Emotion classification using advanced neural networks on sentence-level data

Athapol Ruangkanjanases[1] and Taqwa Hariguna[2]

[1] Department of Commerce, Chulalongkorn Business School, Chulalongkorn University, Bangkok, Thailand
[2] Magister of Computer Sciences, Universitas Amikom Purwokerto, Purwokerto, Indonesia

ABSTRACT

This study explores the efficacy of advanced neural network architectures, including bidirectional long short-term memory (BiLSTM), bidirectional gated recurrent unit (BiGRU), and bidirectional encoder representations from transformers (BERT), for sentence-level emotion classification using a large-scale, imbalanced dataset of 422,746 text samples spanning six emotions. Static embeddings (Global Vectors for Word Representation (GloVe), FastText), trainable embeddings, and contextual embeddings (BERT) with varying dimensionality were evaluated. While BERT achieved the best performance (accuracy: 94.07%, F1-score: 94.05%) due to its dynamic contextual understanding, it required significantly higher computational resources and training time. Class imbalance was addressed using class-weighted loss, with potential for future exploration of oversampling, undersampling, and synthetic data generation. Error analysis revealed frequent misclassifications among semantically overlapping emotions, suggesting opportunities for hybrid embeddings and multimodal integration in future work. These findings highlight the trade-offs between performance and computational cost, providing a robust baseline for scalable emotion classification systems.

Subjects Human-Computer Interaction, Artificial Intelligence, Data Mining and Machine Learning, Natural Language and Speech, Neural Networks
Keywords Emotion classification, BiLSTM, BiGRU, BERT, Pre-trained

Submitted 24 April 2025
Accepted 30 October 2025
Published 9 December 2025

Corresponding author
Taqwa Hariguna,
taqwa@amikompurwokerto.ac.id

Academic editor
Juan Lara

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj-cs.3411**

INTRODUCTION

The rapid advancement of natural language processing (NLP) technologies has significantly enhanced the capacity to analyze and interpret human emotions expressed in text. Emotion classification, a specialized subfield of sentiment analysis, has garnered substantial attention due to its wide-ranging applications, including customer feedback analysis, mental health monitoring, and human–computer interaction (*Fei et al., 2020*; *Huang et al., 2019*). Unlike document-level or word-level analysis, sentence-level emotion classification requires accurate interpretation of emotions from concise, context-dependent expressions—making it particularly relevant for real-time applications such as chatbot interactions, social media monitoring, and early-stage mental health screening.

Despite its potential, emotion classification remains challenging due to the inherent complexity of emotional expressions and the ambiguity of natural language (*Ameer et al., 2022*). Words and phrases may convey different emotions depending on their context, while subtle distinctions between emotions, such as sadness and fear or joy and surprise, further complicate classification (*Gu et al., 2022*; *Wang et al., 2024*). Moreover, imbalanced datasets—where certain emotions are overrepresented and others underrepresented—can bias models and reduce their effectiveness in recognizing minority classes.

Traditional machine learning approaches to emotion classification have primarily relied on lexicon-based methods, which use predefined dictionaries of emotional terms to map text to emotion labels (*Purpura et al., 2019*). While useful in some contexts, these methods struggle with polysemous words and lack the ability to capture contextual nuances. The emergence of supervised learning models such as support vector machines (SVMs) and Naïve Bayes classifiers (*Li et al., 2019*) offered improved performance but relied heavily on manually engineered features, limiting scalability and adaptability (*Xu et al., 2016*).

Recurrent neural network (RNN)-based models, empowered by the introduction of distributed word representations such as Word2Vec, Global Vectors for Word Representation (GloVe), and FastText, have significantly advanced emotion classification (*Giulianelli & Kok, 2018*; *Bandhakavi, Wiratunga & Massie, 2017*). Static embeddings enabled architectures like bidirectional long short-term memory (BiLSTM) and bidirectional gated recurrent unit (BiGRU) networks to model sequential dependencies effectively (*Zhou et al., 2020*). However, static embeddings cannot adapt to sentence-specific context, limiting their performance in nuanced classification tasks (*Alhuzali & Ananiadou, 2021a*).

Transformer-based architectures have further revolutionized the field by introducing contextual embeddings that leverage self-attention mechanisms to capture bidirectional dependencies (*Li et al., 2020*; *Alp Toçoğlu & Alpkocak, 2019*). Models such as BERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), and Generalized Autoregressive Pretraining for Language Understanding (XLNet) dynamically adapt token representations based on surrounding context, achieving state-of-the-art results in various NLP tasks, including emotion classification (*Das, Poria & Bandyopadhyay, 2012*). Nevertheless, their superior performance often comes at the expense of substantial computational resources, raising concerns about scalability in resource-constrained environments.

Recent research has also explored multimodal emotion recognition, integrating textual features with visual and acoustic cues to improve classification accuracy (*Fairuz, Yusliani & Miraswan, 2021*). For example, EEG-based emotion recognition systems using hybrid convolutional neural networks (CNN) and LSTM architectures have demonstrated enhanced detection of subtle affective states, highlighting potential for future cross-modal integration. However, challenges such as data imbalance, overlapping emotional categories, and computational complexity remain significant barriers to broader adoption (*Saputra & Kumar, 2025*).

This study addresses these challenges by evaluating and comparing the performance of advanced neural network architectures—BiLSTM, BiGRU, and BERT—on a large-scale, imbalanced dataset comprising 422,746 labeled sentences across six emotion categories: Joy, Sadness, Anger, Surprise, Fear, and Disgust. We systematically examine the influence of three embedding strategies: static embeddings (GloVe, FastText), trainable embeddings, and contextual embeddings (BERT) (*Mossad et al., 2023*). The objectives are threefold: (1) assess the effectiveness of different architectures in sentence-level emotion classification, (2) analyze the impact of embedding methods on classification performance, and (3) identify persistent challenges—such as class imbalance, semantic overlap, and computational cost—that affect model accuracy (*Tanabe et al., 2020*; *Bareiss, Klinger & Barnes, 2024*). Through this exploration, the study provides actionable insights and establishes a robust baseline for future work in emotion classification and related NLP applications (*Warrier, Arshey & Jency Rena, 2021*; *Banimelhem & Amayreh, 2023*).

MATERIALS AND METHODS

Figure 1 illustrates the systematic process of this study, starting with the data preprocessing stage. This stage involves several critical steps: loading the dataset, performing text cleaning to remove unwanted characters and normalize the text, tokenizing the sentences into individual tokens, and applying padding or truncation to ensure uniform input lengths (*Sonu et al., 2022*). Additionally, label encoding is performed to convert categorical emotion labels into numerical format, making the data compatible with model input requirements (*Ding et al., 2020*). These preprocessing steps are essential to prepare high-quality data that can be effectively utilized by different neural network architectures.

The workflow then proceeds to model selection and configuration. In this phase, three key architectures are implemented: BiLSTM combined with static embeddings (GloVe or FastText), BiGRU with trainable embeddings, and the pre-trained BERT model utilizing contextual embeddings. This diverse selection of embedding methods—static, trainable, and contextual—enables a comprehensive comparison of their respective capabilities in capturing emotional nuances in text. For each model, careful hyperparameter tuning is conducted to optimize performance, including adjustments to learning rates, batch sizes, dropout rates, and optimizer choices (*Durachman & Rahman, 2024*).

Subsequently, the dataset is divided into training, validation, and testing subsets to ensure robust evaluation. The models are then trained iteratively, with regular monitoring of loss and performance metrics to prevent overfitting (*Buchdadi, 2024*). Hyperparameter optimization strategies such as grid search and early stopping are employed to further refine the models' effectiveness.

The evaluation phase assesses the trained models using a suite of metrics: accuracy, precision, recall, and F1-score, offering a multi-dimensional view of model performance across different emotion classes (*Irfan, 2024a*). Special attention is given to analyzing model behavior on minority classes, addressing challenges related to data imbalance (*Koufakou et al., 2023*).

Finally, the results analysis provides a detailed comparative study across the models. This includes assessing the overall impact of different embedding methods on model outcomes and conducting an error analysis to identify common misclassification patterns. Insights drawn from this analysis help to highlight not only the strengths and competitive advantages of each architecture but also their limitations, guiding recommendations for future work and practical applications.

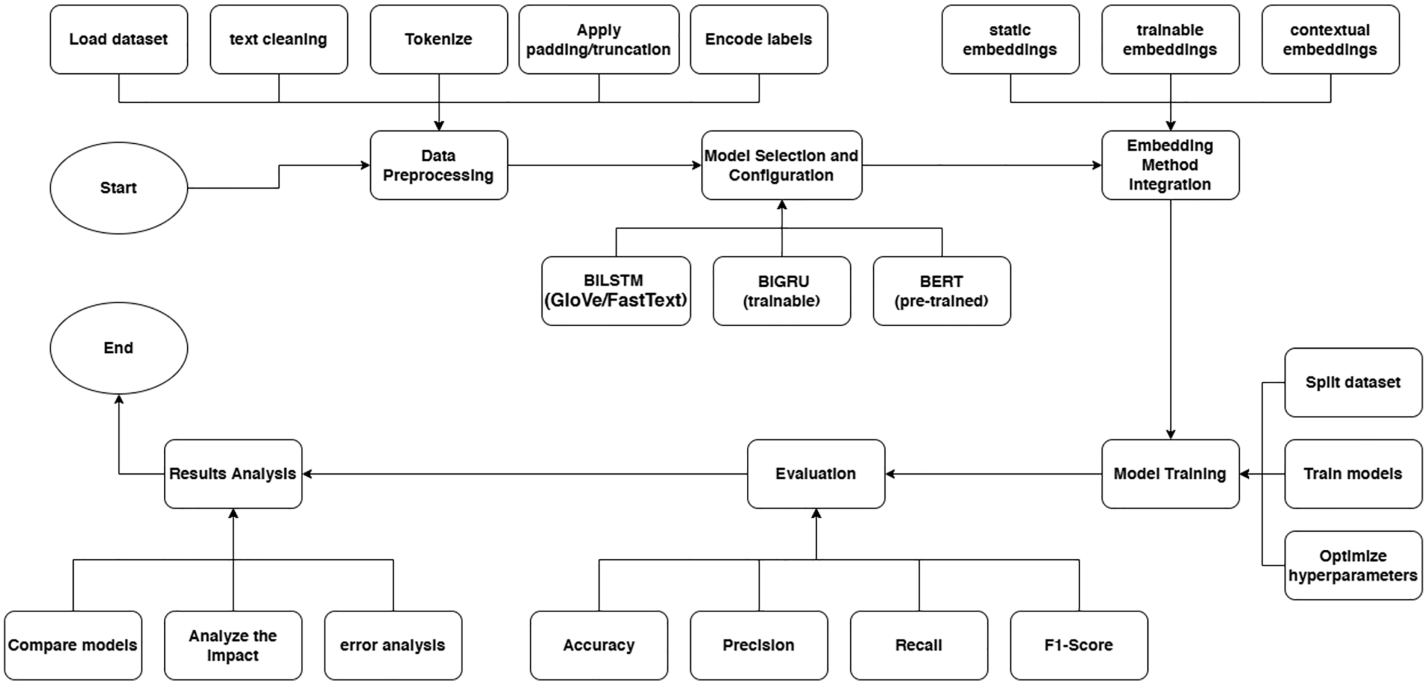

Figure 1 **Research flow.** Full-size DOI: 10.7717/peerj-cs.3411/fig-1

Dataset description

The dataset used in this study contains 422,746 labeled text samples categorized into six emotion classes: Joy, Sadness, Anger, Surprise, Fear, and Disgust. The distribution of the dataset is highly imbalanced, with Joy comprising 33.8% of the total samples and Disgust accounting for only 2.3%. This imbalance introduces challenges in model training and evaluation, as models often prioritize majority classes while underperforming on minority ones (*Wahyuningsih & Chen, 2024*). To mitigate this issue, preprocessing strategies and specialized modeling techniques were applied to ensure balanced performance across all emotion classes (*Sangsawang, 2024a*). The data used in this study can be accessed at this link: https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset.

Preprocessing pipeline

To prepare the dataset for model training, a preprocessing pipeline was implemented. Initially, text cleaning was performed to remove special characters, numbers, and extra spaces, leaving only meaningful textual content (*Hariguna, Sarmini & Azis, 2024*). Sentences were then tokenized into individual tokens, represented as $T = \{t_1, t_2, \ldots, t_m\}$, where m is the number of tokens in the sentence. To ensure consistency, sequences were padded or truncated to a fixed length of 50 tokens. For sequences longer than 50, the first 50 tokens were retained ($T' = T[1:L]$). For shorter sequences, padding tokens were added to achieve the required length ($T' = T + \{<PAD>\}^{L-m}$) (*Henderi et al., 2024*). Finally, categorical emotion labels were encoded into numerical values, mapping each label y to an integer for compatibility with the neural network models (*Fu et al., 2025*).

Embedding methods

In this study, four types of embedding strategies were employed to represent textual data. The first was GloVe embeddings with 100 dimensions (glove.6B.100d.txt), which were pre-trained on large corpora and kept static (non-trainable) during model training. The second was FastText embeddings with 300 dimensions (cc.en.300.bin), which incorporate subword information to better handle out-of-vocabulary words (*Kumar, 2024*). The third approach utilized a trainable embedding layer with 100 dimensions, initialized randomly and updated throughout training to capture task-specific semantic features. Finally, BERT embeddings from the pre-trained bert-base-uncased model were applied, providing contextual representations that dynamically adapt to sentence-level context (*Irfan, 2024b*). It is worth noting that the dimensional differences between GloVe (100D) and FastText (300D) may influence performance outcomes; however, these configurations were retained to preserve the integrity of the original pre-trained models.

Model architectures

Three neural network architectures were evaluated in this study: BiLSTM, BiGRU, and BERT. Each architecture was optimized for emotion classification using specific configurations.

BiLSTM utilized bidirectional long short-term memory cells to capture both forward and backward dependencies in text. The hidden state at time t was computed as Eq. (1)

$$h_t = LSTM(x_t,\ h_t - 1,\ c_t - 1), \tag{1}$$

x_t is the input, h_t is the hidden state, and c_t is the cell state. The final representation was obtained by concatenating forward and backward hidden states use Eq. (2) (*Sukmana & Khairani, 2024*):

$$h_{final} = concat\left(h_{forward},\ h_{backward}\right). \tag{2}$$

BiGRU employed gated recurrent units, which simplified the computations by removing the cell state present in LSTMs. The hidden state was updated as Eq. (3)

$$h_t = GRU(x_t,\ h_t - 1), \tag{3}$$

allowing the model to adapt dynamically to textual inputs (*Sangsawang, 2024b*).

BERT leveraged its transformer-based architecture, which relies on self-attention mechanisms to capture bidirectional contextual information. The self-attention mechanism was formulated as Eq. (4).

$$\frac{alpha_{ij} = \exp\left((WQ\ *\ e_i) *\ \left(WK\ *\ e_j\right)\right)}{sum_{k(\exp((WQ * e_i)*\ (WK * e_k)))}}, \tag{4}$$

$alpha_{ij}$ represents the attention weights between tokens i and j, and $WQ,\ WK,$ and WV are the query, key, and value weight matrices. The final representation of token i was computed as Eq. (5) (*Wahyuningsih & Chen, 2024*):

$$e_i' = sum_{j\left(alpha_{ij}*\left(WV*e_j\right)\right)}. \tag{5}$$

Training configuration and loss function

For model training, specific hyperparameter configurations were applied to ensure consistency and fair comparison across architectures (*Yadav & Hananto, 2024*). The BiLSTM models with GloVe and FastText embeddings, as well as the BiGRU model with trainable embeddings, were trained with a batch size of 64, learning rate of 0.001, and dropout rate of 0.5, each employing 128 hidden units over five epochs. In contrast, the BERT model was fine-tuned using a smaller batch size of 16, a learning rate of 5e−5, and a dropout rate of 0.1, also for five epochs (*Sukmana & Oh, 2024*).

All experiments were conducted on an MSI Cyborg 15 A12V laptop equipped with an Intel Core i7-12650H processor (10 cores), 16 GB DDR5 RAM, 512 GB NVMe SSD storage, and an NVIDIA GeForce RTX 4050 Laptop GPU with 6 GB VRAM, running on Windows 11. Training times varied across models, reflecting differences in computational complexity. On average, BiLSTM-GloVe required approximately 1.8 min per epoch (9 min for five epochs), BiLSTM-FastText 2.1 min per epoch (10.5 min total), and BiGRU 1.6 min per epoch (8 min total). By comparison, BERT was considerably more resource-intensive, averaging 6.5 min per epoch and totaling around 32.5 min for five epochs. These results highlight the trade-off between computational efficiency and predictive performance across different model architectures To train the models, the cross-entropy loss function was employed, which is defined as Eq. (6):

$$L = -\left(\frac{1}{N}\right) * sum\left(i = 1\ to\ N,\ sum\left(j = 1\ to\ k,\ y_{ij} * \log\left(y_{hat_{ij}}\right)\right)\right). \tag{6}$$

Here, y_{ij} is the true label for the *i-th* sample and the *j-th* class, $y_{hat_{ij}}$ is the predicted probability for the same, N is the number of samples, and k is the number of classes.

Evaluation method

The evaluation of model performance in this study was conducted through a rigorous and multi-dimensional approach, employing both quantitative metrics and qualitative error analysis to ensure a comprehensive assessment of emotion classification capabilities. Quantitatively, four primary evaluation metrics were utilized: accuracy, precision, recall, and F1-score. Model performance was assessed using several metrics (*Doan, 2024*). Accuracy was calculated as Eq. (7):

$$Accuracy\ = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN}, \tag{7}$$

$TP,\ TN,\ FP$, and FN are true positives, true negatives, false positives, and false negatives. Precision was defined as Eq. (8):

$$Precision = \frac{TP}{TP + FP}, \text{and recall as } Recall\ = \frac{TP}{TP\ +\ FN}. \tag{8}$$

The F1-score, which balances precision and recall, was computed as Eq. (9).

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{9}$$

Accuracy reflects the proportion of correctly classified instances over the total number of instances, providing an overall measure of model performance (*Wang et al., 2025*). Precision quantifies the proportion of true positive predictions among all positive predictions, thereby assessing the model's specificity in classifying each emotion class. Recall measures the proportion of true positive predictions relative to the actual number of positive instances, capturing the model's sensitivity (*Yadulla et al., 2024*). F1-score, defined as the harmonic mean of precision and recall, was employed to balance the trade-off between these two metrics, particularly in the presence of class imbalance.

During model training, the cross-entropy loss function was adopted to quantify the discrepancy between predicted probabilities and true class labels. To address the inherent class imbalance within the dataset, a class-weighted loss function was implemented, ensuring that minority classes received proportionally higher penalties for misclassification, thereby promoting a more balanced learning process.

Furthermore, confusion matrix analysis was conducted to systematically examine the distribution of correct and incorrect predictions across all emotion classes. This analysis facilitated the identification of prevalent misclassification patterns, particularly between semantically overlapping categories such as "Joy" and "Surprise" or "Sadness" and "Fear".

In addition to these evaluation metrics, hyperparameter optimization played a critical role in enhancing model performance. Strategies such as grid search and early stopping were employed to fine-tune parameters including learning rate, batch size, hidden layer dimensions, and dropout rates, thereby preventing overfitting and ensuring optimal generalization to unseen data (*Pratama, 2024*).

This comprehensive evaluation framework enabled a robust comparison of the BiLSTM, BiGRU, and BERT architectures, providing valuable insights into their respective strengths and limitations in the context of emotion classification.

Regularization technique

To mitigate the risk of overfitting and improve model generalization, dropout layers were incorporated across all neural architectures. For the BiLSTM and BiGRU models, a dropout rate of 0.5 was applied to the hidden layers, effectively preventing co-adaptation of neurons by randomly deactivating 50% of units during training. This technique is particularly useful in recurrent architectures, where long-term dependencies can otherwise lead to overfitting on training data (*Cheng et al., 2025*). In the case of BERT, a smaller dropout rate of 0.1 was used, consistent with standard fine-tuning practices for transformer-based models, as larger dropout values tend to degrade performance in pre-trained transformers.

Other commonly used regularization strategies were not implemented in this study but are acknowledged as promising directions for future work. Label smoothing, which prevents the model from becoming overconfident by redistributing a small portion of the probability mass across incorrect classes, could be beneficial for handling ambiguous emotion categories such as Joy–Surprise and Sadness–Fear. L2 weight decay, another widely adopted technique, applies penalties to large weight values during optimization, helping to reduce variance and improve robustness. Furthermore, text data augmentation methods, such as synonym replacement and back-translation, offer an effective way to expand training data and reduce sensitivity to lexical variations. These methods are particularly relevant in imbalanced datasets, where augmenting minority classes (*e.g.*, Disgust) could help the model better generalize to underrepresented emotions.

Although not applied in the present study due to scope constraints, these regularization techniques remain important avenues for future exploration. Their integration has the potential to further enhance classification accuracy, improve robustness against overfitting, and increase the fairness of predictions across all emotion categories.

Reproducitbility

Both the dataset and the complete source code used in this study are publicly available at the following repository: https://doi.org/10.5281/zenodo.15009245. This ensures transparency and facilitates reproducibility of the experiments, allowing other researchers to replicate and extend the findings presented in this work.

RESULTS

Dataset overview

The dataset used in this study consists of 422,746 text samples distributed across six emotion labels: Joy, Sadness, Anger, Surprise, Fear, and Disgust. Among these, Joy represents the majority class with 33.8% of the samples, while Disgust, the minority class, accounts for only 2.3%. This significant imbalance presents challenges in training robust models, as they often prioritize performance on majority classes while neglecting minority ones.

Addressing this imbalance is crucial for achieving fair and generalized results. Techniques such as oversampling, class-weighted loss functions, or data augmentation could improve minority class performance. The distribution of the dataset is summarized in Table 1, providing a clear overview of the class imbalances.

Preprocessing and feature engineering

To prepare the dataset for modeling, a robust preprocessing pipeline was implemented. This began with text cleaning, removing special characters, numbers, and irrelevant symbols to ensure that the textual content was clean and meaningful. Next, tokenization split sentences into individual tokens, which were subsequently padded or truncated to a fixed length of 50 tokens to standardize input sizes for the neural network models.

Label encoding was applied to convert the categorical emotion labels into numerical values suitable for training. These preprocessing steps ensured uniformity in the data while preserving essential features. The preprocessing steps are summarized in Table 2.

This pipeline provided a solid foundation for applying neural network architectures and ensured compatibility with advanced models such as BERT.

Model configurations and performance

Three neural network architectures were evaluated: BiLSTM, BiGRU, and BERT. BiLSTM was tested with two pre-trained embedding methods: GloVe (100-dimensional) and FastText (300-dimensional). BiGRU was trained with trainable embeddings, allowing it to

Table 1 Dataset overview.

Label	Count	Percentage (%)
Joy	143,067	33.8
Sadness	120,000	28.4
Anger	80,000	18.9
Surprise	40,000	9.5
Fear	30,000	7.0
Disgust	9,679	2.3

Table 2 Preprocessing step.

Step	Description
Text cleaning	Removed symbols, numbers, and special characters.
Tokenization	Split sentences into tokens.
Padding/Truncation	Adjusted token sequences to a fixed length of 50.
Label encoding	Converted categorical labels into numerical values.

Table 3 Model configuration.

Model	Embedding	Hidden size	Layers	Dropout	Optimizer	Learning rate
BiLSTM	GloVe (100 dim)	64	2	0.5	Adam	0.001
BiLSTM	FastText (300 dim)	64	2	0.5	Adam	0.001
BiGRU	Trainable	64	2	0.5	Adam	0.001
BERT	Pre-trained BERT	768	–	–	AdamW	5e–5

Table 4 Model performance metrics.

Model	Embedding	Accuracy (%)	Loss	Precision (%)	Recall (%)	F1-Score (%)
BiLSTM	GloVe (100 dim)	93.16	0.1165	93.10	93.20	93.15
BiLSTM	FastText (300 dim)	93.04	0.1230	93.00	93.10	93.05
BiGRU	Trainable	93.40	0.0981	93.50	93.30	93.40
BERT	Pre-trained BERT	94.07	0.0950	94.10	94.00	94.05

adapt specifically to the dataset. BERT, leveraging its transformer architecture, utilized pre-trained contextual embeddings for enhanced performance. Table 3 outlines the configurations of these models.

The performance metrics of these models are summarized in Table 4.

BERT achieved the best performance across all metrics, emphasizing the effectiveness of contextual embeddings. BiGRU with trainable embeddings outperformed both BiLSTM configurations, demonstrating the importance of fine-tuning embedding layers.

Error analysis

Error analysis revealed specific challenges faced by the models. Misclassifications frequently occurred between Joy and Surprise due to their overlapping semantic contexts.

Table 5 Confusion matrix results.

Actual\Predicted	Joy	Sadness	Anger	Surprise	Fear	Disgust
Joy	14,000	200	150	300	250	100
Sadness	150	13,800	100	250	300	120

Table 6 Hyperparameter tuning results.

Hyperparameter	Value	Accuracy (%)	Loss
Learning rate (BERT)	5e−5	94.07	0.0950
Learning rate (BERT)	1e−4	93.75	0.1100
Dropout (BiGRU)	0.5	93.40	0.0981
Dropout (BiGRU)	0.3	93.10	0.1050

Similarly, Sadness and Fear were often confused, reflecting subtle textual nuances that require advanced contextual understanding.

The confusion matrix, shown in Table 5, highlights these patterns. For instance, the Joy label was misclassified as Surprise in 250 instances, indicating areas for improvement in contextual sensitivity. Addressing these errors may require incorporating external knowledge bases or augmenting the dataset with context-rich examples.

Hyperparameter tuning

Hyperparameter tuning was critical in optimizing the performance of each model. For BERT, adjustments to learning rate and batch size significantly influenced the model's convergence and accuracy. The best performance was achieved with a learning rate of 5e−5, as shown in Table 6. Higher learning rates led to unstable training, while lower rates resulted in slower convergence.

For BiLSTM and BiGRU, the number of hidden units and dropout rates were varied to improve generalization. Increasing the number of hidden units to 128 resulted in slightly improved accuracy but increased computational cost. The addition of dropout (set at 0.5) proved essential in mitigating overfitting, particularly for models with trainable embeddings.

These findings underscore the importance of systematic hyperparameter optimization to unlock the full potential of each model.

Confusion matrix analysis

The confusion matrix for BERT, displayed in Table 7, reveals the model's classification strengths and weaknesses. While BERT excelled in distinguishing well-separated classes such as Joy and Anger, it struggled with overlapping categories like Joy and Surprise or Sadness and Fear.

The matrix highlights areas where the model's understanding of context may need improvement. Augmenting the dataset with more diverse examples and incorporating external knowledge bases could help reduce these errors.

Table 7 Confusion matrix for BERT.

Actual\Predicted	Joy	Sadness	Anger	Surprise	Fear	Disgust
Joy	14,000	200	150	300	250	100
Sadness	150	13,800	100	250	300	120
Anger	120	130	14,100	80	250	220
Surprise	300	220	180	14,350	200	150
Fear	200	300	270	220	13,900	180
Disgust	100	150	200	170	250	14,150

Comparative analysis of embedding methods

The choice of embedding methods played a crucial role in determining the overall performance of the models. Pre-trained embeddings such as GloVe and FastText provided strong initial baselines, particularly for BiLSTM. GloVe, with its 100-dimensional representation, demonstrated effective generalization and computational efficiency, making it suitable for tasks with limited computational resources. On the other hand, FastText's 300-dimensional embeddings utilized sub-word information, allowing the model to handle out-of-vocabulary words effectively. However, both embeddings are static, which limits their ability to adapt dynamically to the specific context of the dataset.

Trainable embeddings, as employed in BiGRU, offered significant flexibility by allowing the model to learn features directly from the dataset. This adaptability resulted in better performance compared to BiLSTM with static embeddings. Trainable embeddings provided the model with domain-specific nuances, which were particularly advantageous for improving accuracy and loss metrics.

BERT's contextual embeddings outperformed all other methods by a significant margin. Unlike static or trainable embeddings, BERT dynamically generates token-level embeddings that consider the entire context of a sentence. This contextual understanding enabled BERT to capture intricate relationships within the text, making it exceptionally well-suited for handling nuanced and overlapping emotional expressions. As shown in Table 8, BERT achieved the highest accuracy (94.07%) and F1-score (94.05%), along with the lowest loss (0.0950), confirming its superiority.

The analysis highlights a clear trend: the more contextual and adaptable the embedding method, the better the model's performance. Static embeddings, while efficient and robust for certain tasks, fail to capture the nuanced relationships in text data. Trainable embeddings bridge this gap by learning domain-specific representations but still lack the global contextual understanding provided by BERT.

Future work could explore combining these embeddings into hybrid models. For instance, using pre-trained embeddings as initialization for trainable layers or combining static embeddings with contextual embeddings could further improve performance. Additionally, contextual embeddings like those in BERT can be fine-tuned on domain-specific corpora to enhance their adaptability, especially for tasks with highly specialized language.

Table 8 Comparison of embedding methods and model performance in emotion classification.

Embedding method	Model	Accuracy (%)	Loss	F1-score (%)	Key strengths
GloVe (100 dim)	BiLSTM	93.16	0.1165	93.15	Effective generalization, efficient size
FastText (300 dim)	BiLSTM	93.04	0.1230	93.05	Handles out-of-vocabulary words well
Trainable embedding	BiGRU	93.40	0.0981	93.40	Adapts dynamically to the dataset
Contextual embedding	BERT	94.07	0.0950	94.05	Captures context dynamically and effectively

DISCUSSION

This study demonstrates the significant advancements achieved through the use of advanced neural networks in emotion classification. The superior performance of BERT highlights the importance of contextual embeddings, particularly in capturing subtle emotional nuances within text (*Guo, 2022*; *Alhuzali & Ananiadou, 2021b*). Unlike traditional static embeddings such as GloVe and FastText, BERT dynamically adapts to contextual variations, allowing for a more accurate representation of emotions in complex textual data (*Mossad et al., 2023*).

However, this performance improvement comes with a trade-off in computational cost. Training BERT required substantially more time and GPU resources, averaging 6.5 min per epoch (~32.5 min for five epochs), compared to less than 10 min in total for BiLSTM and BiGRU. While BERT provides the highest accuracy and F1-score, its computational demands may limit practicality in resource-constrained environments. By contrast, BiGRU achieved competitive performance with significantly lower training times, making it a more efficient option for deployment. Similarly, BiLSTM with GloVe or FastText embeddings, though slightly less accurate, offered the fastest training and lowest resource consumption. These findings emphasize that model choice should consider both predictive performance and resource efficiency, depending on the requirements of real-world applications.

Another limitation observed is the risk of overfitting, particularly for models trained on imbalanced data. While dropout layers were applied to all neural architectures, additional regularization strategies were not employed. Techniques such as label smoothing could reduce overconfidence in predictions, L2 weight decay could constrain excessively large weights, and text data augmentation (*e.g*., synonym replacement or back-translation) could expand training data diversity. Integrating these methods in future work may enhance robustness, particularly for underrepresented classes such as Disgust.

Class imbalance remains a key challenge. In this study, class-weighted loss was used to penalize misclassifications of minority classes. While this approach improved balance to some extent, alternative methods such as oversampling, undersampling, or synthetic data generation (*e.g*., Synthetic Minority Over-sampling Technique (SMOTE), Easy Data Augmentation) were not implemented. Future studies could combine these techniques to further mitigate imbalance, potentially reducing the frequent misclassifications observed between semantically similar categories such as Joy–Surprise and Sadness–Fear.

Another factor that may have influenced results is the difference in embedding dimensionality between GloVe (100D) and FastText (300D). Higher-dimensional embeddings typically capture richer semantic information, which may partially explain FastText's stronger performance compared to GloVe. Although this difference does not alter the main trend—where contextual embeddings outperform static ones—it represents a potential bias in cross-embedding comparisons. Standardizing dimensionality in future experiments would provide a more equitable benchmark.

The results of this study align with prior work demonstrating the superiority of transformer-based models in emotion classification tasks. For instance, *Fei et al. (2020)* reported that transformer architectures achieved notable gains over recurrent networks for multi-label emotion classification, while *Mossad et al. (2023)* showed that BERT-based models significantly outperformed BiLSTM in Arabic text emotion analysis. Our finding that BERT achieved 94.07% accuracy is consistent with these reports, underscoring the robustness of contextual embeddings across languages and datasets. Furthermore, the error patterns observed in this study—particularly the confusion between Joy–Surprise and Sadness–Fear—mirror those described in earlier research (*Gu et al., 2022*; *Tanabe et al., 2020*), suggesting that semantic overlap between closely related emotions remains a persistent challenge regardless of model architecture.

Nevertheless, the generalizability of these findings is limited by the use of a single dataset. While the large scale of the *corpus* ensures internal validity, emotion expression can vary significantly across domains (*e.g*., social media, clinical narratives, customer reviews) and cultures. Evaluating models on multiple datasets or through domain adaptation techniques would strengthen the external validity of the results. Future work should therefore explore cross-domain benchmarks and multilingual corpora to assess the adaptability of different embedding strategies and architectures beyond the specific dataset examined here.

Despite these challenges, the findings of this study are consistent with prior literature reporting the superiority of transformer-based models in emotion classification (*Mossad et al., 2023*; *Wang et al., 2024*). At the same time, the results underscore the value of lighter models such as BiGRU for practical deployment scenarios where efficiency and scalability are critical.

Future research should build upon these insights by integrating additional regularization techniques, employing multimodal datasets, and testing across multiple corpora to improve generalizability. Ensemble learning and domain adaptation also hold promise for enhancing robustness, while hybrid embeddings that combine static, trainable, and contextual representations may yield further gains. By addressing these aspects, future studies can advance the development of emotion classification systems that are not only accurate but also efficient, interpretable, and broadly applicable.

CONCLUSIONS

This study demonstrates the effectiveness of advanced neural network architectures, specifically BiLSTM, BiGRU, and BERT, in the task of emotion classification.

Through comparative analysis, BERT emerged as the most effective model, achieving the highest accuracy and F1-scores across all metrics. Its dynamic contextual embeddings enabled it to outperform both static pre-trained embeddings (GloVe and FastText) and trainable embeddings used in BiGRU. The choice of embedding methods significantly influenced model performance, with trainable and contextual embeddings proving superior in capturing domain-specific nuances and complex linguistic relationships. These results underscore the transformative potential of contextual embeddings in advancing natural language processing tasks.

Despite these advancements, several challenges remain. The dataset's imbalanced distribution, with an overrepresentation of the "Joy" label and an underrepresentation of "Disgust," made achieving consistent performance across all emotion classes difficult. Addressing this imbalance through data augmentation, cost-sensitive learning, or class-weighted loss functions is crucial for future work. Additionally, error analysis highlighted frequent misclassifications between semantically overlapping emotions, such as "Joy" and "Surprise" or "Sadness" and "Fear," suggesting the potential benefits of integrating external knowledge sources or leveraging ensemble techniques to enhance model robustness.

However, this study also has several limitations. The use of a single dataset restricts the generalizability of the findings across different domains and cultures. Furthermore, while BERT achieved the best results, its high computational cost may limit its practical deployment in resource-constrained environments.

Future research should focus on exploring hybrid approaches that combine static, trainable, and contextual embeddings, integrating multimodal data to capture richer emotional contexts, and developing lightweight models to enhance robustness and applicability. This study provides a strong foundation for further exploration in emotion classification and highlights the critical role of model architecture and embedding choice in achieving state-of-the-art results.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Athapol Ruangkanjanases conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Taqwa Hariguna conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data and code are available at Kaggle and Zenodo:

- https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset.

- Taqwa, H. (2025). Emotion Classification Using Neural Networks. Zenodo. https://doi.org/10.5281/zenodo.15009245.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3411#supplemental-information.

REFERENCES

Alhuzali H, Ananiadou S. 2021a. SpanEmo: casting multi-label emotion classification as span-prediction. In: *Proceedings of EACL 2021*, 1573–1584 DOI 10.18653/v1/2021.eacl-main.135.

Alhuzali H, Ananiadou S. 2021b. Improving textual emotion recognition based on intra-and inter-class variations. *IEEE Transactions on Affective Computing* **14(2)**:1297–1307 DOI 10.1109/taffc.2021.3104720.

Alp Toçoğlu M, Alpkocak A. 2019. Lexicon-based emotion analysis in Turkish. *Turkish Journal of Electrical Engineering and Computer Sciences* **27**:1213–1227 DOI 10.3906/elk-1807-41.

Ameer I, Sidorov G, Gómez-Adorno H, Nawab RMA. 2022. Multi-label emotion classification on code-mixed text: data and methods. *IEEE Access* **10(2)**:8779–8789 DOI 10.1109/access.2022.3143819.

Bandhakavi A, Wiratunga N, Massie S. 2017. Lexicon generation for emotion detection from text. *IEEE Intelligent Systems* **32(1)**:102–108 DOI 10.1109/mis.2017.22.

Banimelhem O, Amayreh W. 2023. The performance of ChatGPT in emotion classification. In: *Proceedings of the 2023 14th International Conference on Information and Communication Systems (ICICS)*, 1–4 DOI 10.1109/ICICS60529.2023.10330544.

Bareiss P, Klinger R, Barnes J. 2024. English prompts are better for NLI-based zero-shot emotion classification than target-language prompts. In: *Companion Proceedings of the ACM Web Conference 2024* DOI 10.1145/3589335.3651902.

Buchdadi AD. 2024. Segmenting walmart customers for personalized marketing strategies using MiniBatchKMeans clustering and decision trees: an analysis of purchasing behavior. *Journal of Digital Market and Digital Currency* **1(3)**:204–224 DOI 10.47738/jdmdc.v1i3.18.

Cheng F, Sangsawang T, Pigultong M, Watkraw W. 2025. Data-driven development of an elderly training package using the GCC model. *Journal of Applied Data Sciences* **6(1)**:773–787 DOI 10.47738/jads.v6i1.662.

Das D, Poria S, Bandyopadhyay S. 2012. A classifier-based approach to emotion lexicon construction. *Lecture Notes in Computer Science* **7337**:320–326 DOI 10.1007/978-3-642-31178-9_41.

Ding F, Kang X, Nishide S, Guan Z, Ren F. 2020. A fusion model for multi-label emotion classification based on BERT and topic clustering. In: *Proceedings of the International Conference on Signal Processing and Communication Systems* DOI 10.1117/12.2579255.

Doan ML. 2024. Predicting the success of virtual-themed animated movies using random forest regression. *International Journal Research on Metaverse* **1(3)**:187–198 DOI 10.47738/ijrm.v1i3.16.

Durachman Y, Rahman AWA. 2024. Blockchain and the evolution of decentralized finance navigating growth and vulnerabilities. *Journal of Current Research in Blockchain* **1(3)**:166–177 DOI 10.47738/jcrb.v1i3.20.

Fairuz D, Yusliani N, Miraswan KJ. 2021. Classification of emotions on Twitter using emotion lexicon and Naïve Bayes. *Sriwijaya Journal of Informatics and Applications* **2(2)**:1–12.

Fei H, Ji D, Zhang Y, Ren Y. 2020. Topic-enhanced capsule network for multi-label emotion classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**:1839–1848 DOI 10.1109/taslp.2020.3001390.

Fu Y, Sangsawang T, Pigultong M, Watkraw W. 2025. Utilizing systematic digital platforms and instructional design in health communication: a data-driven approach in China's Curriculum. *Journal of Applied Data Sciences* **6(1)**:695–713 DOI 10.47738/jads.v6i1.651.

Giulianelli M, Kok D. 2018. Semi-supervised emotion lexicon expansion with label propagation. ArXiv DOI 10.48550/arXiv.1708.03910.

Gu M, Kwon J, Jeong J, Kwon S-I. 2022. An emotion classification scheme for English text using natural language processing. In: *Proceedings of the 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 1941–1946 DOI 10.1109/ICTC55196.2022.9952880.

Guo J. 2022. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems* **31(1)**:113–126 DOI 10.1515/jisys-2022-0001.

Hariguna T, Sarmini S, Azis A. 2024. Health and socio-demographic risk factors of childhood stunting: assessing the role of factor interactions through the development of an AI predictive model. *Journal of Applied Data Sciences* **5(4)**:2175–2186 DOI 10.47738/jads.v5i4.612.

Henderi H, Asro A, Sulaiman A, Kurniawan T, Dewi D, AlQudah M. 2024. Utilizing sentiment analysis for reflect and improve education in Indonesia. *Journal of Applied Data Sciences* **6(1)**:189–200 DOI 10.47738/jads.v6i1.527.

Huang C, Trabelsi A, Qin X, Farruque N, Zaiane OR. 2019. Seq2Emo for multi-label emotion classification based on latent variable chains transformation. ArXiv DOI 10.48550/arXiv.1911.02147.

Irfan M. 2024a. Exploring the adoption of metaverse platforms in corporations. *International Journal Research on Metaverse* **1(3)**:212–235 DOI 10.47738/ijrm.v1i3.13.

Irfan M. 2024b. Optimizing publisher revenue in digital marketing using decision trees and random forests. *Journal of Digital Market and Digital Currency* **1(3)**:247–266 DOI 10.47738/jdmdc.v1i3.19.

Koufakou A, Grisales D, Costa de Jesus R, Fox O. 2023. Data augmentation for emotion detection in small imbalanced text data. ArXiv DOI 10.48550/arXiv.2310.17015.

Kumar A. 2024. Decentralizing identity with blockchain technology in digital identity management. *Journal of Current Research in Blockchain* **1(3)**:178–189 DOI 10.47738/jcrb.v1i3.22.

Li Z, Chen X, Xie H, Li Q, Tao X. 2020. EmoChannelAttn: exploring emotional construction towards multi-class emotion classification. In: *Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 242–249 DOI 10.1109/WIIAT50758.2020.00036.

Li R, Lin Z, Fu P, Wang W, Shi G. 2019. EmoMix: building an emotion lexicon for compound emotion analysis. *Lecture Notes in Computer Science* **11689**:353–368 DOI 10.1007/978-3-030-22734-0_26.

Mossad N, Mohamed Y, Fares A, Zaky A. 2023. Arabic text sentiment analysis and emotion classification using transformers. In: *Proceedings of the 2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, 131–137.

Pratama SF. 2024. Analyzing the determinants of user satisfaction and continuous usage intention for digital banking platform in Indonesia: a structural equation modeling approach. *Journal of Digital Market and Digital Currency* **1(3)**:267–285 DOI 10.47738/jdmdc.v1i3.21.

Purpura A, Masiero C, Silvello G, Susto GA. 2019. Supervised lexicon extraction for emotion classification. In: *Companion Proceedings of the 2019 World Wide Web Conference*.

Sangsawang T. 2024a. Investigating the determinants of NFT purchase intention: an integrated model combining the theory of planned behavior and technology acceptance model. *Journal of Current Research in Blockchain* **1(3)**:214–241 DOI 10.47738/jcrb.v1i3.19.

Sangsawang T. 2024b. Predicting ad click-through rates in digital marketing with support vector machines. *Journal of Digital Market and Digital Currency* **1(3)**:225–246 DOI 10.47738/jdmdc.v1i3.20.

Saputra JP, Kumar A. 2025. Emotion detection in railway complaints using deep learning and transformer models: a data mining approach to analyzing public sentiment on twitter. *Journal of Digital Society* **1(2)**:109–122 DOI 10.63913/jds.v1i2.6.

Sonu S, Haque R, Hasanuzzaman M, Stynes P, Pathak P. 2022. Identifying emotions in code-mixed Hindi-English tweets. In: *Proceedings of the 2022 International Conference*.

Sukmana HT, Khairani D. 2024. Study of bitcoin market efficiency using runs test and autocorrelation. *Journal of Current Research in Blockchain* **1(1)**:20–32 DOI 10.47738/jcrb.v1i1.9.

Sukmana HT, Oh LK. 2024. Using K-means clustering to enhance digital marketing with flight ticket search patterns. *Journal of Digital Market and Digital Currency* **1(3)**:286–304 DOI 10.47738/jdmdc.v1i3.22.

Tanabe H, Ogawa T, Kobayashi T, Hayashi Y. 2020. Exploiting narrative context and a priori knowledge of categories in textual emotion classification. In: *Proceedings of COLING 2020*, 5535–5540.

Wahyuningsih T, Chen SC. 2024. Determinants of virtual property prices in decentraland an empirical analysis of market dynamics and cryptocurrency influence. *International Journal Research on Metaverse* **1(2)**:157–171 DOI 10.47738/ijrm.v1i2.12.

Wang K, Jing Z, Su Y, Han Y. 2024. Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. ArXiv DOI 10.48550/arXiv.2403.06108.

Wang Y, Sangsawang T, Vipahasna P, Vipahasna K, Watkraw W. 2025. Applying factor analysis to assess employment competitiveness strategies: a data science perspective. *Journal of Applied Data Sciences* **6(1)**:726–741 DOI 10.47738/jads.v6i1.650.

Warrier GS, Arshey M, Jency Rena NM. 2021. Interpretation and classification of emotions on computational social science. *International Journal of Engineering Research & Technology (IJERT)* **10(2)**:594–598.

Xu H, Zhang F, Wang J, Yang W. 2016. Chinese micro-blog emotion classification by exploiting linguistic features and SVMperf. *Lecture Notes in Computer Science* **8710**:221–236 DOI 10.1007/978-3-319-30319-2_10.

Yadav S, Hananto AR. 2024. Comprehensive analysis of Twitter conversations provides insights into dynamic metaverse discourse trends. *International Journal Research on Metaverse* **1(1)**:1–19 DOI 10.47738/ijrm.v1i1.2.

Yadulla AR, Maturi MH, Nadella GS, Satish S. 2024. Volatility comparison of dogecoin and solana using historical price data analysis for enhanced investment strategies. *Journal of Current Research in Blockchain* **1(2)**:91–111 DOI 10.47738/jcrb.v1i2.13.

Zhou D, Wu S, Wang Q, Xie J, Tu Z, Li M. 2020. Emotion classification by jointly learning to lexiconize and classify. In: *Proceedings of COLING 2020*, 3235–3245.