ComTech: Computer, Mathematics and Engineering Applications, 17(2), December 2026, 117−125 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v17i2.14618 E-ISSN: 2476-907X

Hate Speech Detection Leveraging Emoji Descriptions and BERT-Based Contextual Data Augmentation

Junita Amalia[1]*; Walker Valentinus Simanjuntak[2]; Ruth Marelisa Hutagalung[3]; Lamria Magdalena Tampubolon[4]

[1-4]Department of Information System, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, North Sumatra, Indonesia 22381
[1]junita.amalia@del.ac.id; [2]walker.simanjuntak@del.ac.id;
[3]iss21046@students.del.ac.id; [4]iss21055@students.del.ac.id

Received: 25th October 2025/ **Revised:** 27th February 2026/ **Accepted:** 25th March 2026

How to Cite: Amalia, J., Simanjuntak, W. V., Hutagalung, R. M., & Tampubolon, L. M. (2026). Hate speech detection leveraging emoji descriptions and BERT-based contextual data augmentation. *ComTech: Computer, Mathematics and Engineering Applications, 17*(2), 117−125. https://doi.org/10.21512/comtech.v17i2.14618

Abstract **-** This study evaluates the efficacy of emoji-to-text transformation and contextual data augmentation in improving hate speech detection on social media. Using a unified dataset from OLID, TweetEval, and the Sushil benchmark, we compared Multinomial Naive Bayes (MNB), Bidirectional LSTM (Bi-LSTM), and Multilingual BERT (mBERT). To address class imbalance and semantic loss, we implemented a preprocessing pipeline that converts emojis into textual descriptions and utilized a BERT-based contextual augmenter. Results indicate that textualizing emojis consistently improves performance by providing critical semantic anchors. The mBERT model achieved the highest performance with a Macro F1-score of 0.87 and a hate speech recall of 0.94, outperforming the Bi-LSTM (0.835) and MNB (0.83) baselines. Notably, the Bi-LSTM showed the highest sensitivity to the "Emoji Description" method, with a 5% performance increase compared to the "No Emoji" condition. These findings demonstrate that bridging the gap between visual symbols and textual intent significantly reduces model bias and improves the detection of nuanced toxicity. This research provides an optimized framework for context-aware moderation in imbalanced digital environments.

Keywords**:** hate speech detection, emoji description, mBERT, data augmentation, Bi-LSTM

I. INTRODUCTION

Social media has revolutionized how individuals share perspectives, yet the lack of oversight on these platforms has enabled hate speech to spread rapidly. This trend is dangerous, as it risks causing significant social harm and dividing communities. In today's globalized digital era, developing effective systems to detect and address offensive language is no longer optional; it is a fundamental requirement for protecting public space. Statista data shows that in the first half of 2024, the X platform received over 81 million reports related to harassment and bullying, over 66 million reports related to hateful content, and over 40 million reports concerning violent content (Dixon, 2025). Therefore, a hate speech classification model in English is needed as an initial identification step to prevent hate speech from spreading widely across social media. This also underscores the urgent need for automated systems capable of detecting hate speech with high contextual sensitivity. Hate speech, defined as language that expresses prejudice or incites discrimination against individuals or groups based on attributes such as race, religion, or gender, is a growing concern on social media platforms (Gelber, 2021). However, hate speech presents an ongoing challenge for both individuals and communities, especially due to user anonymity, unrestricted expression, and limited platform regulation.

Besides, social media users have increasingly adopted graphic symbols in their interactions to convey emotional expressions. Emojis should not be viewed merely as playful icons or decorative additions to digital text. Instead, they function as elements of a broader visual communication system with its own expressive rules. The meaning of any emoji is strongly shaped by the social and cultural background of the user, making interpretation highly context-dependent. For instance, the folded hands emoji (🙏) is often understood in Western contexts as a symbol of gratitude, prayer, or pleading. However, in several Asian cultures, the same emoji is more commonly

*Corresponding Author

interpreted as a gesture of respect, greeting, or acknowledgment. This variation highlights how emojis carry culturally embedded meanings that can shift significantly across communities. Such diversity in interpretation reinforces the idea that emojis are not universally understood symbols but culturally mediated expressions whose meanings may change depending on the communicative setting and the users involved (Salsabila et al., 2025, Telaumbanua et al., 2024).

Research titled "Emojis Trash or Treasure: Utilizing Emoji to Aid Hate Speech Detection" (Saikh et al., 2024) , attempted to address this by utilizing Emoji2Vec to process these symbols rather than deleting them. The model achieved an F1-score of 72.8% using text-only features; however, the performance decreased to 69.67% upon the integration of emoji2vec embeddings. However, two significant research gaps remain. First, that study lacked a robust data balancing strategy, justified by the perceived sufficiency of their English and Bengali dataset sizes. Second, while Emoji2Vec assigns static embeddings to emojis, it often fails to capture the fluid, contextual, and sarcastic ways emojis are used in hate speech. This research fills these gaps by proposing a novel baseline: converting emojis into their official textual descriptions. Unlike the static vector approach of Emoji2Vec, emoji descriptions transform visual symbols into natural language tokens, allowing the model to process them through the same contextual lens as the surrounding text. Following the principles of Simple Augmentation as described by Chai & Xie (2025), this study employs contextual text transformation to address data scarcity. By utilizing the nlpaug library with a BERT-base-uncased model, we generate synthetic variations of the minority hate speech class. This prevents the overfitting typically associated with small, imbalanced datasets and ensures that our model can adequately capture the complexities of hateful language and emoji-embedded contexts.

To further address the limitation, this study utilizes one of the traditional machine learning models, which is Naive Bayes as the baseline model. Research (Alaoui & Farhaoui, 2022) using Naive Bayes and tf-idf, reached the highest accuracy value of 93.06% but did not apply data augmentation to see wheter the model performed better. These studies use accuracy as the evaluation metric. From this gap, our research fills the gap by employing the F1-score as the primary evaluation metric while keeping track of the performance using classification reports, because, unlike accuracy, F1-score provides a robust measure of performance. Accuracy can be misleadingly high by over-representing the majority class. The F1 score, as the harmonic mean of precision and recall, ensures that the model is penalized for failing to identify the 'hate' or 'non-hate' class (Christen et al., 2023). Inspired by numerous studies (Amalia & Suyanto, 2024, Rawat et al., 2024; Amalia et al., 2025; Lazareva, 2023), the study will also investigate the role of emoji descriptions in improving RNN and Transformer-based models, focusing on Bi-LSTM and mBERT within English hate speech classification. BERT (Bidirectional Encoder Representations from Transformers) has been one of the most influential models, trained on massive corpora using masked language modeling to capture deep contextual semantics (Devlin et al., 2019).

II. METHODS

The methodology is categorized into the data processing workflow and the model evaluation. The first specifies the data lifecycle, tracing the path from initial dataset input through model training, and the lates is the final classification performance metrics. This study uses three benchmark datasets. Sushil ; A specialized dataset focusing on nuanced offensive content, including bilingual and emoji-heavy instances (Dalavi et al., 2023), TweetEval ; a multi-task benchmark derived from Twitter, specifically utilizing the hate speech identification subset (Barbieri et al., 2020), and OLID ; a widely recognized dataset for offensive language identification in social media, categorized by levels of aggression (Zampieri et al., 2019).

Because the source datasets utilized different annotation schemas, a unification strategy was required to transform multi-level labels into a binary format. OLID Level-A Consolidation data is categorized into OFF (Offensive) and NOT (Not Offensive). For this research binary classification is conducted, so label NOT was mapped to Class 0 (Normal), and label OFF was mapped to Class 1 (Hate). While OLID further breaks down offensive tweets into TIN (Targeted Insult and Threats) and UNT (Untargeted Profanity), this study consolidated both into the Hate class. This decision was made to ensure the model captures the broadest possible spectrum of harmful content, regardless of whether a specific target was identified. The labels from TweetEval and the Sushil dataset were similarly aligned to this binary schema. This process is summarized in Table 1 below.

This dataset will be utilized under two configurations, original text and text augmented with emoji descriptions. To prevent model bias and ensure experimental integrity, we removed duplicate tweets and purged entries with null values in the tweet field. A standardized mapping was applied to consolidate labels into Class 0 (Normal) and Class 1 (Hate). This included stripping whitespace and converting strings like 'hateful', 'hate', and '1' into a unified integer representation. To reduce the dimensionality of the feature space and remove non-informative noise characteristic of Twitter data, we applied a comprehensive cleaning function, noise remove user, remove url, remove Non-ASCII, remove punctuation and number, normalization, tokenization, and stopword filtering.

The core novelty of this research lies in the representation of emojis during text preprocessing. Rather than treating emojis as removable noise or representing them solely through static vector

Table 1 Dataset Label Mapping

Source	Original Label	Modified Label	Description
OLID	NOT	0	Non-offensive content
OLID	OFF (TIN/UNT)	1	Targeted or untargeted offensive language
TweetEval	0 non-hate	0	Neutral or benign tweets
TweetEval	1 hate	1	Explicitly hateful content
Suhsil	normal/non-hate	0	Benign text
Sushil	hateful/hate	1	Toxic content including emojis

Table 2 Text Transformation

Stage	Text
Original Input	@user @user I'm gunna play the shit out of it then bitch slap the both of you for talking smack 😘
Initial Cleaning	im gunna play shit bitch slap talking smack
Emoji Extraction	😘 (Count: 1)
Demojization	:face_blowing_a_kiss: hate
Emoji Description Cleaning	face blowing kiss
Text Combined	im gunna play shit bitch slap talking smack face blowing kiss

embeddings, this study converts emojis into natural language descriptions so that their semantic information can be incorporated into the textual representation. This approach is particularly relevant to social media data, where emojis can provide additional contextual or affective information that may complement the explicit words contained in a tweet. The emoji transformation process consists of extraction, demojization, and textual cleaning. Each emoji is first identified and converted into its corresponding descriptive phrase. The resulting description is then processed using the same normalization and cleaning procedures applied to the original textual content. The cleaned emoji description is finally concatenated with the cleaned tweet to produce the emoji-enriched representation. For example, an emoji such as 😘 is transformed into the phrase "face blowing kiss" and incorporated into the textual sequence rather than being discarded.

Any instances that became empty after preprocessing were removed from the dataset. Such instances may occur when a tweet contains only elements that are eliminated during cleaning, such as URLs, mentions, punctuation, or stopwords. Retaining these instances would provide no meaningful textual features for the classification models and could introduce uninformative training samples. Therefore, removing them ensures that each remaining instance contains textual information that can be processed by the subsequent feature extraction and classification stages. Examples of the complete text transformation process are presented in Table 2.

A significant challenge in hate speech detection is the inherent class imbalance within real-world datasets, where offensive content often represents a minority of the samples. In our initial unified dataset, the distribution of "Normal" (0) and "Hate" (1) tweets was notably skewed (as shown in Figure 1).

To mitigate this, this study implemented a data balancing strategy using the nlpaug library powered by a BERT-base-uncased model (Ma, 2019). Unlike conventional synonym replacement, the implemented approach uses the ContextualWordEmbsAug augmenter with the insert action. In this process, the augmenter inserts a new word into the existing text based on its surrounding linguistic context. The BERT-base-uncased model predicts a suitable word representation from the contextual information of the input sequence, allowing the generated sentence to preserve its general semantic structure while introducing lexical variation. This contextual approach is particularly useful for text classification because the generated samples are not produced solely from predefined synonym relationships, but from the linguistic context of each tweet. The augmentation process can therefore be represented as an input tweet being passed to the contextual word embedding augmenter, followed by contextual word prediction and word insertion, producing a new text sample with the original class label. By increasing the number of samples in the minority class, the resulting dataset

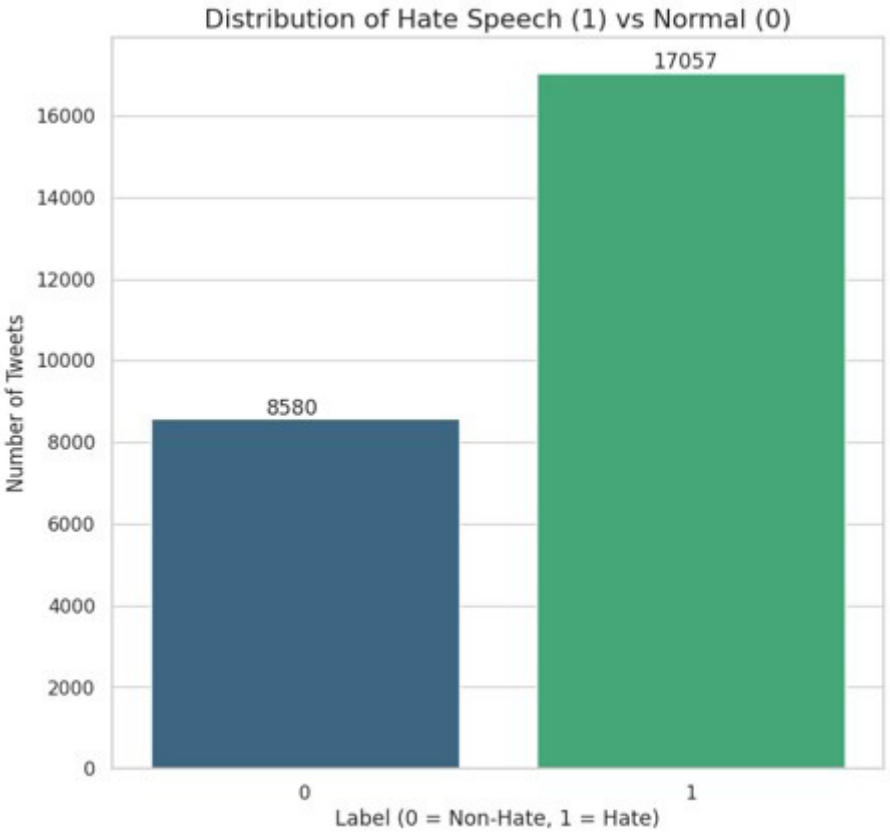


Figure 1 Class Imbalance

provides a more balanced representation between the two target classes.

This research evaluates three distinct architectures to identify the most effective approach for detecting hate speech integrated with emoji descriptions. The primary baseline for this study is the Multinomial Naive Bayes algorithm, a probabilistic classifier based on Bayes' Theorem. It is particularly effective for text classification when combined with TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, as it considers the frequency of word occurrences within a document (Safawi & Shafie, 2024). Term Frequency (TF) quantifies the prominence of a specific token within a document, operating on the premise that words appearing with higher regularity are more indicative of the document's core subject matter. The formula is in Equation (1).

$$TF(t,d) = \frac{ft,d}{\Sigma t' \epsilon d\, ft'd} \quad (1)$$

Conversely, Inverse Document Frequency (IDF) serves as a corrective mechanism that penalizes ubiquitous terms across the entire corpus while amplifying the significance of rare, highly descriptive words. This dual approach ensures that unique identifiers are prioritized over common linguistic fillers. The formula is in Equation (2).

$$idf(t,D) = \frac{N}{|\{d \epsilon D : t \epsilon d\}|} \quad (2)$$

The second model is Bi-LSTM. Bi-LSTM architecture extends the standard LSTM by employing a dual-layer approach that processes input sequences in both forward and backward directions. This bidirectional propagation enables the model to capture comprehensive contextual dependencies by integrating both preceding and succeeding information within a sequence. By learning long-term dependencies without redundant context retention (Handayani & Hasyim, 2024) , Bi-LSTM has established itself as a robust framework for sequential modeling and text classification tasks. Bi-LSTM architecture can be seen in Figure 2.

The primary architecture used in this study is Multilingual BERT (mBERT), specifically the bert-base-multilingual-cased variant. mBERT is a Transformer-based language model pre-trained on the top 104 languages with the largest Wikipedias. It utilizes a shared vocabulary across all languages, allowing it to capture cross-lingual semantics and common linguistic structures. The core novelty of mBERT lies in its Bidirectional Transformer architecture. Unlike the Bi-LSTM, which processes sequences step-by-step, mBERT utilizes a Self-Attention mechanism to weigh the importance of every word in a tweet simultaneously (Aminu et al., 2024, Putra & Wang,

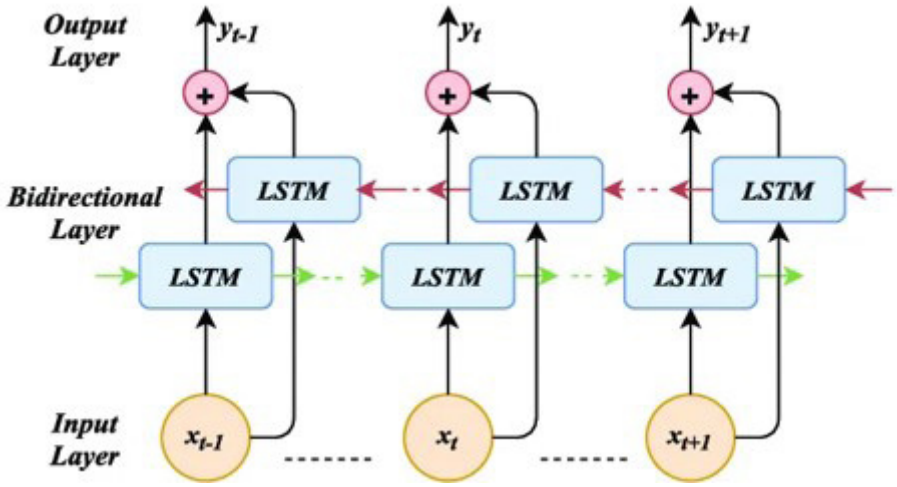


Figure 2 Bi-LSTM architecture
source: https://shorturl.at/dkdTa

2023). The relationship between words is calculated using Query (Q), Key (K), and Value (V) matrices. For a given input embedding, the attention score formula can be seen in Equation (3).

$$Attention(Q,K,V) = softmax(\frac{QKt}{\sqrt{dk}})V \quad (3)$$

Where d_k is the scaling factor (dimension of the keys). This allows the model to understand the context of an emoji description (e.g., "face blowing kiss") by relating it to specific aggressive keywords elsewhere in the tweet.

To rigorously assess the performance of the proposed models, this study utilizes a comprehensive Classification Report alongside the F1-score as the primary evaluation metric. The classification report provides a granular view of model performance by calculating Precision and Recall for each class. Precision measures the accuracy of the "Hate" predictions. The formula is in Equation (4).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall measures the model's ability to find all hateful content. The formula is in Equation (5).

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

The F1-score is the primary metric for this study because it provides a balance between Precision and Recall. This is especially important for the "Hate" class, as failing to detect hate speech (False Negative) and incorrectly flagging a normal tweet (False Positive) are both significant errors. The formula is in Equation (6).

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

To investigate the effect of emoji representation on hate speech classification, the experiments were organized based on two text representation configurations: Original Text and Emoji-Enriched Text. The Original Text configuration uses the cleaned tweet obtained after the standard preprocessing pipeline, without incorporating the semantic descriptions of emojis. In contrast, the Emoji-Enriched Text configuration uses the same cleaned tweet while retaining the semantic information conveyed by emojis through their corresponding natural language descriptions. This design allows the contribution of emoji-related information to be evaluated independently from the general text preprocessing procedure.

Both text representations were evaluated using the three classification architectures described previously: Multinomial Naive Bayes, Bi-LSTM, and mBERT. Consequently, each model was trained and evaluated under both representations to provide a controlled comparison between conventional textual information and textual information enriched with emoji descriptions. The experimental configurations therefore consist of six model-representation combinations: Multinomial Naive Bayes with Original Text, Multinomial Naive Bayes with Emoji-Enriched Text, Bi-LSTM with Original Text, Bi-LSTM with Emoji-Enriched Text, mBERT with Original Text, and mBERT with Emoji-Enriched Text.

The comparison between Original Text and Emoji-Enriched Text serves as the primary experimental factor of this study. Since the two configurations originate from the same underlying tweet content and differ primarily in whether emoji semantics are incorporated into the textual representation, changes in classification performance can be examined in relation

to the additional semantic information provided by emojis. The models were evaluated using the same classification task and performance criteria to ensure that the comparison focused on the effect of the input representation rather than differences in the evaluation procedure.

III. RESULTS AND DISCUSSION

The primary objective of this research was to evaluate the impact of emoji descriptions and contextual data augmentation on hate speech detection across three different architectures. The performance of each model is summarized in Table 3.

MNB showed a marginal increase from 0.810 to 0.830. This 2% gain is primarily attributed to vocabulary expansion. By converting emojis into text (e.g., 😘 to face blowing kiss), the model gains additional "tokens" to associate with specific labels. However, because MNB relies on the Bag-of-Words assumption where word order and context are ignored it merely treats the emoji description as an isolated keyword. It cannot "link" the emoji's meaning to the surrounding threat, explaining why it yielded the smallest improvement.

Model Bi-LSTM exhibited a significant jump from 0.785 to 0.835. This 5% improvement demonstrates that the Bi-LSTM is highly dependent on contextual anchors. Without emojis, the sequences were often sparse or fragmented, leading to a baseline lower than MNB. By introducing descriptions, we provided the Bi-LSTM with a continuous semantic flow. The bidirectional layers could finally establish long-term dependencies, connecting aggressive phrases to the "intensifying" or "sarcastic" nature of the emoji description. Model mBERT achieved the highest peak score, rising from 0.825 to 0.870. This 4.5% improvement (on top of an already high baseline) signifies mBERT's superior self-attention capabilities. While the percentage gain is slightly lower than Bi-LSTM, the absolute score of 0.870 represents a highly refined understanding of digital linguistics. mBERT can attend to the emoji description and the hateful keywords simultaneously, regardless of their distance in the tweet. This allows the model to identify "masked" hate speech where the emoji description acts as a critical signal for sarcasm that traditional preprocessing would have discarded.

The empirical results of this study demonstrate a consistent performance trajectory, the integration of emoji descriptions serves as a catalyst for improving classification accuracy across all architectures. The most notable observation is the performance of mBERT, which achieved the peak Macro F1-score of 0.870. This superiority is attributed to the Self-Attention mechanism, which allows the model to dynamically weigh the importance of emoji-derived textual tokens (e.g., "face blowing kiss") against aggressive verbal tokens, regardless of their spatial distance within the sequence.

A significant finding in this research is the initial underperformance of the Bi-LSTM (0.785) relative to Multinomial Naive Bayes (0.810) in the absence of emoji data. In a state of reduced semantic density where emojis are stripped, the Bi-LSTM struggles to establish robust long-term dependencies due to fragmented sentence structures. Conversely, MNB, operating on a Bag-of-Words (BoW) paradigm, remains resilient to sequence fragmentation as it relies on frequency-based statistical signals rather than contextual flow. However, the inclusion of emoji descriptions reversed this trend. The Bi-LSTM surpassed the MNB baseline. This suggests that while sequential models are highly sensitive to "noise" in sparse data, they possess a superior capacity for feature fusion once the semantic gap is bridged by textualizing visual symbols.

The significant improvement in Macro F1-scores suggests that the BERT-based augmentation effectively mitigated model bias. By synthetically expanding the non-hate (normal) class, we provided the model with a more robust linguistic foundation to recognize benign discourse. This ensures the model does not become 'over-sensitized' to hate speech, a common pitfall where a model incorrectly flags neutral text as hateful simply because it lacks a sufficient internal representation of 'normal' language. Consequently, this balancing strategy cultivates a more nuanced sense of non-hate context, allowing the model to accurately differentiate between true toxicity and harmless social media interactions.

To provide a more granular evaluation of the experimental results, this section presents the classification report for each model, focusing on the scenarios with and without emoji descriptions. We utilize the classification report to analyze the performance through Precision, Recall, and the F1-

Table 3 Model F1-Score Performance

Model	Without Emoji	With Emoji Description
Multinomial Naive Bayes (MNB)	0.810	0.830
Bi-LSTM	0.785	0.835
mBERT	**0.825**	**0.870**

Table 4 Classification Report Without Emojis

Model	Label	Without Emoji	With Emoji Description
		precision	recall
Multinomial Naive Bayes (MNB)	0	0.80	0.84
	1	0.83	0.78
Bi-LSTM	0	0.75	0.86
	1	0.83	0.72
mBERT	0	0.90	0.74
	1	0.78	0.92

Table 5 Classification Report With Emoji Descriptions

Model	Label	Without Emoji	With Emoji Description
		precision	recall
Multinomial Naive Bayes (MNB)	0	0.83	0.83
	1	0.83	0.83
Bi-LSTM	0	0.87	0.78
	1	0.81	0.89
mBERT	0	0.93	0.81
	1	0.83	0.94

Score for both the 'Normal' (Class 0) and 'Hate' (Class 1) categories. Classification report for models without emojis can be seen in Table 4.

While MNB provides a consistent baseline, mBERT is the superior choice for practical implementation. Its ability to reach a 0.92 recall for the "Hate" class ensures that harmful content is rarely missed, fulfilling the primary objective of automated moderation. The inclusion of emoji descriptions was the catalyst that allowed mBERT to reach this level of sensitivity by providing the necessary semantic context. The classification report for models with emoji descriptions can be seen in Table 5.

An F1-score of 0.87 is supported by a nearly perfect recall for hate speech. This means mBERT successfully identifies 94% of all hateful content in the dataset. The high precision for the normal class (0.93) means that when mBERT labels something as "Normal," it is almost certainly correct. The jump to 0.94 recall proves that the Self-Attention mechanism in mBERT is exceptionally skilled at "reading between the lines." By transforming emojis into descriptions, we provided mBERT with the semantic cues needed to catch almost every instance of toxicity, even those masked by sarcasm or indirect language.

Across the three architectures, the results indicate that emoji descriptions provide a consistent positive contribution to hate speech detection, although the magnitude of the improvement varies according to the model's ability to represent contextual information. MNB increased by 0.020 F1-score, while Bi-LSTM and mBERT improved by 0.050 and 0.045, respectively. These results suggest that the benefit of emoji semantic information is not limited to a particular modeling paradigm. However, the larger improvements observed in Bi-LSTM and mBERT indicate that models capable of processing contextual relationships can exploit the additional information more effectively than frequency-based approaches.

Nevertheless, the results should not be interpreted as evidence that emoji descriptions universally resolve contextual ambiguity. Emoji meanings can remain dependent on cultural, conversational, and sarcastic contexts. The findings instead demonstrate that converting emojis into textual semantic descriptions preserves information that would otherwise be lost during conventional text preprocessing.

IV. CONCLUSIONS

This research underscores the pivotal role of emoji descriptions in enhancing the performance of automated hate speech detection systems. Our findings conclude that visual symbols in digital communication carry essential semantic weight that, when converted into textual descriptions, significantly clarifies the intent and nuance of social media discourse. The comparative analysis revealed that while Multinomial Naive Bayes remains a robust baseline for keyword-heavy datasets, it lacks the depth required for complex linguistic patterns. The transition from emoji embeddings to textual descriptions reversed the initial performance degradation, replacing the 3.13% decline

in the baseline with improvements across all models, most notably a 5% surge in Bi-LSTM and a peak F1-score of 0.87 for mBERT. This success is attributed to its self-attention mechanism and the effectiveness of BERT-based data augmentation, which allowed the model to develop a balanced understanding of both hateful and benign contexts without class bias. Unlike the initial experiments using emoji2vec, which resulted in a performance drop from 72.8% to 69.67%, the current approach of using textual emoji descriptions successfully surpassed the baseline, achieving a peak F1-score with mBERT.

While this study demonstrates the efficacy of emoji-to-text transformation, several avenues for future research remain. Due to constraints in computing resources, this study utilized pre-trained architectures without extensive fine-tuning. Future research should explore hyperparameter optimization and the fine-tuning of larger Transformer variants (such as RoBERTa or XLM-R) to push performance boundaries. Implementing advanced transfer learning techniques could allow the model to adapt more effectively to specific social media dialects or slang.

ACKNOWLEDGEMENT

The authors would like to express their our deepest gratitude to Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Institut Teknologi Del for the financial support and facilities provided for this research. We also extend our gratitude to the anonymous reviewers for their insightful comments and suggestions that significantly improved the quality of this manuscript.

AUTHOR CONTRIBUTIONS

Conceived and designed the analysis, J. A., W. V. S., R. M. H. and L. M. A. T.; Collected the data, J. A., W. V. S., R. M. H. and L. M. A. T.; Contributed data or analysis tools, W. V. S.; Performed the analysis, J. A., W. V. S., R. M. H. and L. M. A. T.; Wrote the paper, J. A. and W. V. S.

DATA AVAILABILITY

Data openly available in a public repository that does not issue DOIs. The data that support the findings of the research are openly available in GitHub at https://github.com/bluga404/emoji-hate-speech classification, reference number main.

REFERENCES

Alaoui, S. S., Farhaoui, Y., & Aksasse, B. (2022). Hate speech detection using text mining and machine learning. *International Journal of Decision Support System Technology (IJDSST), 14*(1), 1-20. https://doi.org/10.4018/IJDSST.286680

Amalia, F. S., & Suyanto, Y. (2024). Offensive language and hate speech detection using BERT model. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 18*(4). https://doi.org/10.22146/ijccs.99841

Amalia, J., Tambunan, S. R., Purba, S. E. M., & Simanjuntak, W. V. (2025). Enhancing hate speech detection: leveraging emoji preprocessing with BI-LSTM model. *Journal of Information Systems and Informatics, 7*(2). https://doi.org/10.51519/journalisi.v7i2.1147

Aminu, E. F., Ayobami Ekundayo, Sarkibaka, S. D., Ojerinde, O. A., & Ugwuoke, U. C. (2024). Hate speech detector based on hybridized BERT-attention mechanism and context analyzer. *Journal of Computer Sciences and Informatics, 1*(1). 10.5455/JCSI.20240613103822

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Computer Science, Computation and Language*. https://doi.org/10.48550/arXiv.2010.12421

Chai, Y., & Xie, H. (2025). Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. *Artificial Intelligence Review, 59*. https://doi.org/10.1007/s10462-025-11405-5

Christen, P., Hand, D. J., & Kirielle, N. (2023). A review of the f-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys, 56*(3), 1 – 24. https://doi.org/10.1145/3606367

Dalavi, S. R., Nivelkar, T., Patil, S., & Sawant, A. (2023). Enhancing hate speech detection through emoji-based classification using bi-lstm and glove embeddings. *2023 6th International Conference on Advances in Science and Technology (ICAST)*. https://doi.org/10.1109/ICAST59062.2023.10455077

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (Long and Short Papers)*. https://doi.org/10.18653/v1/N19-1423

Dixon, S. J. (2025, November 10). *Number of social media users worldwide 2030*. Statista. Retrieved February 2, 2026, from https://www.statista.com/forecasts/278414/number-of-worldwide-social-network-users/

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4). https://doi.org/10.1080/13698230.2019.1576006

Handayani, T. P., & Hasyim, W. (2024). Comparative analysis of CNN-RNN models for hatespeech detection incorporating L2 regularization. *International Journal of Engineering, Science, and Information Technology, 4*(1).

Lazareva, S. (2023). Emojis as graphical hate speech markers. *Journal of Language Works, 8*(1). https://tidsskrift.dk/lwo/article/view/138019

Ma, E. (2019). *nlpaug*. Github. https://github.com/makcedward/nlpaug

Putra, C. D., & Wang, H.-C. (2023). Advanced BERT-CNN for hate speech detection. *Procedia Computer Science, 234*. https://doi.org/10.1016/j.procs.2024.02.170

Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wires Computational Statistics, 16*(2). https://doi.org/10.1002/wics.1648

Safawi, N. U. C. M., & Shafie, N. A. (2024). Performance of TF-IDF for text classification reviews on google play store: Shopee. *Journal of Computing Research and Innovation*, *9*(2), 13-22. https://doi.org/10.24191/jcrinn.v9i2.410

Saikh, T., Barman, S., Kumar, H., Sahu, S., & Palit, S. (2024). Emojis trash or treasure: Utilizing emoji to aid hate speech detection. *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. https://aclanthology.org/2024.icon-1.64/

Salsabila, N. S., Rahma, A., Agustin, A., & Nanda, D. (2025). The impact of emoji use on perception differences in online conversations among teenagers. *LITERA Jurnal Bahasa Dan Sastra*.

Telaumbanua, Y. A., Tealumbanua, N. T. N., Halawa, M. D., & Benedikta Gulo. (2024). The use of emojis in language communication on social media platforms. *Journal of English Language and Education, 9*(4). https://doi.org/10.31004/jele.v9i4.524

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 10.18653/v1/N19-1144