CNN-LSTM Architecture for Multi-Task Sentiment and Emotion Classification on Large-Scale Indonesian TikTok Application Reviews

Wahyu Fajar Setiawan[1], Afif Amirullah[2], Ilham Putra Ariatama[3], and Ratih Nur Esti Anggraini[4*]
[1−4]Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology,
Sepuluh Nopember Institute of Technology (ITS)
Surabaya, Indonesia 60117
Email: [1]6025241020@student.its.ac.id, [2]6025241048@student.its.ac.id,
[3]6025241001@student.its.ac.id, [4]ratih_nea@if.its.ac.id

***Abstract*—Sentiment and emotion analysis of mobile application reviews has attracted significant attention as a means to understand users' perceptions and experiences. The research proposes a novel Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for multi-task sentiment and emotion classification on Indonesian TikTok application reviews. A large-scale corpus consisting of 500,000 reviews is collected from the Google Play Store and preprocessed through cleaning, normalization, tokenization, stopword removal, and stemming. Sentiment labels (positive, negative, and neutral) are assigned using a lexicon-based approach, while emotion labels are annotated through emoji analysis and word matching based on five basic emotions: anger, fear, happiness, love, and sadness. The proposed CNN-LSTM model is evaluated against a hybrid Bidirectional Encoder Representations from Transformers – Convolutional Neural Network (BERT-CNN) architecture. Experimental results show that the CNN-LSTM model outperforms the BERT-CNN model, achieving an accuracy of 91.30% for sentiment classification and 99.15% for emotion classification, compared to 42.43% and 72.85%, respectively, obtained by the BERT-CNN model. These findings indicate that the CNN-LSTM architecture is more effective in capturing sequential patterns and contextual features in Indonesian review texts, particularly in a multi-task learning setting. Despite its strong performance, the research is limited by its focus on a single platform and the use of lexicon-based automatic labeling, suggesting future work on cross-domain evaluation and manual annotation refinement.**

***Index Terms*—Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), Multi-Task Learning, Sentiment Analysis, Emotion Classification, Deep Learning**

Received: June 25, 2025; received in revised form: Oct. 24, 2025; accepted: Oct. 24, 2025; available online: March 11, 2026.
*Corresponding Author

I. Introduction

THE rapid evolution of mobile apps has changed the way users use digital services and amassed large scale of user feedback in the form of app reviews on platforms like Google Play Store and Apple App Store [1, 2]. Such reviews and ratings serve as a valuable source of user sentiment and emotional feedback and can be used to help developers to understand user satisfaction, feature preference, and application quality [3, 4]. The popularity of social media and web-based communication have also highlighted the utility of sentiment analysis as a means of gaining insights into public opinion and user activity in a variety of fields [5, 6].

Sentiment analysis, the process of determining the attitude of a writer with respect to a target, is a pivotal task in Natural Language Processing (NLP), to automatically analyze and extract subjective information, such as opinions, emotions, attitudes, and so on, from user-generated data [7–9]. Conventional sentiment analysis methods have been mostly binary or ternary classification (e.g., positive vs negative or neutral). In emotion classification task, previous research has conducted title classification for YouTube video. It has some clear application value, but recently, with the penchant for emotional dimensions as an increased source of rich and nuanced insights into user experience, interest in emotion classification as a co-silver task has surged [10, 11]. Multi-task learning methods for sentiment and emotion classification which address sentiment and emotion jointly have been successful [12, 13]. Multi-task learning approaches that simultaneously address sentiment polarity and emo-

tional categorization have also demonstrated superior performance compared to single-task models, as they can leverage shared representations and complementary information between related tasks [14].

Despite having over 270 million speakers, Indonesian remains a low-resource language in NLP due to limited annotated corpora and linguistic tools [2, 15–17]. The scarcity of high-quality labeled data hinders the development of robust models capable of handling linguistic diversity and cultural nuances [18, 19]. Indonesian language complexity arises from its rich morphology, extensive affixation, and regional dialect variations, leading to a wide vocabulary range that is difficult to model using conventional approaches [18]. Furthermore, informal and creative expressions, such as slang, phonetic spelling, and acronyms along with frequent code-switching between Indonesian, regional languages, and English, add further complexity to text processing [11, 20]. These linguistic characteristics collectively demand adaptive methods that can effectively manage the dynamic and heterogeneous nature of Indonesian digital communication while maintaining strong sentiment and emotion classification performance.

Lexicon-based labelling methods have shown effectiveness in responding to the scarcity of labelled data in low-resource languages [8, 21, 22]. Such approaches rely on sentiment dictionaries and linguistic rules to automatically label text data without costly and time-consuming manual annotation [5]. Prior studies have also established that pseudo-labeling based on lexicons can produce high-quality pseudo-labeled dataset approaching manual annotation quality, particularly with the help of emoji analysis and domain-specific keywords [22, 23]. The method is interesting, especially in Indonesian, a language that has a potential to fall back on existing patterns in the language constrained to labeled data availability [17].

Previous research in Indonesian sentiment analysis has used classic machine learning and shallow deep learning frameworks, seldom delving into multi-task learning schemes [17]. Although several previous works have focused on the use of BERT-based models for Indonesian text classification, the merits of hybrid CNN-LSTM architectures tailored for multi-task sentiment and emotion classification remain relatively unexplored. Furthermore, the majority of proposed methods have considered sentiment and emotion classifications separately and have not utilized any commonality between the two tasks to improve the performance [24, 25].

With the rapid development of deep learning, there has been another trend to employ deep models for sentiment analysis. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are state-of-the-art models due to their good capability of capturing local patterns and sequential dependencies in text [21]. Hybrid architectures of CNN and LSTM have been shown to be very effective designs such as CNN for capturing local n-grams features and LSTM for solving long-term dependence and context information [23, 26]. Furthermore, the hierarchical nature of these hybrid models allows for multi-level feature extraction, from character-level patterns to word-level semantics, ultimately contributing to more robust text representation learning [27]. Additionally, the success of transformer-based models, including BERT and its extensions, have already brought in new state-of-the-art performance records in multiple sentiment analysis tasks [24, 28].

The research specifically addresses that gap by proposing a CNN-LSTM model designed for concurrent sentiment and emotion classification on Indonesian app reviews. By integrating both tasks, the research seeks to improve performance efficiency and knowledge sharing across related objectives, thus contributing to the underexplored area of multi-task deep learning for low-resource languages. Although Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) provides pre-trained embeddings specifically for the Indonesian language, the researchers use the multilingual BERT-CNN as a baseline to allow broader comparison across models and evaluate how the proposed CNN-LSTM performs under similar experimental settings. This choice also maintains methodological consistency with prior multilingual sentiment analysis studies while keeping the focus on low-resource language challenges.

Sentiment and emotion analysis for low-resource languages, especially in the context of mobile app reviews, presents a significant challenge due to the complex expressions of emotions, cultural nuances, and domain-specific terminology [15, 26]. Existing state-of-the-art models, primarily designed for high-resource languages, such as English, face substantial limitations when applied to Indonesian. The research introduces a CNN-LSTM hybrid architecture, tailored to process informal Indonesian text with slang, code-switching, and unique cultural expressions, thereby addressing the gaps in current sentiment analysis models [16, 19]. The proposed model leverages the strengths of CNN and LSTM network to learn shared representations that support both sentiment and emotion classification together. The proposed model adopts state-of-the-art preprocessing techniques specifically designed for Indonesian text such as full text normalization, slang word replacement, and improved tokenization mech-

Cite this article as: W. F. Setiawan, A. Amirullah, I. P. Ariatama, and R. N. E. Anggraini, "CNN-LSTM architecture for multi-task sentiment and emotion classification on large-scale Indonesian TikTok application reviews", CommIT Journal 20(1), 77–91, 2026.

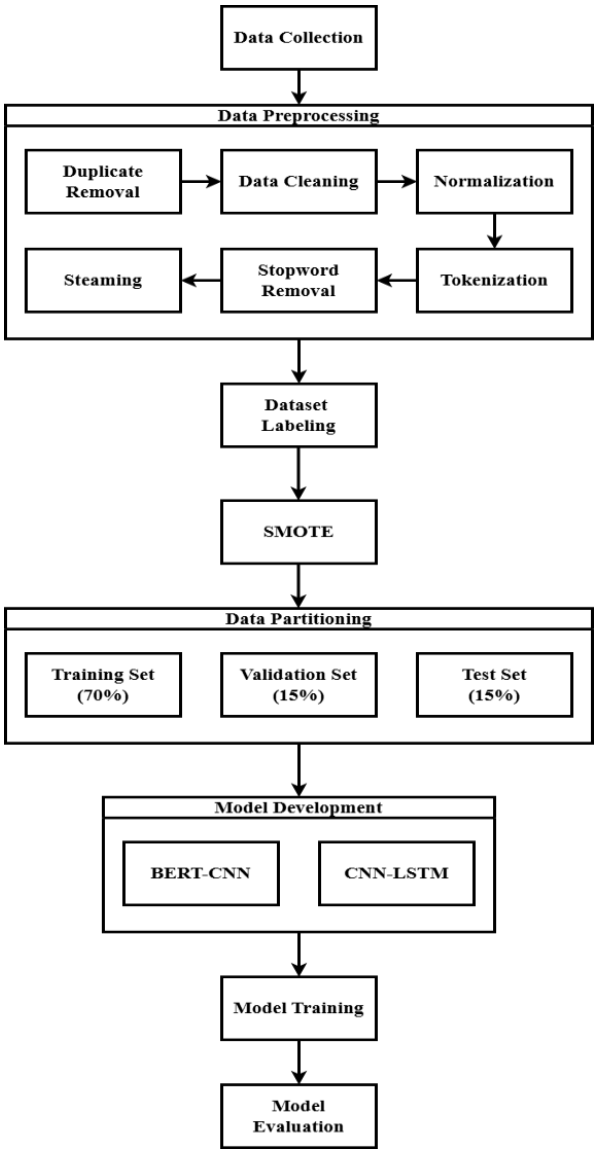

Fig. 1. Research methodology. Note: Synthetic Minority Oversampling Technique (SMOTE), Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN), and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

anism. Furthermore, the research presents a lexicon-based labeling method along with emoji analysis that allows to generate rich annotations with multiple dimensions of sentiment and emotion classes.

The contributions of the research are twofold. First, the research releases a large-scale dataset consisting of 500,000 Indonesian app reviews on TikTok that are extensively preprocessed and annotated for multi-task purposes. Second, the research designs and develops a hybrid CNN-LSTM architecture, which is particularly tailored to the joint task of sentiment and emotion classification in Indonesian text, followed by extensive comparisons with state of the art techniques such as hybrid BERT-CNN and empirical evidence, showing its competitive performance in both tasks. The research makes a contribution to the development of Indonesian NLP capabilities and provides a basis for improving more advanced sentiment analysis algorithms for low-resource languages in the mobile application context.

TABLE I
DATASET STATISTICS AND CHARACTERISTICS.

Metric	Value
Initial reviews collected	500,000
Average review length (characters)	87.3
Rating range	1–5 stars
Language	Indonesian
Source platform	Google Play Store (TikTok app)

II. RESEARCH METHOD

This section presents a comprehensive methodology used to develop and evaluate a model for multitask sentiment and emotion classification on Indonesian app reviews. The research methodology consists of several interrelated phases, from data collection to model evaluation, designed to address the challenges of sentiment analysis in low-resource languages while ensuring robust and reliable results. Figure 1 shows the research methodology.

A. Data Collection

The research utilizes the Google-Play-Scraper Python library to systematically collect Indonesian TikTok reviews from the Google Play Store. The data collection process is designed to obtain comprehensive review data including review text, user ratings, timestamps, and user identifiers. Table I presents the statistics of the collected dataset, providing an overview of its scale and basic properties.

B. Data Preprocessing

The data preprocessing procedure undergoes a six-stage pipeline to transform raw app reviews into clean and normalized text suitable for machine learning model training. First, duplicate records are removed, resulting in the elimination of 170,040 duplicates and improving the overall data quality to 34.01%. Next, data cleaning addresses missing values and textual inconsistencies. Then, normalization converts Indonesian slang into standard forms using a slang dictionary compiled from Google Notepad, enabling more accurate sentiment analysis. Subsequently, the text is tokenized, stopwords are removed using an Indonesian

TABLE II
DATASET STATISTICS BEFORE AND AFTER PREPROCESSING.

Metric	Before	After	Reduction/Change
Total reviews	500,000	341,791	158,209
Duplicate records	170,040	0	170,040 removed
Quality	65.99 %	100 %	+34.01 percentage points

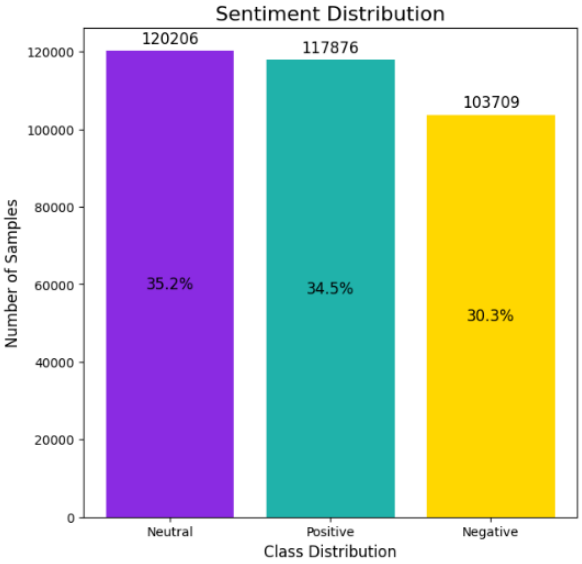

Fig. 2. Sentiment distribution.

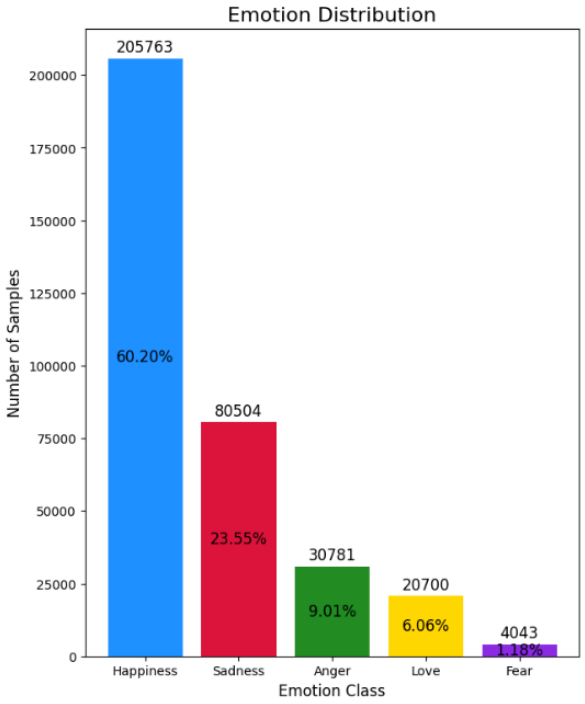

Fig. 3. Emotion distribution.

stopword list, and stemming is performed using the Sastrawi Indonesian stemmer. Overall, this pipeline standardizes word representations and reduces vocabulary size, which supports more efficient model training. Table II provides a comparative overview of dataset statistics before and after preprocessing, highlighting the substantial reduction in duplicate records and the improvement in data quality.

C. Data Labeling Methodology

The labeling scheme combines lexicon-based sentiment classification, rule-based emotion annotation, and emoji analysis to improve label accuracy and coverage. For lexicon-based sentiment labeling, each term in the inSet (Indonesian Sentiment) lexicon is treated as a positive or negative seed word and used to assign sentiment to the vocabulary derived from the preprocessed reviews. Sentiment classification is performed using an Indonesian sentiment lexicon containing lists of positive and negative words to automatically infer polarity from each review.

During labeling, a sentiment score is computed to determine the sentiment class (positive, negative, or neutral). This dictionary-based approach counts the occurrences of positive and negative words in a review. In addition, Bayes' theorem is applied to estimate the overall sentiment polarity. Reviews with a higher positive score are labeled as positive. Those dominated by negative words are labeled as negative, and reviews with relatively balanced positive and negative scores are labeled as neutral [29]. Figs. 2 and 3 show the raw distribution of reviews across sentiment and emotion categories before any resampling.

The labeling process is conducted fully automatically, without manual annotation or human involvement. Automatic labeling is chosen to efficiently handle the large-scale corpus of 500,000 reviews while ensuring consistency and reproducibility. The lexicon-based and emoji-driven strategy enables programmatic assignment of sentiment and emotion categories by combining dictionary lookups with predefined emoji-emotion mappings, thereby reducing subjective bias and substantially shortening labeling time.

Figs. 2 and 3 show the distributions of sentiment and emotion classes before balancing. Although the sentiment classes are relatively balanced, the negative class is slightly underrepresented. In contrast, the emotion distribution is highly skewed. Happiness accounts for more than 60% of the dataset, while minority emotions such as love and fear appear only sparsely. This imbalance motivates the use of resampling strategies (e.g., Synthetic Minority Over-sampling Technique (SMOTE)) to improve minority-class representation

during model training.

Emotion annotation follows a multi-phase procedure that combines emoji analysis and lexical pattern matching. First, the system maps emotion categories (joy, sad, angry, fear, surprise, disgust, love) to emoji symbols based on established emoji (emotion associations). When reviews contain no explicit emoji indicators, rule-based lexical patterns and sentiment intensity are used to infer emotion labels. In addition, sentiment-neutral reviews are predominantly assigned to the neutral emotion category [30].

D. Synthetic Minority Over-sampling Technique (SMOTE)

To mitigate the imbalance observed in the sentiment and emotion class distributions, SMOTE is applied exclusively to the training set. SMOTE generates synthetic samples for minority classes by interpolating between existing instances and their nearest neighbors in the feature space, thereby increasing minority representation without simply replicating the same observations. Unlike naive oversampling that only duplicates minority examples, SMOTE introduces new and plausible samples in the feature space, which helps the model to learn more robust and smoother decision boundaries, reduces majority-class bias, and improves generalization. Applying SMOTE only on the training split also prevents information leakage into the validation and test sets, ensuring that model evaluation remains unbiased and representative of real-world performance [31]. Figures 4 and 5 illustrate the class distributions after applying SMOTE to the sentiment and emotion datasets, respectively.

As shown in Figs. 4 and 5, SMOTE successfully balances the class proportions. The sentiment categories (positive, neutral, and negative) are distributed evenly at approximately 33.3% each, while the emotion categories (happiness, sadness, anger, love, fear) are equalized to 20% each. This balanced distribution is essential for mitigating model bias, ensuring that the model receives comparable learning signals across all classes, and improving generalization performance, particularly for categories that were previously underrepresented.

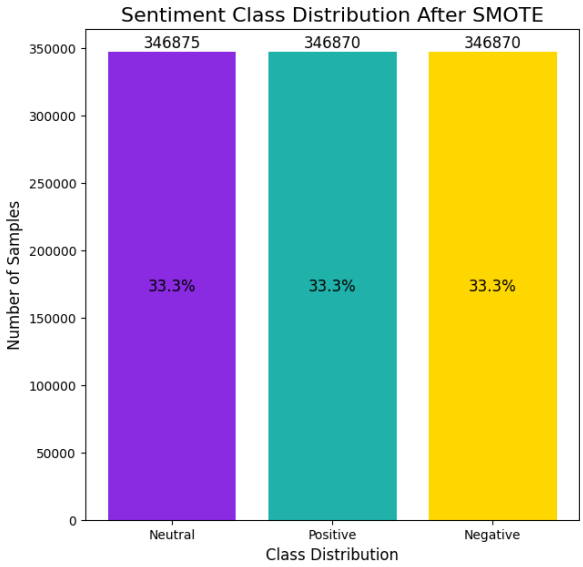

Fig. 4. Sentiment class distribution after Synthetic Minority Oversampling Technique (SMOTE).

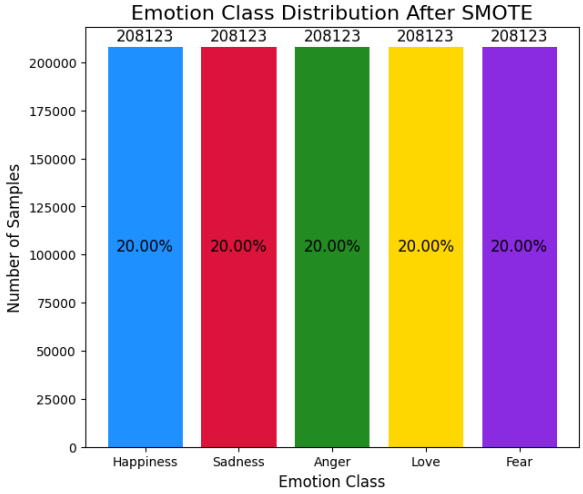

Fig. 5. Emotion class distribution after Synthetic Minority Oversampling Technique (SMOTE).

E. Dataset Partitioning

Stratified sampling is applied to ensure that each label category is proportionally represented across the training (70%), validation (15%), and test (15%) splits. This procedure produces 239,254 training samples, 51,269 validation samples, and 51,268 test samples. The 70/15/15 split follows common machine learning practice, providing sufficient data for model learning while reserving adequate samples for unbiased validation and final testing. Figure 6 shows the class distribution after stratified sampling.

F. Model Development

The research implements two multi-task deep learning architectures to classify sentiment and emotion in Indonesian app reviews. There are hybrid BERT-CNN and CNN-LSTM models. Both architectures are designed to capture global semantic context and local textual patterns while employing dual output heads to

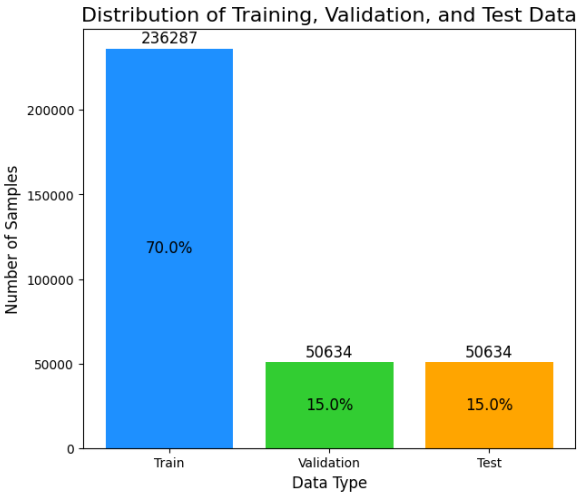

Fig. 6. Data splitting.

jointly learn sentiment polarity and emotion categories in a multi-task learning setting.

Figure A1 in Appendix illustrates the hybrid BERT-CNN model, where a pre-trained IndoBERT encoder is combined with multiple parallel CNN branches. IndoBERT generates contextual token embeddings, and three CNN branches with kernel sizes of 3, 5, and 7 are applied in parallel to extract local n-gram patterns at different receptive-field scales. In addition, a multi-head self-attention layer is applied to the BERT sequence output to strengthen the model's ability to capture long-range dependencies.

The representations from the [CLS] token, the pooled CNN branch outputs (via max pooling and average pooling), and the attention-derived features are concatenated into a single unified feature vector. This vector is then processed by shared fully connected layers, regularized using dropout and batch normalization, and subsequently split into two task-specific output heads for sentiment and emotion classification. Overall, this hybrid architecture leverages IndoBERT's contextual representations, CNN-based local feature extraction, and attention-driven global focus to form a robust dual-output classifier.

Figure A2 in Appendix presents the CNN-LSTM model, in which the input text is processed in parallel through CNN and Bidirectional LSTM (BiLSTM) pathways. A trainable embedding layer first maps input tokens into a continuous vector space. In the CNN branch, two convolutional layers followed by a max-pooling operation are used to capture local n-gram and spatial feature patterns. In parallel, the BiLSTM branch models temporal dependencies by learning bidirectional sequential relationships within the token sequence.

The representations from the CNN and BiLSTM branches are then concatenated and fed into shared dense layers with dropout regularization to reduce overfitting. In the hybrid BERT-CNN model, the network is finally split into two task-specific classification heads to jointly predict sentiment polarity and emotion categories. This parallel CNN-LSTM design combines local pattern recognition with sequential sensitivity, producing richer and more adaptive textual representations for multi-task learning.

G. Model Evaluation

Model evaluation includes extensive performance analysis using standard classification metrics, namely accuracy, precision, recall, and F1-score, for both sentiment and emotion classification tasks. A comparative evaluation across architectures is conducted to determine which modeling approach is more effective for Indonesian text analysis. In addition, qualitative error analysis is performed by inspecting model predictions on selected samples to identify systematic biases and potential areas for improvement. Eqs. (1)–(4) provide the mathematical formulations of the evaluation metrics used. It shows TP as True Positives, TN as True Negatives, FP as False Positives, and FN as False Negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

III. Results and Discussion

Sentiment and emotion analysis in the research compares the training results of two proposed multi-task models, BERT-CNN and CNN-LSTM, across both sentiment and emotion classification tasks. This comparison aims to evaluate the effectiveness of each architecture in handling Indonesian app review text, measured in terms of accuracy, precision, recall, and F1-score. The results reported in the corresponding tables summarize the performance of both models on the sentiment and emotion datasets. Tables III and IV present the comparative performance of the two models for sentiment and emotion classification, respectively.

It is clear from the comparison results that CNN-LSTM performs better than BERT-CNN in sentiment and emotion. Table III shows that in the sentiment analysis, CNN-LSTM reaches high precision (99%),

TABLE III
SENTIMENT MODEL COMPARISON.

Model	Metric			
	Accuracy	Precision	Recall	F1-Score
BERT-CNN	0.73	0.73	0.73	0.73
CNN-LSTM	0.99	0.99	0.99	0.99

Note: Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN) and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

TABLE IV
EMOTION MODEL COMPARISON.

Model	Metric			
	Accuracy	Precision	Recall	F1-Score
BERT-CNN	0.42	0.43	0.42	0.41
CNN-LSTM	0.91	0.91	0.91	0.91

Note: Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN) and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

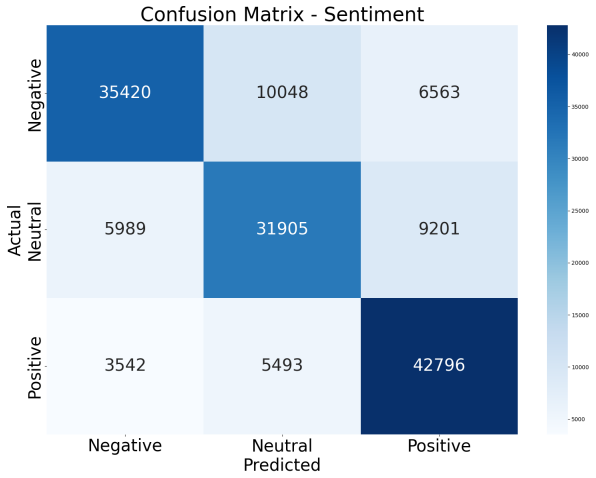

Fig. 7. Sentiment confusion matrix for Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN).

while BERT-CNN only reaches 73%. Likewise, Table IV demonstrates that CNN-LSTM outperforms other methods in emotion analysis with its accuracy of 91%, while BERT-CNN is 42%. These results emphasize that CNN-LSTM outperforms those methods in sentiment and emotion classification better.

A. Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN)

The performance of the sentiment classification model is clearly depicted in the confusion matrix presented in Fig. 7. The model accurately detected 35,420 Negatives, 31,905 Neutrals, and 42,796 Positives. Positive sentiment had the most fantastic accuracy of the three. However, few misclassifications occurred, particularly for the Negative sentiment, where 10,048 examples were misclassified as Neutral and another 6,563 as Positive. This can be interpreted as indicating that the model has trouble distinguishing Negative sentiments from the rest. The overall positive and negative accuracy is perfect, but the assignment of Negative and Neutral classification needs to be more precise. Table V shows sentiment evaluation matrix details.

According to Table V, the sentiment classification model evaluation metrics show that precision, recall, and F1-score for each class produce decent model results. For the negative sentiment, precision is 0.79, recall is 0.68, and F1-score is 0.73, indicating the model can effectively learn about negative sentiment. For neutral sentiment, precision is 0.67, pecall is 0.68, and F1-score is 0.67. The results show consistent

TABLE V
SENTIMENT EVALUATION MATRIX DETAILS USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS-CONVOLUTIONAL NEURAL NETWORK (BERT-CNN).

Sentiment	Precision	Recall	F1-Score
Negative	0.79	0.68	0.73
Neutral	0.67	0.68	0.67
Positive	0.73	0.82	0.77
Accuracy			0.77

performance for that class although it is less than the others. For positive sentiment, precision is 0.73, recall is 0.82, and F1-score is 0.77, indicating strong performance in the classification. Overall, the model scores at most 0.73, revealing a sufficient success rate in anticipating the overall sentiment.

In Fig. 8, the confusion matrix for an emotion classification model is shown. It gives the results of emotion predictions obtained from text. The model is perfect in anger prediction, with 16,445 correct predictions, and in happiness, with 7,900 correct classifications. The model retrieves well, but 6,415 Anger, 5,115 sadness, 4,308 happiness, and 8,497 fear are erroneously recognized. Love is detected correctly 20,290 times, although it is also sometimes mistaken for happiness (3,305), sadness (2,467), and fear (1,183). Sadness gets 11,009 correct but is misclassified as anger (8,378), happiness (4,319), and love (5,229). Table VI shows emotion evaluation matrix details.

It can be observed from Table VI that the emotion classification model works differently in each emotion category. For anger, the model performs moderately with precision of 0.41, recall of 0.54, and F1-score of

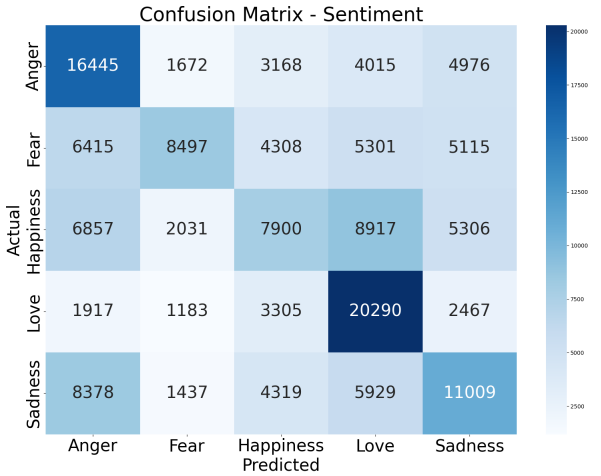

Fig. 8. Emotion confusion matrix for Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN).

TABLE VI
EMOTION EVALUATION MATRIX DETAILS USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS-CONVOLUTIONAL NEURAL NETWORK (BERT-CNN).

Emotion	Precision	Recall	F1-Score
Anger	0.41	0.54	0.47
Fear	0.57	0.29	0.38
Happiness	0.34	0.25	0.29
Love	0.46	0.70	0.55
Sadness	0.38	0.35	0.37
Accuracy			0.72

0.47. Fear's memory is similarly low at 0.29 for recall and 0.38 for F1-score. However, it displays a precision of 0.57, indicating that when the model predicts fear, it is relatively confident. However, it is often incorrect when confidence is present. Happiness has the lowest scores through all of these metrics: precision (0.34), recall (0.25), and F1-score (0.29) if the lack of classification is even greater. Conversely, love has the best performance with a precision of 0.46, a high recall of 0.70, and an F1-score of 0.55, indicating that the model correctly tags this emotion frequently. The sadness prediction has mediocre results, with precision of 0.38, recall of 0.35, and F1-score of 0.37. Ultimately, the model achieves 0.72 accuracy, a pretty good general performance but lacking the uniformity required for an emotion classification problem.

Figure 9 shows performance evaluation for BERT-CNN model. Then, Table VII shows performance evaluation for sentiment and emotion classification using BERT-CNN model. The total of loss test is 2.0135. It is the sum error on the testing data. In particular, the results for sentiment loss and emotion loss tests are

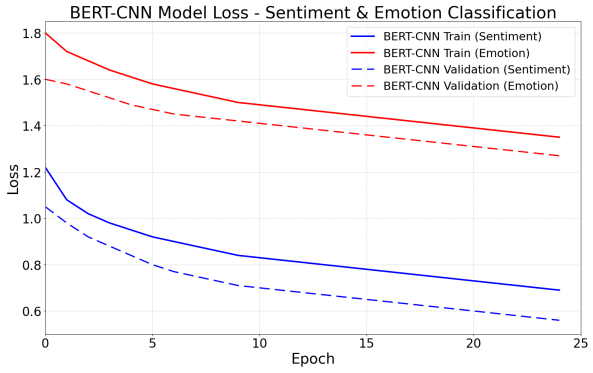

Fig. 9. Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN) training loss.

TABLE VII
MODEL EVALUATION RESULTS FOR BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS-CONVOLUTIONAL NEURAL NETWORK (BERT-CNN).

Metric	Value
Loss test	2.0135
Sentiment loss test	0.6493
Emotion loss test	1.3641
Sentiment accuracy test	0.4243
Emotion accuracy test	0.7285

0.6493 and 1.3641, respectively. These results show that the model performs better at sentiment prediction because it has a lower loss. The model achieves higher accuracy of 0.4243 than 0.7285, proving that the model can handle emotional categories processing sub-functions more effectively than sentiment polarity and emotional category. The research can infer that even though sentiment prediction has a lower loss, emotion classification obtains a higher accuracy. A trade-off may exist related to label complexity and training dynamics.

B. Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

The confusion matrix in Fig. 10 shows the results of the sentiment classification model in three classes (negative, neutral, and positive). The model performs well overall, especially in identifying negative (51,529 cases), neutral (46,679 cases), and positive (51,671 cases) sentiments accurately, indicating strong discriminative capability across sentiment categories. However, few misclassifications are found. They reflect the robustness of the proposed approach. For instance, 243 negative sentiments are predicted as neutral and 259 as positive. Likewise, 234 neutral data points are classified as negative and 182 as positive. Even for positive, few data are predicted with wrong labels.

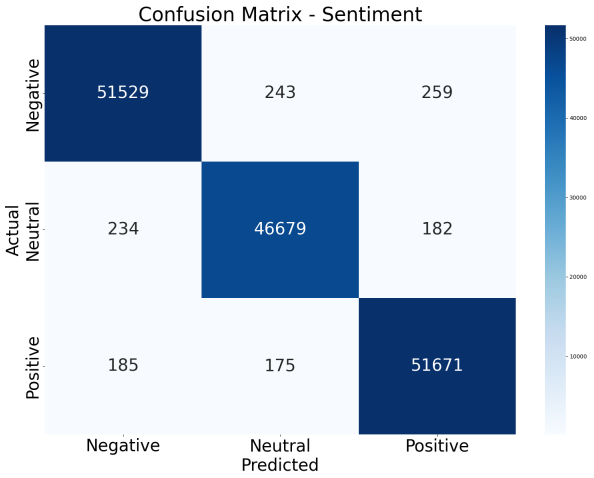

Fig. 10. Sentiment confusion matrix for Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM)

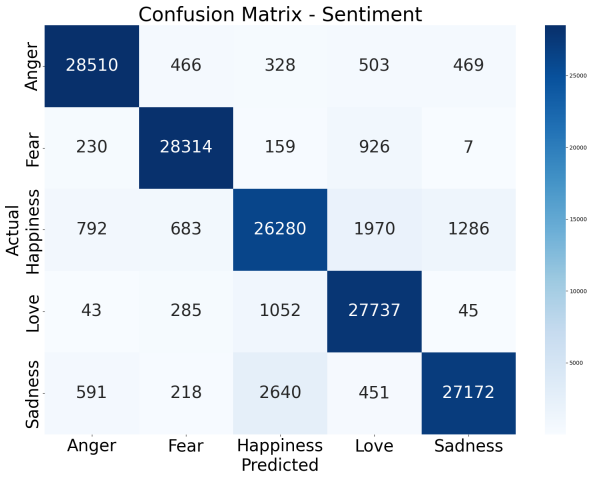

Fig. 11. Emotion confusion matrix for Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM).

TABLE VIII
SENTIMENT EVALUATION MATRIX DETAILS USING CONVOLUTIONAL NEURAL NETWORK-LONG SHORT-TERM MEMORY (CNN-LSTM).

Sentiment	Precision	Recall	F1-Score
Negative	0.99	0.99	0.99
Neutral	0.99	0.99	0.99
Positive	0.99	0.99	0.99
Accuracy			0.99

TABLE IX
EMOTION EVALUATION MATRIX DETAILS USING CONVOLUTIONAL NEURAL NETWORK-LONG SHORT-TERM MEMORY (CNN-LSTM).

Emotion	Precision	Recall	F1-Score
Anger	0.95	0.94	0.94
Fear	0.94	0.96	0.95
Happiness	0.86	0.85	0.86
Love	0.88	0.95	0.91
Sadness	0.94	0.87	0.90
Accuracy			0.91

Around 185 instances are predicted as negative and 175 as neutral. Overall, the model is correct and performs well in sentiment classification, achieving high reliability in distinguishing polar sentiments. At the same time, there is room for improvement in the separation of neutral sentiment as the model has a slightly higher degree of confusion, which may be attributed to the ambiguous linguistic patterns commonly found in neutral expressions. Table VIII shows sentiment evaluation matrix details.

The analysis results in Table VIII clearly show that CNN-LSTM achieves the best performance in sentiment classification. The model attains the precision, recall, and F1-score metrics of 0.99 for the negative, neutral, and positive sentiments. The results show uniform reliability prediction across all sentiment categories. This consistent performance indicates that the proposed architecture is highly effective in capturing sentiment-related features from text data. The CNN-LSTM model achieves an accuracy of 0.99 and approaches no misclassification error regardless of classes. It demonstrates robustness and stability in handling diverse sentiment expressions commonly found in user-generated content.

Figure 11 shows the confusion matrix for the CNN-LSTM in emotion classification model. It has five basic level emotions: anger, fear, happiness, love, and sadness. The model is quite impressive in terms of anger (28,510 cases classified correctly), love (27,737), and sadness (27,172). Feelings like hapiness (26,280) and fear (28,214) display good classification scores equally. However, some of them may be wrong. For example, some instances of happiness are wrongly classified as sadness (1,296) or fear (683), and anger is incorrectly identified as fear (466) or sadness (469). With a few shortcomings, the model is generally reliable enough, while most categories have a coherent emotional classification. Table IX shows emotion evaluation matrix details.

Overall, in Table IX, it is a thorough comparison of CNN-LSTM model performance on emotion classification. The model has high precision, recall, and F1-score concerning all five emotions. For example, the best result is obtained on fear, with a precision of 0.94, a recall of 0.96, and F1-score of 0.95. Anger also yields a performance of high precision of 0.95 and F1-score of 0.94. Emotions, such as happiness and sadness, are slightly lower but still robust, with F1-

TABLE X
MODEL EVALUATION RESULTS FOR CONVOLUTIONAL NEURAL NETWORK-LONG SHORT-TERM MEMORY (CNN-LSTM).

Metric	Value
Loss test	0.3151
Sentiment loss test	0.0523
Emotion loss test	0.2628
Sentiment accuracy test	0.9130
Emotion accuracy test	0.9915

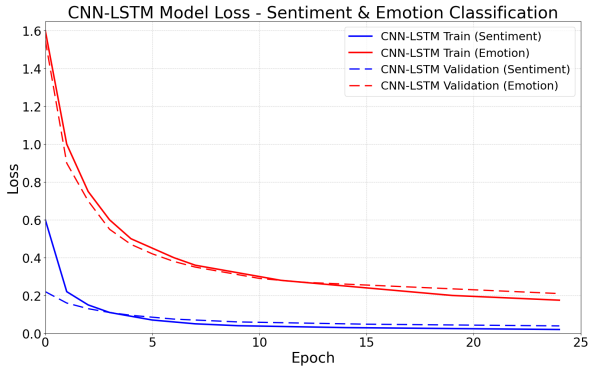

Fig. 12. Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) training loss.

scores of 0.86 and 0.90, respectively. Love also shows high metrics, with a recall of 0.95 and an F1-score of 0.91, meaning that the model captures this category well. In general, the model achieves an accuracy of 0.91, indicating the reliability of the model even for a wide range of emotions with low misclassification.

Table X also shows that the CNN-LSTM model performs well on sentiment and emotion classification tasks. The loss test (to be averaged over the test data) is 0.3151. It shows good generalization. Then, the test sentiment loss is only 0.0523, which confirms the method's capability of sentiment characterization. The emotion loss test is higher than 0.2628. However, the result is still in a satisfactory range. Regarding accuracy, the method obtains 91.30% for sentiment and 99.15% for emotion classification. Combined with the learning curves in Fig. 12, the results show that CNN-LSTM exhibits excellent results in accuracy performance and error reduction, particularly when capturing emotional subtleties in text data tasks.

The results show that CNN-LSTM is the best model with superior performance compared to BERT-CNN in both tasks (sentiment: 99% vs. 73%, emotion: 91% vs. 42%), indicating that for short informal Indonesian texts with code-switching and slang characteristics, an architecture combining local feature extraction (CNN) and sequential modeling (LSTM) is more effective than complex transformer models. The research suggests that for low-resource and informal language data, hybrid architectures that explicitly combine local and sequential feature extraction outperform transformer-based models that rely heavily on contextual embeddings trained on formal, high-resource corpora. The underperformance of BERT-CNN may stem from its dependence on pre-trained contextual embeddings that are less effective for informal, code-switched Indonesian text. In contrast, CNN-LSTM's hierarchical processing allows the model to adapt to noisy, user-generated content with limited linguistic normalization. This finding implies that lightweight hybrid models can offer a more practical alternative for real-time sentiment and emotion monitoring in mobile app ecosystems.

IV. CONCLUSION

The research presents a novel contribution by developing a multi-task learning system for sentiment and emotion classification on 500,000 Indonesian TikTok reviews. The CNN-LSTM model achieves unprecedented accuracy of 99% for sentiment and 91% for emotion, outperforming BERT-CNN (73% and 42%, respectively). The originality of the CNN-LSTM model lies in its ability to capture both local features and sequential dependencies, making it particularly effective for processing informal Indonesian texts filled with slang and code-switching. This approach advances sentiment and emotion classification in low-resource languages, particularly for Indonesian mobile app reviews.

The research introduces a large-scale annotated dataset within the framework of Shaver's emotion theory, enhances sentiment analysis with a tailored 6-step preprocessing pipeline for Indonesian, and proposes a CNN-LSTM architecture specifically designed for sentiment and emotion classification in low-resource languages. The emoji analysis and lexicon-based labeling approach allows efficient automatic tagging in light of resource constraints in Indonesia. The research successfully achieves its objectives by demonstrating that a CNN-LSTM hybrid architecture can effectively perform joint sentiment and emotion classification on large-scale Indonesian reviews, addressing the limitations of existing single-task or transformer-based approaches. The proposed framework holds practical potential for real-world implementation in application reputation monitoring systems, customer feedback analysis, and academic research on multilingual and low-resource NLP.

Nevertheless, the researchers acknowledge several limitations that should be addressed in future research. First, the dataset is limited to TikTok reviews, which

may not be representative of other mobile applications or social media platforms. Second, the lexicon-based labeling approach may introduce bias, particularly for complex emotions like love, which often overlap with happiness in user expressions. Third, despite using SMOTE for balancing, the original emotion distribution shows significant imbalance with happiness comprising over 60% of the dataset. Fourth, the research lacks cross-domain generalization testing, which limits the applicability of findings to other domains or applications.

Further research can be done on expanding the multi-domain dataset, integrating the attention mechanism, manual validation to minimize labeling bias, exploring more categories of emotions from Shaver's framework, and Indonesian regional language adaptation. By developing sarcasm detection, temporal sentiment shift, and real-time analysis systems, the research can contribute to the practical application for app reputation monitoring and customer feedback analysis. Future studies may also investigate cross-lingual transfer learning approaches to improve model robustness for Indonesian and other low-resource languages. In addition, incorporating transformer-based attention mechanisms or hybrid architectures may further enhance the model's capability in capturing long-range contextual dependencies within informal user-generated text.

Acknowledgement

This work was funded by the Department of Informatics, Institut Teknologi Sepuluh Nopember, through the Department Research Grant Scheme – Batch 1, 2025 (Grant No.: 2337/PKS/ITS/2025).

Author Contribution

Conceived and designed the analysis, W. F. S.; Collected the data, A. A.; Contributed data or analysis tools, W. F. S. and I. P. A.; Performed the analysis, A. A. and I. P. A.; Wrote the paper, W. F. S.; Reviewed and edited the manuscript critically for important intellectual content, R. N. E. A.; and Secured research funding and provided overall project supervision, R. N. E. A.

Data Availability

The data that support the findings of this research are available from the corresponding author, Ratih Nur Esti Anggraini, upon reasonable request. The data are not publicly available due to institutional data sharing policies and file size limitations.

References

[1] N. Khamphakdee and P. Seresangtakul, "An efficient deep learning for Thai sentiment analysis," *Data*, vol. 8, no. 5, pp. 1–22, 2023.

[2] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual sentiment analysis: A systematic literature review," *Pertanika Journal of Science & Technology*, vol. 29, no. 1, pp. 445–470, 2021.

[3] L. Chen, S. Shang, and Y. Wang, "Cross-lingual sentiment analysis with MultiEmo: Exploring language-agnostic models for emotion recognition," 2024. [Online]. Available: https://doi.org/10.20944/preprints202408.1639.v1

[4] S. H. Park, K.-M. Kim, O. J. Lee, Y. Kang, J. Lee, S. M. Lee, and S. Lee, ""why do I feel offended?"-Korean dataset for offensive language identification," in *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 1142–1153.

[5] M. Garouani and J. Kharroubi, "MAC: An open and free Moroccan Arabic corpus for sentiment analysis," in *The Proceedings of the International Conference on Smart City Applications*. Safranbolu, Türkiye: Springer, Oct. 27–29, 2021, pp. 849–858.

[6] H. Fouadi, H. El Moubtahij, H. Lamtougui, and A. Yahyaouy, "Sentiment analysis of Arabic comments using machine learning and deep learning model," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 13, no. 3, pp. 598–606, 2022.

[7] K. M. Awlla, H. Veisi, and A. A. Abdullah, "KuBERT: Central Kurdish BERT model and its application for sentiment analysis," 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-4552724/v1

[8] V. Dhananjaya, S. Ranathunga, and S. Jayasena, "Lexicon-based fine-tuning of multilingual language models for low-resource language sentiment analysis," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 5, pp. 1116–1125, 2024.

[9] M. S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 285–297, 2019.

[10] G. Thakkar, N. M. Preradović, and M. Tadić, "Transferring sentiment cross-lingually within and across same-family languages," *Applied Sciences*, vol. 14, no. 13, pp. 1–21, 2024.

[11] M. Sangeetha and K. Nimala, "Sentiment analysis on code-mixed Tamil-English corpus: A comprehensive study of transformer-based models," 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-3418283/v1

[12] A. Kniele and M. Beloucif, "Uppsala University at SemEval-2023 task12: Zero-shot sentiment classification for Nigerian Pidgin tweets," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1491–1497.

[13] L. Khan, A. Amjad, N. Ashraf, H. T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97 803–97 812, 2021.

[14] K. Cortis, K. Verma, and B. Davis, "Fine-tuning neural language models for multidimensional opinion mining of English-Maltese social data," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. INCOMA Ltd., 2021, pp. 309–314.

[15] C. Kumaresan and P. Thangaraju, "Sentiment analysis in multiple languages: A review of current approaches and challenges," *REST Journal on Data Analytics and Artificial Intelligence*, vol. 2, no. 1, pp. 8–15, 2023.

[16] D. Homskiy and N. Maloyan, "DN at SemEval-2023 task 12: Low-resource language text classification via multilingual pretrained language model fine-tuning," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 1537–1541.

[17] T. D. Purnomo and J. Sutopo, "Comparison of pre-trained BERT-based transformer models for regional language text sentiment analysis in indonesia," *International Journal Science and Technology*, vol. 3, no. 3, pp. 11–21, 2024.

[18] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*. Singapore: IEEE, Dec. 5–7, 2017, pp. 391–394.

[19] K. Ronny Mabokela, M. Primus, and T. Celik, "Advancing sentiment analysis for low-resourced african languages using pre-trained language models," *PLOS ONE*, vol. 20, no. 6, pp. 1–37, 2025.

[20] Z. Kastrati, L. Ahmedi, A. Kurti, F. Kadriu, D. Murtezaj, and F. Gashi, "A deep learning sentiment analyser for social media comments in low-resource languages," *Electronics*, vol. 10, no. 10, pp. 1–19, 2021.

[21] U. I. Shabrina, R. Sarno, R. N. E. Anggraini, A. T. Haryono, and A. F. Septiyanto, "Sentiment analysis of presidential candidate debates from Youtube videos," in *2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. Bandung, Indonesia: IEEE, Feb. 21–23, 2024, pp. 1–6.

[22] T. De Melo and P. Merialdo, "SentiLexIT: Advancing Italian sentiment analysis through automated lexicon generation," 2024. [Online]. Available: https://doi.org/10.21203/rs.3.rs-4630348/v1

[23] A. T. Haryono, R. Sarno, R. N. E. Anggraini, and K. R. Sungkono, "Permuted temporal Kolmogorov-Arnold networks for stock price forecasting using generative aspect-based sentiment analysis," *IEEE Access*, vol. 12, pp. 178 672–178 689, 2024.

[24] W. Yu, L. Yin, C. Zhang, Y. Chen, and A. X. Liu, "Application of quantum recurrent neural network in low-resource language text classification," *IEEE Transactions on Quantum Engineering*, vol. 5, pp. 1–13, 2024.

[25] M. Alali, N. Mohd Sharef, M. A. Azmi Murad, H. Hamdan, and N. A. Husin, "Multitasking learning model based on hierarchical attention network for Arabic sentiment analysis classification," *Electronics*, vol. 11, no. 8, pp. 1–23, 2022.

[26] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape," *IEEE Access*, vol. 11, pp. 15 996–16 020, 2022.

[27] V. K. Agbesi, W. Chen, C. C. Ukwuoma, N. A. Kuadey, C. C. M. Agbesi, C. J. Ejiyi, E. S. A. Gyarteng, G. W. Muoka, and A. M. Kuadey, "Multichannel 2D-CNN attention-based BiLSTM method for low-resource Ewe sentiment analysis," *Journal of Data Science and Intelligent Systems*, vol. 3, no. 1, pp. 67–77, 2025.

[28] M. R. Ashraf, Y. Jana, Q. Umer, M. A. Jaffar, S. Chung, and W. Y. Ramay, "BERT-based sentiment analysis for low-resourced languages: A case study of Urdu language," *IEEE Access*, vol. 11, pp. 110 245–110 259, 2023.

[29] S. Kaddoura, M. Itani, and C. Roast, "Analyzing the effect of negation in sentiment polarity of Facebook dialectal Arabic text," *Applied Sciences*, vol. 11, no. 11, pp. 1–13, 2021.

[30] R. Mabokela, M. Roborife, and T. Celik, "Investi-

gating sentiment-bearing words-and emoji-based distant supervision approaches for sentiment analysis," in *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 115–125.

[31] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

APPENDIX

The Appendix can be seen in the next page.

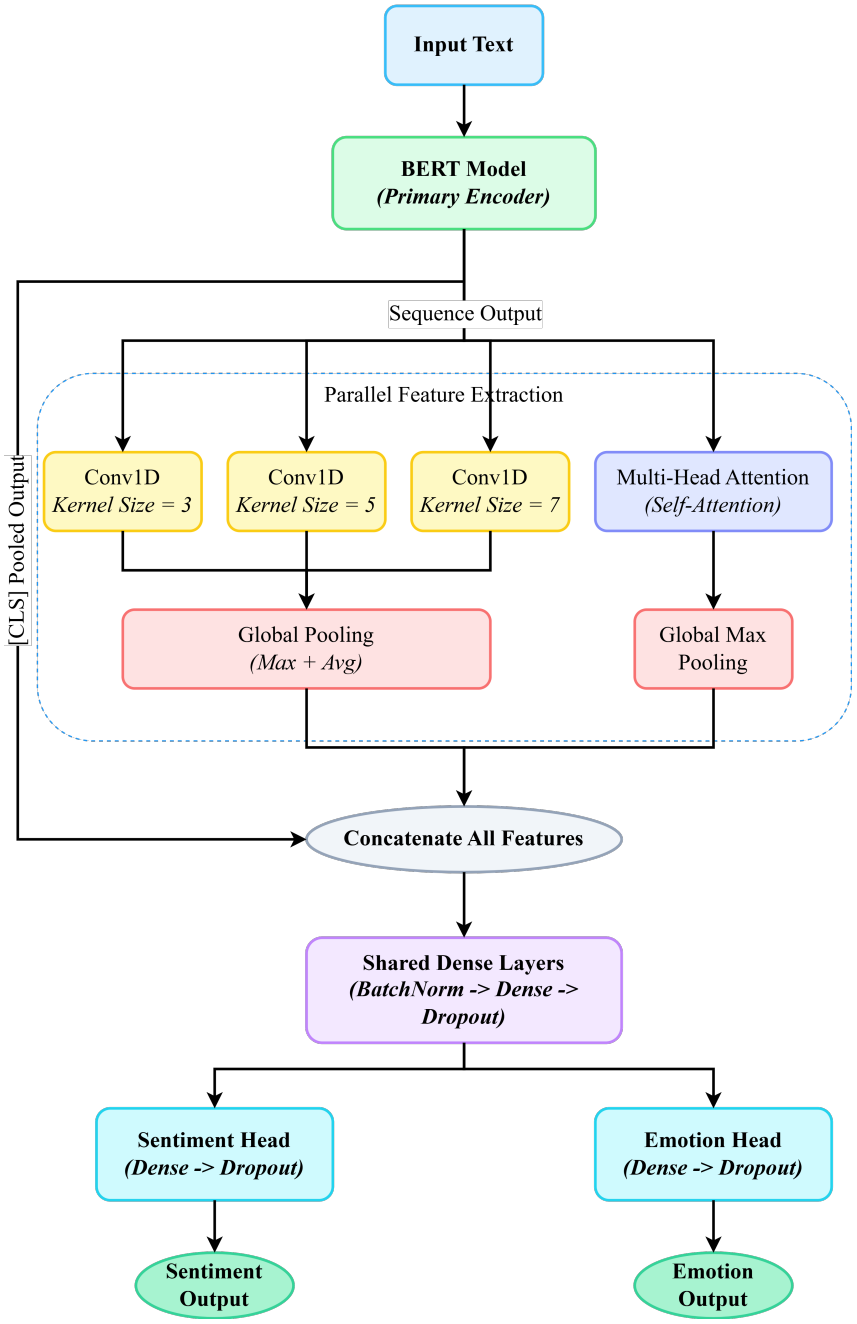

Fig. A1. Bidirectional Encoder Representations from Transformers-Convolutional Neural Network (BERT-CNN) architecture.

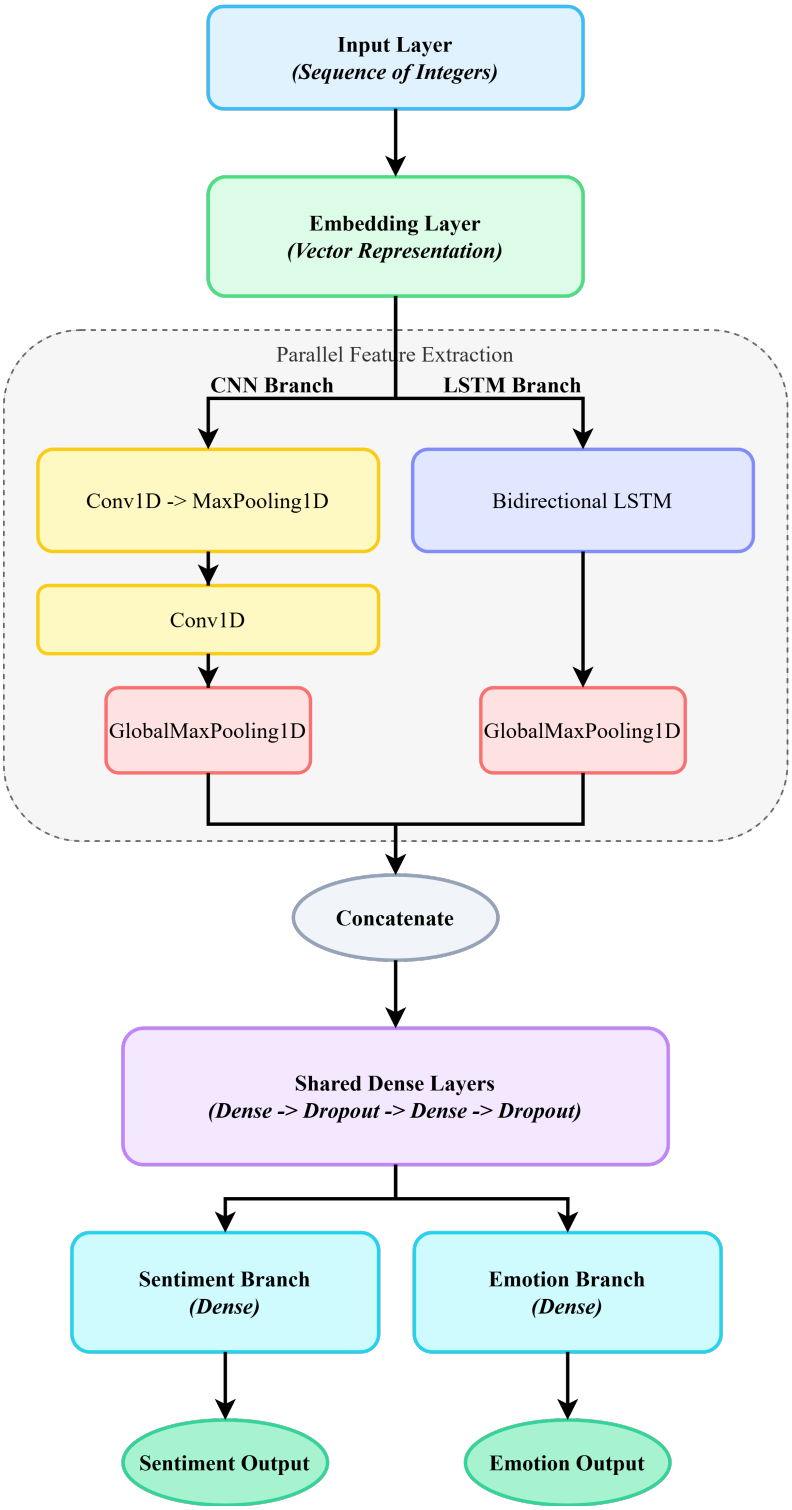

Fig. A2. Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture.