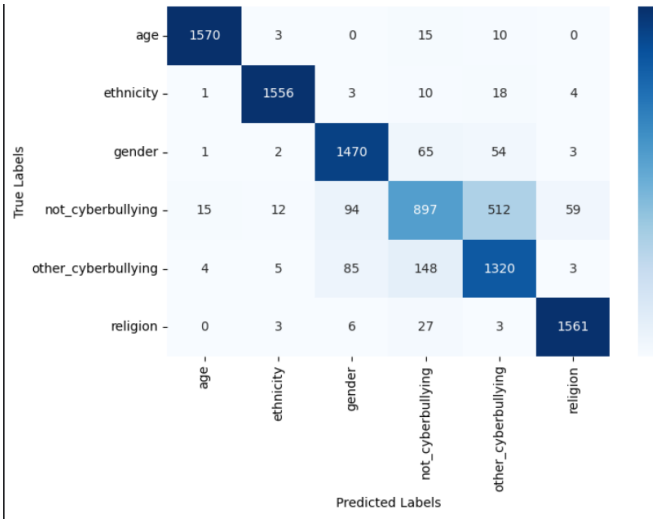

Fig. 4. Confusion Matrix – RoBERTa.

5.5 Why RoBERTa Outperforms DistilBERT

RoBERTa's advantage over DistilBERT is small in aggregate (87.79% vs. 87.19%) but is concentrated in the two hardest classes: RoBERTa's F1-score for not_cyberbullying (0.67) and other_cyberbullying (0.75) both exceed DistilBERT's (0.66 and 0.73). RoBERTa's larger hidden representation (768 dimensions across 12 layers, pretrained on a larger corpus) appears to provide a marginal but consistent benefit specifically for the classes that depend on diffuse contextual cues rather than lexical markers. On the classes dominated by explicit lexical cues (age, ethnicity, religion), the two models are essentially tied.

5.6 Scenario-Based Evaluation

Scenario 1 – Identity-Based Bullying (Age, Ethnicity, Gender, Religion): Both models achieve F1-scores between 0.90 and 0.99 across these four classes, with macro-averaged F1 of approximately 0.96 for RoBERTa and 0.95 for DistilBERT. Misclassifications within this scenario are rare (fewer than 90 instances per class out of approximately 1,600). This scenario represents the case where both models are effectively production-ready.

Scenario 2 – Generalised Harassment (other_cyberbullying only): F1-scores drop to 0.73 (RoBERTa) and 0.75 (DistilBERT). Roughly 16–19% of true other_cyberbullying instances are misclassified as not_cyberbullying, indicating that the model frequently treats generalised harassment as benign content when no attribute-specific cue is present.

Scenario 3 – Ambiguous / Borderline Content (not_cyberbullying): F1-scores of 0.66–0.67, driven by a 29–32% misclassification rate toward other_cyberbullying. The ablation study shows that when this boundary is removed entirely (five-class setting), both models recover to approximately 95% weighted F1.

Fig. 5. Training Steps – DistilBERT.

Fig. 6. Training Steps – RoBERTa.

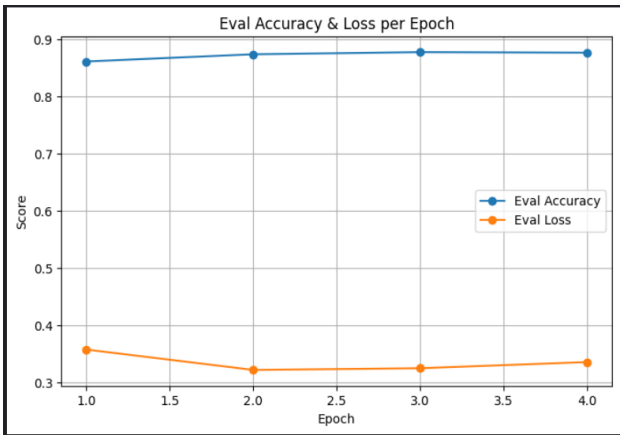

Fig. 7. Evaluation Accuracy and Loss per Epoch – RoBERTa.

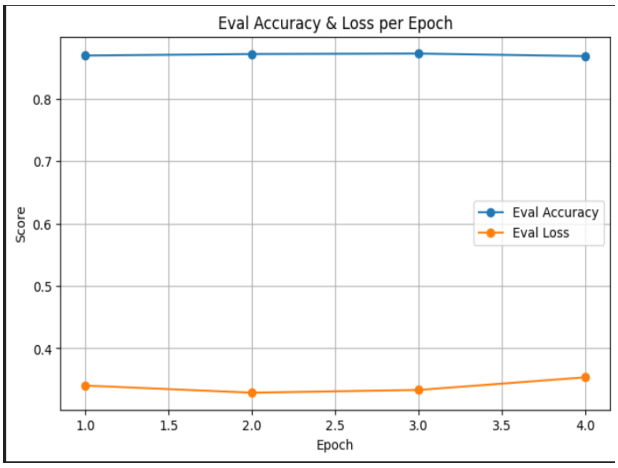

Fig. 8 Evaluation Accuracy and Loss per Epoch – DistilBERT

5.7 Comparison with Prior Work

Table 3 situates the present results against prior cyberbullying detection studies. Direct numerical comparison must be interpreted with caution, since the studies differ in dataset size, class count, and class definitions; nevertheless, the comparison highlights the trade-off between class granularity and reported accuracy.

6. CONCLUSION

This study comprehensively evaluated RoBERTa and DistilBERT on multiclass cyberbullying detection using a challenging six-class dataset that included a novel, ambiguous generalised harassment class. The evaluation encompassed a comparison against prior studies, a computational comparison, a confusion-matrix-based error analysis, and a scenario-based evaluation.

RoBERTa achieved 87.79% accuracy (weighted F1 = 0.88) and DistilBERT 87.19% (weighted F1 = 0.87), with both models achieving near-perfect F1 scores (0.97–0.99) for identity-specific classes (Age, Ethnicity, Religion). DistilBERT achieved this with approximately 47% fewer parameters than RoBERTa, indicating that for identity-attribute cyberbullying, model compression incurs little practical cost.

The ablation study confirmed that the other_cyberbullying class is the primary driver of performance reduction, with both models reaching approximately 95% weighted F1 on the five-class subset, consistent with results reported in prior Bi-LSTM-based studies. While Bi-LSTM models remain competitive for keyword-dominant hate speech, Transformers offer superior semantic nuance for broader content moderation. Future work should focus on improving detection of generalised harassment through larger contextual windows, ensemble methods, targeted pre-training on social media corpora, and cross-dataset evaluation to test generalisation.

7. ACKNOWLEDGMENTS

The authors acknowledge the support of PP Savani University, Kosamba, Surat, India, and the open-source community for providing the Hugging Face Transformers library and the Kaggle Cyberbullying Classification Dataset used in this research.

Table 7. Comparison with previous cyberbullying detection studies

Study	Model	Classes	Accuracy
Reynolds et al. [1]	J48 Decision Tree	Binary	78.5%
Dadvar et al. [2]	SVM + gender features	Binary	N/A (P=0.43)
Wang et al. [4]	Word2Vec+GCN+XGBoost	Binary	86.0%
Aldhyani et al. [5]	Bi-LSTM	5-class	99.0%
Orelaja & Akinola [6]	Bi-LSTM	5-class	93.0%

Study	Model	Classes	Accuracy
This work – 5-class ablation	RoBERTa / DistilBERT	5-class	95.0%
This work – 6-class full	RoBERTa	6-class	87.79%
This work – 6-class full	DistilBERT	6-class	87.19%

On the five-class subset, the Transformer models in this study (95% accuracy) sit between the figures reported by Orelaja and Akinola [6] (93%) and Aldhyani et al. [5] (99%), despite using a different dataset and class composition. The drop to approximately 87–88% once other_cyberbullying is reintroduced is not visible in any of the five-class baselines, since none of the compared studies model a generalised harassment category.

8. REFERENCES

[1] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 241–244, IEEE, 2011.

[2] M. Dadvar, F. M. de Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), pp. 23–25, Universiteit Gent, 2012.

[3] A. Bhatt, "Cyberbullying Classification Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification. [Accessed: 2024].

[4] J. Wang, K. Fu, and C. T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in 2020 IEEE International Conference on Big Data, pp. 1699–1708, IEEE, 2020.

[5] T. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, "Cyberbullying identification system based on deep learning algorithms," Electronics, vol. 11, no. 20, p. 3273, 2022.

[6] A. Orelaja and O. Akinola, "Attribute-specific cyberbullying detection using artificial intelligence," Journal of Electronic and Information Systems, vol. 6, no. 1, pp. 10–21, 2024.

[7] J. Patil, V. Patil, K. Prajapati, D. Patel, S. Trivedi, and R. Patel, "Enhanced depression detection on social media using advanced machine learning and linguistic analysis techniques," in International Conference on Intelligent Computing and Communication, pp. 263–275, Singapore: Springer Nature Singapore, 2024.

[8] H. Kacha, D. Bhikadiya, K. Sharma, and J. Patil, "Comparative analysis of visual motion and multimodal strategies in depression recognition," International Journal of Computer Applications, vol. 187, no. 58, pp. 58–64, 2025.

[9] J. Patil, K. Prajapati, D. Patel, R. Chauhan, and M. Patel, "A review of transforming AI for depression detection: Transformer model dominance, multimodal approaches, and future pathways," in International Conference on Computing and Machine Learning, pp. 87–106, Singapore: Springer Nature Singapore, 2025.

[10] J. Patil and J. Sheth, "Deep learning and machine learning approaches for the classification of personality traits," in Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2022, pp. 139–146, Singapore: Springer Nature Singapore, 2022.

[11] J. Patil and J. Sheth, "Comparative study of data sources, features, and approaches for automatic personality classification from text," International Journal of Computer Applications (IJCA), vol. 174, no. 10, 2021.

[12] J. Patil and J. Sheth, "Data preparation and quality challenges for the personality recognition in Indian languages using machine learning and deep learning approaches," Journal of IoT in Social, Mobile, Analytics, and Cloud, vol. 4, no. 1, pp. 33–40, 2022.

[13] M. J. A. Patil and M. P. B. Godhwani, "Review of name entity recognition in Marathi language," IJSART, vol. 2, no. 6, 2016.

[14] S. Verma, J. A. Patil, and I. Tamhankar, "Integrating natural language processing with CGANs to generate customized, realistic traffic scenarios for autonomous vehicle training," in 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE), pp. 1–5, IEEE, 2024.

[15] D. Pandya, A. Jadeja, M. A. Khan, S. B. Trivedi, M. A. Ramnath, and B. P. Satish, "Significance of sentiment analysis with text-based mining approach," in International Conference on Emerging Trends in Expert Applications & Security, pp. 315–323, Singapore: Springer Nature Singapore, 2024.

[16] D. D. Pandya, A. Jadeja, S. Degadwala, and D. Vyas, "Ensemble learning based enzyme family classification using n-gram feature," in 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1386–1392, IEEE, 2022.

[17] D. D. Pandya, A. Jadeja, S. Degadwala, and D. Vyas, "Diagnostic criteria for depression based on both static and dynamic visual features," in 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1–1, IEEE, 2023.