www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(4), 2026, 105-114. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i4.1528

A Comprehensive Comparative Study of Toxic Language Detection in Roman Urdu Using Hybrid Machine Learning, Deep Learning, and LoRA-Optimized Transformer Models

Abdul Jabbar [1], Muhammad Jawad Nadeem[2,✉], Muhammad Zain[3], Nisar Hussain[3], Amna Qasim[3], Momina Hafeez[3], and Grigori Sidorov[3]

[1] University of Okara, Punjab, Pakistan
[2] Univeristy of Agriculture Faisalabad, Punjab, Pakistan
[3] Instituto Politécnico Nacional, Centro de Investigación en Computación, (CIC-IPN), Mexico City, Mexico
Email address(es): abduljabbar17643@gmail.com, jawad.nadeem241@gmail.com, zainburaaq@gmail.com, nhussain2022@cic.ipn.mx, aqasim2300@alumno.ipn.mx, mominahafeez7@gmail.com, sidorov@cic.ipn.mx
✉ Corresponding author: Muhammad Jawad Nadeem
✉

Abstract. The rapid growth of social media has intensified the spread of toxic language, abusive comments, and harmful online discourse, creating serious challenges for digital safety and automated moderation. This study presents a comprehensive framework for toxic language detection in Roman Urdu using machine learning, deep learning, and transformer-based approaches. First, we do the preprocessing and normalization, then text features were extracted using TF-IDF and Count Vectorizer with unigram, bigram, trigram, and combined n-gram settings. The classification models included Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, AdaBoost, XGBoost, Convolutional Neural Network, Bidirectional Long Short-Term Memory, and LLaMA 3 fine-tuned with Low-Rank Adaptation. Experimental findings show that transformer-based fine-tuning produced the strongest results, with LLaMA 3 + LoRA achieving the highest F1-score of 96.78%, outperforming all baseline machine learning and deep learning models. Among conventional methods, XGBoost and Support Vector Machine demonstrated strong and stable performance, while CNN emerged as the best-performing deep learning baseline. The results confirm that parameter-efficient transformer adaptation is highly effective for toxic language detection in Roman Urdu and provides a scalable solution for content moderation in low-resource multilingual settings.

Keywords: Toxic language detection, Roman Urdu, machine learning, deep learning, LLaMA 3, LoRA.

Article information

Received: Jan 16, 2026
Accepted: Jul 31, 2026

1 Introduction

Online social platforms have transformed the way people communicate, share opinions, and participate in public discussion. However, the same platforms have also become major channels for toxic language, verbal abuse, harassment, hate, and aggressive interactions. Toxic language harms digital communities, discourages participation, affects psychological well-being, and increases the need for intelligent moderation systems capable of identifying harmful content automatically.

Toxic language detection has therefore become an important task in natural language processing. Considerable progress has been made in English and other high-resource languages using machine learning, deep learning, and transformer-based methods (Caselli et al., 2020). Yet, for low-resource and code-mixed languages such as Roman Urdu, toxic language detection remains significantly underexplored. Roman Urdu is widely used across YouTube, Facebook, WhatsApp, and other online platforms by South Asian users, but it lacks a standardized writing system. The same word may appear in multiple spellings, and users often mix Roman Urdu with English expressions, abbreviations, slang, and phonetic spellings. These characteristics make Roman Urdu challenging for automated text classification (Clarke et al., 2023).

Another major challenge is the informal nature of user-generated content. Toxic expressions are often written using short forms, repeated letters, misspellings, or indirect insulting patterns. In many cases, the toxicity is contextual rather than purely lexical. This makes it difficult for basic feature-based systems to detect harmful language reliably. As a result, there is a strong need for a comparative framework that evaluates both traditional and modern NLP models for toxic language detection in Roman Urdu (Ansari et al., 2024).

This study addresses that gap by presenting a comprehensive comparative analysis of machine learning, deep learning, and transformer-based models for Roman Urdu toxic language detection. A large annotated dataset was collected from YouTube comments and processed through a multi-stage pipeline involving cleaning, normalization, vectorization, model training, tuning, and performance evaluation (García-Díaz et al., 2023). Along with classical models such as Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine, this work also evaluates Decision Tree, AdaBoost, and XGBoost. For deep learning, CNN and Bi-LSTM were used, while LLaMA 3 was fine-tuned with Low-Rank Adaptation to provide a parameter-efficient transformer-based solution (Hussain et al., 2026).

The main objective of this study is to determine which modeling strategy is most effective for identifying toxic Roman Urdu comments. The study also aims to establish a strong baseline for future research in low-resource toxic language detection.

1.1. Main Contributions

The major contributions of this study are as follows:

1. A comprehensive Roman Urdu toxic language detection framework is presented using machine learning, deep learning, and transformer-based models.
2. A manually annotated YouTube comment dataset was developed for binary classification into toxic and non-toxic categories.
3. In addition to standard machine learning models, three extra classifiers were evaluated: Decision Tree, AdaBoost, and XGBoost.
4. LLaMA 3 was fine-tuned using LoRA to achieve high performance with reduced trainable parameter cost.
5. A comparative evaluation was conducted using Accuracy, Precision, Recall, and F1-score to determine the most effective modeling strategy for Roman Urdu toxic language detection.

2. Related Work

Toxic language and offensive language detection have received growing attention in recent years due to the widespread presence of abusive content on social media. Early studies mainly relied on machine learning algorithms such as Support Vector Machine, Naive Bayes, Logistic Regression, and Random Forest using lexical features, bag-of-words, and TF-IDF representations. These methods produced strong baseline results and proved effective for datasets where toxicity could be identified through prominent keywords and phrases (Al-Harigy et al., 2025).

With the rise of deep learning, models such as CNN, LSTM, and Bi-LSTM became common in harmful-language detection tasks. CNN-based architectures were particularly effective in extracting local phrase patterns, while recurrent models captured contextual dependencies across sequences (Zain et al., 2025). These models performed better than traditional machine learning approaches in cases where the meaning of toxicity depended on neighboring words and sentence-level context (Aklouche et al., 2024).

More recently, transformer-based models have significantly improved performance in abusive-language detection, hate speech detection, and cyberbullying classification. Pre-trained language models such as BERT and related variants demonstrated strong contextual understanding and improved generalization across noisy social media text (Naseeb et al., 2025). Parameter-efficient fine-tuning methods such as LoRA further increased the practicality of large language models by reducing the computational burden of training while maintaining strong classification performance (Althobaiti, 2022).

Despite these developments, Roman Urdu remains a low-resource language variety with limited benchmark studies (Hussain, Qasim, Mehak, Zain, Sidorov, et al., 2025). The lack of orthographic standardization, frequent code-mixing, informal spelling variation, and inconsistent grammar make Roman Urdu particularly difficult to process. Existing studies in Roman Urdu harmful-language detection have shown that both machine learning and deep learning models can provide promising performance, but transformer-based methods appear to offer superior contextual understanding. However, a wider comparison covering classical algorithms, neural architectures, and LoRA-optimized large language models for toxic language detection in Roman Urdu has remained limited (Hussain, Qasim, Mehak, Zain, Hafeez, et al., 2025).

This study extends the literature by presenting a broad comparative benchmark and showing how modern parameter-efficient transformer models outperform traditional approaches while still remaining practical for real-world deployment.

3. Proposed Methodology

3.1. Dataset Collection and Annotation

A large-scale dataset of Roman Urdu comments was collected from YouTube news and public discussion channels. YouTube was selected because it provides naturally occurring user-generated text containing diverse opinions, informal expressions, sarcasm, direct insults, and toxic remarks. The collected corpus consisted of 46,025 Roman Urdu comments, out of which 24,026 comments were labeled as toxic and 21,999 comments were labeled as non-toxic, which is balanced.

The annotation process was carried out manually by native Roman Urdu speakers to ensure contextual correctness. Each comment was reviewed according to predefined guidelines that categorized text as toxic when it contained abusive, insulting, hateful, degrading, threatening, humiliating, or strongly offensive language directed at individuals or groups. Comments that did not contain such harmful content were labeled as non-toxic.

To improve annotation quality, each comment was independently reviewed by three annotators. In cases of disagreement, final labeling was decided through majority voting and consensus discussion. The annotation consistency was measured using Cohen's Kappa, which yielded a score of 0.84, indicating strong agreement among annotators.

3.2. Data Preprocessing

Because Roman Urdu is highly inconsistent in writing style, a detailed preprocessing pipeline was applied before training the models. The following steps were performed:

1. **Punctuation Removal**. Unnecessary punctuation symbols such as commas, exclamation marks, question marks, and other special characters were removed to reduce noise.
2. **Extra Space Removal.** Irregular spacing in comments was normalized to ensure consistent tokenization.
3. **Digit Removal.** Digits and unrelated numeric patterns were removed unless they formed meaningful toxic expressions.
4. **Hyperlink Removal.** URLs and hyperlinks were removed because they did not contribute meaningful linguistic information for toxicity classification.
5. **Text Normalization.** Spelling variations, elongated characters, slang abbreviations, and alternate Roman Urdu forms were normalized. For example, repeated character sequences and informal spellings were reduced to their normalized forms.
6. **Stopword Filtering.** A custom Roman Urdu stopword list was applied. However, words carrying toxic meaning or contextual emphasis were preserved when necessary.

3.3. Feature Engineering

The following textual features were considered during experiments:

- Unigram,
- Bigram,
- Trigram,
- Combined Uni + Bi + Tri.

As values for these features we used TF and TF-IDF values, obtained using TF-IDF Vectorizer and Count Vectorizer.

These feature settings enabled a comparative analysis of individual word patterns, phrase-level combinations, and richer lexical context.

3.4. Classification Models

The following models were used in the experiments.

For traditional machine learning models, we used:

- Naive Bayes,
- Logistic Regression,
- Random Forest,
- Support Vector Machines,
- Decision Tree,
- AdaBoost,
- XGBoost.

For deep learning models we used:

- Convolutional Neural Network,

- Bidirectional Long Short-Term Memory.

Finally, for Transformer-Based Model we used LLaMA 3 fine-tuned with Low-Rank Adaptation.

3.5. Hyperparameter Tuning

Hyperparameter tuning was conducted to optimize model performance. Grid search was used for machine learning models, while deep learning models were optimized using validation-based tuning of learning rate, dropout, batch size, and number of epochs. LLaMA 3 was fine-tuned using LoRA to reduce training cost while adapting the model to the Roman Urdu toxicity classification task.

3.6. Evaluation Metrics

The models were evaluated using the following standard classification metrics:

- Accuracy,
- Precision,
- Recall,
- F1-score.

These metrics were selected to provide a balanced assessment of classification quality, especially because F1-score is particularly important in harmful-language detection tasks.

4. Proposed Framework

The proposed framework begins with Roman Urdu comment collection from YouTube, followed by annotation into toxic and non-toxic categories. The text is then cleaned and normalized through preprocessing. Next, lexical features are extracted for machine learning models, while sequence-based and transformer-based models are trained directly on tokenized text. Hyperparameter tuning is performed to optimize performance, and the trained models are finally evaluated using standard classification metrics. The complete workflow is illustrated in the methodology figure.

5. Results and Discussion

5.1. Overall Performance Comparison

Table 1 presents the overall performance of all models on Roman Urdu toxic language detection.

Table 1: *Overall model performance for toxic language detection in Roman Urdu.*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	82.03	82.14	81.72	81.90
Logistic Regression	91.63	91.84	91.26	91.52

Random Forest	89.44	89.72	88.95	89.28
Support Vector Machine	94.60	94.86	94.31	94.52
Decision Tree	86.24	86.43	85.97	86.11
AdaBoost	90.05	90.12	89.75	89.93
XGBoost	95.36	95.44	95.07	95.21
CNN	95.74	95.91	95.42	95.63
Bi-LSTM	93.17	93.28	92.81	93.04
LLaMA 3 + LoRA	**96.81**	**96.82**	**96.74**	**96.78**

The results clearly show that LLaMA 3 + LoRA achieved the best performance among all evaluated models, reaching an F1-score of 96.78%. This indicates that transformer-based contextual modeling is highly effective for Roman Urdu toxic language detection. The model successfully captured semantic context, implicit toxicity, and code-mixed language patterns better than classical and deep learning baselines.

Among machine learning models, XGBoost achieved the best performance with an F1-score of 95.21%, followed closely by Support Vector Machine with 94.52%. These results demonstrate that well-optimized machine learning models remain highly competitive for Roman Urdu text classification, particularly when combined with effective feature extraction.

Among deep learning architectures, CNN outperformed Bi-LSTM, achieving an F1-score of 95.63%. This suggests that local phrase-level toxic patterns are highly informative in Roman Urdu comments and that convolutional filters can capture them efficiently.

5.2. Performance of Machine Learning Models Across n-Gram Features

Table 2 presents the performance of machine learning models across different n-gram settings.

Table 2: *F1-score (%) of machine learning models across feature representations.*

Model	**Unigram**	**Bigram**	**Trigram**	**Uni + Bi + Tri**
Naive Bayes	79.84	80.92	74.63	81.90

Logistic Regression	90.76	87.11	82.54	91.52
Random Forest	88.31	84.72	79.65	89.28
Support Vector Machine	94.12	90.28	85.94	94.52
Decision Tree	84.75	81.43	76.92	86.11
AdaBoost	88.43	86.22	82.04	89.93
XGBoost	94.63	91.14	86.48	95.21

The results show that the combined Uni + Bi + Tri feature set consistently produced the strongest results for most machine learning models. Unigrams alone already provided strong performance, especially for SVM and XGBoost, indicating that many toxic comments can be identified through individual lexical cues. However, combining multiple n-gram levels improved discrimination by incorporating richer phrase-level patterns.

Trigram performance was generally lower than unigram and combined features. This suggests that overly specific phrase patterns introduced sparsity and reduced generalization, particularly in noisy Roman Urdu text.

5.3. Comparative Analysis

The overall performance pattern shows three clear trends.

First, transformer-based fine-tuning achieved the best results. LLaMA 3 + LoRA outperformed all other models because it captured deeper contextual information, handled code-mixed language more effectively, and modeled subtle toxic intent beyond keyword matching.

Second, CNN performed better than Bi-LSTM, which indicates that Roman Urdu toxicity often depends on short contextual phrases rather than long-distance sequential dependencies alone.

Third, machine learning models such as XGBoost and SVM remained highly competitive. Their strong performance suggests that Roman Urdu toxic language contains sufficiently distinctive lexical patterns that can still be modeled effectively using sparse text representations.

5.4. Error Analysis

Despite the strong results, some challenges remained. Misclassifications mainly occurred in the following situations:

- Indirect toxicity expressed through sarcasm,
- Code-mixed comments combining Roman Urdu and English,
- Comments with highly irregular spelling or elongated forms,
- Context-dependent phrases that appeared aggressive but were not truly toxic,

– Non-toxic emotional statements that included strong negative vocabulary.

Naive Bayes and Decision Tree produced more false positives because they relied heavily on surface lexical signals. Bi-LSTM sometimes failed in very short toxic comments where phrase-level cues were more important than sequence modeling. In contrast, LLaMA 3 + LoRA showed the most balanced behavior and reduced both false positives and false negatives.

6. Conclusion

This study presented a comprehensive comparative framework for toxic language detection in Roman Urdu using machine learning, deep learning, and transformer-based methods. A dataset of 46,025 Roman Urdu YouTube comments was collected and manually annotated into toxic and non-toxic classes. After preprocessing and feature extraction, ten different models were trained and evaluated.

The experimental findings showed that LLaMA 3 + LoRA achieved the best performance with an F1-score of 96.78%, followed by CNN, XGBoost, and Support Vector Machine. The results confirm that transformer-based parameter-efficient fine-tuning is highly effective for low-resource toxic language detection, especially in noisy and code-mixed social media environments such as Roman Urdu.

At the same time, the strong performance of XGBoost and SVM demonstrates that classical machine learning models remain useful baselines and can still offer robust classification quality when computational resources are limited. Overall, this study establishes a strong benchmark for future research in Roman Urdu toxic language detection and contributes toward the development of scalable automated moderation systems.

7. Future Work

Future research can extend this work in several directions. A multi-class toxicity framework could be developed to distinguish profanity, hate speech, harassment, threats, and identity-based abuse. Cross-lingual transfer learning could also be explored for Urdu, Roman Urdu, and English mixed-text environments. In addition, explainability techniques such as SHAP, attention analysis, and feature attribution can be applied to improve model transparency. Finally, lightweight deployment frameworks may be designed to support real-time moderation in social media platforms and mobile applications.

References

Aklouche, B., Bazine, Y., & Ghalia-Bououchma, Z. (2024). Offensive language and hate speech detection using transformers and ensemble learning approaches. *Computación y Sistemas, 28*(3), 1031–1039. https://doi.org/10.13053/CyS-28-3-4724

Al-Harigy, L. M., Al-Nuaim, H. A., Moradpoor, N., & Tan, Z. (2025). Towards a cyberbullying detection approach: Fine-tuned contrastive self-supervised learning for data augmentation. *International Journal of Data Science and Analytics, 19*, 469–490. https://doi.org/10.1007/s41060-024-00607-9

Almaliki, M., Almars, A. M., Gad, I., & Atlam, E.-S. (2023). ABMM: Arabic BERT-mini model for hate-speech detection on social media. *Electronics, 12*(4), Article 1048. https://doi.org/10.3390/electronics12041048

Althobaiti, M. J. (2022). BERT-based approach to Arabic hate speech and offensive language detection in Twitter: Exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications, 13*(5), 972–980. https://doi.org/10.14569/IJACSA.2022.01305109

Altinel, A. B., Baydogmus, G. K., Sahin, S., & Gurbuz, M. Z. (2024). So-haTRed: A novel hybrid system for Turkish hate speech detection in social media with ensemble deep learning improved by BERT and clustered-graph networks. *IEEE Access, 12*, 86252–86270. https://doi.org/10.1109/ACCESS.2024.3415350

Ansari, G., Kaur, P., & Saxena, C. (2024). Data augmentation for improving explainability of hate speech detection. *Arabian Journal for Science and Engineering, 49*, 3609–3621. https://doi.org/10.1007/s13369-023-08100-4

Asif, M., Al-Razgan, M., Ali, Y. A., & Yunrong, L. (2024). Graph convolution networks for social media trolls detection use deep feature extraction. *Journal of Cloud Computing, 13*, Article 33. https://doi.org/10.1186/s13677-024-00600-4

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). *HateBERT: Retraining BERT for abusive language detection in English* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2010.12472

Clarke, C., Hall, M., Mittal, G., Yu, Y., Sajeev, S., Mars, J., & Chen, M. (2023). *Rule by example: Harnessing logical rules for explainable hate speech detection* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.12935

García-Díaz, J. A., Pan, R., & Valencia-García, R. (2023). Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in Spanish and English. *Mathematics, 11*(24), Article 5004. https://doi.org/10.3390/math11245004

Hussain, N., Qasim, A., Liaquat, F., Mehak, G., Meque, A. G. M., Usman, M., Sidorov, G., & Gelbukh, A. (2026). Toward bias-aware and efficient offensive language detection using QLoRA-optimized LLaMA and GPT models. In L. Martínez-Villaseñor, R. A. Vázquez, & G. Ochoa-Ruiz (Eds.), *Advances in soft computing: MICAI 2025* (Lecture Notes in Computer Science, Vol. 16221, pp. 206–217). Springer. https://doi.org/10.1007/978-3-032-09037-9_17

Hussain, N., Qasim, A., Mehak, G., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2025). ORUD-Detect: A comprehensive approach to offensive language detection in Roman Urdu using hybrid machine learning–deep learning models with embedding techniques. *Information, 16*(2), Article 139. https://doi.org/10.3390/info16020139

Hussain, N., Qasim, A., Mehak, G., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2025). Hybrid machine learning and deep learning approaches for insult detection in Roman Urdu text. *AI, 6*(2), Article 33. https://doi.org/10.3390/ai6020033

Hussain, N., Qasim, A., Mehak, G., Zain, M., Hafeez, M., & Sidorov, G. (2025). *Fine-tuning large language models with QLoRA for offensive language detection in Roman Urdu-English code-mixed text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.03683

Hussain, N., Qasim, A., Mehak, G., Zain, M., Sidorov, G., Gelbukh, A., & Kolesnikova, O. (2025). Multi-level depression severity detection with deep transformers and enhanced machine learning techniques. *AI, 6*(7), Article 157. https://doi.org/10.3390/ai6070157

Hussain, N., Qasim, A., Zain, M., & Sidorov, G. (2025). Secure data provenance and semantic interoperability model for big data in IoT: A healthcare perspective. *Environment-Behaviour Proceedings Journal, 10*(SI42), 103–108. https://doi.org/10.21834/e-bpj.v10iSI42.7747

Naseeb, A., Zain, M., Hussain, N., Qasim, A., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). Machine learning- and deep learning-based multi-model system for hate speech detection on Facebook. *Algorithms, 18*(6), Article 331. https://doi.org/10.3390/a18060331

Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformer-based models. *Information, 16*(2), Article 114. https://doi.org/10.3390/info16020114

Shaheen, M., Awan, S. M., Hussain, N., & Gondal, Z. A. (2019). Sentiment analysis on mobile phone reviews using supervised learning techniques. *International Journal of Modern Education and Computer Science, 11*(7), 32–43. https://doi.org/10.5815/ijmecs.2019.07.04

Zain, M., Hussain, N., Qasim, A., Mehak, G., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). RU-OLD: A comprehensive analysis of offensive language detection in Roman Urdu using hybrid machine learning, deep learning, and transformer models. *Algorithms, 18*(7), Article 396. https://doi.org/10.3390/a18070396