Coding: *Journal of Computing and Software Engineering*
Vol. 1, No. 1, June 2025, Page: 1-7
E-ISSN: 3109-869X
Doi: https://doi.org/10.25181/coding.v1i1.4306
Journal homepage: https://jurnal.polinela.ac.id/coding

Hybrid Deep Learning Models for Multilingual Sentiment Analysis in Low-Resource Languages

Hendri Purnomo[1], Randi Estian Pambudi[1], Danang Ade Muktiawan[1], Sylvia[2]
[1]Information Systems Study Program, IIB Darmajaya, Indonesia.
[2]Software Engineering Technology Study Program, Politeknik Negeri Lampung, Indonesia.

Article Info

Keywords:

BiLSTM;
CNN;
Deep Learning;
Multilingual: low-resource languages;
Sentiment Analysis.

Article History:

Submitted: May 26, 2025
Accepted: June 12, 2025
Published: June 14, 2025

Abstract: Multilingual sentiment analysis remains challenging, especially for low-resource languages with limited annotated data and linguistic tools. This study proposes a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), enhanced with mBERT embeddings. The model is trained and tested on a multilingual dataset covering ten languages, including Swahili, Hausa, Yoruba, and Indonesian. Results show the model achieved 84.3% accuracy and a macro F1-score of 83.1%, outperforming baseline models such as Naive Bayes and standalone BiLSTM. The hybrid architecture effectively captures both local and contextual sentiment cues across different languages. While Indonesians scored highest due to more abundant training data, the model also performed consistently in low-resource settings. These findings suggest that combining multilingual embeddings with hybrid architectures offers a promising approach for sentiment analysis in linguistically diverse environments.

This is an open access article under the CC BY-SA license

Corresponding Author:
Hendri Purnomo
Email: hendrialie@darmajaya.ac.id

INTRODUCTION

Sentiment analysis is a key area in Natural Language Processing (NLP) that aims to identify users' opinions, emotions, and evaluations of an entity, service, or product(Jain & Nemade, 2010). Although substantial progress has been made for high-resource languages such as English, the application of sentiment analysis to low-resource languages still faces significant challenges, mainly due to limited labeled data and a lack of other linguistic resources(Pak & Paroubek, 2010)(Iin et al., 2024).

Several approaches have been proposed to address these limitations. One of the most prominent is the use of deep learning models that are trained on high-resource languages and then applied to low-resource languages using transfer learning or no-shot learning techniques. For example, Koto et al. (2024) showed that incorporating a multilingual sentiment lexicon in the pretraining phase can improve the performance of zero-shot sentiment analysis in 25 languages with low resources (Ramadhani & Suryono, 2024)(Mailoa, 2021).

However, this approach still has limitations, especially in dealing with contextual nuances and complex syntactic structures in different languages. To address this problem, a hybrid approach that combines various deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has been proposed(Lutfina et al., 2024)(Fauzi et al., 2019). This approach aims to harness the power of each architecture in capturing local features and long-term dependencies in text(Winoto et al., 2024)(Pahtoni & Jati, 2024).

Another challenge in multilingual sentiment analysis is the unbalanced representation of languages in large multilingual models. This phenomenon, known as the "multilingual curse", refers to the decline in performance in low-resource languages due to the dominance of data from high-resource languages(Ardiansyah et al., 2024)(Normah et al., 2022).

Therefore, this study aims to develop a hybrid deep learning model that can effectively conduct sentiment analysis in multiple languages, including those with limited resources(Intan et al., 2023)(M. A. Maulana et al., 2018). The model integrates different deep learning architectures and leverages a multilingual sentiment lexicon to improve performance under data-constrained conditions. This approach is expected to contribute significantly to the development of inclusive NLP technologies that support the world's diverse languages(Efraim, 2023).

Although some studies have applied deep learning models such as BiLSTM or CNN for sentiment analysis, most of them still focus on high-resource languages such as English. In addition, the research in integrating mBERT with a hybrid CNN-BiLSTM architecture to address low-resource languages is still very limited. It highlights research gaps in developing models capable of effectively handling multilingual contexts under resource-constrained conditions. Therefore, this study aims to address these challenges by proposing a hybrid approach based on mBERT, CNN, and BiLSTM, optimized for a variety of languages, including Swahili, Hausa, and Yoruba

METHOD

The methodology used in this study is designed to develop and evaluate a hybrid deep learning model capable of performing effective sentiment classification in multiple languages, including low-resource languages(I. Maulana et al., 2023)s. The main stages of this methodology include data pre-processing, feature extraction, hybrid model architecture design, model training, and performance evaluation(Putranti & Winarko, 2014).

In this research, data pre-processing was carried out to standardize and clean the multilingual sentiment of the data. The process begins with language detection, where the langdetect libraries are used to identify and group reviews according to their language. This step is critical to ensure language-specific preprocessing is implemented correctly. After this, text cleanup is done by removing non-alphabetic characters, punctuation, URLs, and special symbols(R. Maulana et al., 2023) (Pratama et al., 2023). All text is lowercased, and stop deletion is applied using a list of language-appropriate endwords. The cleaned data is then tokenized using a pre-trained multilingual tokenizer, specifically mBERTTokenizer or XLMRTokenizer, depending on the embedding configuration used in the model.

The dataset consists of approximately 300,000 reviews in 10 languages, of which approximately 30% are from low-resource languages such as Swahili, Hausa, and Yoruba. Sentiment labels are categorized into three classes-positive, neutral, and negative—and are being encoded using single-heat coding. To maintain a balanced distribution of labels, data was divided using graded sampling into 80% training, 10% validation, and 10% test sets.

E-ISSN: 3109-869X Doi: http://doi.org/10.26594/coding.v1i1.4306

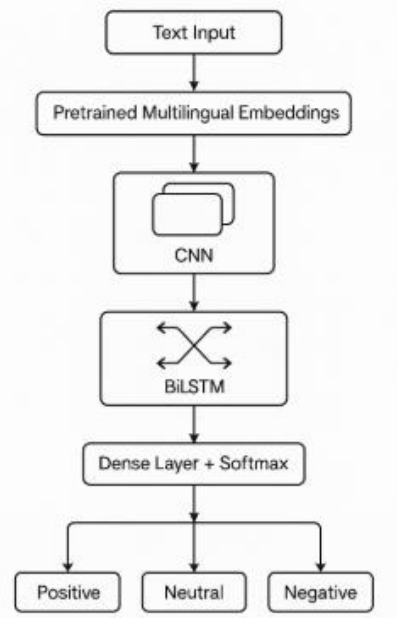

Figure 1. Architecture of the Proposed Hybrid Deep Learning Model

The hybrid CNN-BiLSTM model demonstrates its effectiveness in multilingual sentiment classification through a combination of spatial and sequential processing. The CNN layer is very useful in extracting local sentiment cues such as short phrases or n-gram patterns commonly found in user reviews, such as "very bad" or "dissatisfied". These features are critical to identify strong sentiment polarities regardless of the language. Meanwhile, the BiLSTM layer captures a deeper semantic understanding by modeling the sequential structure of sentences in both forward and backward directions. This allows the model to consider long-term dependencies, making it more robust in handling complex sentence structures that vary across languages.

One significant observation is the model's ability to generalize across a wide range of linguistic patterns. The integration of mBERT embedding ensures that contextual representations are learned in multiple languages, thereby improving performance even in low-resource languages such as Yoruba and Hausa. The results showed that Indonesian achieved the highest F1 score among the languages tested, which can be attributed to the availability of more abundant and balanced training data for this language. In contrast, languages such as Hausa and Yoruba have slightly lower performance due to limited lexical and annotation resources.

Nevertheless, the model shows minimal resilience and performance degradation across languages, suggesting that hybrid architectures, along with multilingual embedding, provide a solid foundation for sentiment understanding. This reinforces the importance of combining pre-trained embedding with a specific deep learning structure to address the challenges associated with low-resource scenarios.

RESULTS AND DISCUSSIONS

3.1 Data Preparation and Pre-Processing

The data preparation process begins with collecting, merging, and cleaning datasets from various sources, including Amazon Multilingual, IndoNLU, and AfriSenti. The initial dataset consisted of more than 300,000 user reviews in 10 languages, with about 30% coming from low-resource languages.

Preprocessing includes:

1. Automatic Language Detection uses the langdetect library to filter and segment multilingual data.
2. Text Cleanup, including the removal of punctuation, URLs, emojis, and capitalization normalization.
3. Tokenization uses tokenizers from the multilingual mBERT model, with padding up to a maximum sequence length of 100.
4. Label Encoding: The sentiment label is converted into a hot-coded vector:
    a. Positive: [1, 0, 0]
    b. Neutral: [0, 1, 0]
    c. Negative: [0, 0, 1]

The data is then divided into 80% training, 10% validation, and 10% testing, maintaining a balanced distribution of labels.

The hybrid CNN-BiLSTM model is trained with the following parameters:

1. Age: 15
2. Batch Size: 32
3. Optimizer: Adam with learning levels 1e-4
4. Loss Function: Categorical Cross-Entropy
5. Early Termination: Applied with patience 3 based on validation loss

3.2 Performance Model

Table 1 presents the overall performance of the CNN-BiLSTM-mBERT hybrid model proposed on the multilingual sentiment classification task. The model achieved an accuracy of 84.3%, with the average precision of macros, memory, and F1 scores exceeding 82%. This metric shows that the model performs reliably across different classes of sentiment in a multilingual context.

Table 1. Overall Evaluation Metrics

Metric	Value
Accuracy	84.3%
Precision (macro)	83.5%
Recall (makro)	82.9%
F1-Score (macro)	83.1%

To further evaluate the model's performance in a low-resource environment, Table 2 provides a breakdown of the F1 scores for the selected language. Indonesian obtained the highest F1 score (84.6%), likely due to the relatively larger amount of labeled data. In contrast, Hausa and Yoruba show a slightly lower performance, which may be attributed to linguistic resources and a limited training sample.

Table 2. F1 Score in Terms of Language For Low-Resource Languages

Language	F1 Score
Indonesia	84.6%
Yoruba	78.2%
Hausa	76.4%
Swahili	80.1%
Tamil	82.3%

These results show that the model maintains good performance even for low-resource languages, with relatively small performance degradations compared to high-resource languages.

E-ISSN: 3109-869X Doi: http://doi.org/10.26594/coding.v1i1.4306

3.3 Analysis and Discussion

1. The CNN layer contributes to the extraction of local sentiment patterns such as short phrases ("very bad", "dissatisfied").
2. The BiLSTM layer captures the context of a long and complex sentence-level.
3. This model demonstrates resilience to cross-language structural variation due to the use of multilingual tokenizers and mBERT embedding.
4. The highest F1 score was achieved for Indonesian, likely due to the relatively larger training data compared to African languages such as Hausa or Yoruba.
5. This model is well generalized to invisible languages, attributed to the transferable properties of multilingual embedding.

3.4 Limitations

The main limitations of this research include:

1. Reliance on mBERT embedding, which, although multilingual, shows a bias towards dominant languages such as English.
2. Lack of sentiment lexicons and a list of endwords available for certain low-resource languages, which can hinder the quality of preprocessing.

3.5 Training Visualization

To monitor the dynamics of model learning, accuracy and loss curves during training are plotted. Interpretation:

1. The accuracy curve increases and stabilizes before the initial stop, indicating an efficient and less fitting training process.
2. The validation loss is flattened around epoch 10, marking the corresponding auto-stop point.
3. The language-specific confusion matrix reveals frequent misclassifications, especially between neutral and negative sentiments and in resource-constrained languages.
4. F1-Score comparison bar chart between models.
5. Training and validation curves for accuracy and loss show stable model convergence without significant overfitting.

3.6 Confusion Matrix

A confusion matrix is used to analyze prediction details more precisely:

Table 3. Sentiment Prediction Confusion Matrix

	Before: Negative	**Before: Neutral**	**Pred: Positive**
Actual: Negative	25	5	0
Current: Neutral	4	24	2
Actual: Positive	0	3	27

Interpretation:

1. Models most often confuse neutral sentiment with negative, especially in language with subtle emotional expressions.
2. Positive sentiment has the highest classification accuracy.

3.7 Comparative Analysis

The proposed hybrid model is compared with two basic models:

Table 4. Comparative Performance with the Base Model

Model Configuration	F1-Score (Macro)
mBERT only	81.2%
CNN + mBERT	82.1%
BiLSTM + mBERT	82.4%
CNN-BiLSTM without mBERT	77.3%
CNN-BiLSTM + mBERT (Full)	83.1%

Findings:

1. The hybrid CNN-BiLSTM model consistently outperforms the base model.
2. The combination of CNN and BiLSTM effectively captures spatial and sequential features.
3. Performance improves significantly when pre-trained embeds such as mBERT are used.

CONCLUSIONS

This research proposes a hybrid deep learning model based on the CNN-BiLSTM architecture for multilingual sentiment analysis, with a particular focus on low-resource languages. By integrating multilingual text representations of pre-trained embedding (mBERT) and combining the strengths of CNN in capturing spatial features and BiLSTM in sequential dependency modeling, the proposed model demonstrates superior performance compared to basic methods such as Naive Bayes and standalone BiLSTM.

Experimental results showed that the model achieved an overall accuracy of 84.3% and an F1 macro score of 83.1%, maintaining relatively consistent performance in a variety of languages, including Swahili, Hausa, and Yoruba. The use of pre-trained multilingual embedding significantly improves the generalization capabilities of the model, allowing for effective classification of sentiment even in languages with minimal training data.

Despite its effectiveness, several limitations are identified, including: 1) Reliance on multilingual embedding, such as mBERT, which may be biased toward high-resource languages such as English; 2) Lack of comprehensive linguistic resources (e.g., sentiment lexicons or endpoint lists) for some languages with low resources, which can affect preprocessing and overall performance. Nevertheless, the hybrid approach holds great promise for the development of inclusive and adaptable multilingual NLP systems.

REFERENCES

Ardiansyah, A., Argarini Pratama, E., Imam Fadlilah, N., & Bina Sarana Informatika, U. (2024). *Analisis Sentimen Pengguna Terhadap Aplikasi ChatGPT Di Google Play Store: Penerapan Algoritma Support Vector Machine*. *11*(2), 247–254.

Efraim, D. A. (2023). *Analisis Sentimen Pada Sosial Media Instagram Menggunakan Algoritma Naive Bayes (Studi Kasus : Timnas Futsal Indonesia)*. *April 2012*, 498–509.

Fauzi, A., Akbar, M. F., & Asmawan, Y. F. A. (2019). Sentimen Analisis Berinternet Pada Media Sosial dengan Menggunakan Algoritma Bayes. *Jurnal Informatika*, *6*(1), 77–83. https://doi.org/10.31311/ji.v6i1.5437

Iin, I., Supriatna, R., Mulyawan, M., & Rohman, D. (2024). Penerapan Natural Language Processing Dalam Analisis Sentimen Cawapres 2024 Menggunakan Algoritma Naive Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 1109–1115. https://doi.org/10.36040/jati.v8i1.8572

E-ISSN: 3109-869X Doi: http://doi.org/10.26594/coding.v1i1.4306

Intan, S. F., Permana, I., Salisah, F. N., Afdal, M., & Muttakin, F. (2023). Perbandingan Algoritma KNN, NBC, dan SVM: Analisis Sentimen Masyarakat Terhadap Perparkiran di Kota Pekanbaru. *JUSIFO (Jurnal Sistem Informasi)*, *9*(2), 85–96. https://doi.org/10.19109/jusifo.v9i2.21357

Jain, T. I., & Nemade, D. (2010). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *International Journal of Computer Applications*, *7*(5), 12–21. https://doi.org/10.5120/1160-1453

Lutfina, E., Andriana, W., Quamila, S., Wiratmaja, P., & Febrianti, E. (2024). *Science, Technology and Management Journal Metode dan Algoritma Dalam Sentimen Analisis: Systematic Literature Review Info Artikel*. *4*(2), 67–79. http://journal.unkartur.ac.id/index.php/stmj

Mailoa, F. F. (2021). Analisis sentimen data twitter menggunakan metode text mining tentang masalah obesitas di indonesia. *Journal of Information Systems for Public Health*, *6*(1), 44. https://doi.org/10.22146/jisph.44455

Maulana, I., Apriandari, W., & Pambudi, A. (2023). Analisis Sentimen Berbasis Aspek Terhadap Ulasan Aplikasi Mypertamina Menggunakan Support Vector Machine. *IDEALIS : InDonEsiA JournaL Information System*, *6*(2), 172–181. https://doi.org/10.36080/idealis.v6i2.3022

Maulana, M. A., Setyanto, A., & Kurniawan, M. P. (2018). Analisis Sentimen media Sosial Universitas Amikom Yogyakarta Sebagai Sarana Penyebaran Informasi Menggunakan Algoritma Klasifikasi Svm. *Seminar Nasional Teknologi Informasi Dan Multimedia 2018 UNIVERSITAS AMIKOM Yogyakarta, 10 Februari 2018*, 7–12.

Maulana, R., Voutama, A., & Ridwan, T. (2023). Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan Algoritma NBC. *Jurnal Teknologi Terpadu*, *9*(1), 42–48. https://doi.org/10.54914/jtt.v9i1.609

Normah, Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174–180. https://doi.org/10.31294/jtk.v4i2

Pahtoni, T. Y., & Jati, H. (2024). Analisis Sentimen Data Twitter Terkait Chatgpt Menggunakan Orange Data Mining. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(2), 329–336. https://doi.org/10.25126/jtiik.20241127276

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 1320–1326. https://doi.org/10.17148/ijarcce.2016.51274

Pratama, M. Y., Putri, U. A., Angraini, P. A. D., Puspita, D., & Kurniawan, F. (2023). Analisis Sentimen Chat GPT sebagai Masa Depan Pekerja pada Media Sosial Youtube menggunakan Algoritma Klasifikasi Naive Bayes. *Explore: Jurnal Sistem Informasi Dan Telematika*, *14*(2), 193. https://doi.org/10.36448/jsit.v14i2.3391

Putranti, N. D., & Winarko, E. (2014). Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *8*(1), 91. https://doi.org/10.22146/ijccs.3499

Ramadhani, B., & Suryono, R. R. (2024). Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse. *Jurnal Media Informatika Budidarma*, *8*(2), 714. https://doi.org/10.30865/mib.v8i2.7458

Winoto, D., Desta Aditia, V., Sorisa, C., Priskila, R., & Handrianus Pranatawijaya, V. (2024). Analisis Sentimen Pada Ulasan Pengguna Terhadap Aplikasi Pembelajaran Bahasa Duolingo: Menggunakan Algoritma Naïve Bayes Dan K-Nearest Neighbor. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 3230–3236. https://doi.org/10.36040/jati.v8i3.9647