International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

Lightweight Transformer-Based Cyberbullying Detection for English, Malayalam, and Manglish Social Media Texts

Clive Lawrence Xavier[1], C Vishnu Mohan[2]
[1]Scholar, Department of Computer Science, Sacred Heart College Kochi, India.
[2]Assistant Professor, Department of Computer Science, Sacred Heart College Kochi, India.
Emails: *clivelawrence17@gmail.com[1], vishnumohan@shcollege.ac.in[2]*

Abstract
Detecting cyberbullying in multilingual and informal online environments remains a significant challenge, particularly for low-resource languages such as Malayalam and for mixed-script variants like Manglish. This work proposes a lightweight, deployment-oriented framework for cyberbullying detection that aims to classify English, Malayalam, and Manglish social media text with high accuracy and efficiency. The design centers on a single multilingual transformer encoder—IndicBERT v2—adapted using Low-Rank Adaptation (LoRA) to reduce computational requirements while preserving strong representational capacity. A combined dataset, envisioned to include publicly available internet text, existing corpora, and generative AI–augmented samples, is planned to support broad linguistic and contextual coverage. Minimal preprocessing and native SentencePiece tokenization are incorporated into the design to retain natural text characteristics across diverse languages. The proposed system is intended to output binary bullying predictions alongside interpretable indicators such as class-level confidence scores and attention-based token importance. Further optimizations, including INT8 quantization and ONNX/TFLite export, are outlined to facilitate efficient real-time use on resource-constrained devices. Overall, this work presents a scalable and practical model design for cyberbullying detection in linguistically diverse and code-mixed social media environments.
Keywords: *Cyberbullying detection; NLP; Malayalam; Manglish; IndicBERT v2*

1. Introduction
The rapid proliferation of social networking sites like Facebook, Instagram, Twitter (now X), TikTok, and Reddit has transformed how individuals connect, communicate, and share ideas. However, this digital transformation has also given rise to cyberbullying — the intentional and repeated use of digital communication to harm, intimidate, or humiliate others. Cyberbullying can manifest through direct insults, threats, exclusion, rumor-spreading, and sharing of harmful content. Research has consistently linked cyberbullying to severe psychological consequences, including depression, anxiety, selfharm tendencies, and in extreme cases, suicide. Unlike face-to-face bullying, cyberbullying transcends geographical boundaries and time constraints, allowing perpetrators to target victims at any time. The anonymity and lack of immediate consequences online further embolden aggressors. The volume of data generated on social media makes human moderation insufficient. As a result, automated detection systems powered by NLP have become critical to timely intervention and prevention. NLP-based cyberbullying detection leverages linguistic cues, semantic analysis, and contextual understanding to differentiate between benign conversation and harmful intent. Advances in ML and deep learning have significantly improved detection capabilities. However, the diversity of languages, slang, and cultural nuances presents persistent challenges. This paper reviews literature from 2013 to 2025, highlighting key developments, methodologies, datasets, and research gaps in this domain.

2. Related Works
2.1. Surveys and Taxonomies
Several surveys provide structured overviews of cyberbullying detection methodologies. Elsafoury et al. [1] categorize approaches into lexicon-based, classical machine learning, deep learning, and transfer learning, highlighting that lexicon-based methods are interpretable but limited in scalability, while deep learning methods achieve higher accuracy

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

but demand large annotated datasets. They further emphasize the role of linguistic, syntactic, semantic, and multimodal features, while noting challenges such as dataset imbalance and annotation inconsistencies. Mahmud et al. [3] extend these insights by focusing on low-resource and dialectal languages, discussing how transliteration, cross-lingual embeddings, and transfer learning can support detection in Hinglish, Manglish, or Arabic dialects. Their study stresses that detection systems trained exclusively on English underperform in multilingual contexts, leaving many global communities underprotected. Collectively, these surveys reveal both the strengths and limitations of mainstream approaches while underlining the need for taxonomies that incorporate multilingual, multimodal, and culturally adaptive perspectives.

2.2. Machine Learning and Deep Learning Approaches

Traditional machine learning techniques form the baseline for cyberbullying detection. Islam et al. [5] demonstrate the utility of classifiers such as Support Vector Machines (SVM) and Logistic Regression trained on TF-IDF features, showing competitive accuracy but vulnerability to noisy and informal social media text. Desai et al. [14] further explore Decision Trees and Random Forests, emphasizing that n-gram features and sentiment scores improve detection, though performance declines in heavily obfuscated or code-mixed contexts. Deep learning models provide significant advances in this regard. Teoh and Varathan [11] show that transfer learning architectures such as BERT outperform traditional approaches by capturing contextual semantics, though with higher computational requirements. Sayed et al. [17] extend this line of work by incorporating multi-class classification, differentiating abusive content based on categories like race, gender, and religion, thereby enabling more nuanced moderation. Similarly, Nikitha et al. [10] investigate bilingual settings, demonstrating that preprocessing strategies such as transliteration handling and multilingual embeddings are critical for detecting harmful content in languages like Hinglish. Together, these studies highlight the shift from handcrafted features toward context-aware deep models, with transformers offering state-of-the-art performance when resources permit.

2.3. Ensemble and Hybrid Models

Ensemble and hybrid methods have emerged as effective strategies for improving robustness. Kumar et al. [12] and Muneer et al. [13] show that stacking ensembles with transformer models enhance detection accuracy and reduce false negatives by leveraging complementary strengths of multiple classifiers. Perera and Fernando [6] propose thematic classification within ensemble frameworks, enabling differentiation between context-specific abuse such as political, cultural, or personal attacks, thereby improving interpretability and user trust. Pradheep et al. [9] advance multimodal detection by combining textual and image features, a crucial development for identifying harmful memes where abusive intent is both visual and textual. These ensemble and hybrid approaches consistently outperform individual models, though they introduce higher computational costs and reduced interpretability. Nevertheless, their ability to capture subtle, context-specific, and multimodal patterns makes them highly relevant for realworld deployment

2.4. Feature Engineering and Obfuscation Handling

Effective preprocessing and feature engineering remain indispensable, particularly given the prevalence of obfuscation in social media text. Zhang et al. [7] demonstrate that combining user behavior features, such as posting frequency and network centrality, with linguistic cues enhances early detection in Chinese social networks. Shekhar and Venkatesan [8] introduce a bag-of-phonetic-codes approach to address deliberate obfuscation, such as replacing letters with symbols or phonetic equivalents, thereby improving recall on noisy text. Similarly, Abdrakhmanov et al. [15] show that lexical and character-level features, including character n-grams, are effective against spelling variations and low-resource conditions, serving as preprocessing filters for downstream models. These studies collectively underscore that robust feature engineering—spanning behavioral signals, phonetic encodings, and character-level representations—remains a critical step in handling adversarial and noisy online environments.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

2.5. Real-Time, Cultural, and Ethical Perspectives

Research has increasingly shifted toward addressing practical, cultural, and ethical considerations in cyberbullying detection. Li et al. [2] investigate real-time detection challenges, showing that asynchronous message flows and random delays can significantly impair streaming-based moderation, and propose dynamic frameworks for improved responsiveness. Shah et al. [4] highlight the need for culturally adaptive detection, arguing that sensitivity thresholds should vary across regions to account for community norms and prevent over-flagging. Carter [16], from a social science perspective, emphasizes the psychological toll of cyberbullying on undergraduate students and stresses that detection systems must be paired with ethical, user-centered interventions. Collectively, these studies demonstrate that technological effectiveness alone is insufficient. Real-time responsiveness, cultural awareness, and ethical responsibility are equally vital in building trustworthy and impactful cyberbullying detection systems.

3. Methodology Trends

Across studies, common steps in the cyberbullying detection pipeline include:

- **Data Collection**: Using datasets from Twitter, Facebook, Instagram, YouTube comments, or custom crawlers.
- **Preprocessing**:Tokenization, stopword removal, lemmatization, handling slang/obfuscation.
- **Feature Extraction** — TF-IDF, word embeddings (Word2Vec, GloVe, BERT embeddings), sentiment features.
- **Model Training:** Traditional ML (SVM, Logistic Regression), deep learning (CNN, LSTM), transformer models (BERT, DistilBERT), or ensemble learning.
- **Evaluation:** Metrics include Accuracy, Precision, Recall, F1-score; cross-validation ensures robustness.

4. Challenges and Research Gaps

As summarized in Table I, existing approaches excel in specific domains but often lack generalization across platforms, languages, and cultural contexts. Despite progress, several challenges remain:

- **Sarcasm and Figurative Language:** Most models misclassify sarcastic remarks as non-bullying.
- **Low-Resource Languages:** Datasets for many regional languages are scarce.
- **Code-Mixed Content:** Social media often blends multiple languages in one post.
- **Multimodality:** Limited integration of images, videos, and text for holistic detection.
- **Real-Time Detection:** High computational cost of deep models hinders deployment.
- **Explainability:** Black-box models reduce trust among users and moderators.
- **Ethics and Privacy:** Risk of bias and misuse of detection systems, shown in Table 1.

Table I Summary of Reviewed Cyberbullying Detection Studies

Reference	Dataset	Method	Lang.	Remarks
[1] Elsafoury et al.	50+ public datasets	Survey of ML, DL, TL	Multi	Taxonomy of features/methods; notes data imbalance issues
[2] Li et al.	Simulated streams	Delay-compensation filtering	N/A	Useful for streaming moderation; not NLP-focused
[3] Mahmud et al.	Bengali, Arabic, Hinglish	Transfer learning	Low-resource	Effective for dialect/code-mix; lacks large datasets
[4] Shah et al.	Twitter, Instagram	ML + cultural	Multi	Adjusts detection by culture;

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

		adaptation		manual calibration needed
[5] Islam et al.	Twitter	TF-IDF + SVM, LR	Eng.	Good baseline accuracy; poor sarcasm detection
[6] Perera & Fernando	Facebook	Theme-specific ML	Eng.	Detects bullying themes; needs better generalization
[7] Zhang et al.	Weibo	Behavior + text features	Chinese	Social patterns boost detection; limited to China
[8] Shekhar & Venkatesan	Twitter	Bag-of-phonetic-codes	Eng.	Handles obfuscated profanity; fails on semantic attacks
[9] Pradheep et al.	FB/Twitter multimodal	Text + image fusion	Eng.	Captures abusive memes; high annotation cost
[10] Nikitha et al.	Social networks	NLP + ML bilingual	Eng./Hing.	Works on transliterated text; limited scope
[11] Teoh & Varathan	Public datasets	ML vs BERT, RoBERTa	Eng.	TL better than ML; high compute cost
[12] Kumar et al.	Twitter	Ensemble classifiers	Eng.	Robust detection; complex architecture
[13] Muneer et al.	Twitter/Facebook	Stacking + BERT	Eng.	High F1-score; needs large GPUs
[14] Desai et al.	Twitter/Facebook	NB, DT, RF	Eng.	Good with feature engineering; weak on noisy text
[15] Abdrakhmanov et al.	Twitter/Facebook	Offensive lang. ML	Eng.	Good pre-filter; no intent classification
[16] Carter	Survey	Questionnaire study	Eng.	Focuses on coping strategies; no automation
[17] Sayed et al.	Twitter/Reddit	NLP + multi-class ML	Eng.	Labels target group; needs larger dataset

5. Proposed Methodology

The proposed system is a lightweight, multilingual cyberbullying detection framework designed for English, Malayalam, and Manglish social media text. The approach emphasizes efficiency, compactness, and real-time deployability while maintaining strong classification performance. A single multilingual transformer encoder, IndicBERT v2, is fine-tuned using Low-Rank Adaptation (LoRA) to keep the model size small and training cost minimal.

5.1. Data Collection

The dataset is constructed from three primary sources:

- Publicly available internet text, including social media posts, comments, and discussion threads.
- Existing datasets from platforms such as Kaggle, containing English, Malayalam, and Romanized-Malayalam (Manglish) text.
- Synthetic and augmented samples generated using generative AI to ensure balanced representation of abusive and non-abusive content across all target languages.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

The compiled dataset includes direct insults, profanity, threats, harassment, sarcasm, informal slang, and code-mixed text. All samples are labeled as either cyberbullying or non-cyberbullying. A sample of the collected dataset, including English, Malayalam, and Manglish text instances, is shown in Sample entries from the multilingual dataset (English, Malayalam, and Manglish), Shown in Table 2.

Table 2 Sample Entries

Text	Label	Language
Sugamano	0	Manglish
Hope you are doing well	0	English
Shut up, you snake	1	English
ചത്തൂടെ നിനക്ക്	1	Malayalam
Ith kathum bro, all the best	0	Malayalam

5.2. Preprocessing

Preprocessing is intentionally kept minimal to preserve the natural characteristics of social media text. The following steps are applied:

- Basic cleaning (removal of URLs and excessive special characters)
- No transliteration of Manglish
- No handcrafted feature extraction
- Tokenization performed exclusively using the IndicBERT v2 SentencePiece tokenizer. This lightweight preprocessing ensures that the model directly learns from realistic multilingual and code-mixed patterns.

5.3. Model Architecture and LoRA Fine-Tuning

The system uses a single multilingual backbone, IndicBERT v2, which handles English, Malayalam, and Manglish text within one unified architecture. To keep the model size below deployment constraints, LoRA modules are used:

- The base model weights remain frozen.
- Only small LoRA low-rank matrices are trained.
- A simple classification layer predicts bullying vs. non-bullying.

This design maintains model compactness (final size < 50 MB after quantization) while achieving effective multilingual performance. As shown in Fig. 1, the proposed system follows a streamlined pipeline beginning with multilingual data collection, followed by minimal preprocessing and tokenization using the IndicBERT v2 tokenizer. The processed text is then passed through the IndicBERT v2 encoder with LoRA fine-tuning, after which the classification layer predicts whether the content is cyberbullying or non-cyberbullying. Finally, confidence scores and attention-based token importance are generated to provide lightweight explainability before deployment.

5.4. Explainability

To improve interpretability without increasing model size, the system incorporates lightweight, intrinsic explainability mechanisms. Two forms of explanation are provided:

- **Class Confidence Score:** Softmax probabilities for each class indicate prediction certainty.
- **Attention-Based Token Importance:** The top tokens with the highest attention scores in the final transformer layer are highlighted as the most influential for the prediction.

These explanation signals require no additional external models or computational overhead, Figure 1.

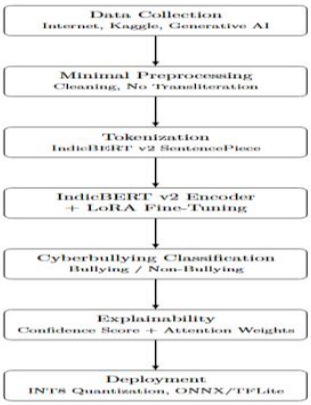

Figure 1 Pipeline of the Proposed Multilingual Cyberbullying Detection System.

5.5. Training and Optimization

The proposed model will be trained using cross-entropy loss with LoRA rank and learning rate optimized through validation. After fine-tuning:

- INT8 quantization is applied to reduce size and latency.
- The model is exported to ONNX/TFLite formats for efficient deployment on CPU or mobile

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4499-4505
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0661

devices.

This ensures fast and resource-efficient inference.

5.6. Evaluation

The system will be evaluated on stratified multilingual datasets covering English, Malayalam, Manglish, and code-mixed samples. Evaluation metrics include accuracy, precision, recall, F1-score, latency, memory footprint, and the clarity of attention-based explanations. The results will validate the model's ability to handle diverse linguistic patterns while remaining lightweight.

5.7. Research Priorities

Future enhancements will focus on:

- Expanding generative AI–based data augmentation for low-resource languages.
- Enhancing robustness to emerging slang and evolving abusive expressions.
- Improving the granularity of intrinsic interpretability methods.
- Exploring multimodal extensions combining text, emojis, and metadata.

Conclusion

A compact multilingual cyberbullying detection system design was developed using IndicBERT v2 with LoRA-based fine-tuning. The framework supports English, Malayalam, and Manglish text with minimal preprocessing and efficient inference. Lightweight interpretability is achieved through class confidence scores and attention-based token importance without increasing model complexity. Future work will expand linguistic coverage, improve explanation depth, and extend the system to multimodal environments.

References

[1]. Elsafoury, F., Pervez, Z., Katsigiannis, S., and Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. IEEE Access, 9, 106788–106817. https://doi.org/10.1109/ACCESS.2021.3098979

[2]. Li, Z., Zhang, H., Mu, D., and Guo, L. (2016). Random time delay effect on out-of-sequence measurements. IEEE Access, 4, 7862–7873. https://doi.org/10.1109/ACCESS.2016.2610098

[3]. Mahmud, T., Ptaszynski, M., Eronen, J., and Masui, F. (2023). Cyberbullying Detection for Low-resource Languages and Dialects: Review of the State of the Art. Feedback. arXiv preprint.

[4]. Shah, V., Sinha, A., Navalkar, N., Gupta, S., Gonsalves, P., and Malik, A. (2023). ML and natural language processing: Cyberbullying detection system for safer and culturally adaptive digital communities. Journal of Sensor and Internet of Things, 7(2), 115–126. https://doi.org/10.2478/jsiot-2023-0020

[5]. Islam, M. M., Uddin, M. A., Rahman, R., Akhter, A., and Acharjee, U. K. (2021). Cyberbullying detection on social media platform: Machine learning based approach. Jagannath University Journal of Science, 10(1), 67–77.

[6]. Perera, A., and Fernando, P. (2024). Cyberbullying detection system on social media using supervised machine learning. Procedia Computer Science, 234. https://doi.org/10.1016/j.procs.2024.06.200

[7]. Zhang, P., Gao, Y., and Chen, S. (n.d.). Detect Chinese cyber bullying by analyzing user behaviors and language patterns. Shanghai Jiao Tong University.

[8]. Shekhar, A., and Venkatesan, M. (n.d.). A bag-of-phonetic-codes model for cyber bullying detection in Twitter. National Institute of Technology, Karnataka.

[9]. Pradheep, T., Yogeshwaran, T., Sheeba, J. I., and Devaneyan, S. P. (2017). Automatic multimodel cyberbullying detection from social networks. Proceedings of the International Conference on Intelligent Computing Systems (ICICS 2017). Elsevier SSRN eLibrary – Journal of Information Systems and eBusiness Network.

[10]. Nikitha, G. S., Shenoy, A., Chaturya, K., Latha, J. C., and Shree, J. M. (2024). Detection of cyberbullying using NLP and machine learning in social networks for bi-language. International Journal of Scientific Research and Engineering Trends, 10(1). ISSN 2395-566X.

[11]. Teoh, H. T., and Varathan, K. D. (2023).

Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11, 49955–49968. https://doi.org/10.1109/ACCESS.2023.3275130

[12]. Kumar, Y. J. N., Vanapatla, R. R., Pinamoni, V. K., Kandukuri, J., Almusawi, M., Aravinda, K., Kansal, L., and Kalra, R. (2024). Detecting cyberbullying in social media using text analysis and ensemble techniques. E3S Web of Conferences, 507, 01069. https://doi.org/10.1051/e3sconf/202450701069

[13]. Muneer, A., Alwadain, A., Ragab, M. G., and Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. Information, 14(8), 467. https://doi.org/10.3390/info14080467

[14]. Desai, A., Kalaskar, S., Kumbhar, O., and Dhumal, R. (2021). Cyber bullying detection on social media using machine learning. ITM Web of Conferences, 40, 03038. https://doi.org/10.1051/itmconf/20214003038

[15]. Abdrakhmanov, R., Kenesbayev, S. M., Berkimbayev, K., Toikenov, G., Abdrashova, E., Alchinbayeva, O., and Ydyrys, A. (2024). Offensive language detection on social media using machine learning. International Journal of Advanced Computer Science and Applications, 15(5). https://doi.org/10.14569/IJACSA.2024.0150520

[16]. Carter, M. A. (2013). Protecting oneself from cyber bullying on social media sites – A study of undergraduate students. Procedia – Social and Behavioral Sciences, 93, 1229–1235. https://doi.org/10.1016/j.sbspro.2013.10.020

[17]. Sayed, F. R., Elnashar, E. H., and Omara, F. A. (2025). Cyberbullying detection in social media using natural language processing. Scientific African, 21, e02713. https://doi.org/10.1016/j.sciaf.2025.e02713