International Journal Science and Technology
IJST Vol 3 No. 3 | November 2024 | ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 11-21

COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA

Taufiq Dwi Purnomo[1*], Joko Sutopo[2]
Informatics, Faculty of Science & Technology, University of Technology Yogyakarta, Indonesia

Article History
Received : November 2024
Revised : November 2024
Accepted : November 2024
Published : November 2024

Corresponding author*:
taufiq.5210411277@student.uty.ac.id

No. Contact:
+628995275067

Cite This Article:
Taufiq Dwi Purnomo and Joko Sutopo, "COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA ", IJST, vol. 3, no. 3, pp. 11–21, Nov. 2024.

DOI:
doi.org/10.56127/ijst.v3i3.1739

Abstract: This study compared the performance of eight pre-trained BERT-based models for sentiment analysis across ten regional languages in Indonesia. The objective was to identify the most effective model for analyzing sentiment in low-resource Indonesian languages, given the increasing need for automated sentiment analysis tools. The study utilized the NusaX dataset and evaluated the performance of IndoBERT (IndoNLU), IndoBERT (IndoLEM), Multilingual BERT, and NusaBERT, each in both base and large variants. Model performance was assessed using the F1-score metric. The results indicated that models pre-trained on Indonesian data, specifically IndoBERT (IndoNLU) and NusaBERT, generally outperformed the multilingual BERT and IndoBERT (IndoLEM) models. IndoBERT-large (IndoNLU) achieved the highest overall F1-score of 0.9353. Performance varied across the different regional languages. Javanese, Minangkabau, and Banjar consistently showed high F1 scores, while Batak Toba proved more challenging for all models. Notably, NusaBERT-base underperformed compared to IndoBERT-base (IndoNLU) across all languages, despite being retrained on Indonesian regional languages. This research provides valuable insights into the suitability of different pre-trained BERT models for sentiment analysis in Indonesian regional languages.

Keywords: Sentiment Analysis, BERT, Indonesian Regional Languages, IndoBERT, NusaBERT

INTRODUCTION

With the rapid development of the internet and social media, a large amount of text data containing emotional information has been generated. Sentiment analysis, as a technique for identifying and processing emotional tendencies, has become an important research direction in natural language processing (NLP) [1]. Sentiment analysis, also often referred to as opinion mining, is an automated process that aims to understand, extract, and process text data to obtain the information contained in an opinion statement. The goal of sentiment analysis is to assess the opinion or tendency of a person's view towards an issue or object, whether it is positive, negative, or neutral [2].

In previous research, traditional methods have been widely used in sentiment analysis, such as Naïve Bayes [3], Support Vector Machine (SVM) [4], decision tree, and random forest [5]. With the significant development of Deep Learning-based methods, sentiment analysis has experienced significant improvements in terms of accuracy and context understanding. With the advantages offered by Deep Learning-based methods, many researchers have adopted them in sentiment analysis cases, such as the use of the Artificial Neural Network (ANN) method [6]. Along with the development of ANN, various deep learning architectures such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) have emerged, which allow models to handle sequential data and maintain context within the text. Research conducted by Raza [7] compared three methods, namely RNN, LSTM, and GRU, in the case of cloud computing customer sentiment analysis, with results showing that GRU is the best method in that case. Further developments such as Transfomer models [8], like Bidirectional Encoder Representations from Transformers (BERT) [9] and Generative Pre-trained Transformer (GPT) [10], further enhance capabilities in sentiment analysis, enabling a better understanding of nuances and emotions in more complex and diverse texts.

Indonesia, as a country with extraordinary linguistic and cultural richness, has more than 700 regional languages used by people in various regions [11]. This diversity, while reflecting cultural richness, presents significant challenges in sentiment analysis. Most of these regional languages are low-resource languages with very limited availability of digital data [12], considering that most research related to sentiment analysis in Indonesia is still dominated by Indonesian-language text [13], [14], [15].

BERT has proven effective in various NLP tasks, including sentiment analysis. Pre-trained BERT models can be fine-tuned with labeled data for specific languages. BERT was originally developed for language understanding in English, but it has been adapted and improved for various other languages, including Indonesian. Several efforts have been made to develop BERT models for Indonesian, such as IndoBERT [16], [17] and NusaBERT [18]. These models are trained on large Indonesian text data and have shown good performance in various language understanding tasks, one of which is sentiment analysis.

Several studies have implemented Indonesian BERT models for specific use cases. Research by Nugroho K [19] demonstrates the performance of BERT-Base multilingual and IndoBERT-Base in sentiment analysis on user reviews of the 2020 Google Play Best Apps, with IndoBERT-Base achieving the highest model accuracy of 84%. Additionally, research by Geni L [20] on sentiment analysis of Twitter data leading up to the 2024 presidential election shows that IndoBERT large-p1 achieved an accuracy of 83.5%. Furthermore, research by Basbeth F [21] on classifying emotions in sentences using IndoBERT resulted in an accuracy of 83%.

Although BERT models for Indonesian have shown good performance in sentiment analysis tasks for Indonesian sentences, the performance of these models on sentiment analysis tasks for various regional languages in Indonesia remains unknown. Given the diversity of regional languages in Indonesia, which have their own complexities and nuances, it is necessary to evaluate the ability of these BERT models to understand and analyze sentiment in regional language texts in Indonesia.

The core problem that this research aims to address is the lack of comprehensive evaluation and comparison of pre-trained BERT-based models for sentiment analysis in Indonesian regional languages. While BERT models have been adapted for Indonesian [16], [17], [18], their effectiveness in analyzing sentiments expressed in the diverse regional languages of Indonesia remains unexplored. This gap in knowledge is significant because it hinders the development of accurate and culturally sensitive NLP tools for these languages. By comparing the performance of various BERT models across multiple regional languages, this research seeks to identify which models are most effective for each language and whether there are significant variations in performance across different languages and models. This information is crucial for developing more accurate and tailored NLP tools for Indonesia's linguistically diverse population.

This research aims to compare the performance of several pre-trained BERT-based Transformer models, including IndoBERT (IndoNLU) [16], IndoBERT (IndoLEM) [17], Multilingual BERT [22], and NusaBERT [18] in analyzing the sentiment of text in 10 Indonesian regional languages including Acehnese, Balinese, Banjar, Bugis, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Batak Toba. We will use the NusaX sentiment dataset [23] and evaluate performance using the F1-score. This comparative analysis will provide valuable insights into the suitability and effectiveness of different BERT models for Indonesian regional language sentiment analysis, informing future research and development in this critical area.

RESEARCH METHOD

The comparison process in this research is carried out through several structured and systematic stages, as illustrated in Figure 1. The process begins with the use of the NusaX dataset [23], which contains sentiment data for 10 regional languages in Indonesia. Before using the data, a tokenizer selection stage is performed to determine the appropriate tokenization method, considering that each study uses different tokenization for the BERT model. Next, the data from the NusaX dataset is processed through the tokenization stage based on the selected tokenizer. The tokenized data is then divided into two parts: training data and test data. Concurrently with the data preparation process, the BERT model selection stage is also carried out. The BERT model selection is also based on the corresponding tokenizer. The selected model is then fine-tuned using the training data. After the fine-tuning process is complete, the model's performance is evaluated using the test data, measuring its performance with the F1-score metric. This workflow ensures that the model is trained with representative data and evaluated with unseen data. This process is performed on all the pre-trained BERT models being tested.

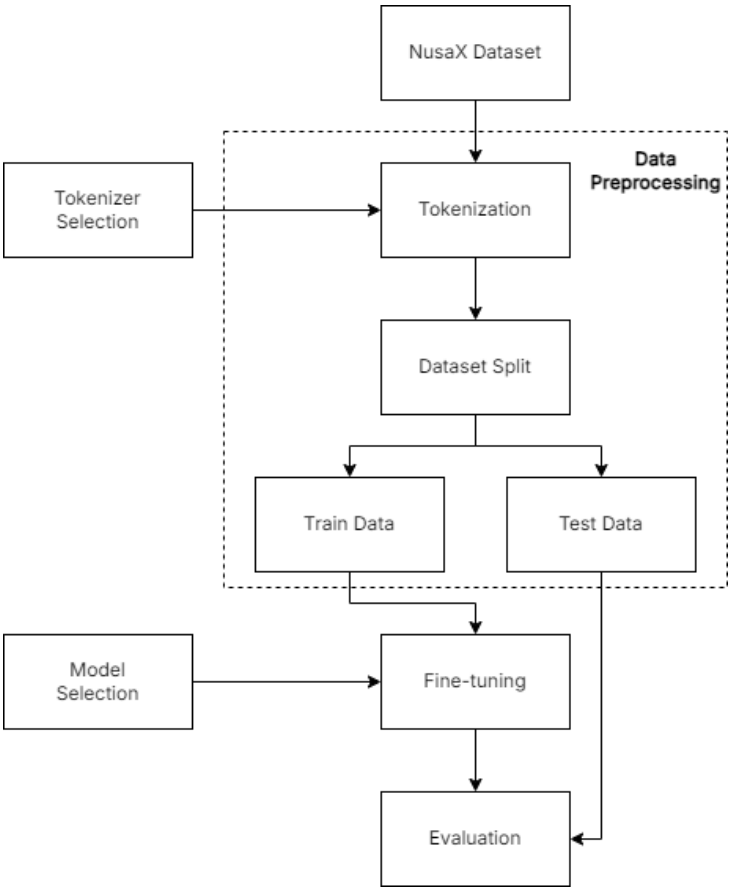

Figure 1. Research Method Workflow

.

Dataset

The dataset used in this study is NusaX, which provides resources for regional language sentiment analysis in Indonesia [23]. NusaX covers sentiment data from ten regional languages: Acehnese (ace), Balinese (ban), Banjarese (bjn), Buginese (bug), Madurese (mad), Minangkabau (min), Javanese (jav), Ngaju (nij), Sundanese (sun), and Toba Batak (bbc). This dataset was created through the translation of SmSA [24], an existing Indonesian sentiment analysis dataset, using skilled bilingual speakers, which helps ensure cultural relevance across the represented languages [23]. As shown in Table 1, each language dataset within NusaX contains a total of 1,000 samples, distributed across three sentiment categories: 383 negative sentiments, 239 neutral sentiments, and 378 positive sentiments. This balanced distribution, coupled with NusaX's diverse linguistic coverage, makes it a suitable benchmark for evaluating and comparing the performance of pre-trained BERT models for sentiment analysis in the context of Indonesian regional languages.

Table 1. NusaX Dataset Label Distribution

Language	Negative	Neutral	Positive
ace	383	239	378
ban	383	239	378
bjn	383	239	378
bug	383	239	378
mad	383	239	378
min	383	239	378
jav	383	239	378
nij	383	239	378
sun	383	239	378
bbc	383	239	378

Table 2 shows examples of the sentiment text data from each of the ten Indonesian regional languages represented in the dataset.

Table 2. Examples of NusaX Sentiment Sata

Text	Language	Label
Guna shazam: ngat teupeue lagu-lagu ajayeb nyang jiputa supe gocar	ace	neutral
Pelayan lan pemilik sane negak ring kasir nenten komunikatif lan tusing bisa ngemaan rekomendasi sane porsi menu sane ka pesen. Rasane standar. Parkirne meweh.	ban	negative
Amun ka sini manyiapakan parut kusung lah karena di sini kunsipnya kawa makan sapuasnya. Puas banar makan di sini, sasuai ja wan haraganya. Ulun katujuannya nasi goreng, nyaman banar.	bjn	positive
Iya' lauwitai maega pappake indosat nakennai attappereng pulsa, kasi' na.	bug	negative
Engkok rua la pendet se aghigire. Dhina la engkok beih se ngalak a pole paketanna ben engkok e kerema via ekspedisi laenna bhei. Kecewa sara la keberempa kalena.	mad	negative
Katiko nio manyantap variasi makanan, piliahannyo salalu Hanamasa. Lokasi tampeknyo cukuik lamak. Piliahan makanannyo banyak, dari mulai makanan ringan, baka-bakaan, abuihan hinggo makanan panutuik. Sangaik indak mangacewaan.	min	positive
Warung mangan nduwe suasana sing nyaman karo gaya kampung lawas. Luwih sekeco diparani wektu sore sampek dina wengi ning kahanan sing luwih romantis karo cahyaning remeng-remeng saka obor lan geni unggun. Panganan sing disajikne rena-rena, mulai panganan indonesia sampek panganan eropa serba ana. Cocok sanget kanggo ngentekne wektu karo kanca lak keluarga.	jav	positive
Pas wayah tuh mawi gawian visual telu ratus jahawen puluh derajat hapan samsung! Numunlah langkah mudah hong video tuh.	nij	neutral
Tina segi letak / lokasi mah gampil dipilarianana. Bangunan jeung eusina sae, komo aya taman alit di lebetna nu dieusian ku gentong nu caina ngocor. Nyieun suasana na tambih romantis jeung pikaresepeun. Menu nu disadiakeun lumayan variatif jeung tina segi rasa lumayan raos.	sun	positive
Bapakku i ma sahalak parkarejo ni net tv.	bbc	neutral

Data Preprocessing

Since the NusaX dataset was previously cleaned by Winata G [23], we only performed tokenization and data splitting.

Tokenization

In Natural Language Processing (NLP), tokenization is a fundamental step that breaks down text into smaller units called tokens [25]. In more modern approaches, tokenization methods such as WordPiece tokenization [26] and SentencePiece with Byte Pair Encoding and Unigram [27] are used.

BERT models require input text to be tokenized into smaller units using model-specific tokenizers, necessitating separate tokenization steps for each model to ensure proper input formatting. Two special tokens play crucial roles in BERT's architecture: the [CLS] token, prepended to the beginning of every input sequence and used as a representation of the entire sentence or sequence for classification tasks, and the [SEP] token, which separates two sentences in tasks where BERT processes pairs of sentences, indicating the boundary between them [9]. These tokenization practices and special tokens are fundamental to BERT's ability to understand and process text effectively across various natural language processing tasks.

IndoBERT (IndoNLU) uses SentencePiece with Byte Pair Encoding (BPE) for tokenization, with a vocabulary size of 30.522 for the IndoBERT model variant and 30.000 for the IndoBERT-lite variant [16]. IndoBERT (IndoLEM) uses WordPiece tokenization with a vocabulary size of 31.923 [17], while NusaBERT employs WordPiece with a vocabulary size of 32.032 [18]. Finally, Multilingual BERT (mBERT) uses WordPiece tokenization with a vocabulary size of 105.879 [22]. Table 3 shows examples of tokenization results from each BERT research.

Table 3. Examples of Tokenization Result

Text	Tokenizer	Token
Angel banget mercoyoi wong sing wis tau khianat	IndoBERT (IndoNLU)	['[CLS]', 'angel', 'banget', 'merc', '##oyo', '##i', 'wong', 'sing', 'wis', 'tau', 'kh', '##ianat', '[SEP]']
	IndoBERT (IndoLEM)	['[CLS]', 'angel', 'banget', 'merc', '##oyo', '##i', 'wong', 'sing', 'wis', 'tau', 'kh', '##iana', '##t', '[SEP]']
	NusaBERT	['[CLS]', 'angel', 'banget', 'merc', '##oyo', '##i', 'wong', 'sing', 'wis', 'tau', 'kh', '##ianat', '[SEP]']
	Multilingual BERT	['[CLS]', 'angel', 'bang', '##et', 'merc', '##oy', '##oi', 'wong', 'sing', 'wis', 'tau', 'khi', '##anat', '[SEP]']

Data Splitting

In this study, we employed an 80/20 data splitting strategy. Specifically, 80% of the NusaX dataset was allocated for training the pre-trained BERT models. The remaining 20% of the data was reserved as test data, used to evaluate the performance of the fine-tuned models on unseen examples, as measured by the F1-score. This resulted in a training set of 8.000 data points and a test set of 2.000 data points.

BERT Models

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning model based on the encoder transformer architecture, specifically designed to understand the context in a sentence by considering the words on its left and right [9]. BERT is structured in components namely Input Embedding, Positional Encoding, Normalization layer, Multi-head Attention, Feed Forward, and output classification layer. Figure 2 shows the basic BERT architecture of the model.

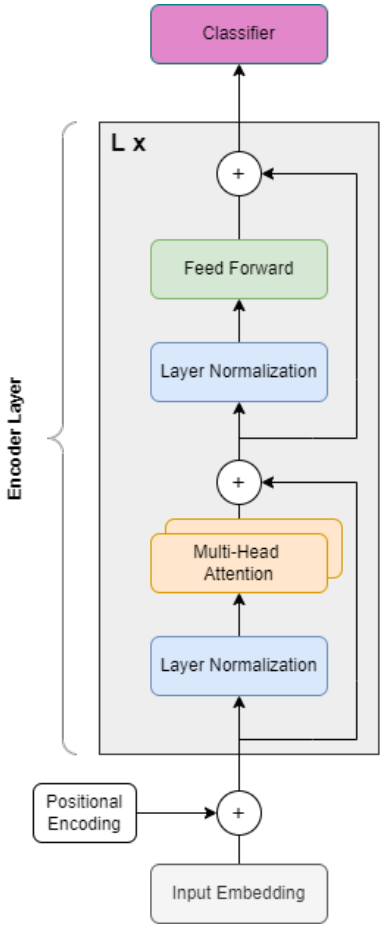

Figure 2. BERT Architecture

The Input Embeddings component serves to capture semantic information about each word and feeds it into the model. The input embedding layer is a combination of token embedding and segment embedding.

The Inpur embedding is then added with Positional Encoding to provide information about the position of each word. After the Input embedding component is the encoder layer. BERT consists of multiple stacked encoder layers (Lx represents the number of layers). Within the Encoder Layer is the Multi-Head Attention component. This is the core component of the transformer architecture. It allows the model to attend to different parts of the input sequence and learn the relationships between words. There is also a Feed forward component that serves to perform non-linear transformations on the representation generated by multi-head attention and a normalization layer component to stabilize training and improve performance. The last component is the classification layer which in this study contains a layer to predict positive, negative, or neutral sentiment labels.

IndoBERT (IndoNLU)

IndoBERT (IndoNLU) was trained using a dataset called Indo4B, which consists of around 4 billion words (~23 GB) and 250 million sentences in Indonesian. The Indo4B dataset was collected from various sources such as online news, social media, Wikipedia, online articles, and subtitle texts [16]. IndoBERT offers two main variants: IndoBERT-lite and IndoBERT-base. IndoBERT-lite is based on the ALBERT (A Lite BERT) model [28], which has fewer parameters compared to IndoBERT-base.

IndoBERT-base has two models: IndoBERT-base and IndoBERT-large. IndoBERT-base has 768 hidden units, 12 layers, and 12 attention heads in each layer, with a total of 124.443.651 parameters. Meanwhile, IndoBERT-large has 1024 hidden units, 24 layers, and 16 attention heads in each layer, with a total of 335.144.963 parameters.

IndoBERT-lite also has two models: IndoBERT-lite-base and IndoBERT-lite-large. IndoBERT-lite-base has 768 hidden units, 12 layers, and 12 attention heads in each layer, with a total of 11.685.891 parameters. Meanwhile, IndoBERT-lite-large has 1024 hidden units, 24 layers, and 16 attention heads in each layer, with a total of 17.687.043 parameters. Although both lite models have the same hidden units, layers, and attention heads as the base version, what makes the lite version smaller is the ALBERT architecture, where each layer in the lite version has the same parameter value, known as Cross-layer Parameter Sharing [28].

IndoBERT (IndoLEM)

IndoBERT (IndoLEM) was trained using a text dataset containing approximately 220 million words in Indonesian. The training data came from three main sources: Indonesian Wikipedia with 74 million words, news articles from Indonesian newspapers such as Kompas, Tempo, and Liputan6 with 55 million words, and the Indonesian Web Corpus with 90 million words [17].

IndoBERT (IndoLEM) has only one model variant, which is IndoBERT-base, with 768 hidden units, 12 layers, and 12 attention heads in each layer, totaling 110.560.515 parameters.

Multilingual BERT

Multilingual BERT (mBERT) is a pre-trained model capable of handling 104 languages, including Indonesian. It's trained on a massive corpus of Wikipedia text across these languages. The advantage of multilingual models is their ability to generalize across languages, offering the potential for zero-shot cross-lingual transfer learning [22].

Multilingual BERT has 768 hidden units, 12 layers, and 12 attention heads in each layer, totaling 167.358.723 parameters.

NusaBERT

NusaBERT is a language model built upon IndoBERT (IndoNLU) [16], aimed at addressing the linguistic diversity in Indonesia. NusaBERT continues the pre-training of IndoBERT (IndoNLU) on multilingual text data that includes regional languages of Indonesia [18]. The pre-training text data for NusaBERT consists of 13 languages, including Indonesian, Javanese, Sundanese, Acehnese, Malay, Minangkabau, Banjar, Balinese, Gorontalo, Banyumasan, Bugis, Nias, and Tetum. The data comes from various sources, including CulturaX [29], Wikipedia, and subset of NLLB [30].

NusaBERT offers two models: NusaBERT base and NusaBERT large. The NusaBERT base model features 768 hidden units, 12 layers, and 12 attention heads per layer, totaling 110.644.227 parameters. The NusaBERT large model has 1024 hidden units, 24 layers, and 16 attention heads per layer, totaling 336.691.203 parameters.

Hyperparameter Configuration

Hyperparameters are critical components in the development and optimization of machine learning models, as they control the behavior of the training process [31]. Unlike model parameters, which are learned

during training, hyperparameters must be predefined and tuned to achieve optimal performance. Table 4 presents the hyperparameters used in our fine-tuning process.

Table 4. Hyperparameters for Fine-tuning

No	Hyperparameter	Value
1	Epoch	5
2	Batch Size	64
3	Optimizer	AdamW
4	Initial Learning Rate	3e-5
5	Weight Decay	0.01
6	Loss Function	Cross Entropy Loss
7	Learning Rate Scheduler	Cosine Scheduler
8	Computation	Auto Mixed Precision (FP16 and FP32)
9	Max Sequence Length	128

The model was fine-tuned using selected hyperparameters to optimize performance. We set the number of epochs to 5 with batch size of 64. For optimization, we used the AdamW optimizer [32], which is an improved version of Adam that incorporates weight decay.

The initial learning rate was set to 3e-5, a relatively small value to ensure fine adjustments to the pre-trained weights. To prevent overfitting, we applied a weight decay of 0.01. The loss function used was Cross Entropy Loss, which is standard for classification tasks. To manage the learning rate during training, we implemented a Cosine Scheduler [33], which gradually decreases the learning rate following a cosine curve.

To maximize computational efficiency, we utilized Auto Mixed Precision [34], combining FP16 and FP32 calculations. The maximum sequence length was set to 128 tokens.

Metric Evaluation

The performance measurement of each model uses the F1-score metric. The F1-score provides a single score that balances both precision and recall. The F1-score is a performance metric bounded between 0 and 1. A score of 1 represents perfect precision and recall, indicating optimal model performance. Conversely, a score of 0 signifies the worst possible performance. The F1-score formula can be expressed as shown in Equation 1.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (1)$$

Where $Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$, $Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$. Precision is the fraction of correctly predicted positive instances out of all instances predicted as positive. Recall is the fraction of correctly predicted positive instances out of all actual positive instances.

RESULT AND DISCUSSION

In this section, we present and analyze the performance of 8 pre-trained BERT-based models in sentiment analysis across ten regional languages in Indonesia. We compare the effectiveness of the models using the NusaX dataset. Our evaluation is based on the F1-score metric. We start with an overview of the overall performance, followed by a detailed analysis of language-specific results, architecture impact, and pre-training dataset impact.

Overall Performance Comparison

IndoBERT-large (IndoNLU) achieved the highest overall F1 score of 0.93534, followed by IndoBERT-base (IndoNLU) at 0.93190 and NusaBERT-large at 0.93134. The multilingual BERT (mBERT) model showed competitive performance with the IndoBERT-base (IndoLEM) model with an average F1 score of mBERT of 0.897975 and IndoBERT-base (IndoLEM) of 0.89753. The lite version of IndoBERT models from IndoNLU showed good performance despite the small model size, with IndoBERT-lite-large achieving an average F1 score of 0.9015, while IndoBERT-lite-base achieved a score of 0.88383. Meanwhile,

NusaBERT-base showed a lower average performance of 0.8888 compared to the other base models. Figure 3 shows the overall f1 score values across the 8 models.

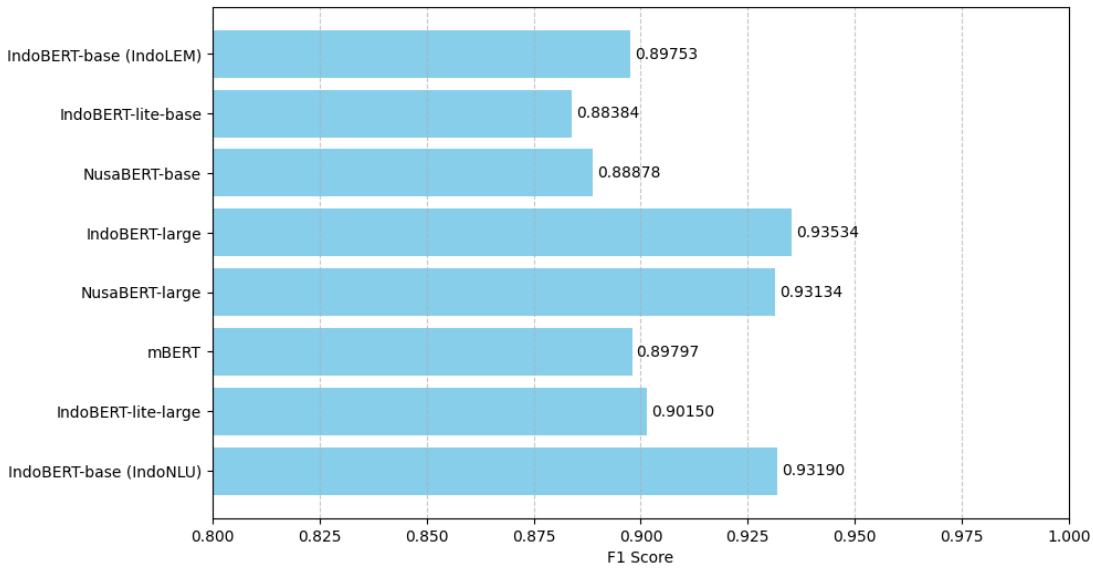

Figure 3. Overall Performance of Each Model

Performance Analysis by Language

Table 5. F1-score of All Models in Each Language

Model	ace	ban	bbc	bjn	bug	jav	mad	min	nij	sun
IndoBERT-base (IndoLEM)	0.8789	0.8900	0.8210	0.9352	0.8558	0.9397	0.9299	0.9353	0.8790	0.9104
IndoBERT-base (IndoNLU)	0.9149	0.9147	**0.8807**	0.9749	**0.9007**	0.9699	0.9350	**0.9851**	0.8838	0.9601
IndoBERT-large	**0.9544**	0.9097	0.8497	0.9599	0.8904	**0.9849**	**0.9450**	0.9800	0.9045	**0.9751**
IndoBERT-lite-base	0.9098	0.8848	0.8098	0.9397	0.8046	0.9349	0.9200	0.9047	0.8333	0.8955
IndoBERT-lite-large	0.8845	0.9147	0.8109	0.9500	0.8406	0.9497	0.9249	0.9550	0.8640	0.9207
NusaBERT-base	0.9040	0.8945	0.7802	0.9400	0.8254	0.9547	0.8851	0.9449	0.8288	0.9301
NusaBERT-large	0.9297	**0.9196**	0.8349	**0.9800**	0.8606	**0.9849**	0.9447	0.9700	**0.9193**	0.9701
mBERT	0.8751	0.9094	0.8307	0.9297	0.8607	0.9101	0.9196	0.9350	0.8792	0.9303

Table 5 shows the performance of all models in each language. Javanese (jav) consistently showed high F1 values in most models, with IndoBERT-large and NusaBERT-large models achieving the same F1 value of 0.9849. Minangkabau (min) also showed strong performance, with IndoBERT-base (IndoNLU) achieving an F1 value of 0.9851. In contrast, Batak Toba (bbc) proved to be the most challenging language for all models, with F1 values ranging from 0.7802 (NusaBERT base) to 0.8807 (IndoBERT IndoNLU base).

Banjar (bjn) and Sundanese (sun) also showed high performance in most models, with NusaBERT-large achieving the highest F1 value of 0.9800 for Banjar and IndoBERT-large achieving 0.9751 for Sundanese. Acehnese (ace) shows more variability, with IndoBERT-large performing very well (0.9544) while the other models have lower scores.

For Balinese (ban), most models performed relatively well, with scores consistently above 0.88, and NusaBERT-large leading the way with a score of 0.9196. Bugis (bug) proved more challenging, with scores

ranging from 0.8046 (IndoBERT-lite-base) to 0.9007 (IndoBERT-base IndoNLU). Madurese (mad) performed well overall, with NusaBERT-large achieving the highest score of 0.9450.

Ngaju (nij) showed more variation in model performance, with values ranging from 0.8288 (NusaBERT-base) to 0.9193 (NusaBERT-large). Interestingly, the performance in Nias improves significantly with larger models, as seen in the jump from NusaBERT-base to NusaBERT-large.

Surprisingly, NusaBERT-base, which is a model based on IndoBERT-base (IndoNLU) that has been retrained with local Indonesian languages, did not perform as well as expected. NusaBERT-base consistently performed poorly compared to IndoBERT-base (IndoNLU) across all languages, and in some cases, even performed worse than mBERT which was not specifically trained for Indonesian languages. For example, in Toba Batak (BBC), NusaBERT-base had the lowest score (0.7802) among all models, much lower than IndoBERT-base (IndoNLU) which reached 0.8807.

Overall, IndoBERT (IndoNLU) and NusaBERT, especially the larger version, tended to outperform mBERT in most languages, indicating that models pre-trained on Indonesian and Indonesian regional languages have a significant advantage in handling comprehension of Indonesian regional languages. However, the unexpected performance of the NusaBERT-base model, which is expected to show high performance since it has been trained with Indonesian local language data, shows often lower performance than the other models.

CONCLUSION

This study compares the performance of eight pre-trained BERT-based models for sentiment analysis across ten regional languages in Indonesia using the NusaX dataset. The results show that models specifically pre-trained on Indonesian data, specifically IndoBERT (IndoNLU) and NusaBERT, generally outperform multilingual BERT models in most languages. IndoBERT-large (IndoNLU) achieved the overall highest F1 value of 0.93534, indicating the importance of language-specific pre-training. However, performance varied significantly across different regional languages, with Javanese, Minangkabau, and Banjar consistently showing high F1 values, while Toba Batak proved more challenging for all models. Interestingly, the NusaBERT base, despite being retrained with regional languages in Indonesia, performed lower compared to the IndoBERT (IndoNLU) base across all languages.

REFERENCES

[1] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis," 2024. doi: https://doi.org/10.48550/arXiv.2403.08217.

[2] A. Halim and A. Safuwan, "Analisis Sentimen Opini Warganet Twitter Terhadap Tes Screening Genose Pendeteksi Virus Covid-19 Menggunakan Metode Na\"\ive Bayes Berbasis Particle Swarm Optimization," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 5, no. 1, pp. 170–178, 2023, doi: https://doi.org/10.51401/jinteks.v5i1.2229.

[3] Z. Li, R. Li, and G. Jin, "Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary," *IEEE Access*, vol. 8, pp. 75073–75084, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2986582.

[4] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 2022, pp. 163–167. doi: https://doi.org/10.1109/ISMODE53584.2022.9742906.

[5] M. Aufar, R. Andreswari, and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 2020, pp. 1–7. doi: https://doi.org/10.1109/ICoDSA50139.2020.9213078.

[6] A. Thakkar, D. Mungra, A. Agrawal, and K. Chaudhari, "Improving the Performance of Sentiment Analysis Using Enhanced Preprocessing Technique and Artificial Neural Network," *IEEE Trans Affect Comput*, vol. 13, no. 4, pp. 1771–1782, 2022, doi: https://doi.org/10.1109/TAFFC.2022.3206891.

[7] M. R. Raza, W. Hussain, and J. M. Merigó, "Cloud Sentiment Accuracy Comparison using RNN, LSTM and GRU," in *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2021, pp. 1–5. doi: https://doi.org/10.1109/ASYU52992.2021.9599044.

[8] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. doi: https://doi.org/10.48550/arXiv.1706.03762.

[9] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, doi: https://doi.org/10.48550/arXiv.1810.04805.
[10] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," 2020. doi: https://doi.org/10.48550/arXiv.2005.14165.
[11] S. Zein, *Language Policy in Superdiverse Indonesia*. New York : Routledge, 2020. | Series: Routledge studies in sociolinguistics: Routledge, 2020. doi: https://doi.org/10.4324/9780429019739.
[12] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," 2022. doi: https://doi.org/10.48550/arXiv.2203.13357.
[13] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl Soft Comput*, vol. 151, p. 111112, 2024, doi: https://doi.org/10.1016/j.asoc.2023.111112.
[14] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment analysis in Indonesian healthcare applications using IndoBERT approach," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, pp. 113–117, 2023, doi: https://dx.doi.org/10.14569/IJACSA.2023.0140813.
[15] P. Subarkah, P. Arsi, D. I. S. Saputra, A. Aminuddin, Berlilana, and N. Hermanto, "Indonesian Police in the Twitterverse: A Sentiment Analysis Perspectives," in *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2023, pp. 76–81. doi: https://doi.org/10.1109/ICITISEE58992.2023.10405357.
[16] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," 2020. doi: https://doi.org/10.48550/arXiv.2009.05387.
[17] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020. doi: https://doi.org/10.48550/arXiv.2011.00677.
[18] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," 2024. doi: https://doi.org/10.48550/arXiv.2403.01817.
[19] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "Bert fine-tuning for sentiment analysis on indonesian mobile apps reviews," in *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, 2021, pp. 258–264. doi: https://doi.org/10.48550/arXiv.2107.06802.
[20] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 746–757, 2023, doi: http://dx.doi.org/10.26555/jiteki.v9i3.26490.
[21] F. Basbeth and D. H. Fudholi, "Klasifikasi Emosi Pada Data Text Bahasa Indonesia Menggunakan Algoritma BERT, RoBERTa, dan Distil-BERT," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, pp. 1160–1170, 2024, doi: http://dx.doi.org/10.30865/mib.v8i2.7472.
[22] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. doi: https://doi.org/10.18653/v1/P19-1493.
[23] G. I. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 815–834. doi: https://doi.org/10.18653/v1/2023.eacl-main.57.
[24] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving Bi-LSTM Performance for Indonesian Sentiment Analysis Using Paragraph Vector," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2019, pp. 1–5. doi: https://doi.org/10.1109/ICAICTA.2019.8904199.
[25] G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*, H. van Halteren, Ed., Dordrecht: Springer Netherlands, 1999, pp. 117–133. doi: 10.1007/978-94-015-9273-4_9.
[26] Y. Wu *et al.*, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," 2016. doi: https://doi.org/10.48550/arXiv.1609.08144.
[27] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds.,

Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. doi: https://doi.org/10.18653/v1/D18-2012.

[28] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," 2020. doi: https://doi.org/10.48550/arXiv.1909.11942.

[29] T. Nguyen *et al.*, "CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages," 2023. doi: https://doi.org/10.48550/arXiv.2309.09400.

[30] N. Team *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation," 2022. doi: https://doi.org/10.48550/arXiv.2207.04672.

[31] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, doi: https://doi.org/10.1016/j.neucom.2020.07.061.

[32] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," 2019. doi: https://doi.org/10.48550/arXiv.1711.05101.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," 2017. doi: https://doi.org/10.48550/arXiv.1608.03983.

[34] P. Micikevicius *et al.*, "Mixed Precision Training," 2018. doi: https://doi.org/10.48550/arXiv.1710.03740.