A Hybrid Classification Model Based on BERT for Multi-Class Sentiment Analysis on Twitter

Shofwatul Uyun, Rizqi Praimadi Rosalin, Luky Vianika Sari, Hanny Handayani Sucinta
Informatics Department, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Indonesia.

ARTICLE INFO

Article history:

Received February 04, 2025
Revised March 31, 2025
Accepted April 24, 2025

Keywords:

Sentiment Analysis;
BERT;
LTSM;
CNN;
Emotion Classification

ABSTRACT

Social media is one of the media to convey opinions and sentiments. Sentiment analysis is an important tool for researchers and business people to understand user emotions efficiently and accurately. Choosing the right classification model has a significant impact on sentiment classification performance. However, the diversity of model architectures and training techniques poses its own challenges. In addition, relying on a single classification model often causes noise, bias, data imbalance, and limitations in handling data variations effectively. This study proposes a hybrid classification model where BERT is the baseline. Furthermore, BERT will be hybridized using LSTM, and BERT is hybridized with CNN to improve sentiment analysis on Twitter social media data. The hybrid approach aims to reduce the limitations of a single model classifier by increasing model effectiveness, reducing bias, and optimizing the model on imbalanced data. The following are the steps in this study, data preprocessing, data balancing, tokenization, model training, and performance evaluation. Three models were trained: the baseline BERT model, the BERT-CNN hybrid, and the BERT-LSTM hybrid. Model performance was assessed using accuracy, precision, recall, and F1 score. Experimental results show that the baseline BERT model achieves an accuracy of 91.45%, while BERT-LSTM achieves 91.60%, and BERT-CNN achieves the highest accuracy of 91.80%. However, further analysis is needed to determine whether these improvements are statistically significant and whether the hybrid model offers additional benefits beyond accuracy, such as remembering underrepresented sentiment categories.

Corresponding Author:

Luky Vianika Sari, Informatics Department, Universitas Islam Negri Sunan Kalijaga Yogyakarta, DI Yogyakarta 55281, Indonesia
Email: 23206052010@uin-suka.ac.id

1. INTRODUCTION

The rapid growth of social media has made platforms like Twitter a primary medium for users to express their opinions, emotions, and reactions to various events [1]. With an estimated 415 million monthly active users in 2023 [2], Twitter generates a huge and diverse volume of comments every day [3]. Sentiment analysis is widely used to extract meaningful insights from user-generated content [4], benefiting businesses [5] and researchers [6]. Various deep learning models, such as BERT, LSTM, and CNN, have been employed for sentiment classification [7]. Each model has unique strengths and limitations in handling informal and dynamic social media language [8]. Some common challenges in sentiment classification using a single deep learning model include noise [9], bias [10], imbalance dataset [11], and suboptimal accuracy [12]. These issues are mainly caused by the complexity of social media texts, which include informal language [13], slang [14], and varied sentence structures [15]. One effective approach to addressing this challenge is the implementation of hybrid deep learning models, which integrate multiple architectures to enhance classification performance. Additionally, to mitigate the risk of overfitting caused by data instability, this study employs random sampling techniques. The combination of hybrid models and random sampling is therefore recommended as a strategy to improve classification accuracy and generalization [16]. While BERT remains a state-of-the-art model for sentiment classification [17], the need for more robust and adaptive models continues to increase, highlighting the importance of further advancements in deep learning architectures and optimization techniques [18]. BERT and other deep learning models, such as LSTM and CNN, are widely used for sentiment analysis, each with distinct architectures, advantages, and limitations. BERT, a transformer-based model, excels in capturing contextual relationships between words, making it highly effective for natural language understanding [19]. However, its performance can be affected by the complexity and variability of social media language [20], including slang [21], abbreviations [22], and misspellings [23]. This limitation is particularly significant in sentiment classification, where informal and context-dependent expressions are prevalent. LSTM, on the other hand, specializes in processing sequential data, making it effective in capturing long-term dependencies in text [24]. CNN, known for feature extraction capabilities, has demonstrated strong performance in text classification tasks by identifying key patterns within word sequences [25]. According to Talaat [26], a hybrid model integrating BERT with BiLSTM and BiGRU enhances sentiment classification by capturing deeper textual dependencies. Similarly, Tan et al. [27] demonstrated that a hybrid approach combining RoBERTa, LSTM, BiLSTM, and GRU improves contextual representation and mitigates the issue of data imbalance and found that integrating RoBERTa with LSTM increased sentiment classification accuracy to 91.37% on the Twitter dataset. Considering the strengths and limitations of these models, this study analyzes the performance of hybrid BERT-LSTM and BERT-CNN models for multi-class sentiment classification on social media. The objective is to evaluate whether these hybrid models can address BERT’s limitations in understanding informal language and enhance sentiment classification accuracy. The effectiveness of hybrid models is demonstrated by their ability to improve classification accuracy, precision, recall, and F1 score across different datasets. Other studies conducted by Dang [28] have also shown that sentiment classification performance is greatly influenced by model architecture and ability to handle complex social media texts. Although BERT is widely used for its contextual understanding, it has limitations in effectively classifying short and informal texts [29]. Therefore, hybrid models, such as BERT-CNN and BERT-LSTM, selected for to address this challenge by combining BERT's language representation capabilities with CNN's pattern recognition power and LSTM's sequential processing capabilities, so it is expected resulting in improved sentiment classification performance. The goal of these hybrid models is to reduce classification errors, improve generalization, and achieve higher accuracy compared to using only one model. We propose a hybrid deep learning approach for sentiment classification on social media by integrating BERT, CNN, and LSTM to enhance classification performance. BERT serves as the primary model, leveraging its ability to understand deep contextual relationships within text. CNN acts as a parallel classification model, efficiently capturing spatial patterns and short-range dependencies in sentiment expressions. Meanwhile, LSTM captures long-term dependencies, helping the model understand sequential relationships and improve sentiment classification.

The main contributions of this study are summarized as follows:

- We propose a hybrid BERT-based sentiment classification model, where BERT, CNN, and LSTM work together as classification components to improve sentiment analysis in social media texts.
- BERT and CNN function as classification models, where BERT captures contextual semantics, while CNN classifies sentiment based on spatial feature representations. LSTM further enhances the classification by modeling sequential dependencies, making the system more robust for long and complex texts.
- We systematically analyze the effectiveness of different hybrid architectures, comparing multiple configurations of BERT, CNN, and LSTM to determine the most effective approach for sentiment classification.
- The model performance is evaluated using accuracy, precision, recall, and F1-score, providing insights into how hybrid deep learning improves sentiment classification compared to standalone models.
- To address the issue of imbalanced datasets, we apply random sampling techniques during preprocessing, ensuring a more balanced representation across sentiment classes and improving model generalization

The structure of this paper consists of several sections: Section I presents the introduction, Section II describes the proposed methodology, Section III and Section IV present the experimental results and discussion, and Section V provides conclusions and future research directions, followed by acknowledgments.

2. MATERIAL AND METHODS

2.1. Research Dataset

The dataset used in this study is sourced from Hugging Face (dair-ai/emotion), consisting of 20,000 samples, divided into three subsets: 16,000 for training, 2,000 for validation, and 2,000 for testing. This dataset contains text labeled into six emotion categories, as shown in Table 1.

Table 1. label of the emotional category data set

Label	Emotion
0	Sadness
1	Joy
2	Love
3	Anger
4	Fear
5	Surprise

Table 1 presents the labels of the emotional category dataset used in this study. The dataset was split using the *stratified sampling* method to ensure that each subset maintains a balanced distribution of emotion categories [30]. This approach prevents any particular category from being overrepresented, which could introduce bias in model training [31]. By maintaining an even distribution, the model can learn the patterns of each emotion more effectively and improve its generalization to new data [32].

2.2. Research Flow

This study proposes the use of a Transformer-based model, namely BERT, as well as a hybrid model that combines BERT with LSTM and BERT with CNN for sentiment analysis in English comment text on Twitter. In general, this study involves five main stages in building and evaluating the sentiment analysis model, including pre-processing of data, data cleansing, data balancing, tokenization of BERT, initialization of the model, and evaluation of the model depicted in Fig. 1 [33].

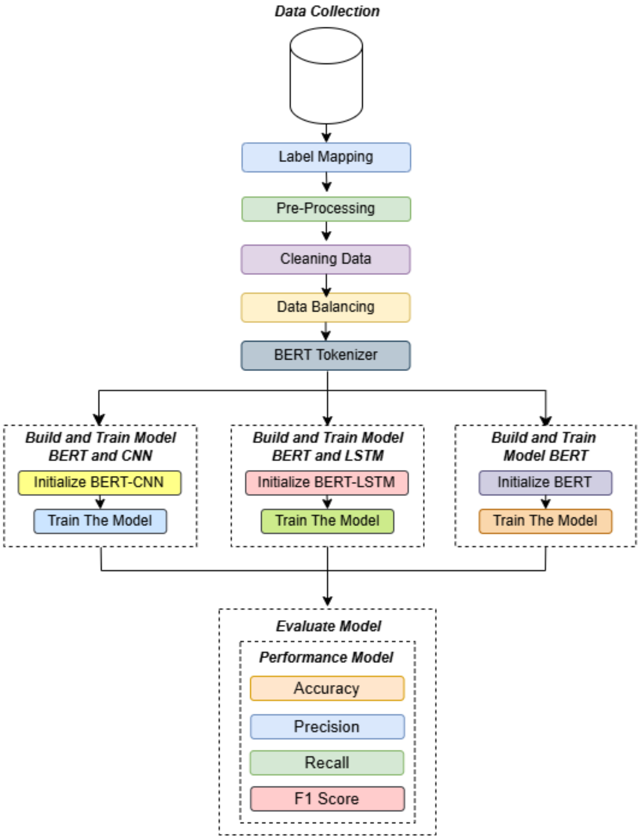

Fig. 1. Visualisation of the research flow

Fig. 1 illustrates a structured workflow for sentiment analysis using various BERT-based models, beginning with data collection, followed by label mapping to standardize emotion categories. The pre-processing stage refines textual data by removing irrelevant elements, while data balancing ensures fair model performance. The BERT tokenizer then converts text into numerical representations. The training phase explores three approaches: BERT-CNN, which integrates convolutional layers for feature extraction; BERT-LSTM, which captures sequential dependencies using a Long Short-Term Memory (LSTM) network; and a baseline BERT model without additional layers. Model evaluation is conducted using key performance metrics, including accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment of each model's effectiveness in sentiment classification.

2.3. Pre-processing

In the pre-processing stage, text cleanup is done by removing irrelevant elements such as URLs, hashtags, and mentions to prevent interference in sentiment analysis [34]. Emojis were converted to text [35], and slang was normalized to retain their meaning for models [36]. Stopwords were removed using the NLTK list of stopwords while considering their relevance in emotion analysis [37]. Furthermore, text normalization is applied by converting uppercase to lowercase and removing special characters. The final step is tokenization using the WordPiece tokenizer BERT, which breaks text into subword tokens to help the model understand rare words [38], while the pre-processing technique seen in Table 2.

Table 2. Pre-processing Techniques

Step	Technique Used	Purpose
URL, mention, and hashtag removal	Regex pattern matching	Remove irrelevant elements
Emoji conversion to text	Emoji-to-text mapping	Preserve emotional meaning in text
Slang normalization	Slang dictionary replacement	Convert informal words into a standard format
Stopword removal	NLTK stopword filtering	Remove common words that do not carry specific meaning
Text normalization	Lowercasing & special character removal	Ensure a consistent text format
Tokenization	WordPiece tokenizer (BERT)	Convert text into tokens suitable for the model

Table 2 outlines the preprocessing steps to improve data quality for emotion classification. Irrelevant elements like URLs, mentions, and hashtags are removed, while emojis are converted into text to retain their meaning. Slang words are replaced with standard terms, and common words that do not add value are filtered out. Text is then normalized by converting it to lowercase and removing special characters. Finally, tokenization breaks the text into smaller units to make it suitable for model training. These steps ensure that the data is clean, consistent, and ready for accurate emotion classification.

2.4. Data Balancing

Data balancing is necessary to ensure that the model does not favor majority classes while ignoring minority classes, thereby improving its ability to generalize across all categories of emotions. An unbalanced dataset can lead to skewed predictions [39], where models perform well on emotions that occur frequently but struggle to accurately classify underrepresented emotions. Therefore, the distribution of the initial classes in this study dataset shows a significant imbalance, where the likes and sorrows classes have a much larger number of samples compared to other classes, especially the surprises that have the smallest sample count. This imbalance can lead to bias in model training [40], where models tend to be more accurate in recognizing dominant emotions but less accurate in identifying emotions with fewer samples. The following is a visualization of the class distribution in the training data before implementing the balancing process depicted in Fig. 2.

Fig. 2 illustrates the distribution of data before data balancing, it can be seen that each data is unbalanced. Therefore, to overcome this imbalance, several methods were applied, including calculating class weights using compute_class_weight, which was incorporated into the loss function to reduce the effect of the imbalance [41]. In addition, a Weighted Random Sampler is used to ensure that each batch maintains a more balanced class distribution during training [42].

2.5. BERT Tokenizer

After the preprocessing and data balancing stages, the next step is tokenization using the BERT Tokenizer (*bert-base-uncased*) to convert text into a format that can be understood by the BERT model. The tokenization process begins by determining the maximum token length through an analysis of the token length distribution from the first 1000 samples. Based on this analysis, the *MAX_LEN* value is set to ensure an optimal token length, capturing essential information without excessive length. Tokenization is performed using the batch_encode_plus function, which produces two main outputs: (1) *input_ids*, a numerical representation of the text based on BERT's vocabulary, and (2) *attention_mask*, an indicator of tokens that should be attended to by the model. Additionally, labels are converted into tensors for model processing. The following is an example of tokenization for the *cyberbullying* sentence: "You are so ugly and stupid!" depicted in Table 3.

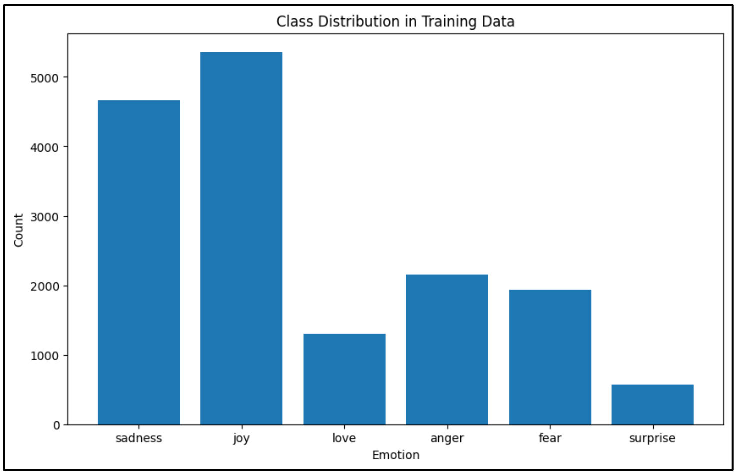

Fig. 2. Visualization of Class Distribution in Training Data

Table 3. Tokenization Process

Original Text	BERT Tokenization	input_ids
You are so ugly and stupid!	[‘you’, ‘are’, ‘so’, ‘ugly’, ‘and’, ‘stupid’, ‘!’]	[2017, 2024, 2061, 6372, 1998, 5236, 999]

Table 3 shows a tokenization process that ensures that cyberbullying text is represented in a BERT-understandable format, thus enabling the model to capture semantic relationships between words and improve accuracy in detecting cyberbullying speech.

2.6. Model Building and Training

2.6.1. Baseline Model BERT

The BERT (Bidirectional Encoder Representations from Transformers) model is a transformer-based model that uses a two-way approach to understand the context of the text [43]. In this study, the BERT (bert-base-uncased) model was used to understand the context of the text before it was classified into six categories of emotions. The model structure consists of a BERT layer, followed by a dropout to prevent overfitting, and then a fully connected layer with activation and normalization of ReLU (LayerNorm) for more stable training. A BERT freezing option is also available so that only the classification section is trained [44]. Although it is powerful in understanding the meaning of words, this model is less than optimal in capturing the order of words in long texts [45]. Mathematically, the BERT model is calculated using the following formula depicted in formula (1):

$$Attention\ (Q, K, V)\ =\ softmax\left(\frac{QK^T}{\sqrt{dk}}\right) \tag{1}$$

Here, Q, K, and V represent the query, key, and value, respectively, while dk is the dimension of the key. The softmax function is used to give weight to words based on their relevance to other words. Thus, BERT is able to understand the global context in the text, which makes it very effective in the task of emotion analysis.

2.6.2. Hybrid Model BERT and LSTM

Hybrid Model combining BERT with BiDirectional LSTM is done to produce a better model in understanding word sequences [46]. After BERT converts text into vectors, the results are processed by a two-layer LSTM, which reads the text from two directions to capture deeper meaning. *Dropout layers* are used to avoid overfitting before the results enter the fully connected layer for emotion classification. This model is superior in understanding long-term context, but requires more computation [47].

2.6.3. Hybrid Model BERT and CNN

A hybrid BERT model with CNN is performed to extract patterns in text [48]. After BERT generates word vectors, the results are processed by *convolutional layers* with different filter sizes (2, 3, and 4) to capture different patterns. The *max-pooling layer* helps select important features before being passed to the *fully* *connected layer* for classification. This model is faster and more efficient for short texts, but is less optimal in understanding word relationships in long sentences [49].

2.7. Evaluation

At this stage, three models, namely BERT, BERT-LSTM and BERT-CNN, were evaluated using validation data to measure the performance of emotional sentiment analysis [50]. The evaluation was conducted using acrobatic metrics, precision, memory, and F1 scores to get a more complete picture of each model's ability to analyze emotions in Twitter comments [51]. However, given the class imbalances in the dataset, accuracy alone may not be a reliable metric, as it can be misleading if one class dominates the dataset. Therefore, macro average F1 scores are prioritized to ensure a balanced evaluation across all categories of emotions. To reinforce the findings, statistical significance testing was performed to assess whether the differences in performance between models were meaningful. Finally, a comparative analysis is carried out to determine which model shows the best ability to recognize and classify the emotions formulated in formula (2)-(5).

$$Accuracy = \frac{TP - TN}{TP + TN + FP + FN} * 100\% \tag{2}$$

$$Precision = \frac{TP}{TP + FP} * 100\% \tag{3}$$

$$Recall = \frac{TP}{TP + FN} * 100\% \tag{4}$$

$$F1\,Score = \frac{2x(PrecisionxRecall)}{(Precision + Recall)} * 100\% \tag{5}$$

3. RESULTS AND DISCUSSION

3.1. Pre-processing

The dataset used in this study is the HuggingFace dataset (https://huggingface.co/datasets/dair-ai/emotion/), which contains Twitter comment data with six categories of emotions. The pre-processing stages include the transformation of the text into lowercase letters, the removal of unimportant words (stop words), the simplification of the word form using lemmatization, and the removal of numbers and punctuation depicted in Table 4.

Table 4. Pre-processing data results

Before Preprocessing	After Preprocessing
"i didnt feel humiliated"	"didnt feel humiliated"
"im grabbing a minute to post i feel greedy wrong"	"im grabbing minute post feel greedy wrong"
"i feel like a faithful servant"	"feel like faithful servant"
"im updating my blog because i feel shitty"	"im updating blog feel shitty"

Table 4 illustrates the impact of the pre-processing steps on the dataset by showing examples of text before and after processing. The removal of stop words, such as "I" and "a," helps eliminate common but non-informative words, while lemmatization ensures that words retain their base forms without altering their meanings. Additionally, transforming text into lowercase ensures uniformity, reducing inconsistencies caused by case variations. The elimination of numbers and punctuation further refines the text, making it more suitable for tokenization and model training. These steps collectively enhance data quality, ensuring that only meaningful linguistic patterns are retained for sentiment classification.

3.2. Data Balancing

Class imbalance is a major challenge in this study, particularly in the *Surprise* and *Love* categories, which have significantly fewer samples compared to other categories. To address this issue, the *class weighting* method was applied, where class weights are calculated proportionally based on the sample distribution in the training data. Here are the calculated weights of classes presented in Table 5.

Table 5 present the higher weights indicate that a class has fewer samples, prompting the model to pay more attention to it during training. Surprise has the highest weight (4.6620), indicating that it is the least represented category in the dataset, followed by *Love* (2.0450). In contrast, *Sadness* (0.5715) and *Joy* (0.4973) have the lowest weights due to their larger sample sizes. The class weighting method improves the model's sensitivity to underrepresented classes without reducing the number of majority-class samples. However, this approach does not introduce real data variation for minority classes, posing a potential risk of bias in generalization.

Table 5. The Sum of The Weights of Each Class

Emotion Class	Class Weight
Sadness	0.5715
Joy	0.4973
Love	20.450
Anger	12.351
Fear	13.767
Surprise	46.620

3.3. Main Findings of the Present Study

This study evaluates three BERT-based models, namely BERT, BERT+LSTM, and BERT+CNN, in the task of classifying emotions based on six categories: Sadness, Joy, Love, Anger, Fear, and Surprise. The evaluation results show that BERT+LSTM has the highest accuracy of 91.60%, followed by BERT+CNN (91.80%), and BERT (91.45%). The following table summarizes the evaluation results based on test loss, accuracy, and macro average precision, recall, and F1-score for each model depicted in Table 6:

Table 6. Evaluation Results of BERT-Based Models for Emotion Classification

Model	Test Loss	Test Accuracy
BERT	0.1487	91.45%
BERT+LSTM	0.1446	91.60%
BERT+CNN	0.1569	91.80%

Table 6 can be seen that BERT+CNN has the highest accuracy (91.80%), although its test loss is higher than other models. This shows that BERT+CNN is superior in predicting categories correctly, but may experience slight overfitting compared to other models. In addition, the macro average precision, recall, and F1-score values are relatively balanced across all models, indicating that the three models are able to maintain consistent performance across emotion categories.

Fig. 3 show the validation accuracy graph further supports the evaluation results by illustrating the learning trends of the three models across training epochs. Initially, all models exhibit rapid accuracy improvement, with BERT+LSTM showing the steepest increase. However, after reaching its peak at the second epoch, BERT+LSTM experiences a slight decline, indicating potential overfitting. In contrast, BERT+CNN and BERT demonstrate more stable accuracy trends, with BERT achieving the highest validation accuracy in the final epoch. These findings reinforce that while BERT+LSTM enhances recall for minority categories, its generalization capability may be slightly weaker. Meanwhile, BERT+CNN offers a balance between accuracy and stability, making it a suitable choice for robust emotion classification.

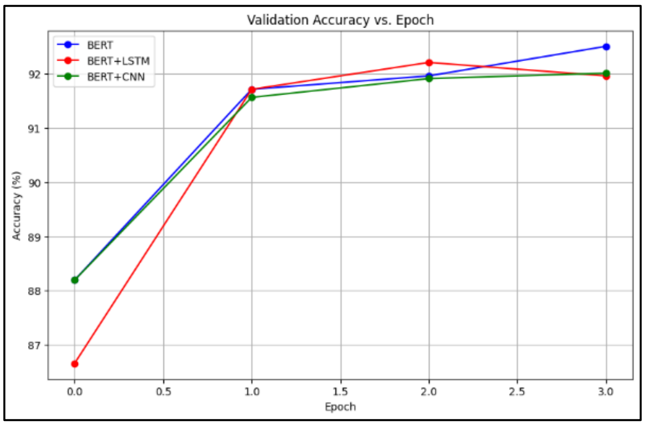

Fig. 3. Visualization of Validation Accuracy vs. Epoch

3.4. Comparison with Previous Studies

The results of this study showed that BERT-CNN had the highest accuracy of 91.80%, followed by BERT-LSTM (91.60%) and baseline BERT (91.45%). Compared to the previous study, the

RoBERTa+BiLSTM method in SST-2 reached 93.2%, while CNN+BiLSTM in Amazon Product Reviews reached 94.5%. Although the results of this study were slightly lower, the difference may be due to variations in the datasets and pre-processing techniques used. The advantage of this approach lies in the utilization of BERT's contextual representations, which are more effective than the doc2vec method (92.3%). However, unlike some studies that applied attention or transfer learning mechanisms, this study still relied on the basic architecture of BERT without additional accuracy-enhancing techniques. In addition, the use of a single dataset without cross-dataset validation can limit the generalization of the model. The following is a comparison of the research conducted by the current researcher with the previous research described in Table 7.

Table 7. Previous Research Comparison

Researcher (Year)	**Methods**	**Dataset**	**Best Accuracy**
2022 [52]	doc2vec + Deep Learning + Attention Mechanism	IMDB Sentiment Dataset	92.3%
2024 [53]	CNN + BiLSTM	Amazon Product Reviews	94.5%
2024 [54]	RoBERTa + BiLSTM	SST-2 (Stanford Sentiment Treebank)	93.2%
2024 [55]	Transfer Learning + Dependency Parsing	Laptop and Restaurant Reviews (SemEval 2014)	89.6%
Purposed Framework	BERT	HuggingFace	91.45%
	BERT + LSTM		91.60%
	BERT + CNN		91.80%

Table 7 show competitive values, further exploration of fine-tuning, hyperparameter optimization, or integration of attention mechanisms could improve the accuracy and generalization of the model. In addition, implementing cross-dataset validation could provide a more comprehensive evaluation of the model performance, ensuring its robustness across domains. Future research could also explore hybrid approaches, such as combining transformer-based architectures with advanced feature selection techniques or leveraging ensemble learning to improve classification performance. These improvements would help reduce dataset-specific biases and further optimize the model for real-world applications.

3.5. Implication and Explanation of Findings

The graphs in Fig. 4 illustrate the progression of Training Loss and Validation Loss throughout the training process for the three models: BERT, BERT+LSTM, and BERT+CNN. These visualizations provide insight into the learning dynamics of each model, highlighting their convergence behavior, potential overfitting or underfitting issues, and overall stability during training. By analyzing these loss curves, we can assess how effectively each model generalizes to unseen data and determine which architecture achieves better optimization. Additionally, comparing the loss trends among the models allows us to evaluate the impact of incorporating LSTM and CNN layers on the BERT framework, particularly in terms of learning efficiency and model complexity. Here is a visualization of the image seen in Fig. 4.

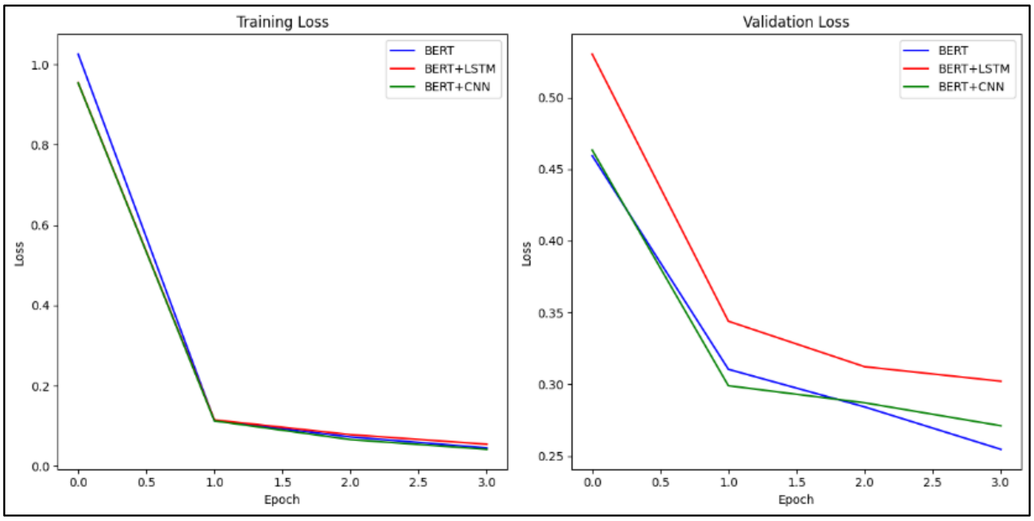

Fig. 4. Visualization of Model Performance in Training and Validation

Fig. 4 shows that the Training Loss decreases rapidly at the beginning and stabilizes, indicating that the model has successfully learned. However, in the Validation Loss, BERT+CNN has the smallest value, indicating better generalization than BERT+LSTM, which has a higher loss and is potentially overfitting. This shows that BERT+CNN is more stable in validation and more effective in handling new data. Therefore, if the main goal is to get a more stable model with better generalization, BERT+CNN is the optimal choice. Conversely, if using BERT+LSTM, hyperparameter adjustments are needed, such as adding dropout or data augmentation, to reduce overfitting.

Fig. 5 shows the performance of the BERT+CNN model in classifying six emotion categories based on precision, recall, and F1-score metrics. The Sadness and Joy categories have high and balanced scores across all metrics, indicating that the model can recognize these two emotions well. In contrast, the Surprise category has lower precision than recall, indicating that the model often misclassifies this emotion as another category. The Love category also shows a significant difference between precision and recall, which could be due to limited data or similarity to other categories. Overall, these results show that while BERT+CNN excels in general emotion classification, its performance can vary depending on the amount of data and the complexity of each emotion category.

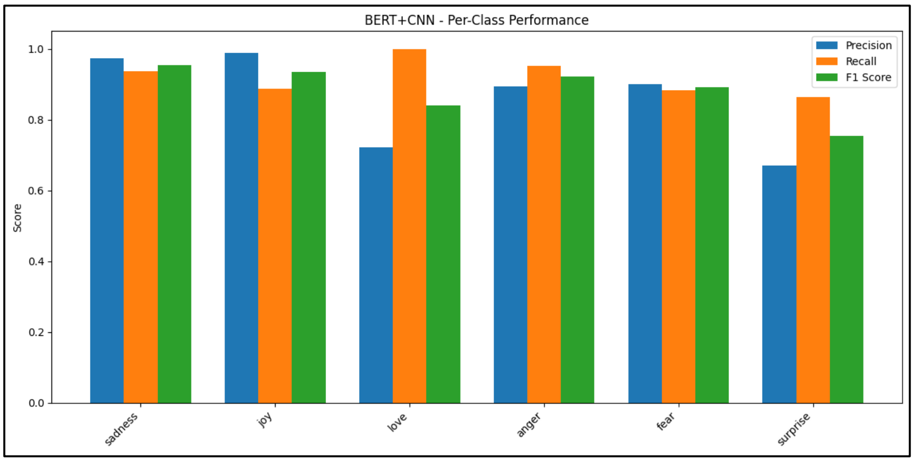

Fig. 5. Visualization of Model BERT and CNN Per-Class Performance

3.6. Strengths and Limitations

This study has several advantages, especially in the use of a BERT-based model that has been proven effective in understanding linguistic context in emotion classification tasks. The combination of BERT with additional architectures such as LSTM and CNN also provides insight into how sequential processing and spatial features can improve classification performance. In addition, the evaluation of metrics including precision, recall, and F1-score allows for a more in-depth analysis of model performance across different emotion categories. However, this study also has several limitations. One of them is the class imbalance in the dataset, which has an impact on model generalization. Categories with fewer data, such as Surprise and Love, tend to have lower precision and recall scores than more dominant categories, such as Sadness and Joy. This indicates a potential model bias in classifying low-frequency categories, which can decrease overall accuracy and affect the interpretation of evaluation metrics, especially F1-score. In addition, this study only uses one dataset from HuggingFace without external validation or cross-dataset testing, which may limit the model's generalization to data from different sources. Without validation on other datasets, it is difficult to ascertain whether the model can maintain its performance in real-world environments with a wider variety of languages and emotional expressions. Therefore, in real-world applications, the model may need to be fine-tuned or retrained on more representative datasets to improve its reliability across different communication and cultural contexts.

3.7. Computational Efficiency and Feasibility

This study compares the computational efficiency of BERT, BERT+LSTM, and BERT+CNN in emotion classification. BERT+CNN has the fastest execution due to more efficient feature extraction, while BERT+LSTM is better at capturing sequential context but requires more time and memory. Standard BERT is balanced in performance and efficiency. For resource-constrained environments, BERT+CNN is more feasible, and optimizations such as pruning and quantization can reduce the computational burden without significant loss of accuracy. The following is a description of computational efficiency and feasibility described in Table 8.

Table 8. Computational Cost Comparison of BERT-Based Models

Model	Execution Time (s)	Memory Usage (GB)	Strengths	Weaknesses
BERT	1.5	4.8	Strong baseline, balanced performance	Lacks optimized feature extraction
BERT+LSTM	2.1	6.5	Captures sequential dependencies well	Slow due to sequential processing
BERT+CNN	1.2	5.2	Fastest inference, efficient classification	Potential loss of sequential context

Table 8 shows that BERT+CNN has an advantage in computational efficiency, offering the fastest execution time and relatively lower memory usage, making it a suitable choice for real-time applications. In contrast, BERT+LSTM excels in capturing long-range dependencies, which can improve contextual understanding but comes at the cost of increased execution time and memory consumption due to its sequential nature. The standard BERT model, while providing a balanced performance, lacks specialized mechanisms for feature extraction, which may limit its effectiveness in capturing complex text patterns. These findings highlight the trade-offs between speed, resource consumption, and contextual understanding in choosing the most appropriate model for sentiment classification tasks.

4. CONCLUSION

This study evaluates the performance of BERT, BERT-LSTM, and BERT-CNN for sentiment classification using a dataset from HuggingFace. The experimental results indicate that BERT-CNN achieved the highest accuracy (91.80%), followed by BERT-LSTM (91.60%), and the baseline BERT model (91.45%). These findings highlight the potential of CNN-based architectures in enhancing classification from contextual embeddings, contributing to the theoretical understanding of hybrid deep learning models in sentiment analysis. Despite its promising results, this study has several limitations. The class imbalance in the dataset may have introduced bias, affecting the generalizability of the model. Additionally, the use of a single dataset without external validation limits the robustness of the findings. Future studies should explore cross-dataset validation, data augmentation techniques, or ensemble learning methods to improve model adaptability and accuracy. This research contributes to the field by demonstrating the comparative advantages of hybrid BERT architectures in sentiment classification. The insights gained from this study can inform future improvements in deep learning-based text classification, particularly in optimizing computational efficiency without compromising accuracy. For future research, we suggest investigating fine-tuning techniques, attention mechanisms, and lightweight transformer models to enhance both accuracy and efficiency. Furthermore, expanding the study to include multilingual datasets and real-world applications will strengthen the model's practicality and applicability across diverse domains.

REFERENCES

[1] K. Chakraborty, S. Bhattacharyya, R. Bag, "A Survey of Sentiment Analysis from Social Media Data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450-464, 2020, https://doi.org/10.1109/TCSS.2019.2956957.

[2] A. Yadav, M. Alahmar, A. Singh, K. Sharma, R. Agrawal, C. B. Sharma, "Analyzing User Behavior in Social Media through Big Data Analytics," *IEEE International Conference on ICT in Business Industry & Government (ICTBIG),* pp. 1–5, 2023, https://doi.org/10.1109/ICTBIG59752.2023.10456112.

[3] Simon Kemp, "Twitter Users, Stats, Data & Trends." [Online]. Available: https://datareportal-com.translate.goog/essential-twitter-stats?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc.

[4] G. Rasool, A. Pathania, "Reading between the lines: untwining online user-generated content using sentiment analysis," *J. Res. Interact. Mark*, vol. 15, no. 3, pp. 401–418, 2021, https://doi.org/10.1108/JRIM-03-2020-0045.

[5] A. R. Abas, I. Elhenawy, M. Zidan, M. Othman,"Aspect-based sentiment analysis on social media comments (twitter): the attributes of service robots in the hotel and restaurant industry," *J. Qual. Assur. Hosp. Tour*, pp. 1–26, 2024, https://doi.org/10.1080/1528008X.2024.2386590.

[6] C. J. Hartmann, M. Heitmann, C. Siebert, "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Mark*, vol. 40, no. 1, pp. 75–87, 2023, https://doi.org/10.1016/j.ijresmar.2022.05.005.

[7] M. Abas, A. R., Elhenawy, I., Zidan, M., & Othman, "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text," *Comput. Mater. Contin*, vol. 71, no. 2, 2022, https://doi.org/10.32604/cmc.2022.021671.

[8] J. Hartmann, M. Heitmann, C. Sieber, "Usability evaluation of a nursing information system by applying cognitive walkthrough method," *Int. J. Med. Inform*, vol. 152, p. 104459, 2021, https://doi.org/10.1016/j.ijmedinf.2021.104459.

[9] N. Raghunathan, K. Saravanakumar, "Challenges and issues in sentiment analysis: A comprehensive survey," *IEEE Access*, vol. 11, 2023, https://doi.org/69626-69642.

[10] M. Wankhade, A. C. S. Rao, C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev*, vol. 55, no. 7, pp. 5731–5780, 2022, https://doi.org/10.1007/s10462-022-10144-1.

[11] R. Obiedat *et al.,* "Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, https://doi.org/10.1109/ACCESS.2022.3149482.

[12] J. Hartmann *et al.,* "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Mark.*, vol. 40, no. 1, pp. 75–87, 2023, https://doi.org/10.1016/j.ijresmar.2022.05.005.

[13] M. F. R. A. Bakar, N. Idris, L. Shuib, N. Khamis, "Sentiment analysis of noisy Malay text: state of art, challenges and future work," *IEEE Access*, vol. 8, pp. 24687–24696, 2020, https://doi.org/10.1109/ACCESS.2020.2968955.

[14] L. R. Sultan, "An Enhanced Emotion Classification Scheme for Twits Based on Deep Learning Approach," *Rev. d'Intelligence Artif*, vol. 37, no. 5, p. 1203, 2023, https://doi.org/10.18280/ria.370512.

[15] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, "Deep learning--based text classification: a comprehensive review," *ACM Comput. Surv*, vol. 54, no. 3, pp. 1–40, 2021, https://doi.org/10.1145/3439726.

[16] M. Celik, O. Inik, "Development of hybrid models based on deep learning and optimized machine learning algorithms for brain tumor Multi-Classification," *Expert Syst. Appl*, no. 122159, p. 238, 2024, https://doi.org/10.1016/j.eswa.2023.122159.

[17] J. H Joloudari *et al.,* "BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets," *Soc. Netw.Anal. Min*, vol. 13, no. 1, p. 99, 2023, https://doi.org/10.1007/s13278-023-01102-y.

[18] W. X. Zhao, J. Liu, R. Ren, J. R. Wenn, "Dense text retrieval based on pretrained language models: A survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 1–60, 2024, https://doi.org/10.1145/3637870.

[19] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, B. Alshemaimri, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 6, pp. 1–49, 2025, https://doi.org/10.1007/s10462-025-11162-5.

[20] A. S. Alammary, "Investigating the impact of pretraining corpora on the performance of Arabic BERT models," *J. Supercomput.*, vol. 81, no. 1, p. 187, 2025, https://doi.org/10.1007/s11227-024-06698-2.

[21] M. Khazeni, M. Heydari, A. Albadvi, "Persian Slang Text Conversion to Formal and Deep Learning of Persian Short Texts on Social Media for Sentiment Classification," *arXiv Prepr. arXiv*, 2024, https://doi.org/10.22061/jecei.2024.10745.731.

[22] F. Miletić, S. S. im Walde, "A systematic search for compound semantics in pretrained BERT architectures," *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguist*, pp. 1499–1512, 2023, https://doi.org/10.18653/v1/2023.eacl-main.110.

[23] G. Sperduti, A. Moreo, "Misspellings in Natural Language Processing: A survey," *arXiv Prepr. arXiv*, 2025, https://doi.org/10.48550/arXiv.2501.16836.

[24] D. Tsirmpas, I. Gkionis, G. T. Papadopoulos, I. Mademlis, "Neural natural language processing for long texts: A survey on classification and summarization," *Eng. Appl. Artif. Intell*, p. 133, 2024, https://doi.org/10.1016/j.engappai.2024.108231.

[25] Y. He, "BERT-CNN-BiLSTM: A Hybrid Deep Learning Model for Accurate Sentiment Analysis," *IEEE 5th Int. Conf. Power, Intell. Comput. Syst.*, pp. 921–926, 2023, https://doi.org/10.1109/ICPICS58376.2023.10235335.

[26] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J. Big Data*, vol. 10, no. 1, 2023, https://doi.org/10.1186/s40537-023-00781-w.

[27] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, no. July, pp. 103694–103704, 2022, https://doi.org/10.1109/ACCESS.2022.3210182.

[28] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, 2021, https://doi.org/10.1155/2021/9986920.

[29] F. A. Acheampong, H. Nunoo-Mensah, W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artif. Intell. Rev*, vol. 54, no. 8, pp. 5789–5829, 2021, https://doi.org/10.1007/s10462-021-09958-2.

[30] A. Onan, K. F. Balbal, "Improving Turkish text sentiment classification through task-specific and universal transformations: an ensemble data augmentation approach," *IEEE Access*, vol. 12, pp. 4413–4458, 2024, https://doi.org/10.1109/ACCESS.2024.3349971.

[31] M. Shah, N. Sureja, "A comprehensive review of bias in deep learning models: Methods, impacts, and future directions," *Arch. Comput. Methods Eng*, vol. 32, no. 1, pp. 255–267, 2025, https://doi.org/10.1007/s11831-024-10134-2.

[32] J Wang *et al.,* "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng*, vol. 35, no. 8, pp. 8052–8072, 2022, https://doi.org/10.1109/TKDE.2022.3178128.

[33] S. Ramakrishnan and L. D. Dhinesh Babu, ""Enhancing Twitter Sentiment Analysis using Attention-based BiLSTM and BERT Embedding," *9th Int. Conf. Smart Comput. Commun*, pp. 36–40, 2023, https://doi.org/10.1109/ICSCC59169.2023.10335010.

[34] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, pp. 509–553, 2023, https://doi.org/10.1017/S1351324922000213.

[35] D. Muhamediyeva, N. Niyozmatova, N. Turgunova, S. Ungalov, N. Almuradova, "Classification of Emoji in Text

Documents of Users in Social Networks Using Machine Learning," *IEEE. 2025 6th Int. Conf. Mob. Comput. Sustain. Informatics*, pp. 1491–1496, 2025, https://doi.org/10.1109/ICMCSI64620.2025.10883250.

[36] N. Merayo, "Applying machine learning to assess emotional reactions to video game content streamed on Spanish Twitch channels," *Comput. Speech Lang*, p. 88, 2024, https://doi.org/10.1016/j.csl.2024.101651.

[37] D. Bino, V. Dhanalakshmi, P. K. Udupi, "Sentiment Analysis and Machine Learning for Tourism Feedback Data Analysis: An Overview of Trends, Techniques, and Applications," *AI Technol. Pers. Sustain. Tour*, pp. 215-252., 2025, https://doi.org/10.4018/979-8-3693-5678-4.ch009.

[38] P. Lauren, "Improving subword embeddings in large language models using morphological information," *Artif. Intell. Mach. Learn. Convolutional Neural Networks Large Lang. Model*, vol. 1, p. 333, 2024, https://doi.org/10.1515/9783111344126-015.

[39] J. Li, Y. Tao, H. Cong, E. Zhu, T. Cai, "Predicting liver cancers using skewed epidemiological data," *Artif. Intell. Med*, vol. 124, no. 102234, 2022, https://doi.org/10.1016/j.artmed.2021.102234.

[40] A. R. Chłopowiec *et al.,* "Counteracting data bias and class imbalance—towards a useful and reliable retinal disease recognition system," *Diagnostics*, vol. 13, no. 11, p. 1904, 2023, https://doi.org/10.3390/diagnostics13111904.

[41] I. Araf, A. Idri, I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev*, vol. 54, no. 4, p. 80, 2024, https://doi.org/10.1007/s10462-023-10652-8.

[42] G Citovsky *et al.,* "Batch active learning at scale.," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 11933–11944, 2021, https://proceedings.neurips.cc/paper/2021/hash/64254db8396e404d9223914a0bd355d2-Abstract.html.

[43] D. Alomari, I. Ahmad, "Exploring Character Trigrams for Robust Arabic Text Classification: A Comparative Analysis in the Face of Vocabulary Expansion and Misspelled Words," *IEEE Access*, vol. 12, pp. 57103–57116, 2024, https://doi.org/10.1109/ACCESS.2024.3390048.

[44] B. Elizalde, S. Deshmukh, M. Al Ismail, H. Wang, "Clap learning audio concepts from natural language supervision," *ICASSP 2023-2023 IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 1–5, 2023, https://doi.org/10.1109/ICASSP49357.2023.10095889.

[45] M. Apidianaki, "From word types to tokens and back: A survey of approaches to word meaning representation and interpretation," *Comput. Linguist*, vol. 49, no. 2, pp. 465–523, 2023, https://doi.org/10.1162/coli_a_00474.

[46] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, K. M. Lim, "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, https://doi.org/10.1109/ACCESS.2022.3162614.

[47] A. K. Kalusivalingam, A. Sharma, N. Patel, V. Singh, "Leveraging BERT and LSTM for Enhanced Natural Language Processing in Clinical Data Analysis," *Int. J. AI ML*, vol. 2, no. 3, 2021, https://doi.org/10.1177/14727978251322656.

[48] X. Chen, P. Cong, S. Lv, "A long-text classification method of Chinese news based on BERT and CNN," *IEEE Access*, no. 10, pp. 34046–34057, 2022, https://doi.org/10.1109/ACCESS.2022.3162614.

[49] S. Chen, "Semantic relationship extraction of English long sentences and quality optimization of machine translation based on BERT model," *J. Comput. Methods Sci. Eng,* p. 14727978251322656, 2025, https://doi.org/14727978251322656.

[50] S. Almlawi, J. Fang, J. LiEnhancing, "Sentiment Analysis Using MCNN-BRNN Model with BERT," *3rd Int. Conf. Electron. Inf. Eng. Comput. Commun*, pp. 574–579, 2023, https://doi.org/10.1109/EIECC60864.2023.10456641.

[51] T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala, N. V. Donthiboina, "Exploring the Effectiveness of BERT for Sentiment Analysis on Large-Scale Social Media Data," *3rd Int. Conf. Intell. Technol*, pp. 1–4, 2023, https://doi.org/10.1109/CONIT59222.2023.10205600.

[52] Y. Zhou, Q. Zhang, D. Wang, and X. Gu, "Text Sentiment Analysis Based on a New Hybrid Network Model," *Genet. Res. (Camb),* p. 6774320, 2022, https://doi.org/10.1155/2022/6774320.

[53] S. Susandri, S. Defit, and M. Tajuddin, "Enhancing Text Sentiment Classification with Hybrid CNN-BiLSTM Model on WhatsApp Group," *J. Adv. Inf. Technol*, vol. 15, no. 3, pp. 355–363, 2024, https://doi.org/10.12720/jait.15.3.355-363.

[54] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis," *arXiv preprint arXiv:2406.00367,* 2024, [Online]. Available: http://arxiv.org/abs/2406.00367.

[55] G. Negi, R. Sarkar, O. Zayed, and P. Buitelaar, "A Hybrid Approach To Aspect Based Sentiment Analysis Using Transfer Learning," *2024 Jt. Int. Conf. Comput. Linguist. Lang. Resour. Eval. Lr. 2024 - Main Conf. Proc,* pp. 647–658, 2024, https://doi.org/10.48550/arXiv.2403.17254.