Hate Speech Detection of Manglish (Malay + English) in X (Twitter) Using XLM-RoBERTa and XLNet

Farisha Binti Azmi[1], Normaisharah Mamat[2], Nur Azaliah Abu Bakar[3], Siti Maherah Hussin[4]

Faculty of Artificial Intelligence, Universiti Teknologi Malaysia
farisha98@graduate.utm.my , normaisharah@utm.my, azaliah@utm.my, sitimaherah@utm.my

Article history

Received:
11 November 2025

Received in revised form:
18 November 2025

Accepted:
1 December 2025

Published online:
26 December 2025

*Corresponding author
normaisharah@utm.my

Abstract

This study explores hate speech detection in Manglish, a code-mixed language of Malay and English widely used among Malaysian social media users. The main objective is to develop and evaluate deep learning-based models capable of identifying hate speech in Manglish tweets. A dataset of 9,241 manually annotated tweets was collected from X (formerly Twitter) and processed using the Malaya NLP library for code-mixing detection. Two state-of-the-art transformer-based models, which are XLM-RoBERTa and XLNet, were fine-tuned under both imbalanced and upsampled training conditions. Evaluation metrics, including precision, recall, F1-score, accuracy, and evaluation loss, were used to assess model performance. Results indicate that XLNet achieved the highest F1-score and fastest inference time under imbalanced conditions, while XLM-RoBERTa demonstrated stronger generalization with lower evaluation loss. After upsampling, both models improved significantly, achieving balanced performance across both classes. This research contributes a novel annotated Manglish dataset and highlights the importance of context-aware multilingual models for hate speech detection in code-mixed social media posts.

Keywords: *Hate speech detection, Manglish, XLM-RoBERTa, XLNet, deep learning, social media, natural language processing*

1. Introduction

The exponential growth of social-media platforms such as X (Twitter), Facebook, and TikTok has transformed online communication, enabling widespread expression and connectivity. However, this also intensifies the spread of hate speech messages that promote hostility or discrimination based on race, religion, gender, or nationality [1]. In multilingual societies like Malaysia, the challenge is compounded by Manglish, a colloquial mixture of Malay and English. The irregular syntax, informal spelling, and constant code-switching of Manglish hinder the performance of traditional Natural Language Processing (NLP) systems trained on monolingual data.

* Corresponding author. normaisharah@utm.my

Existing hate-speech detection systems primarily target English or standard Malay datasets [2]–[4]. Consequently, harmful mixed-language expressions often evade detection, contributing to the unchecked circulation of offensive discourse online.

This research improved previous studies by addressing the linguistic and contextual limitations that have hindered effective hate-speech detection in Malaysia's multilingual landscape. While prior works have primarily focused on standard Malay or English corpora using conventional or shallow transformer architectures, this study pioneers the application of XLM-RoBERTa and XLNet, two high-performing multilingual models that are specifically for Manglish. The research not only contributes a new manually annotated Manglish dataset but also provides an in-depth comparative evaluation of model generalization, runtime efficiency, and sensitivity to class imbalance, and thus, offers empirical insights into how advanced transformers can be adapted to detect nuanced hate speech in low-resource, culturally hybrid linguistic environments

2. Literature Review

Early studies on hate-speech detection predominantly relied on classical machine-learning algorithms such as Support Vector Machines (SVM) and Naïve Bayes, which used lexical and statistical features such as TF-IDF and n-grams [1]. Although effective in detecting explicit profanity, these models lacked contextual understanding and failed to interpret subtle or implicit hate expressions. Later, deep-learning models particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) introduced hierarchical feature learning that improved performance in text classification tasks [2]. However, these architectures struggled with long-range dependencies, language ambiguity, and multilingual code-switching common in social media data.

The introduction of transformer architectures revolutionized Natural Language Processing (NLP) by leveraging self-attention mechanisms [3]. This innovation allowed models to capture global dependencies and bidirectional context efficiently. Multilingual transformer variants, such as XLM-RoBERTa and XLNet, extended this paradigm by pretraining on massive multilingual corpora, enabling robust cross-lingual transfer [4], [5]. XLM-RoBERTa, trained on 2.5 TB of CommonCrawl data, demonstrated strong performance across more than 100 languages, while XLNet used permutation-based pretraining to capture bidirectional relationships without the limitations of masked token prediction.

Recent studies have explored the application of these multilingual transformers to hate-speech detection. Röttger et al. [6] introduced Multilingual HateCheck, a benchmark suite that tests hate-speech models across 10 languages, exposing model weaknesses in cross-lingual generalization. Similarly, Maity et al. [7] proposed a deep-learning framework for Malay hate speech using transformer-based embeddings, reporting improvements of up to 12% over traditional classifiers. Haber et al. [8] demonstrated the effectiveness of multilingual transformers for low-resource social media text from Singapore, while Ng et al. [9] developed

SGHateCheck, a diagnostic framework to evaluate functional reliability in multilingual hate-speech detection systems.

Other researchers have focused on cross-lingual adaptation and transfer learning to improve detection in under-resourced settings. El-Alami et al. [10] fine-tuned XLM-RoBERTa for multilingual offensive-language detection, achieving substantial gains in precision for low-resource languages. Rehman et al. [11] proposed a user-aware multilingual model that improved generalization across unseen datasets. In addition, Vardhan et al. [12] highlighted the importance of code-mixed embeddings for improving model performance in Hindi–English hate-speech classification, demonstrating that language-specific fine-tuning enhanced contextual representation.

Large-scale projects such as Multi3Hate [13] and HateSpeech2025 [14] addressed multilingual and multimodal challenges by integrating text, visual, and cultural cues, indicating that model robustness depends on exposure to diverse linguistic patterns. Yadav et al. [15] compared mBERT and XLM-RoBERTa on Hinglish data, showing that multilingual models handle informal syntax better than monolingual architectures.

Despite these advancements, there is a scarcity of work addressing Manglish, a unique code-mixed blend of Malay and English used in Malaysia's digital communication. The informal spelling, phonetic variability, and cultural semantics of Manglish pose significant challenges to existing multilingual models. Therefore, this research contributes by constructing a manually annotated Manglish hate-speech dataset and performing a comparative analysis of XLM-RoBERTa and XLNet in terms of generalization, runtime efficiency, and class-imbalance handling. This work fills a crucial gap in multilingual hate-speech detection by evaluating advanced transformer models on a previously unstudied, linguistically complex variant of Malay–English code-mixing.

3. Methodology

This research followed a systematic workflow encompassing data collection, preprocessing, manual annotation, exploratory analysis, model setup, training, and evaluation. The overall framework aimed to compare cross-lingual contextual understanding (XLM-RoBERTa) with autoregressive sequence modeling (XLNet) to determine which is better suited for detecting hate speech within a highly informal, code-mixed Malaysian corpus. Both models were trained under identical experimental conditions to ensure fair comparison.

3.1. Data Collection

Tweets were collected via TwitterAPI.io between April and May 2025. Due to the lack of geolocation metadata, the Malaya NLP library was used to detect tweets with both Malay and English components. After filtering for text length and removing duplicates, a final corpus of 9,241 Manglish tweets was curated.

3.2. Data Cleaning and Annotation

Data cleaning involved removing URLs, emojis, hashtags, and special characters using Pandas. Five bilingual annotators manually labeled tweets as hate speech or non-hate speech following strict guidelines adapted from [9]. Hate speech was defined as language inciting hostility based on identity (race, religion, or gender). Majority voting determined final labels, with a Fleiss' Kappa score of 0.644, indicating substantial inter-annotator agreement.

3.3. Exploratory Data Analysis

The dataset exhibited strong class imbalance (1,210 hate speech vs. 8,031 non-hate speech). Hate tweets were generally shorter and contained more identity-based slurs. Word clouds and frequency analysis highlighted the linguistic diversity of Manglish expressions. Custom stopwords were developed using Malaya's get_stopwords(), retaining negators (e.g., tak, tidak) and adding colloquial markers (lah, tu).

Figure 1 illustrates the distribution of character and word counts across the two classes, showing that hate-speech tweets are typically shorter and more abrupt, often containing direct insults or identity-based slurs. Non-hate tweets tend to be longer, expressing criticism or commentary in a more neutral tone.

To further understand the lexical patterns, Figure 2 presents word clouds for each class. It is evident that hate-speech tweets frequently include aggressive terms referencing ethnicity or religion, while non-hate tweets highlight general conversational words or neutral verbs. This visualization emphasizes that context and tone, rather than specific keywords alone, determine whether a tweet qualifies as hate speech.

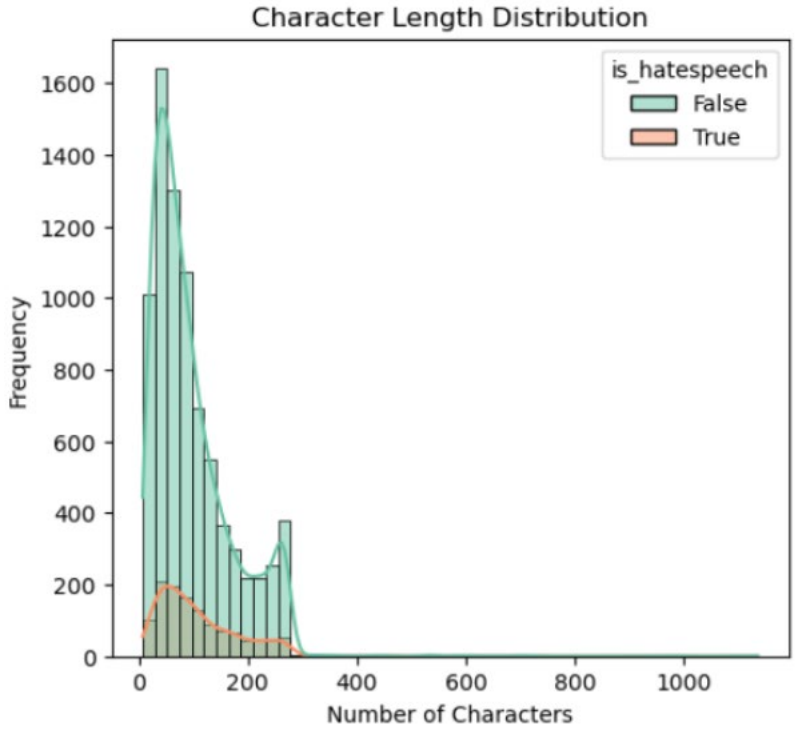

Figure 1. Spread in character and word counts in hate and non-hate speech

Figure 2. The common terms used in each class (hate and non-hate) visualized into word cloud

3.4. Embedding and Model Setup

Two multilingual transformer models were used:

- XLM-RoBERTa [8]: A robust cross-lingual model trained on massive multilingual data.
- XLNet [9]: An autoregressive model using permutation-based pretraining for improved contextual dependency.

Both models were fine-tuned for binary classification using their [CLS] output tokens. Training was conducted for four epochs with a batch size of 16 and an 80/20 stratified split between training and test sets. Two experimental setups were tested: (1) imbalanced training data, and (2) upsampled data using random duplication of the minority class.

3.5. Evaluation Metrics

Performance was measured using accuracy, precision, recall, F1-score, macro-F1, evaluation loss, and runtime (samples/sec). These metrics collectively evaluated prediction accuracy, minority-class sensitivity, and computational efficiency.

4. Results And Discussion

The comparative evaluation of XLM-RoBERTa and XLNet focused on their ability to generalize across imbalanced data and adapt after class balancing. Quantitative metrics were supplemented with runtime and loss analysis to determine each model's trade-off between accuracy and speed. The overall findings show that transformer architectures are effective even for noisy, mixed-language social-media text when combined with minimal preprocessing and fine-tuning.

4.1. Performance Without Class Balancing

When trained on imbalanced data, both models showed a noticeable bias toward the non-hate class. As displayed in Table 1, XLM-RoBERTa achieved the lowest evaluation loss (0.4173), indicating stronger generalization and robustness, while XLNet outperformed in inference efficiency, processing 271 samples/sec compared to XLM-RoBERTa's 193 samples/sec.

In Table 2, the F1-score for the hate-speech class remains low (0.40 for XLM-RoBERTa, 0.42 for XLNet), confirming the challenge of detecting minority-class samples under data imbalance. However, the marginally higher F1 for XLNet indicates slightly better adaptability to sparse class representation. Table 3 presents the macro performance across both classes, where XLNet recorded higher precision (0.72) but slightly lower overall accuracy (0.86).

These results suggest that XLNet's permutation-based training enhances contextual flexibility, enabling it to detect implicit hate speech despite limited examples. Conversely, XLM-RoBERTa's strength lies in stable generalization with lower evaluation loss, making it less prone to overfitting.

Table 1. Evaluation Performance Without Class Balancing

Model	Eval Loss	Runtime (s)	Samples/sec	Steps/sec
XLM-RoBERTa	0.4173	9.58	192.99	12.11
XLNet	0.4413	6.82	271.23	17.02

Table 2. F1-Scores by Class (Imbalanced)

Model	F1 Non-Hate	F1 Hate
XLM-RoBERTa	0.92	0.40
XLNet	0.92	0.42

Table 3. Macro Metrics (Imbalanced)

Model	Precision	Recall	F1-Score	Accuracy
XLM-RoBERTa	0.68	0.65	0.66	0.87
XLNet	0.72	0.64	0.67	0.86

4.2. Performance with Class Balancing (Upsampling)

After random upsampling, both models exhibited significant performance gains, as reflected in Table 4 and Table 5. Evaluation loss dropped substantially (to 0.2604 for XLM-RoBERTa and 0.2961 for XLNet), demonstrating the effectiveness of class rebalancing in improving minority-class learning. In Table 5, both models achieved near-perfect scores (Precision, Recall, and F1 ≈ 0.94), indicating excellent classification balance. XLM-RoBERTa maintained the lowest evaluation loss, suggesting more stable optimization, while XLNet achieved the fastest runtime even after balancing (191.6 samples/sec). The convergence of performance metrics between models underscores that data representation and class balance play a greater role than architectural differences once the dataset is normalized.

Table 4. Evaluation Performance with Upsampling

Model	Eval Loss	Runtime (s)	Samples/sec	Steps/sec
XLM-RoBERTa	0.2604	11.65	275.77	17.25
XLNet	0.2961	16.77	191.59	11.99

Table 5. Macro Metrics with Upsampling

Model	Precision	Recall	F1-Score	Accuracy
XLM-RoBERTa	0.94	0.93	0.93	0.93
XLNet	0.94	0.94	0.94	0.94

Both models achieved near-equivalent performance after upsampling, demonstrating the effectiveness of class rebalancing in improving model sensitivity. XLNet reached the highest F1-score (0.94), while XLM-RoBERTa maintained the lowest evaluation loss, confirming its cross-lingual robustness.

5. Conclusion and Recommendation

This study investigated the performance of XLM-RoBERTa and XLNet for hate speech detection in Manglish tweets. Without balancing, XLNet showed superior speed and recall, making it suitable for real-time applications. XLM-RoBERTa excelled in generalization and cross-lingual transfer, especially after data balancing. Both models significantly improved following upsampling, achieving over 0.93 F1-scores.

Future work will focus on mitigating potential overfitting by employing stratified sampling, increasing dataset diversity, and exploring hybrid architectures combining XLNet's contextual strength with XLM-RoBERTa's multilingual

capabilities. This research contributes valuable insights into multilingual hate speech detection and the linguistic complexities of Malaysian social media.

Acknowledgments

The authors would like to express their highest gratitude to the Universiti Teknologi Malaysia for providing financial support for this research through Potential Academic Staff Grant Q.K130000.2757.04K12.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

[1] A. M. Isa, S. Ahmad, and N. M. Diah, "Detecting offensive Malay language comments on YouTube using Support Vector Machine (SVM) and Naïve Bayes (NB) model," *J. Positive School Psychology*, vol. 6, no. 3, pp. 8548–8560, 2022.

[2] A. Albladi, A. Alhothali, and K. Moria, "Hate speech detection using large language models: A comprehensive review," *IEEE Access*, vol. 13, pp. 1–18, 2025.

[3] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[4] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," *arXiv preprint* arXiv:1911.02116, 2019.

[5] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5753–5763, 2019.

[6] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, and H. Margetts, "Multilingual HateCheck: Functional tests for multilingual hate-speech detection models," *Workshop on Online Abuse and Harms (WOAH)*, ACL, 2022.

[7] K. Maity, S. Bhattacharya, S. Saha, and M. Seera, "A deep learning framework for the detection of Malay hate speech," *IEEE Access*, vol. 11, pp. 79542–79552, 2023.

[8] J. Haber et al., "Improving the detection of multilingual online attacks with rich social media data from Singapore," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 12705–12721.

[9] R. C. Ng, N. Prakash, M. S. Hee, K. T. W. Choo, and R. K. W. Lee, "SGHateCheck: Functional tests for detecting hate speech in low-resource languages of Singapore," *arXiv preprint* arXiv:2405.01842, 2024.

[10] F.-Z. El-Alami, S. O. El Alaoui, and N. E. Nahnahi, "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6048–6056, 2022.

[11] M. Z. U. Rehman, S. Mehta, K. Singh, K. Kaushik, and N. Kumar, "User-aware multilingual abusive content detection in social media," *Inf. Process. Manag.*, vol. 60, no. 5, p. 103450, 2023.

[12] S. Vardhan, A. Kumar, and M. Yadav, "On importance of code-mixed embeddings for hate speech detection," *arXiv preprint* arXiv:2402.01842, 2024.

[13] "Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech," *arXiv preprint* arXiv:2408.01754, 2024.

[14] A. Albladi et al., "HateSpeech2025: A large multilingual benchmark for hate-speech detection across languages," *MDPI Appl. Sci.*, vol. 15, no. 4, pp. 1102–1120, 2025.

[15] A. K. Yadav, M. Kumar, A. Kumar, S. Kusum, and D. Yadav, "Hate-speech recognition in multilingual text: Hinglish documents," *Int. J. Inf. Technol.*, vol. 15, no. 3, pp. 1319–1331, 2023.