Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 9, No. 3, September 2023, pp. 746-757
ISSN: 2338-3070, DOI: 10.26555/jiteki.v9i3.26490

Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models

Lenggo Geni [1,2], Evi Yulianti [1], Dana Indra Sensuse [1]
[1] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
[2] Ministry of Health, Republic of Indonesia

ARTICLE INFO

Article history:

Received June 18, 2023
Revised August 03, 2023
Published August 08, 2023

Keywords:

Sentiment analysis;
Twitter;
Indonesian Election;
Transformers;
IndoBERT

ABSTRACT

General election is one of the crucial moments for a democratic country, e.g., Indonesia. Good election preparation can increase people's participation in the general election. In this study, we conduct a sentiment analysis of Indonesian public opinion on the upcoming 2024 election using Twitter data and IndoBERT model. This study is aimed at helping the government and related institutions to understand public perception. Therefore, they could obtain valuable insights to better prepare for elections, including evaluating the election policies, developing campaign strategies, increasing voter engagement, addressing issues and conflicts, and increasing transparency and public trust. The main contribution of this study is threefold: (i) the application of state-of-the-art transformer-based model IndoBERT for sentiment analysis on political domain; (ii) the empirical evaluation of IndoBERT model against machine learning and lexicon-based models; and (iii) the new dataset creation for sentiment analysis in political domain. Our Twitter data shows that Indonesian public mostly reacts neutrally (83.7%) towards the upcoming 2024 election. Then, the experimental results demonstrate that IndoBERT large-p1 is the best-performing model that achieves an accuracy of 83.5%. It improves our baseline systems by 48.5% and 46.49% for TextBlob, 2.5% and 14.49% for Multinomial Naïve Bayes, and 3.5% and 13.49% for Support Vector Machine in terms of accuracy and F-1 score, respectively.

Corresponding Author:

Lenggo Geni, Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
Email: lenggo.geni11@ui.ac.id

1. INTRODUCTION

The 2024 elections in Indonesia are scheduled to be held simultaneously on February 14, 2024. Candidates for president, vice president, as well as representatives for the DPR (*Dewan Perwakilan Rakyat* or Regional Representatives Council), DPD (*Dewan Perwakilan Daerah* or House of Representatives), and DPRD (*Dewan Perwakilan Rakyat Daerah* or Legislatives Council), in the province and regency levels, will be chosen by the Indonesian people. It is crucial for government to comprehend how general public think of the preparations for Indonesia's upcoming elections in 2024. This comprehension allows the election administrators and some related political parties, such as KPU (*Komisi Pemilihan Umum* or the General Elections Commission), BAWASLU (*Badan Pengawas Pemilihan Umum* or the Election Supervisory Body), and DKPP *(Dewan Kehormatan Penyelenggara Pemilu* or the Election Dispute Resolution Body), to understand voter preferences, detect new issues, increase transparency, and enhance public participation and communication. In general, this information can help the interested political parties to make better plans and strategies for more successful elections in 2024, such as by making better socialization to the public to improve people's sentiments that may results in increasing voter participation.

Sentiment analysis is a Natural Language Processing (NLP) task that analyzes and extracts the sentiments expressed by authors to identify feelings, perspectives, and emotions to classify them as positive, neutral, or negative [1], [2]. Sentiment analysis is beneficial to learn public comments or opinions [3]. More specifically for political research, sentiment analysis is useful for understanding public opinion and attitudes around political figures, policies, parties, or events. For example, government can gain better insight about voting awareness, candidate electability, and other emerging political issues. With regards to the upcoming Indonesia's 2024 elections, sentiment analysis can be performed to discover how public perceives the elections.

Twitter social media platform has become a target for digging up big data about sentiments of the general elections. In Indonesia, it becomes more potential to exploit Twitter data because this country has a lot of Twitter users. According to the Search Logistics data [4], the number of Twitter users from Indonesia reached 18.45 million. Indonesia ranks 5th in the highest number of Twitter users in the world, following the United States, Japan, India, and Brazil. Twitter has become a powerful channel for expressing people's opinions. Therefore, it is potential to exploit Twitter data to explore people's sentiments and opinions about various topics, including politic.

BERT (Bidirectional Encoder Representations from Transformers) [5] is a language model proposed in 2018 by Devlin *et al.* that is built upon transformer models. BERT is a contextual language model that performs superior in many NLP tasks. It employes bi-directional learning, which learns the context from left to right and from right-to-left. Therefore, it can better capture the context of text, which enables it to generate more accurate text representation. BERT has been used in some current researches on various tasks [6], [7], including sentiment analysis [5]-[11]. IndoBERT is a pre-trained BERT for the Indonesian language that uses a large size of Indonesian corpus consisting of four billion words in the pre-training process [12], but another research also pre-trained IndoBERT using a smaller corpus [13].

This study investigates Indonesian public opinion on the upcoming 2024 election using Twitter data and IndoBERT model. This study is aimed at helping the government and related institutions to understand public perception. Therefore they could obtain valuable insights to better prepare for elections, including evaluating policy, developing campaign strategies, increasing voter engagement, addressing issues and conflicts, and increasing transparency and public trust**.** Because IndoBERT can capture the semantic of text better (as the result of bi-directional learning), then we hypothesize that the IndoBERT model may better recognize the difference between positive, negative, or neutral views on the Indonesian tweets data compared with the traditional sentiment analysis models, such as machine learning and lexicon-based models.

Some previous studies in other countries have conducted sentiment analysis on election-related issues through social media using various methods. Rita *et al.* [14] analyzed tweets during an electoral period of the general elections of the United Kingdom in 2019 using machine learning (Naïve Bayes, Support Vector Machine, Decision Tree) and deep learning. Chaudhry *et al.* [15] analyzed public sentiment in Twitter before, during, and after the 2020 United States Presidential election towards the presidential candidates (Biden and Trump) using Naïve Bayes machine learning algorithm and TF-IDF features. Yavari *et al.* [16] also analyzed public sentiment in the 2020 United States Presidential election, but they utilized aging estimation method using the proportion of positive messages rate to negative messages rate. Macrohon *et al.* [17] performed sentiment analysis on the 2022 Philippine presidential election using Multinomial Naïve Bayes. Schmidt *et al.* [18] analyzed the sentiments of political parties in Germany during the 2021 election year using a transformer-based model, BERT. Our work is different to the works in [14]-[17] in which we analyze public sentiments election-related issues using Indonesian twitter data and BERT-based model. Similar to Schmidt *et al.* [18], we also used transformer-based model BERT, but our model is pretrained on Indonesian large corpus, i.e., IndoBERT, while Schmidt *et al.* used German BERT model.

Some previous studies have also analyzed sentiment related to election in Indonesia. Budiharto *et al.* [19], Habibi *et al.* [20], and Buntoro *et al.* [21] investigated the general elections of Indonesia in 2019. Budiharto *et al.* [19] simply used TextBlob; Habibi *et al.* [20] used Naïve Bayes; and Buntoro *et al.* [21] utilized both Naïve Bayes and SVM classifiers for predicting the sentiments in Tweet data. In general, they still used common lexicon-based and machine learning-based methods. In contrast to them, in this work we examine public sentiments before the Indonesia's elections in 2024 using state-of-the-art transformed-based model IndoBERT. A recent work by Mandhasiya *et al.* [22] has predicted the voices of Indonesian public in the 2024 Indonesia's presidential election using IndoBERT model. However, they used YouTube conversation data instead of Twitter data. This distinguish between our work and their work. To the best of our knowledge, none of the previous work on sentiment analysis has investigated the sentiment analysis on the 2024 Indonesian presidential election using Twitter data and IndoBERT model.

The contribution of this study is as follows: (i) We performed sentiment analysis on public perception about the upcoming Indonesia's 2024 election using a state-of-the-art transformer-based model that was pretrained on Indonesian corpus, i.e., IndoBERT; (ii) We performed empirical evaluation on the effectiveness of the transformer-based model against machine learning and lexicon-based models; (iii) We built a dataset of sentiment analysis on the Indonesia's 2024 election consisting of approximately 1K tweets by performing manual human annotation. This research can enrich knowledge about the application of the state-of-the-art IndoBERT language model in sentiment analysis on political domain.

This paper is organized as follows. Section 2 explains our research methodology, including the dataset collection and processing and our sentiment analysis method. Furthermore, this section also presents our experiment. Section 3 presents the results and discussion. Section 4 concludes this study and highlights some potential avenues for future work.

2. METHODS

The research flowchart, which includes four IndoBERT models and three baseline models is shown in Fig. 1. This study begins with the data collection process on Twitter. Then, we implement the IndoBERT-based models by fine-tuning the pre-trained IndoBERT [12]. The effectiveness of the resulting models are then compared againsts machine learning models, i.e., Multinomial Naïve Bayes, and SVM (Support Vector Machine) methods, and a lexicon-based model (TextBlob). We examine the performance of our models in analyzing public sentiments towards the Indonesia's 2024 election preparations. The details of each step in our research methodology are explained in the following subsections.

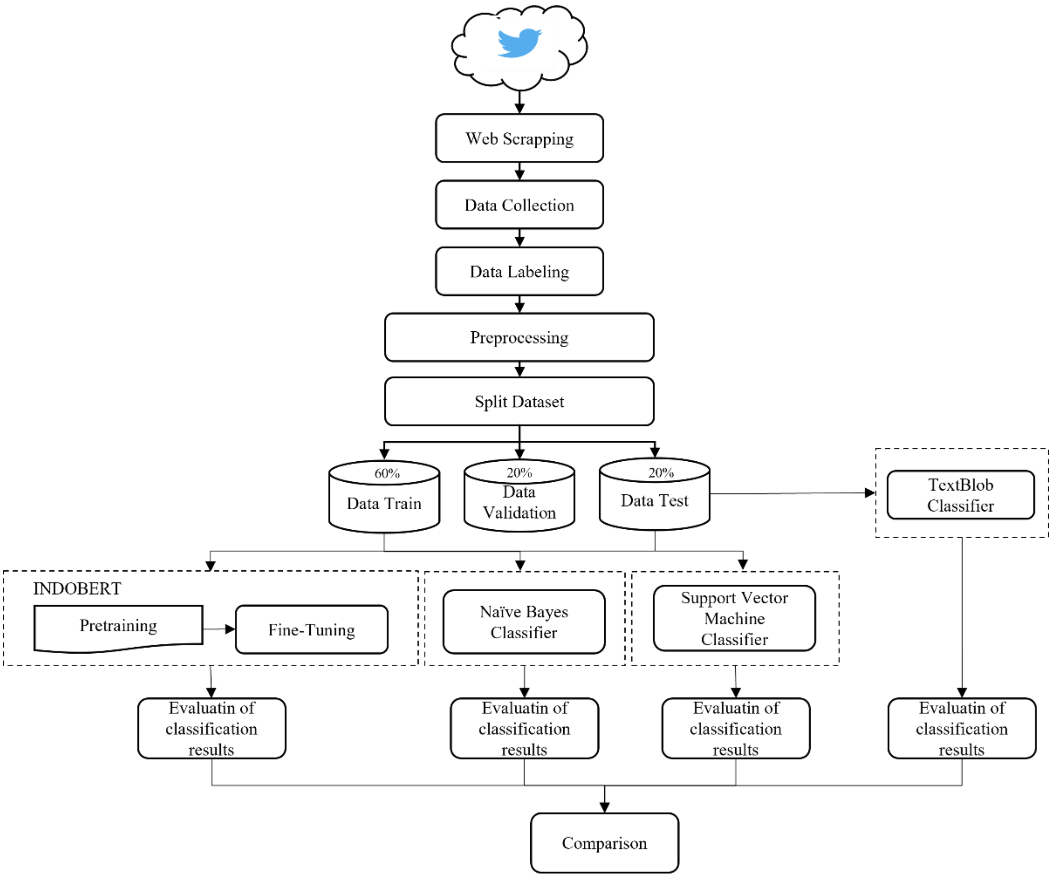

Fig. 1. Research Flowchart

2.1. Data Collection

We collected 1,000 tweets from Twitter containing keywords related to the 2024 elections in Indonesia, which will be held in February 2024. The data crawling process used Tweepy, a Python-based tool to retrieve data from Twitter [23]. The query "*pemilu*" ("general election") was used to retrieve the related tweets. We retrieved tweets posted using the Indonesian language since the beginning of January until the end of March 2023, and the retweets were excluded from the results.

2.2. Data Labeling

In this process, we perform human annotation to label the tweets with the sentiment labels. We asked two annotators, who are college students, to give label to the tweets. The 1000 tweets were split for two annotators, and there was an overlap of 100 tweets to be labeled by both annotators. These overlapping tweets are needed to compute the agreement between two annotators. To sum up, each annotator was asked to annotate 600 tweets, with 100 of them were annotated by both annotators.

Data labeling is done based on our annotation guidelines to ensure the quality of the annotation results. We followed guidelines by Mozetič *et al.* [24] for instructing the annotators to label each tweet as positive, neutral, or negative. The annotation guideline is important to guide the annotation process so that the annotators can have similar perspectives on the charactersitics of postitive, negative, and neutral sentiments in our Twitter data. Therefore, it can reduce the subjectivity issue in the annotation process because the annotation is performed according to the given guidelines.

Based on the agreement evaluation on the overlapped annotation on 100 tweets, we obtain the Kappa score of 0.8668, which indicates a strong agreement between annotators [25]. It confirms that the annotation results between the two annotators are mostly consistent and show they have similar perspectives about the definitions of positive, negative, and neutral. It may also indicate that they have complied with the annotation criteria specified in the guidelines. The example of data labeling is shown in Table 1.

Table 1. Example of Data Labelling Results

Tweets	Annotation
MARI WUJUDKAN PEMILU 2024 YANG BERSIH DAN ADIL (LET'S MAKE TRUTHFUL AND FAIR 2024 ELECTIONS)	positive
Parpol-parpol berlomba menguatkan narasi politik untuk menguatkan elektabilitas menjelang pemilu. Parpol mana yang lebih unggul memainkan narasi politiknya saat ini? (Political parties are competing to strengthen political narratives to strengthen electability ahead of elections. Which political party is superior in playing its current political narrative?)	neutral
PERCAYA atau Tidak, Polemik Pemilu Proporsional Tertutup Hanyalah Pesanan (BELIEVE it or Not, Closed Proportional Election Polemics Are Just Orders)	negative

Based on Table 1, the tweet *"MARI WUJUDKAN PEMILU 2024 YANG BERSIH DAN ADIL"* (LET'S MAKE TRUTHFUL AND FAIR 2024 ELECTIONS) is labelled as positive sentiment. This sentence contains positive support, solicitation, or appeals for the 2024 Election activities. This sentence focuses on a positive hope to have fair elections. Next, in the tweet *"Parpol-parpol berlomba menguatkan narasi politik untuk menguatkan elektabilitas menjelang pemilu. Parpol mana yang lebih unggul memainkan narasi politiknya saat ini?"* (Political parties are competing to strengthen political narratives to strengthen electability ahead of elections. Which political party is superior in playing its current political narrative?) is given a neutral annotation. This sentence focuses on certain objects, people or parties. There are no positive or negative calls for implementing the 2024 elections. Furthermore, the tweet *"PERCAYA atau Tidak, Polemik Pemilu Proportional Tertutup Hanyalah Pesanan"* (BELIEVE it or Not, Closed Proportional Election Polemics Are Just Orders) is given a negative annotation. This sentence indicates having negative sentiments or criticism of certain situations, such as manipulation or orders for closed proportional election polemics. In addition, the tweet also implicitly contains negative connotations related to the 2024 election.

2.3. Data Preprocessing

Data preprocessing is carried out before tweets are analyzed. Data preprocessing combines several procedures to clean and turn raw data into an understandable format [26]. The preprocessing in this study includes case folding, data cleaning, and tokenization. Case folding is generally used in text processing and natural language processing to refer to the standard process of converting letters in text into a uniform form [27], such as changing letters to lowercase or uppercase. This study uses case folding to lowercase all letters in the dataset to make generalizations, so that similar words written in uppercase or lowercase letters will be considered the same.

Data cleaning is an important step to reduce the noisy tweets in the dataset. Our data cleaning process includes removing duplicate data, links, usernames (@usernames), hashtags (#), numbers, symbols, extra spaces, and punctuation [28], [29]. This study uses regular expressions to perform data cleaning. Tokenization is a technique used to separate each tweet into words, punctuation, and other meaningful expressions according to the rules of the language being utilized [29]. In our case, tokens are in the word levels. The dataset was divided into three splits (training, validation, and testing data) using 60%: 20%: 20% proportion as in [30]. The statistics of our dataset is shown in Table 2.

Based on Table 2, most of the sentiments towards the Indonesia's 2024 election in our Twitter dataset are neutral (83.7%). Positive tweets contain likes, support, and invitations for the upcoming elections for individuals or political parties. In contrast, negative tweets contain dislike or unsupportive comments which may discredit individuals or political party candidates. Additionally, many tweets contain critics to the current government, which are not explicitly related to Indonesia's upcoming 2024 elections. For this kind of tweets, they belong to neutral, following the annotation guideline. Our dataset of Indonesian general elections 2024 (ige2024) has been made available for research purposes at https://github.com/ir-nlp-csui/ige2024.

Table 2. The Statistics of Our Dataset

Sentiment	Data Training	Data Validation	Data Testing	Total
Positive	39	21	21	81
Negative	46	16	20	82
Neutral	515	163	159	837
Total	600	200	200	1000

2.4. IndoBERT model

The aim of using IndoBERT as our sentiment analysis method is to produce accurate and linguistically rich representations of Indonesian text. IndoBERT is a BERT-based model that was pretrained using a large Indonesian corpus (Indo4B corpus) [12]. Therefore, the architecture of IndoBERT is the same as BERT, which consist of a stack of Transformer encoders. IndoBERT has two variants of architectures: IndoBERT-base and IndoBERT-large. Both vary in the number of layers of transformer, attention heads, and parameters. IndoBERT-base has 12 layers of transformers, 12 attention heads, and 110 million parameters. Whereas IndoBERT-large has 24 layers of transformers, 16 attention heads, and 340 million parameters [5]. Compared to the IndoBERT-base, IndoBERT-large commonly provides higher precision [5], since it can better capture the semantics of text as a results of utilizing a higher number of neural network parameters. However, training data with IndoBERT-large usually takes longer than IndoBERT-base [5]. The basic architectures of both IndoBERT types are shown in Fig. 2.

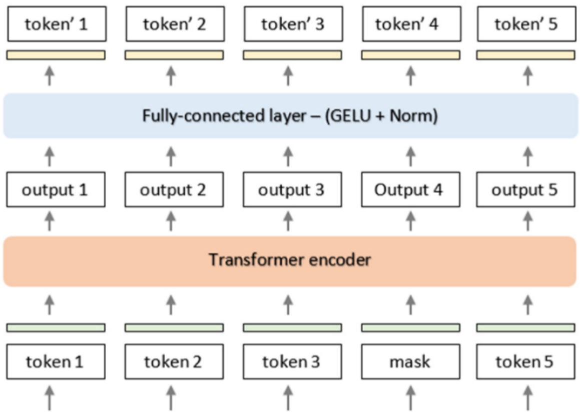

Fig. 2. IndoBERT Architecture [31]

In general, a BERT model (including IndoBERT) is trained using MLM (Masked Language Model) and NSP (Next Sentence Prediction) tasks to accurately capture the text representation in bidirectional way [13]. After the IndoBERT model is pretrained, it can be fine-tuned using our specific dataset by adding an output layer that is specific to our task. In the fine-tuning process, the weights that has been learned during pretraining will be updated based on the input and output in our specific dataset. Therefore, the resulting model is more accurate to perform our specific task. Similar to this basic concept of the use of BERT-based model, we perform fine-tuning of the IndoBERT models using our Twitter dataset to update the weight values learned during pre-training process. We experimented with some versions of IndoBERT, including IndoBERT base-p1 (indobenchmark/indobert-base-p1), IndoBERT base-p2 (indobenchmark/indobert-base-p2), IndoBERT large-p1 (indobenchmark/indobert-large-p1), and IndoBERT large-p2 (indobenchmark/indobert-large-p2). The phase 1 model uses training data with a maximum sequence length of 128, while the phase 2 model uses a maximum sequence length of 384. In addition, they also differ in the hyperparameters applied during the pretraining process: number of epoch and batch sizes.

3. EXPERIMENT

3.1. Experiment Setting

Hyperparameters are parameters that need to be established prior to forming of a neural network, such as batch size, learning rate, and dropout rate [29]. The hyperparameter configuration will significantly affect how well a deep learning model performs [32]. Hyperparameters must be set before the pretraining and finetuning processes. To perform fine-tuning the IndoBERT models, we uses 8 batch sizes and 3 epochs. Batch size is the total sample data used for a backward and forward pass, whereas epoch is the amount of times training is done using training data. It means that the model will process 8 data samples in one iteration and will see the entire training dataset 3 times during the training process.

3.2. Baseline Methods

We employ two machine learning models and one lexicon-based model as our baseline systems. The purpose of the baselines is to evaluate the effectiveness of IndoBERT models to solve our task.

- **TextBlob Classifier**

TextBlob is a well-known lexicon-based model for sentiment analysis [33]. The data scientists prefer using TextBlob library since it is faster and more compact [34]. TextBlob offers some functionalities, such as finding polarity and subjectivity of text [35]. Textblob has a database that allows the text to be compared with it to obtain the polarity and subjectivity scores. The value of subjectivity indicates the level of facts or opinions in the data. The subjectivity value ranges between 0 and 1, with the higher value indicates that the text contains more opinion, while the lower value indicates that the text contains more fact [36]. A few previous work have utilized TextBlob to perform sentiment analysis [19].

- **Naïve Bayes Classifier**

The Naïve Bayes classifier classifies data as having a positive or negative sentiment using the concept of probability at the token level. The entire dataset is split into sentences, followed by individual tokens. Then the polarity of the tokens determines the overall attitude of the sentence and categorizes it as a positive, negative, or neutral statement [37]. Multinomial Naïve Bayes expands the use of the Naïve Bayes algorithm. It is a frequency-based model presented for text classification in which word counts represent data and implements Naïve Bayes for data distributed multinomially. Multinomial Naïve Bayes uses probabilistic measures and presupposes the independence of the variables. It is characterized by the fact that the appearance of one feature does not alter the likelihood of the occurrence of the other feature that falls under that category [38]. Naïve Bayes has been shown to be effective in some previous works on sentiment analysis [20], [39], [40]. We employ CountVectorizer to extract TF (Term Frequency) features for each token in the tweets.

- **Support Vector Machine Classifier**

The Support Vector Machine (SVM) classifier is a novel machine learning technique from the most effective implementation of statistical learning theory [37]. SVM has been shown to be effective in some previous works on sentiment analysis [21], [41], [42], [43]. We employ CountVectorizer to extract TF (Term Frequency) features for each token in the tweets.

3.3. Evaluation

After obtaining the classification results, we must evaluate the effectiveness of the models. Two evaluation metrics employed in this study are accuracy and F-1 measures [44]. In order to compute these metrics, we first need to compute True Positive (TP), False Positive FP), True Negative (TN), and False Negative (FN) values [11]. TP means the number of data that is correctly classified as positive, FP is the number of data that is incorrectly classified as positive, TN means the number of correct data that is classified as negative, and FN implies the number of data that is incorrectly classified as negative.

Accuracy is an evaluation metric that measures the extent to which a classification model can correctly predict all cases in a dataset. The ratio of accurate predictions to all other predictions is known as accuracy. Accuracy shows how well the model can make correct predictions overall. In (1) describes the formula to compute accuracy scores [44].

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

Meanwhile, F1 score is an evaluation metric that combines precision and recall, providing a more holistic picture of the performance of a detection or classification model. This metric is useful when the balance between Precision and Recall is critical. Precision and Recall are two evaluation metrics used in measuring the performance of classification models. These two metrics are often used to evaluate machine learning models, especially when dealing with class imbalance problems [45]. Precision measures the degree to which the model provides relevant or true positive results. Precision is the ratio between the number of true positives (predictions that are true positives) and the total number of positive predictions, including true positives and false positives. Recall measures the extent to which the model can find or retrieve all existing positive cases. Recall is calculated as the ratio between true positives and the total number of positive cases in the dataset.

Equations (2), (3), and (4) respectively describes the formulas to compute Precision, Recall, and F-1 measures [44].

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

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

$$\text{Fscore} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4. RESULTS AND DISCUSSION

4.1. Performance of the classification models

Our study compares four IndoBERT models, including IndoBERT base-p1, IndoBERT base-p2, IndoBERT large-p1, and IndoBERT large-p2, to lexicon-based model and machine learning algorithms (SVM and Multinomial Naïve Bayes). Based on some of previous studies, these lexicon-based and machine-learning baselines have been shown to achieve good accuracy [19], [20], [39], [40], [21], [41], [42], [43]. Table 3 shows the result of models performance.

Table 3. The result of models performance

Model		Accuracy	F1-Score
TextBlob [19]		0.3500	0.4200
Multinomial Naïve Bayes [20], [39], [40]		0.8100	0.7400
SVM [21], [41], [42], [43]		0.8000	0.7500
IndoBERT	base-p1	0.8150	0.8843
	base-p2	0.8200	0.8642
	large-p1	0.8350	0.8849
	large-p2	0.8200	0.8796

As shown in Table 3, our proposed method improves the performance of all baseline methods that have been shown to perform well in previous studies. IndoBERT large-p1 model has the highest accuracy (83.5%) in predicting public sentiments on the upcoming 2024 election in Indonesia. It outperforms the machine learning baselines by up to 14.49%, and the lexicon-based model by up to 46.49%, in terms of F-1 score. This result shows that IndoBERT large-p1 model is more effective than machine learning and lexicon-based models. The 2nd top performing models are IndoBERT large-p2 and IndoBERT base-p2 models that gains F-1 score of 82%. The effectiveness of IndoBERT is caused by the more accurate text representation generated by IndoBERT, which makes it better to distinguish the characteristics of text with positive, negative, and neutral sentiments. Further analysis is presented later in section 4.2 (qualitative analysis).

Fig. 3 shows the the confusion matrix of the best performing model in our experiment, i.e., IndoBERT large-p1 model. The confusion matrix provides information about the extent to which the model can classify data correctly and the number of misclassifications resulted by the model. It appears from the figure that IndoBERT large-p1 model can correctly predict the sentiments for 159 tweets out of 200 tweets in our dataset. The misclassification cases are incorrectly predicting the neutral tweets as negative (18 cases) or positive (15 cases).

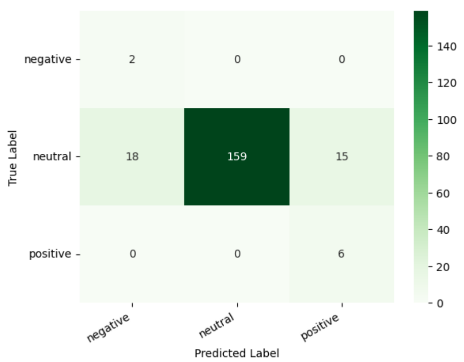

Fig. 3. Confusion matrix of IndoBERT large-p1 model

The misclassification cases may be caused by the imbalance in dataset. As we can see from Table 2, more than 80% tweets in our dataset have neutral sentiments. This makes the model still needs more knowledge to be able to better understand the characteristics of positive and negative tweets.

4.2. Qualitative Analysis

We analyze a few results of sentiment labels outputted by all models employed in our experiment. Table 4 shows an example of tweet that takes advantage of the use of IndoBERT-large as the classification model. The ground truh label for this tweet is negative, because the text in this tweet expresses a movement to torment the people to make it easier to buy people's voices with money in the elections. This statement reflects unethical and manipulative actions in the context of elections. Therefore, the sentiment in the text leans towards negativity. All baseline models could not predict correctly the sentiments of this tweet by predicting neutral sentiments as the results. It is because these models do not have enough context to recognize the negative sentiments contained in the text. In contrast to them, IndoBERT-large can outerform all the baseline models by correctly predicting the sentiment of this tweet as negative. We can see that there is a strength of the IndoBERT large-p1 model in predicting sentiment compared to all other baselines in this study, such as TextBlob, Naïve Bayes, and SVM.

Table 4. The example of Tweet demonstrating the advantages of IndoBERT classification predicition

Tweet data (1)	actual	base-p1	large-p1	TextBlob	Naïve Bayes	SVM
@HelmiFelis_ gerakan menyengsarakan rakyat agar jelang pemilu lebih mudah dibeli suaranya dgn duit gitu loh (the movement to torment the people so that ahead of the election, it is easier to buy their voices with money)	neg	neut	Neg	neut	neut	neut

For further semantic analysis, the word clouds of frequent words in the neutral tweets are generated after stopwords removal. Fig. 4 illustrates the results. Recall that 83.7% tweets in our dataset belongs to neutral tweets, therefore, from the figure we can obtain a better insight on what Indonesian public mostly discusses about the upcoming elections on social media Twitter. The 10 most frequent words appearing in the neutral tweets are: *partai* (party), *capres* (president candidate), *jelang* (ahead of the election), *politik* (politic), *presiden* (presidents), *prabowo, subianto* (Prabowo Subianto is the name of a political figure who may be running for president in the upcoming election), *dekade* (decade), *sistem* (system), and *agusyudhoyono* (Agus Yudhoyono is the name of a political figure who may be running for president in the upcoming election). We can see that most of these words are common or not dominant words. In other words, they are neutral words about elections. This is actually consistent with the sentiments of tweets from which the word clouds are generated, i.e., neutral sentiments. However, the word clouds also displayed a few specific or dominant words denoting some political figures (such as Prabowo Subianto, Agus Yudhoyono, Anies, and Jokowi) and some political parties (such as PKS and PDIP, PAN, and Gerindra). This implies that these political figures as well as political parties are those that attract public attention for the Indonesia's 2024 elections, therefore they are mostly discussed in our Twitter data. Note that Prabowo Subianto, Agus Yudhoyono, and Anies Baswedan are political figures who are known to be running for president or vice president in the upcoming elections.

Fig. 4. Word Cloud for Neutral Sentiment

We also looked at word clouds for the positive and negative sentiment classes. For positive tweets, there are only a few dominant words, for example, words "ahy" or "agus yudhoyono" which reflects that this political figure is perceived positively by public for the upcoming 2024 election. Other words are common words, such as *"pemilu jujur"* (honest election), *"partai"* (party), *"capres"* (presidential candidate), *"jelang"* (ahead of the election), and *"politik"* (politics). Then for negative tweets, the most frequent words are *"rezim"* (regime), *"rakyat"* (citizen), *"muncul"* (appear), *"politik"* (politics), and *"kecurangan"* (fraud), reflecting that many people's tweets against the regime and anticipate deception ahead of the upcoming 2024 elections.

4.3. Research Limitation

There are some limitations in this study. One of them is related to data imbalance. We found that 83.7% tweets in our dataset contain neutral sentiments, while only 8.1% tweets with positive sentiment, and 8.2% tweets with negative sentiment. This data limitation can affect the accuracy of the classification model, so some efforts are needed to balance the data. To overcome this problem, we conducted an initial experiment using the SMOTE (Synthetic Minority Over-sampling Technique) [46] to balance the data. SMOTE is an oversampling method to overcome the problem of class compatibility in datasets, especially whenthere are more samples in the majority class and few samples in the minority class. Table 5 displays the results of our initial experiments using the SMOTE technique.

Table 5. The result of models performance

Model		Accuracy	F1-Score
TextBlob [19]		0.3500	0.4200
Multinomial Naïve Bayes [20], [39], [40]		0.3900	0.4400
SVM [21], [41], [42], [43]		0.8000	0.7000
IndoBERT	base-p1	0.5400	0.5019
	base-p2	0.4100	0.3433
	large-p1	0.7950	0.8858
	large-p2	0.7800	0.8633

Based on Table 5, the results of balancing the data using the SMOTE oversampling method are still unsatisfactory. It could not improve the results using original dataset, but it actually reduces the model's performance. It can be seen that there is a decrease in accuracy and F1-score in almost every model used in this study when we compare the results presented in Table 5 with the those presented in Table 4 earlier.

5. CONCLUSION AND FUTURE WORK

In this study, we perform sentiment analysis on Indonesian public opinion for the upcoming election in 2024 using Twitter data. We propose to use IndoBERT, which is a BERT model pre-trained using a huge amount of Indonesian corpus, to predict the sentiment of tweets. Four variant of IndoBERT models were developed: IndoBERT-large-p1, IndoBERT-large-p2, IndoBERT-base-p1, and IndoBERT-base-p2. Our results show that IndoBERT, with a large-p1 model, is superior in detecting the sentiments before the Indonesia's 2024 elections compared to other machine learning and lexicon-based models tested in this work. It achieves an accuracy of 83.5%, and outperforms the machine learning models by up to 14.49%, and the lexicon-based model by up to 46.49%, in terms of F-1 score. Furthermore, most of the public sentiment towards the issue of preparing for the 2024 election is found to be neutral (83.7%).

Overall, this study contributes to the knowledge on the application of state-of-the-art transformer-based model, IndoBERT, in sentiment analysis on political domain. Here, the effectiveness of IndoBERT model was compared against lexicon-based (TextBlob) and machine learning-based models (Naïve Bayes and SVM). In addition, a new dataset for sentiment analysis consisting of around 1K tweets has also been created in this work that can be utilized by other researches for future studies. The Indonesian general elections 2024 dataset built in this study can be downloaded at https://github.com/ir-nlp-csui/ige2024.

One of the limitations in this study is data imbalance. The number of tweets in each class in our dataset is imbalance, as it is very dominated with neutral tweets. So, an effective technique to balance the dataset is needed to improve the accuracy of the prediction models. To overcome this issue, we have performed an initial experiment using an oversampling method SMOTE. The results, however, is still unsatisfactory as it rather decreases the accuracy of the models. Therefore, investigating effective methods for data balancing or data augmentation can be explored for future work. In addition, future research can be conducted to use more datasets with extended data collection periods or to explore other transformer-based models, such as Roberta, XLM-R, T5, etc, to increase the accuracy of the sentiment prediction.

We also have some recommendations for government and some interested parties to increase the success of the upcoming Indonesian elections based on the results of this study. The government's official general election accounts on social media needs to combat misinformation and to ensure the dissemination of accurate information regarding preparations for the upcoming election. Further, they need to make better socialization on the upcoming elections by providing informative content about the election process, election procedures, and the importance of active participation. They can also collaborate with fact-checking agencies to verify and ensure the accuracy of information before sharing it with the public. In addition, a quick response is needed from the accounts of the general election organizers in Indonesia, such as the KPU, BAWASLU, and DKPP, for general questions about the upcoming elections shortly because these accounts will become the public's trusted media for obtaining accurate information, as well as to overcome the spread of negative content related to the elections, which causes hoaxes or fake news.

Acknowledgment

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Q2 2022 (Grant No. NKB-571/UN2.RST/HKP.05.00/2022).

REFERENCES

[1] M. P. Abraham, "Feature Based Sentiment Analysis of Mobile Product Reviews using Machine Learning Techniques," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2289–2296, 2020, https://doi.org/10.30534/ijatcse/2020/210922020.

[2] S. Sonare and M. Kamble, "Sentiment Analysis in polished Product Based Inspections data using existing supervised machine learning approach," in *IEEE International Conference on Technology, Research, and Innovation for Betterment of Society, TRIBES,* 2021. https://doi.org/10.1109/TRIBES52498.2021.9751663.

[3] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," *IEEE Access*, vol. 8, pp. 14630–14641, 2020, ttps://doi.org/10.1109/ACCESS.2019.2963702.

[4] M. Woodward, "Twitter User Statistics: How Many People Use Twitter In 2023?," *Search Logistics*, 2023. https://www.searchlogistics.com/learn/statistics/twitter-user-statistics/ (accessed May 30, 2023).

[5] A. Patel, P. Oza, and S. Agrawal, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," *Procedia Comput Sci*, vol. 218, pp. 2459–2467, 2023, https://doi.org/10.1016/j.procs.2023.01.221.

[6] 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, https://doi.org/10.3390/bdcc6030077.

[7] A. Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax News Detection Analysis using IndoBERT Deep Learning Methodology," in *10th International Conference on Information and Communication Technology, ICoICT*, pp. 368–373, 2022, https://doi.org/10.1109/ICoICT55009.2022.9914902.

[8] C. Lohith, H. Chandramouli, U. Balasingam, and S. Arun Kumar, "Aspect Oriented Sentiment Analysis on Customer Reviews on Restaurant Using the LDA and BERT Method," *SN Comput Sci*, vol. 4, no. 4, p. 399, 2023, https://doi.org/10.1007/s42979-022-01634-8.

[9] N. Khairu 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 (IJECE)*, vol. 13, no. 5, pp. 5641–5652, 2023, https://doi.org/10.11591/ijece.v13i5.pp5641-5652.

[10] O. A. C. Gunawan *et al.*, "Sentiment Analysis towards Face-to-Face School Activities during the COVID-19 pandemic in Indonesia," in *4th International Conference on Cybernetics and Intelligent System, ICORIS*, pp. 1-7, 2022. https://doi.org/10.1109/ICORIS56080.2022.10031482.

[11] Fransiscus and A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method," *International Journal of Engineering Trends and Technology*, vol. 70, no. 12, pp. 281–288, 2022, https://doi.org/10.14445/22315381/IJETT-V70I12P226.

[12] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10thInternational Joint Conference on Natural Language Processing*, pp. 843–857, 2020, https://doi.org/10.48550/arXiv.2009.05387.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 757-770, 2020, https://doi.org/10.18653/v1/2020.coling-main.66.

[14] P. Rita, N. António, and A. P. Afonso, "Social media discourse and voting decisions influence: sentiment analysis in tweets during an electoral period," *Soc Netw Anal Min*, vol. 13, no. 1, 2023, https://doi.org/10.1007/s13278-023-01048-1.

[15] H. N. Chaudhry *et al.*, "Sentiment analysis of before and after elections: Twitter data of U.S. election 2020," *Electronics (Switzerland)*, vol. 10, no. 17, 2021, https://doi.org/10.3390/electronics10172082.

[16] A. Yavari, H. Hassanpour, B. R. Cami, and M. Mahdavi, "Election Prediction Based on Sentiment Analysis using Twitter Data," *International Journal of Engineering, Transactions B: Applications*, vol. 35, no. 2, pp. 372–379, 2022, https://doi.org/10.5829/IJE.2022.35.02B.13.
[17] J. J. E. Macrohon, C. N. Villavicencio, X. A. Inbaraj, and J. H. Jeng, "A Semi-Supervised Approach to Sentiment Analysis of Tweets during the 2022 Philippine Presidential Election," *Information (Switzerland)*, vol. 13, no. 10, 2022, https://doi.org/10.3390/info13100484.
[18] T. Schmidt, J. Fehle, M. Weissenbacher, J. Richter, P. Gottschalk, and C. Wolff, "Sentiment Analysis on Twitter for the Major German Parties during the 2021 German Federal Election," *Proceedings of the 18th Confereence on Natural Language Processing (KONVENS 2022),* pp. 74-87, 2022. https://aclanthology.org/2022.konvens-1.9.
[19] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J Big Data*, vol. 5, no. 1, 2018, https://doi.org/10.1186/s40537-018-0164-1.
[20] M. N. Habibi and Sunjana, "Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis," *International Journal of Modern Education and Computer Science*, vol. 11, no. 11, pp. 22–30, 2019, https://doi.org/10.5815/ijmecs.2019.11.04.
[21] G. A. Buntoro, R. Arifin, G. N. Syaifuddiin, A. Selamat, O. Krejcar, and H. Fujita, "Implementation of a Machine Learning Algorithm for Sentiment Analysis of Indonesia's 2019 Presidential Election," *IIUM Engineering Journal*, vol. 22, no. 1, pp. 78–92, 2021, https://doi.org/10.31436/iiumej.v22i1.1532.
[22] D. G. Mandhasiya, H. Murfi, A. Bustamam, and P. Anki, "Evaluation of Machine Learning Performance Based on BERT Data Representation with LSTM Model to Conduct Sentiment Analysis in Indonesian for Predicting Voices of Social Media Users in the 2024 Indonesia Presidential Election," in *ICOIACT 2022 - 5th International Conference on Information and Communications Technology: A New Way to Make AI Useful for Everyone in the New Normal Era, Proceeding*, pp. 441–446, 2022, https://doi.org/10.1109/ICOIACT55506.2022.9972206.
[23] S. Kunal, A. Saha, A. Varma, and V. Tiwari, "Textual Dissection of Live Twitter Reviews using Naive Bayes," in *Procedia Computer Science*, pp. 307–313, 2018, https://doi.org/10.1016/j.procs.2018.05.182.
[24] I. Mozetič, M. Grčar, and J. Smailović, "Multilingual twitter sentiment classification: The role of human annotators," *PLoS One*, vol. 11, no. 5, 2016, https://doi.org/10.1371/journal.pone.0155036.
[25] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: the kappa statistic," *Fam. Med.*, vol. 37 no 5, pp. 360–3, 2005. https://pubmed.ncbi.nlm.nih.gov/15883903/.
[26] 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, https://doi.org/10.1109/TAFFC.2022.3206891.
[27] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer," in *3rd East Indonesia Conference on Computer and Information Technology, EIConCIT*, pp. 122–127, 2021, https://doi.org/10.1109/EIConCIT50028.2021.9431852.
[28] T. M. Fagbola and S. C. Thakur, "Lexicon-based Bot-aware Public Emotion Mining and Sentiment Analysis of the Nigerian 2019 Presidential Election on Twitter," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019. https://doi.org/10.14569/IJACSA.2019.0101047.
[29] D. Fimoza, A. Amalia, and T. Henny Febriana Harumy, "Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT," in *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2021 - Proceedings*, pp. 27–34, 2021, https://doi.org/10.1109/DATABIA53375.2021.9650096.
[30] B. S. Abunasser, M. Rasheed, J. Al-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Breast Cancer Detection and Classification using Deep Learning Xception Algorithm," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, https://doi.org/10.14569/IJACSA.2022.0130729.
[31] C. H. Miranda, G. Sanchez-Torres, and D. Salcedo, "Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture," *Data (Basel)*, vol. 8, no. 6, p. 96, 2023, https://doi.org/10.3390/data8060096.
[32] Y. Fan, Y. Zhang, B. Guo, X. Luo, Q. Peng, and Z. Jin, "A Hybrid Sparrow Search Algorithm of the Hyperparameter Optimization in Deep Learning," *Mathematics*, vol. 10, no. 16, 2022, https://doi.org/10.3390/math10163019.
[33] D. Sarkar. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Bangalore: Apress, 2019, https://doi.org/10.1007/978-1-4842-4354-1.
[34] W. Aljedaani *et al.*, "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry," *Knowl Based Syst*, vol. 255, 2022, https://doi.org/10.1016/j.knosys.2022.109780.
[35] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," *Soc Netw Anal Min*, vol. 5, no. 1, pp. 1–25, 2015, https://doi.org/10.1007/s13278-015-0253-5.
[36] S. Sidharta, Y. Heryadi, L. Lukas, W. Suparta, and Y. Arifin, "Sentiment Analysis of Hashtag Activism on Social Media Twitter," in *ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era*, pp. 73–77, 2023, https://doi.org/10.1109/ICCoSITE57641.2023.10127833.
[37] D. Shah, A. Singh, and S. S. Prasad, "Sentimental Analysis Using Supervised Learning Algorithms," in *Proceedings - 3rd International Conference on Computation, Automation and Knowledge Management, ICCAKM,* pp. 1-6, 2022, https://doi.org/10.1109/ICCAKM54721.2022.9990320.
[38] W. Chebil, M. Wedyan, M. Alazab, R. Alturki, and O. Elshaweesh, "Improving Semantic Information Retrieval Using Multinomial Naive Bayes Classifier and Bayesian Networks," *Information (Switzerland)*, vol. 14, no. 5, 2023, https://doi.org/10.3390/info14050272.

[39] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter," in *Fourth International Conference on Informatics and Computing (ICIC)*, pp. 1-5, 2019, https://doi.org/10.1109/ICIC47613.2019.8985884.
[40] R. Patel and K. Passi, "Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning," *IoT*, vol. 1, no. 2, pp. 218–239, 2020, https://doi.org/10.3390/iot1020014.
[41] S. Nabil, J. Elbouhdidi, and M. Y. Chkouri, "Sentiment Analysis of Amazon's Reviews Using Machine Learning Algorithms," *J Theor Appl Inf Technol*, vol. 99, p. 22, 2021, http://www.jatit.org/volumes/Vol99No22/30Vol99No22.pdf.
[42] H. Adamu, M. J. Bin Mat Jiran, K. H. Gan, and N. H. Samsudin, "Text Analytics on Twitter Text-based Public Sentiment for Covid-19 Vaccine: A Machine Learning Approach," in *3rd IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET*, pp. 1-6, 2021. https://doi.org/10.1109/IICAIET51634.2021.9573866.
[43] M. Alzyout, E. Al Bashabsheh, H. Najadat, and A. Alaiad, "Sentiment Analysis of Arabic Tweets about Violence against Women using Machine Learning," in *2021 12th International Conference on Information and Communication Systems, ICICS*, pp. 171–176, 2021, https://doi.org/10.1109/ICICS52457.2021.9464600.
[44] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," in *Procedia Computer Science*, pp. 341–348, 2019, https://doi.org/10.1016/j.procs.2019.05.008.
[45] P. Karthika, Dr. R. Murugeswari, and Mrs. R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1-5, 2019. https://doi.org/10.1109/INCOS45849.2019.8951367.
[46] G. Yang and L. Qicheng, "An over Sampling Method of Unbalanced Data Based on Ant Colony Clustering," *IEEE Access*, vol. 9, pp. 130990–130996, 2021, https://doi.org/10.1109/ACCESS.2021.3114443.

BIOGRAPHY OF AUTHORS

Lenggo Geni is pursuing a postgraduate's degree in Magister of Computer Science at Universitas Indonesia. She works at the Ministry of Health of the Republic of Indonesia as an employee in the Inspectorate General Unit and currently doing a study assignment. She is interested in researching Social Media, Data Mining, Sentiment Analysis, Knowledge Management, e-Government, and e-Health. Orcid: http://orcid.org/0000-0002-8468-9858. Email: lenggo.geni11@ui.ac.id.

Evi Yulianti is a lecturer and researcher at the Faculty of Computer Science, Universitas Indonesia. She received the B. Comp. Sc. degree from Universitas Indonesia in 2010, the dual M.Comp.Sc. degree from Universitas Indonesia and Royal Melbourne Institute of Technology University in 2013, and the Ph. D. degree from Royal Melbourne Institute of Technology University in 2018. Her research interests include information retrieval and natural language processing. Orcid: http://orcid.org/0000-0003-1951-4696. Email: evi.y@cs.ui.ac.id.

Dana Indra Sensuse is a professor and researcher at the Faculty of Computer Science, Universitas Indonesia. He received his Ph. D. degree from the University Of Toronto in 2004. His research interests are e-Government, Smart City, e-Commerce/e-Business, IT Government, Data Mining, Knowledge Management, and IT Plan. Orcid: http://orcid.org/0000-0002-0012-8552. Email: dana@cs.ui.ac.id.