OPEN ACCESS
Indo-JC
Indonesian Journal on Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 9, Issue. 2, August 2024. pp. 61-72
doi:10.34818/indojc.2024.9.2.934

Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates

Primanda Sayarizki[#1], Hasmawati [*2], Hani Nurrahmi [*3]

 # School of Computing, Telkom University
Bandung, Indonesia
[1] primanda@student.telkomunivesity.ac.id
[2] hasmawati@telkomuniversity.ac.id
[3] haninurrahmi@telkomunversity.ac.id

Abstract

In this modern era, Indonesian society widely utilizes social media, particularly Twitter, as a means to express their opinions. Every day, various opinions of Indonesian citizens are disseminated on this platform, including their views on prospective presidential candidates for the year 2024. Analyzing public opinions regarding prospective presidential candidates in 2024 is crucial to understanding the sentiment of the people toward these candidates, especially in the context of the upcoming presidential election. Such sentiment analysis can be conducted using deep learning techniques such as IndoBERT to acquire knowledge regarding the classification of sentiments as positive, neutral, or negative. IndoBERT is employed to generate vector representations that encapsulate the meaning of tokens, words, phrases, or texts. These representation vectors can then be input into a classification model to perform sentiment analysis. The sentiment classification model undergoes testing with a diverse set of tweets in the test dataset, which represent a wide range of public opinions. The evaluation results indicate an overall accuracy rate of 80%, with precision rates of 62% for negative sentiment, 81% for neutral sentiment, and 85% for positive sentiment. Additionally, the recall rates for each sentiment are 64% for negative, 81% for neutral, and 84% for positive, with corresponding F1-scores of 63%, 81%, and 85%, respectively.

Keywords: Sentiment Analysis, Indonesian Presidential Candidates, Presidential Election, IndoBERT, Deep Learning, Twitter, Public Opinion, Classification Model

Abstrak

Pada era modern ini, masyarakat Indonesia secara luas menggunakan media sosial, khususnya Twitter, sebagai sarana untuk menyuarakan pendapat mereka. Setiap hari, berbagai opini warga Indonesia tersebar di platform ini, termasuk pandangan mereka tentang calon presiden untuk tahun 2024. Menganalisis opini publik mengenai calon presiden tahun 2024 sangat penting untuk memahami sentimen masyarakat terhadap para calon tersebut, terutama dalam konteks pemilihan presiden yang akan datang. Analisis sentimen semacam itu dapat dilakukan menggunakan teknik pembelajaran mendalam seperti IndoBERT untuk memperoleh pengetahuan mengenai klasifikasi sentimen sebagai positif, netral, atau negatif. IndoBERT digunakan untuk menghasilkan representasi vektor yang merangkum makna token, kata, frasa, atau teks. Representasi vektor ini kemudian dapat dimasukkan ke dalam model klasifikasi untuk melakukan analisis sentimen. Model klasifikasi sentimen diuji dengan berbagai tweet dalam dataset uji, yang mewakili beragam opini publik. Hasil evaluasi menunjukkan tingkat akurasi keseluruhan sebesar 80%, dengan tingkat presisi sebesar 62% untuk sentimen negatif, 81% untuk sentimen netral, dan 85% untuk sentimen positif. Selain itu, tingkat recall untuk setiap sentimen adalah 64% untuk negatif, 81% untuk netral, dan 84% untuk positif, dengan skor F1 masing-masing sebesar 63%, 81%, dan 85%.

Received on April, 2024. Accepted on August, 2024

Kata Kunci: Analisis Sentimen, Calon Presiden Indonesia, Pemilihan Presiden, IndoBERT, Deep Learning, Twitter, Opini Publik, Model Klasifikasi

I. Introduction

The electoral procedure to choose the President and Vice President of Indonesia for the term 2024-2029 is set to occur during the 2024 Indonesian Presidential Election, slated for Wednesday, February 14, 2024. Based on this, it is anticipated that there will be a plethora of public opinions regarding the Indonesian presidential candidates for 2024 on social media platforms. The impact of social media, notably platforms like Twitter, is progressively becoming pivotal in molding public sentiment and impacting voter choices. [1]. With the abundance of public opinions expressed on Twitter, valuable insights into the current sentiments toward the Indonesian presidential candidates for 2024 can be obtained [2]. Hence, the objective of this study is to collect and analyze tweets from Twitter users to gauge public sentiment regarding the 2024 Indonesian presidential candidates.

This sentiment analysis aims to understand the public's views on three presidential candidates who will compete in the 2024 presidential election: Prabowo Subianto, Anies Baswedan, and Ganjar Pranowo. In terms of positive, negative, or neutral sentiment, this analysis will be used to assess the candidates' image in the eyes of Twitter users, determining the extent to which they are accepted or rejected by the public. IndoBERT is able to capture the meaning of Indonesian text more effectively. Consequently, it is plausible to infer that the IndoBERT model could exhibit enhanced effectiveness in distinguishing between positive, negative, or neutral sentiments in Indonesian tweet data, as opposed to conventional sentiment analysis models like machine learning-based or lexicon-based models [3].

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained model that performs word embedding in Natural Language Processing, where each word is transformed into a set of numerical vectors using the Transformer architecture [4]. Meanwhile, IndoBERT is a variant of the pre-trained BERT model specifically developed using the Indonesian language corpus. Throughout the training phase of BERT, it undergoes training using Masked Language Modeling (MLM) and Next Sentence Prediction techniques. These methods enable BERT to acquire proficiency in understanding language and its contextual nuances [5]. Sentiment analysis using IndoBERT can demonstrate better results compared to regular BERT, which uses a multilingual corpus and other pre-trained models [6]. Subsequently, IndoBERT has demonstrated superior efficacy when compared to K-Nearest Neighbors (KNN), Support Vector Machines (SVM), naive Bayes, decision trees, and random forest models [7].

Based on the exposition above, this research will conduct sentiment analysis on potential candidates for the Indonesian Presidency in 2024 on Twitter using the IndoBERT model. The dataset used will consist of tweets from Twitter users, collected using the Tweet Harvest library based on Python, with three keywords: "Prabowo Subianto," "Anies Baswedan," and "Ganjar Pranowo." The performance evaluation of the model will be calculated using the confusion matrix method. With this research, it is hoped that new methodologies for sentiment analysis specific to the context of Indonesian presidential candidates in 2024 using IndoBERT can be developed and evaluated, demonstrating the superiority of this model.

II. Literature Review

Several studies related to sentiment analysis of prospective presidential candidates in Indonesia in 2024 have been conducted by previous researchers, including research by Hananto and his team [8]. Sentiment analysis on Twitter was assessed employing three classification algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Naïve Bayes (NB). The findings indicated accuracies of 79.57%, 77.21%, and 55.80%, correspondingly.

Another study by Rizki et al. [9] utilized the Long Short Term Memory (LSTM) method for sentiment analysis of prospective presidential candidates in Indonesia on Twitter, achieving an accuracy rate of 76%. Furthermore, research on sentiment classification using the IndoBERT model has also been conducted. Siti et al. [10] focused on public opinions related to COVID-19 vaccination in Indonesia using a tweet dataset. They implemented three models, namely IndoBERT, IndoBERTweet, and CNN-LSTM. The results showed that IndoBERT was able to predict positive sentiment by 80%, while IndoBERTweet and CNN-LSTM achieved 68% and 53%, respectively.

In addition, Herlina et al. [5] conducted sentiment analysis research with fine-tuning using IndoBERT and R-CNN. The results showed an accuracy rate of 95.16%, precision of 94.05%, recall of 92.74%, and F1 score of 93.27%. A recent study by Nuzulul et al. [11] explored text classification on customer reviews in Indonesia. They used IndoBERT as the feature representation and CNN-XGBoost as the classification algorithm, achieving significant improvements in accuracy and F1 score compared to the baseline Word2Vec-CNN-XGBoost.

Overall, these studies reflect the latest developments in applying various sentiment analysis techniques, including the use of language models such as IndoBERT, to understand the views and opinions of the Indonesian public on prospective presidential candidates in 2024.

A. Sentiment Analysis

Sentiment analysis involves the automated assessment and categorization of emotions expressed in text-based data, including written feedback and social media updates. It falls under the umbrella of Natural Language Processing (NLP) and employs machine learning techniques to classify the emotional sentiment conveyed in textual content [12]. This technique involves categorizing information into positive, negative, or neutral sentiments at the sentence, document, or feature level [13]. Overall, sentiment analysis serves as a valuable tool for gathering information, making decisions based on opinions, and gaining insights into public sentiment [14].

The sentiment analysis process begins with selecting data. Next, the data is extracted, often formatted into textual data using scraping techniques. Subsequently, the textual data is analyzed to identify and classify the expressed emotions or opinions. Sentiment analysis can be conducted through diverse methodologies, including corpus-based strategies, dictionary-driven methods, and machine learning techniques.

B. Bert and IndoBERT

BERT is a deep learning model used to represent words contextually in pre-training for Natural Language Processing (NLP). There are two stages in BERT's workflow: pre-training and fine-tuning [15]. During pre-training, words are trained using tasks such as Masked Language Model (MLM) and Next Sentence Prediction (NSP). Subsequently, during fine-tuning, more specific tasks like sentiment analysis classification are performed. BERT, as its acronym suggests - Bidirectional Encoder Representations from Transformers, focuses on the encoding process and generates language models. The structure of the BERT model consists of bidirectional encoder transformers with multiple layers. BERT solely employs encoder stacks within Transformers and does not involve decoder stacks in its architecture, hence BERT concentrates solely on understanding the representation or context of the input. BERT is anticipated to depict contextual understanding of words through two approaches: referencing the preceding word by detailing from right to left, and also considering the succeeding word by composing predictions from left to right in a specific sequence [16].

IndoBERT is a BERT model pre-trained using an extensive and sanitized dataset of Bahasa Indonesia (Indo4B). This dataset comprises content sourced from multiple platforms, including social media, blogs, news articles, and websites [6]. IndoBERT undergoes training using a corpus of more than 220 million Indonesian words. This training dataset is sourced from three primary channels: Indonesian Wikipedia, contributing 74 million words, articles from prominent sources like Kompas, Tempo, and Liputan6, totaling 55 million words, and the Indonesia Web Corpus, comprising 90 million words [17].

Just like BERT, IndoBERT follows a two-stage process for classification: pre-training and fine-tuning. In the pre-training phase, the model learns from unlabeled data, while in the fine-tuning phase, it initializes with pre-training parameters and adapts further using labeled data [18]. An illustration of the pre-training and fine-tuned stages is shown in the figure 1.

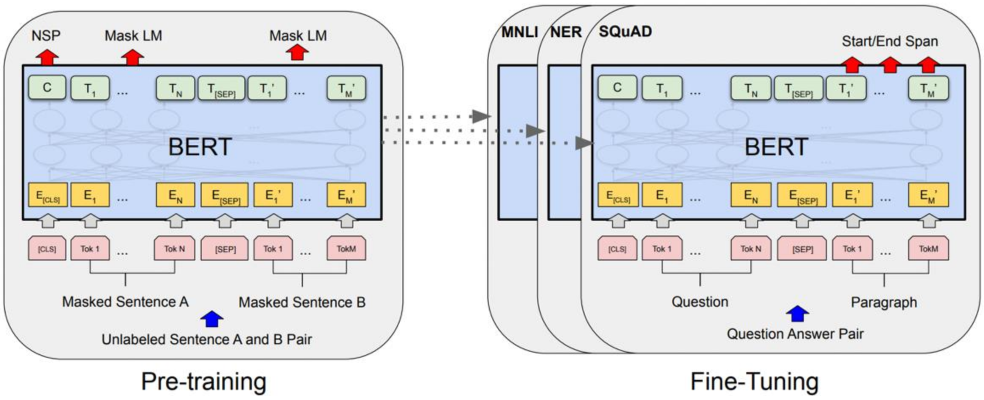

Fig. 1. The illustration of the pre-training and fine-tuning stages [19]

The pre-training and fine-tuning stages in the BERT model have architectural similarities, with differences mainly in the output layers. In fine-tuning, the model starts with pre-trained parameters and fine-tunes them using labeled data from specific tasks. Unique tokens like [CLS] and [SEP] are appended to inputs, aiding BERT in utilizing its pre-trained knowledge for task-specific adjustments during fine-tuning [19].

In the pre-training phase, depicted on the left side of the figure, BERT is trained on two primary tasks: the Masked Language Model (Mask LM) and Next Sentence Prediction (NSP). For the Mask LM task, certain words in a sentence are randomly masked, and the model is trained to predict these masked words. This is represented by the green boxes labeled "Mask LM." The NSP task involves training the model to predict whether a given sentence B follows sentence A in the original text, illustrated by the red box labeled "NSP." During pre-training, BERT processes unlabeled sentence pairs (Sentence A and Sentence B), where Sentence A and B could be either contiguous sentences or not, as part of the NSP task, and masked sentences A and B for the Mask LM task.

On the right side of the figure, the fine-tuning phase is shown. After the pre-training, BERT undergoes further fine-tuning for specific downstream tasks such as Multi-genre Natural Language Inference (MNLI/NLI), Named Entity Recognition (NER), and Stanford Question Answering Dataset (SQuAD). Each of these tasks involves training the model on labeled data pertinent to the task. For instance, MNLI/NLI involves predicting the relationship between pairs of sentences, NER involves identifying and categorizing entities within a text, and SQuAD involves predicting the start and end span of answers to questions based on a given paragraph.

III. RESEARCH METHOD

The system design for implementing the IndoBERT deep learning model in sentiment analysis classification related to the 2024 Indonesian Presidential candidates will be outlined in this section. There are six main processes in designing this system, namely: data preparation, data pre-processing, data labeling, splitting dataset, model adaptation (fine-tuning), and evaluation. Figure 2 depicts the flowchart design of the sentiment analysis system for the 2024 Indonesian Presidential candidates using IndoBERT.

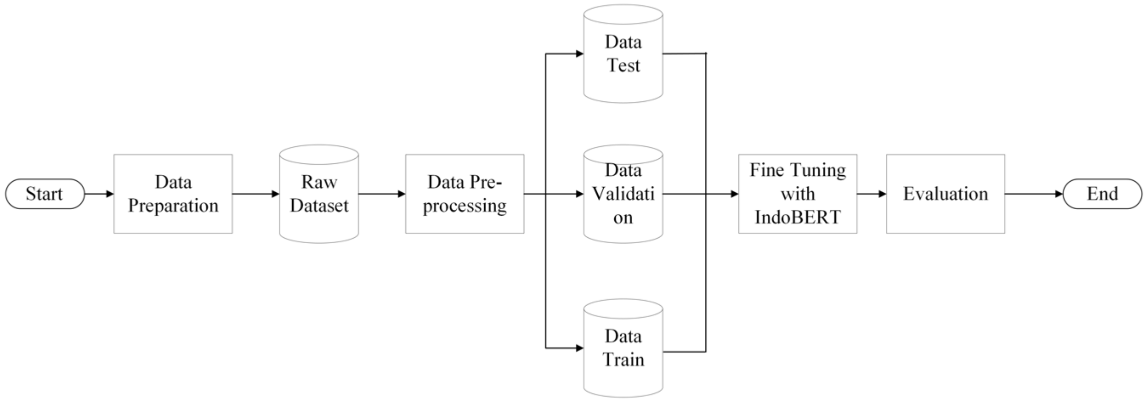

Fig. 2. The system design

A. Data Preparation

Data preparation begins by extracting data from the Twitter platform using crawling techniques with the Tweet Harvest library. Data collection is conducted using the names of the three 2024 Indonesian presidential candidates: "Anies Baswedan," "Prabowo Subianto," and "Ganjar Pranowo". The data was collected before the general election day, specifically from January 7th to February 13th, 2024. This approach ensures that the collected raw data from tweets encompasses a diverse range of user perspectives and responses to the three presidential candidates.

After the crawling process is completed, the data is consolidated and organized into a Comma Separated Values (CSV) format. This process ensures that the dataset reflects the variety of opinions and responses from various sources regarding the three presidential candidates.

B. Data Pre-processing

In the data preprocessing process, several methods will be applied, such as case folding, noise removal, and tokenization.

1) Case folding. Case folding is an important process in documents to maintain text consistency by converting all letters to lowercase. The aim is to create uniformity within the document, ensuring that every letter in the text has the same format.
2) Noise removal. Noise removal refers to the process of eliminating irrelevant or unwanted elements from data. It involves filtering out unnecessary or irrelevant information that may interfere with the analysis or understanding of the data.
3) Slang word removal. Slang word removal in preprocessing is the process of eliminating or replacing slang words with their standard or formal counterparts in text.

C. Data Labeling

During the data labeling phase, the data will be manually labeled by three students. This task will be carried out by three students responsible for classifying each tweet into positive, negative, or neutral categories related to each presidential candidate, namely Prabowo Subianto, Anies Baswedan, and Ganjar Pranowo. The labeling by three students is intended to avoid subjectivity towards any presidential candidate. The result of the labeling by the three students will be a single label for each tweet, based on the majority vote among the three labeling students. If the majority vote is tied, the result will be labeled as "Neutral."

D. Splitting Dataset

Splitting the dataset is an essential procedure in machine learning research, especially in training and assessing predictive models. This process entails dividing the available dataset into distinct subsets, typically including training, validation, and test sets. The training set, comprising the largest portion of the dataset, is utilized to train the model by exposing it to labeled examples. Meanwhile, the validation set is employed to fine-tune the model's hyperparameters and evaluate its performance during training, guarding against overfitting by providing an independent dataset for assessment. Lastly, the test set is utilized to evaluate the model's ability to generalize to unseen data, offering an unbiased estimation of its performance in real-world scenarios. Ensuring proper dataset splitting is crucial to develop a robust and reliable model capable of making accurate predictions on new, unseen data.

E. Fine-tuning with IndoBERT

Fine-tuning with IndoBERT for sentiment analysis downstream tasks involves adapting the pre-trained IndoBERT model to a specific sentiment analysis task. This process typically entails initializing the model with pre-trained parameters and then fine-tuning it using labeled data from the sentiment analysis task. The fine-tuned model learns to extract sentiment-related features from the input text and make predictions based on the specific sentiment labels provided in the training data.

F. Evaluation Model

During the model evaluation phase, the fine-tuned model's effectiveness will be gauged using pertinent metrics like accuracy, precision, recall, and F1-score. These metrics serve to measure how accurately the model categorizes sentiments associated with presidential candidates.

IV. Results And Discussion

In this research, we utilized the IndoBERT model to conduct sentiment analysis classification concerning the 2024 Indonesian presidential candidates on Twitter. The dataset, which has undergone pre-processing and sentiment labeling, serves as the input for this model. The research follows the following stages: a. data preparation, b. data pre-processing, c. data labeling, d. fine-tuning with IndoBERT, and e. model evaluation.

A) Data Preparation

In this stage, we have successfully collected a dataset comprising 3051 rows of tweets containing the keywords "Anies Baswedan," "Prabowo Subianto," and "Ganjar Pranowo". This dataset serves as the primary input for the model.

B) Data Pre-processing

1) Case folding

This process involves converting all the letter characters in the text to either lowercase or uppercase, thereby eliminating capitalization differences. For example, words written with both uppercase and lowercase letters, such as "Data" and "data," will be transformed into a uniform form, in this case, possibly "data" or "DATA." The application of the case folding technique is illustrated in table I.

TABLE I
The Aplication of the Case Folding

No	Before	After
1	@olvaholvah @aniesbaswedan @ganjarpranowo @prabowo Ini yg disebut 'Malam Pembantaian' kpd Sang Penculik bukan debat..!	@aniesbaswedan @ganjarpranowo @prabowo ini yg disebut 'malam pembantaian' kpd sang penculik bukan debat..!

2	@yusuf_dumdum @ganjarpranowo layak duduk di posisi menteri pertahanan RI	@ganjarpranowo layak duduk di posisi menteri pertahanan ri
3	@Gus_Raharjo @ganjarpranowo Lagi meluk apa ya wir	@ganjarpranowo lagi meluk apa ya wir
4	@ganjarpranowo Sukses selalu Pak Ganjar! Perjalanan dari kos-kosan sampai kantor pertama emang luar biasa.	@ganjarpranowo sukses selalu pak ganjar! perjalanan dari kos-kosan sampai kantor pertama emang luar biasa.
5	On top of that semalam @ganjarpranowo juara kontrol emosi. Memang orang yg awalnya kurang diperhitungkan seringkali memberi kejutan.	on top of that semalam @ganjarpranowo juara kontrol emosi. memang orang yg awalnya kurang diperhitungkan seringkali memberi kejutan.

2) Noise removal
In this stage, noise removal is performed, such as special characters, punctuation marks, or other elements that do not contribute significantly to the analysis being eliminated. The application of the noise removal technique is illustrated in table II.

TABLE II
The Aplication of the Noise Removal

No	Before	After
1	@aniesbaswedan @ganjarpranowo @prabowo ini yg disebut 'malam pembantaian' kpd sang penculik bukan debat..!	aniesbaswedan ganjarpranowo prabowo ini yg disebut malam pembantaian kpd sang penculik bukan debat
2	@ganjarpranowo layak duduk di posisi menteri pertahanan ri	ganjarpranowo layak duduk di posisi menteri pertahanan ri
3	@gus_raharjo @ganjarpranowo lagi meluk apa ya wir	ganjarpranowo lagi meluk apa ya wir
4	@ganjarpranowo sukses selalu pak ganjar! perjalanan dari kos-kosan sampai kantor pertama emang luar biasa.	ganjarpranowo sukses selalu pak ganjar perjalanan dari koskosan sampai kantor pertama emang luar biasa
5	on top of that semalam @ganjarpranowo juara kontrol emosi. memang orang yg awalnya kurang diperhitungkan seringkali memberi kejutan.	on top of that semalam ganjarpranowo juara kontrol emosi memang orang yg awalnya kurang diperhitungkan seringkali memberi kejutan

3) Slang word removal
In this stage, slang words or abbreviated words in the text are replaced with their formal counterparts. The application of the slang word removal technique is illustrated in table III.

TABLE III
The Aplication of the Slang Word Removal

No	Before	After
1	aniesbaswedan ganjarpranowo prabowo ini yg disebut malam pembantaian kpd sang penculik bukan debat	aniesbaswedan ganjarpranowo prabowo ini yang disebut malam pembantaian kepada sang penculik bukan debat

2	ganjarpranowo layak duduk di posisi menteri pertahanan ri	ganjarpranowo layak duduk di posisi menteri pertahanan ri
3	ganjarpranowo lagi meluk apa ya wir	ganjarpranowo lagi meluk apa ya wir
4	ganjarpranowo sukses selalu pak ganjar perjalanan dari koskosan sampai kantor pertama emang luar biasa	ganjarpranowo sukses selalu bapak ganjar perjalanan dari koskosan sampai kantor pertama memang luar biasa
5	on top of that semalam ganjarpranowo juara kontrol emosi memang orang yg awalnya kurang diperhitungkan seringkali memberi kejutan	selain itu semalam ganjarpranowo juara kontrol emosi memang orang yang awalnya kurang diperhitungkan seringkali memberi kejutan

C) Splitting Dataset

Upon completion of dataset preparation, the data will undergo a division into three segments. Specifically, 80% will be allocated for training purposes, while 10% will be designated for validation and an additional 10% for testing

D) Fine-tuning with IndoBERT

The proportion of the dataset is shown in figure 3. It consists of 15.3% of the data labeled as negative, followed by positive and neutral labels at 43.6% and 41.2%, respectively.

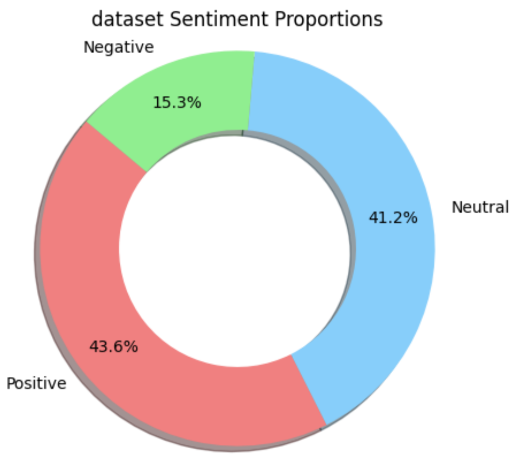

Fig. 3. The proportion of the dataset

Before training, the dataset is divided into 80% training data, with the remaining 10% each for validation and testing data.

In the fine-tuning process, the labels are trained with a learning rate parameter of 0.000005 and for 5 epochs. After testing on the test data, the evaluation results show an overall accuracy rate of 80%, with precision rates of 62% for negative sentiment, 81% for neutral sentiment, and 85% for positive sentiment. Additionally, the recall rates for each sentiment are 64% for negative, 81% for neutral, and 84% for positive, with corresponding F1-scores of 63%, 81%, and 85%, the detailed results can be seen in table V.

TABLE V
EVALUATION RESULT

Sentiment	Precision	Recall	F1-Score
Negative	62%	64%	63%
Neutral	81%	81%	81%
Positive	85%	84%	85%

E) Evaluation Model

Next, the confusion matrix is utilized to examine the testing results. As depicted in figure 4, it can be observed that the model correctly predicted 30 negative data, while 108 data were accurately predicted as neutral, and 106 positive data were predicted correctly.

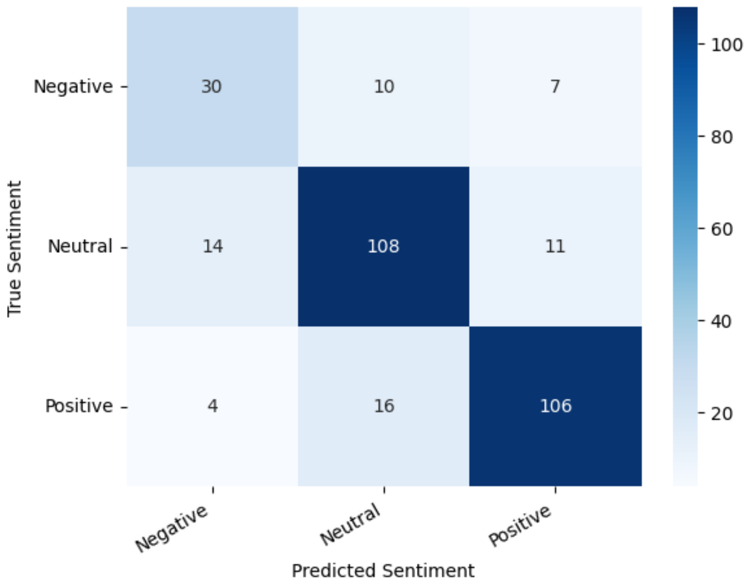

Fig. 4. Confusion matrix data testing result

There are some incorrect predictions, which are likely due to the imbalance in the data. As shown in Figure 3, the proportion of negative labels is much smaller than that of positive and neutral labels. This imbalance may cause the model to lean towards predicting the majority class more often, resulting in less sensitivity to cases that are underrepresented in the data.

The evaluation of the model also revealed some interesting patterns of prediction errors. One example is when a tweet praises one presidential candidate while criticizing another in the same sentence. Such situations pose challenges for the model in determining the overall sentiment of the tweet. For instance, in the sentence 'anies cerdas prabowo tulus ganjar bangsat', there is praise for Anies Baswedan and Prabowo Subianto, but also derogatory remarks about Ganjar Pranowo. The model tends to struggle in classifying the overall sentiment as positive or negative. This highlights the importance of considering context and nuances in sentiment analysis on social media. These findings underscore the need for the development of more sophisticated models that are sensitive to nuances in text to enhance the quality of sentiment prediction in more complex situations.

V. CONCLUSION

In summary, this investigation has showcased the efficacy of employing the IndoBERT model for sentiment analysis classification associated with the 2024 Indonesian presidential candidates on Twitter. Through meticulous data preparation, pre-processing, labeling, and fine-tuning processes, we achieved an overall accuracy rate of 80%. Precision rates for negative sentiment were observed at 62%, with rates of 81% for neutral sentiment and 85% for positive sentiment, complemented by corresponding recall rates of 64%, 81%, and 84%, respectively. Despite these promising results, our evaluation also unveiled intriguing patterns of prediction errors, particularly evident in instances where tweets contained mixed sentiments, such as praising one candidate while criticizing another in the same sentence. This complexity underscores the challenges inherent in sentiment analysis, especially within the dynamic landscape of political discourse on social media platforms. Addressing such challenges presents opportunities for refining sentiment analysis techniques and advancing the sophistication of NLP models. Future research efforts should prioritize addressing data imbalance challenges and exploring alternative methodologies to augment the precision and dependability of sentiment analysis in intricate and multifaceted scenarios.

ACKNOWLEDGMENT

The author would like to express gratitude to Telkom University for its support and opportunity provided for this research. Also, appreciation goes to previous researchers who have inspired the development of the concept and methodology of this study. Thanks also to all parties involved in providing both material and non-material support in this research.

REFERENCES

[1] Endrik, A. Nugroho and A. T. Zy, "Penerapan Algoritma Naive Bayes dan PSO pada Analisis Sentimen Kandidat Calon Presiden 2024," vol. 7, no. 3, August 2023.

[2] L. Geni, E. Yulianti and D. S. Indra, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using Bert Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika,* vol. 9, no. 3, 2023.

[3] S. Imron, E. I. Setiawan and J. Santoso, "Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN," *INSYST: Journal of Intelligent System and Computation* , vol. 5, no. 1, 2023.

[4] C. N. Zempi, A. Kuswanti and S. Maryam, ANALISIS PERAN MEDIA SOSIAL DALAM PEMBENTUKAN PENGETAHUAN POLITIK MASYARAKAT, vol. 6, Jurnal Ilmu Komunikasi, 2023.

[5] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, F. A. Dwiyanto and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *Ilkom Jurnal Ilmiah,* vol. 14, no. 3, pp. 348-354, 2022.

[6] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar and A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist.*, 2020.

[7] K. S. Nugroho, A. Y. Sukmadewa, H. W. DW, F. A. Bachtiar and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *SIET '21: 6th International Conference on Sustainable Information Engineering and Technology,* 2021.

[8] A. L. Hananto, A. P. Nardilasari, A. Fauzi, A. Hananto, B. Priyatna and A. Y. Rahman, "Best Algorithm in Sentiment Analysis of Presidential Election in Indonesia on Twitter Authors," *International Journal of Intelligent Systems and Applications in Engineering,* vol. 11, no. 6s, p. 473–481, 2023.

[9] M. Rizki, M. F. Hidayattullah and D. I. Af'idah, "Klasifikasi Opini Publik di Twitter Terhadap Bakal Calon Presiden Indonesia Tahun 2024 Menggunakan LSTM Secara Realtime Berbasis Website," *INFOTEKMESIN,* vol. 14, no. 2, p. 285–295, 2023.

[10] S. Saadah, K. M. Auditama, A. A. Fattahila, F. I. Amorokhman, A. Aditsania and A. A. Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," *JURNAL RESTI(Rekayasa Sistem dan Teknologi Informasi),* vol. 6, no. 4, pp. 648-655, 2022.

[11] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering,* vol. 13, no. 5, pp. 5641-5652, 2023.

[12] H. Zhu, "Sentiment analysis of 2021 Canadian election tweets," in *International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2022),* Chongqing, 2023.

[13] J. Yadav, "Sentiment Analysis on Social Media," 9 January 2023.

[14] M. Wankhade, A. C. S. Rao and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review,* vol. 55, no. 7, p. 5731–5780, 2022.

[15] D. Rothman, Transformers for Natural Language Processing, Birmingham,Mumbai: Packt Publishing Ltd, 2021.

[16] L. F. Simanjuntak, R. Mahendra and E. Yulianti, "We Know You Are Living in Bali: Location Prediction of Twitter Users Using BERT Language Model," *Big Data and Cognitive Computing* , vol. 6, no. 3, 2022.

[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," in *The 28th International Conference on Computational Linguistics*, 2020.

[18] A. Awalina, F. A. Bachtiar and F. Utaminingrum, "Perbandingan Pretrained Model Transformer pada Deteksi Ulasan Palsu," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 9, no. 3, pp. 597-604, 2022.

[19] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

I. INTRODUCTION

BLOOD pressure estimation using photoplethysmogram (PPG) has become one of the most important aspects of medical innovation today. According to C. El-Hajj [1], the PPG approach allows non-invasive estimation of blood pressure by observing two waveforms obtained from signals, either from two different anatomical areas or through a combination of PPG signals and electrocardiograms. (EKG). Erika Abe [2] explains that the concentration of light reflected from the skin varies according to the changes in hemoglobin due to variations in blood vessel capacity. Near infrared light (600nm - 1000nm) is effective for PPG estimates because it can penetrate the skin and reach the blood circulation system. Sensors that emit near-infrared light, then measure the concentration of reflected light using a phototransistor. In this study, a combination of numerical and experimental methods was used to extract PPG signals from video. Andreia V.

Moço [3] uses the Monte Carlo approach to simulate the PPG amplitude spectrum and perform normalizations to scale pulse waves. Serj Haddad [4] stated that this process allows for accurate PPG signal reception from various parts of the human body. Blood pressure estimates of PPG signals are performed through radial wave analysis (PWA) using an algorithm that extracts morphological features from the PPG signal, mapping these features to systolic blood pressure (SBP) and diastolic (DBP) values with the XGBoost model. Thus, PPG technology provides a continuous method for estimating blood pressure based on signals from different parts of the body. The hope of this study is to produce the most effective blood pressure detection by manipulating the signal length detected using XGBoost, comparing each scenario using MSE and RMSE parameters.Content of Introduction Section.

In introduction, the context of the study and state the precise objective should be explained. An Introduction should contain the following three parts: background, the problem, and proposed solution. In explaining the background, authors have to make clear what the context is. Ideally, authors should give an idea of the state-of-the art of the field the report is about. The problem also need to described clearly so that readers able to understand why they should proceed reading. Authors also need to describe the proposed solution so that readers able to point out what are the novel aspects of authors work. Authors should place the paper in proper context by citing relevant papers.

II. LITERATURE REVIEW

First, Mentioned in the S Sun study [5] on estimating blood pressure using PPG mentioned photoplethysmography (PPG) as one of the techniques for assessing blood pressure. An optical sensor that radiates light into the skin can take light that is reflected or transmitted. PPG monitors variations in local blood volume in tissues at remote locations, such as feet, forehead, and fingers. PPG and arterial blood pressure signals have a comparable morphology as a result of the relationship between blood vessel volume and pressure. Therefore, to extract blood pressure from PPG signals, researchers have suggested a number of characteristics. Introduced by Erika Abe [2] in her study of application through infrared light but the basic theory is to calculate the depth of the light intensity reflected from the skin determined by the amount of hemoglobin change caused by variations in vascular capacity.In the study of Naresh Vempala [6] on the TOI method, the face form is used as a feature in estimating blood pressure but has a shortcoming that is to burden the processing that affects the specifications related to the device used. In the study of Bhanupriya and Neelam [7] on denoising emphasizes that PPG signals obtained from infrared do not have to be used directly but rather pass several steps.

Denoising PPG signal is the process of separating a desired signal component from a noisy or damaged signal, allowing for a simple and accurate diagnosis. This technique should recognize and eliminate noise in the data without endangering the results of unidentified objects.The use of a filter in the Aytrik Bose study [8] as a denoising tool in this study was divided into two methods: high-pass and low-pass as a motion artifact remover. This filter is called a low-pas filter because only the low-frequency component of the image will flow through it. You might think of a highpass filter as a filter that allows a higher frequency portion to pass through. In Xingbin Zhan's [9] study of the GBRT model (gradient boosting regression tree) that has a role in estimates has