Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 2, April 2025, Page. 803-818
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4015

Comparison ff Sentiment Labeling Using Textblob, Vader, and Flair in Public Opinion Analysis Post-2024 Presidential Inauguration with IndoBERT

Khoerul Anam[1], Kusnawi[*2]

[1,2]Informatics, Computer Science Faculty, Universitas Amikom Yogyakarta, Indonesia

Email: [2]khusnawi@amikom.ac.id

Received : Oct 14, 2024; Revised : Nov 7, 2024; Accepted : Nov 11, 2024; Published : Apr 26, 2025

Abstract

The results of the 2024 Indonesian presidential election decided that Prabowo Subianto and Gibran Rakabuming Raka became the elected pair of Indonesian presidential and vice-presidential candidates in 2024. The pair's election triggered various public reactions, especially on social media platforms. Some social media platforms provided diverse opinions, indicating a wide variety of views on this issue. This research aims to analyze public opinion after the election of the 2024 Indonesian president by comparing sentiment using TextBlob, VADER (Valence Aware Dictionary and sEntiment Reasoner), and Flair. Training and testing are done with the IndoBERT model to determine the most effective sentiment labeling. This research starts by collecting text data from social media X, YouTube, and Instagram, then preprocessing, translating, and labeling data using three libraries, training, and testing using IndoBERT. The results of training and testing data show that Flair has an accuracy of 81.29%, TextBlob has an accuracy of 73.35%, and VADER has an accuracy of 74.86%. From the accuracy results obtained, it can be concluded that labeling using Flair provides the greatest accuracy of the others because the Flair labeling process uses deep learning and contextual embedding techniques.

Keywords : *Flair, IndoBERT, Presidential Election, Sentimen Analysis, Textblob, Vader.*

1. INTRODUCTION

The Indonesian presidential election in 2024 is one of the crucial moments to choose Indonesia's leader. In a democratic country, this election manifests the Indonesian people exercising their right to vote to elect leaders and representatives whom the people trust [1]. The vote count results have decided on one elected president and vice president, who will be inaugurated in October 2024. The time between the announcement and the inauguration of the elected president and vice president has led to many public opinions about the condition of the continuity of the Indonesian nation after the presidential inauguration. One of the means to express these opinions is through social media [2].

The number of social media platforms provides a variety of public opinions regarding the country's condition after the Indonesian president's inauguration. Many people give positive opinions regarding the country's condition after the inauguration. However, some give negative or neutral opinions. Many people have high hopes for the elected couple, so the country becomes more advanced and better. In addition, many also give negative opinions regarding the conditions after the inauguration because they are not satisfied with the election results. The variety of public opinions from social media can be collected and used as data for various activities, one of which is sentiment analysis. [3].

Sentiment analysis is the process of identifying the polarity of textual sentiment, whether the text represents a positive, negative, or neutral sentiment [4]. The purpose of sentiment analysis is to understand and provide insight into public opinion or a trend, aid research, or inform decision-making.

[5]. Sentiment analysis is also used in various fields, such as marketing, politics, health, etc [6]. In the process, sentiment analysis contains four main steps: preprocessing, feature extraction, classification, and interpretation of results [7]. Each of these steps has a critical role to play in ensuring the accuracy and relevance of the analysis results, thus providing a clear picture of how people respond to current issues [8].

Sentiment analysis of the country's state following the presidential inauguration is important as it can provide a clear picture of people's reactions and public perceptions of the new leadership. The data from various social media can be used to understand how the president's policies and actions affect people's expectations and trust. in addition, this analysis can also help identify issues of public concern as well as potential challenges faced by the administration. Thus, a deep understanding of the post-inauguration conditions is important to assess the direction and effectiveness of the leadership that will take place.

Community responses are collected and processed to become data for sentiment analysis [9]. The data will be subjected to various stages before obtaining valid results [10]. One of the stages is labeling the data into positive, negative, and neutral based on the responses given. Many techniques can be used to label data, such as a manual approach or a library that can automate the process [11]. Using libraries in data labeling is very useful because it can speed up the process, reduce workload, and increase time efficiency, especially on large and voluminous data.

Currently, many pre-trained libraries can be used to label text data. Some of them are Natural Language Processing (NLP) libraries such as TextBlob [12] , VADER (Valence Aware Dictionary and sEntiment Reasoner), and Flair. Labeling using the TextBlob library is easy because it can restore the polarity and subjectivity of sentences [13]. Labeling using VADER was chosen because this library is designed to understand text directly, especially on social media. VADER also considers capital and lowercase letters, emoticons, emphasis, and slang in providing sentiment [14]. Flair can understand and analyze sentiment bi-directionally and consider where words appear to enable accurate representation; besides that, Flair is also easy to use because it already supports a variety of pre-trained models [15].

Dataset results labeled with sentiment using the library will then be tested for accuracy and analyzed using the IndoBERT model. IndoBERT is a model that has a BERT-based architecture (Bidirectional Encoder Representations from Transformers). This model is known for its ability to understand the context of words in sentences to provide more accurate analysis results [16]. IndoBERT is a deep learning model designed to effectively understand the context of words by reading words in sentences bidirectionally. It is a model based on transformer architecture with self-attention mechanism as its main component. IndoBERT is a model that has been trained with 220 million Indonesian words. BERT is designed to understand the context of words in sentences in two directions to improve understanding of the meaning and relationship between words [17].

Some labeling techniques have advantages and disadvantages. So, how do TextBlob, VADER, and Flair perform in labeling the sentiment of public opinion on social media after the 2024 Presidential Inauguration, analyzed using IndoBERT? This research aims to compare the performance of TextBlob, VADER, and Flair labeling. The sentiment labeling results will be trained and tested using the IndoBERT model. It is hoped that this research can significantly contribute in several ways, especially in determining the effective labeling between TextBlob, VADER, and Flair for sentiment analysis by training and testing the labeling results using IndoBERT. In addition, this research also provides an understanding of public opinion on the country's condition after the presidential inauguration.

2. METHOD

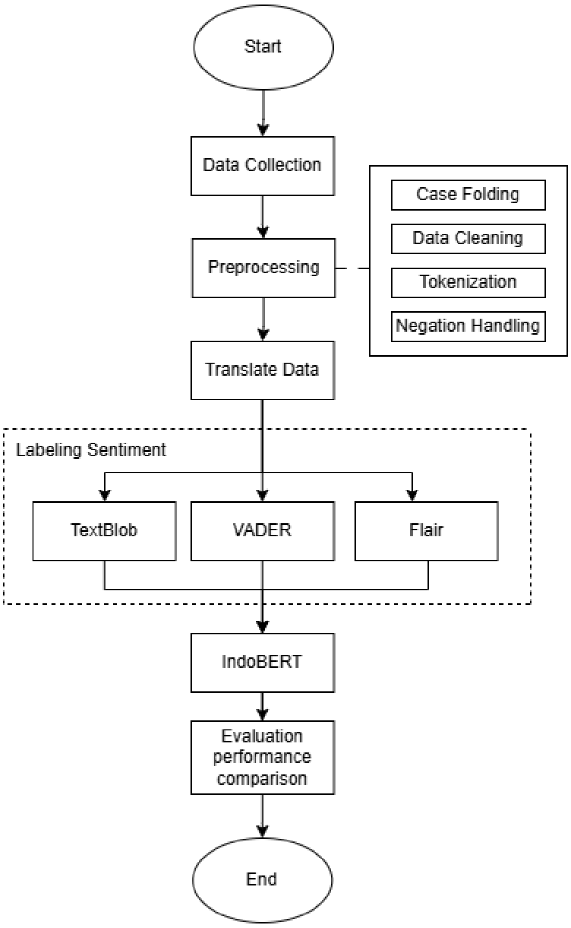


Figure 1. Research Flow

Based on figure 1, this research starts with data collection, preprocessing, translation, and data labeling using TextBlob, VADER, and Flair, which will be trained and tested using the IndoBERT model. The evaluation process is carried out by calculating the accuracy, precision, recall, and F1-score values to determine the performance of each library [18].

2.1. Data Collection

This research uses data from X social media, YouTube, and Instagram comments. Text data collection in the form of tweets on the X platform using crawling techniques using the X API with the keyword "presidential inauguration" from March to July. Data collection on YouTube and Instagram comments uses scrapping techniques, such as a Google extension called Data Scrapper. The scrapping results from the three platforms were combined and focused on the tweet section of platform X and the comments section of the YouTube and Instagram platforms.

2.2. Preprocessing dan Translate

The raw data from each social media platform will proceed to the preprocessing stage. Preprocessing text data is very important to transform raw data, cleaning it from meaningless and unstructured words into a structured format ready for analysis [19][20]. The preprocessing stage in this research includes several steps: case folding, cleaning, tokenization, and negation handling. Applying preprocessing steps appropriately will increase the accuracy of the analysis [21].

2.2.1. CaseFolding

Case folding is the process of converting the entire text to lowercase [22]. This process is done because there is no consistency in using capital and lowercase letters in text data [23]. In this research, using the pandas library in performing the case folding process.

2.2.2. Data Cleaning

Data cleaning is a stage carried out to reduce noise in documents. Noise that is removed includes mentions (@username), numbers, links, punctuation marks, symbols, and other characters that must be removed [24]. In this research, using regular expression and string library in cleaning process.

2.2.3. Tokenization

Tokenization is a tokenization step that breaks text into words [25] which allows a more in-depth analysis of each word in the text. The result of tokenization is a series of tokens in each sentence. In this research, using the nltk library library in performing the tokenization process.

2.2.4. Negation Handling

The negation handling stage aims to identify and manage negation in sentences, which can affect the meaning and polarity of the analyzed text. Negation can change the meaning of a positive statement to a negative one or vice versa. Therefore, proper negation handling can make a significant contribution to improving accuracy [26]. In this research, using the pandas library in performing the negation handling process.

2.2.5. Translate Data

The result of preprocessing produces clean data. This data will be translated from Indonesian to English. This step is done because the following process is sentiment labeling on each data. This labeling will be done using the TextBlob, VADER, and Flair libraries, which are known to label data effectively in English text.

2.3. Labeling

The data that has been translated, the next stage is labeling the text data. This labeling aims to determine whether the text contains positive, negative, or neutral sentiments. In this research, the data will be grouped into three sentiment categories: positive, negative, and neutral. This labeling will be done automatically using TextBlob, VADER, and Flair.

2.3.1. TextBlob

TextBlob is a Python library that provides a simple API (Application Programming Integration) in NLP tasks, one of which is sentiment analysis. [27]. Because this library is practical on English text, it is necessary to translate the text from Indonesian to English. TextBlob is a lexicon-based method that can be categorized into 3 sentiments: positive, negative, and neutral. This category is based on the calculation of the polarity value; if the polarity value is more than 0 (polarity > 0), it will produce a positive sentiment; if the polarity value is equal to 0 (polarity = 0), it will produce a neutral sentiment, and if the polarity value is less than 0 (polarity < 0) it will produce a negative sentiment [28].

2.3.2. VADER

VADER is a lexicon-based method suitable for analyzing social media data because it assigns sentiment scores to words and phrases, thus categorizing sentiment as positive, negative, or neutral [29]. VADER considers the entire text, including emphasis, emoticons, negation, or amplifying words, in determining sentiment. Determining negative, positive, and neutral sentiments is based on compound values ranging from -1 (very negative) to +1 (very positive) [30].

2.3.3. Flair

Flair is an NLP library that utilizes deep learning and contextual embedding techniques [31]. Flair can provide more accurate results than lexicon-based or other methods because Flair can consider the relationship between words in a sentence [32]. In addition, Flair has a pre-trained model that can be used to facilitate the sentiment labeling process.

2.4. Split Data

Split data is dividing data into different subsets [33]. The processed data will then be divided into subsets: training data, validation data, and test data. This division is done with the aim of model development and evaluation. Training data helps to develop the model, validation data helps to check the model's performance, and test data helps to assess the performance of the developed model. The data split consists of 80 % for training, 10 % for validation, and 10 % for data testing. In stratification sampling, this is the case that guarantees that sentiment class distribution is maintained during the data distribution.

2.5. IndoBERT

IndoBERT is a BERT architecture-based model designed explicitly for Indonesians. BERT is a deep learning model designed to effectively understand word context by reading words in sentences bidirectionally [29]. It is a model based on transformer architecture [30] with a self-attention mechanism as its main component.

In this research, the data collected is text data from social media platforms and is Indonesian text, so the IndoBERT model is an ideal choice. IndoBERT is a model trained using Indonesian language data sets [34]. IndoBERT training uses over 220 million words from the Indonesian Wikipedia, online news articles, and Indonesian corpus. IndoBERT has several variations, such as indobert-base, indobert-large, indobert-lite-base, and indobert-lite-large.

Both IndoBERT and BERT have stages of sentiment analysis. 1) Tokenization is the stage of converting a sentence into a series of tokens. At this stage, a unique token [CLS] is also added at the beginning of the sentence, and a token [SEP] is used to separate the sentence [31]. Furthermore, each token is mapped to a numeric ID. A tokenizer from IndoBERT is used in this process. 2) Padding and truncation is the process of determining the maximum length of a sentence (e.g., 128). Sentences with a length of more than 128 tokens will be truncated, while sentences with a length of less than 128 will be padded to the maximum length. 3) Embeddings consist of token embeddings, segment embeddings, and position embeddings. Token embeddings convert each token into a numeric vector. Segment embeddings label the sentences to distinguish one from another. Position embeddings mark the position of each token in the sequence. 4) Transformers process the input in the transformer layer. This layer uses a self-attention mechanism to consider and understand the relationship between all tokens so that the model can understand the word's context. The result will be passed to the feed-forward to generate a new token representation. The token [CLS] will represent the final result of the transformer process.
5) Fine-tuning uses IndoBERT variation of indobert-base-p1 to maximize the analysis performance [35]. This model was chosen due to its ability to understand context and meaning in Indonesian effectively. The performance of the model is also improved through hyperparameter settings, such as the use of the AdamW optimizer, a learning rate of 1e-5, an epsilon value of 1e-5, a weight decay value of 0.001, a number of epochs of 10, the use of early stopping, and the determination of a batch size value of 32.

2.6. Evaluation

This evaluation stage is crucial in assessing how well the analysis results were obtained. Here, a confusion matrix is used during the evaluation process. The confusion matrix is a measure of a classifier's accuracy that is defined concerning four outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The results of confusion matrices are also used to derive evaluation metrics such as accuracy, precision, recall, and f1-score measurement of the model being evaluated. [36].

1. Accuracy

In evaluating a model's classification capability, accuracy measures the performance such a model achieves. The accuracy metric is the percentage of all correct predictions across the target dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

2. Precision

Precision is the value of correct positive predictions to all positive predictions.

$$Precission = \frac{TP}{TP+FP} \qquad (2)$$

3. Recall

Recall merupakan nilai dari data positif yang berhasil diprediksi benar dari seluruh data positif

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

4. F1 – Score

The F1-score accounts for both recall and precision and takes them to be averaged.

$$F1 - score = 2 \times \frac{P \times R}{P+R} \qquad (4)$$

3. RESULT

3.1. Data Collection

In this study, data was sourced from 3 social media platforms: Twitter (now known as X), YouTube comments, and Instagram comments. X tweet data was collected from March to July with the keyword "presidential inauguration." YouTube comment data was collected from several videos that were relevant to the theme of the presidential inauguration and were videos that came from trusted online news accounts. Instagram comment data was taken from several posts relevant to the theme of the presidential inauguration and came from trusted online news media accounts. After collecting data from each platform, it is selected in the tweets and comments section. The final result of the data collection and selection process resulted in a dataset consisting of 5,283 data.

3.2. Preprocessing and Translate

In this research, preprocessing includes several stages: case folding, data cleaning, tokenization, and negation handling. Furthermore, the results of negation handling will be translated from Indonesian to English. The preprocessing and translation process can be seen in Table 1.

Table 1. Preprocessing Process

Process	Before	After
Case Folding	@tvOneNews Selama Periode 2024-2029 apakah akan terjadi 2x pelantikan presiden? Tidak layak utk ditunggu.	@tvonenews selama periode 2024-2029 apakah akan terjadi 2x pelantikan presiden? tidak layak utk ditunggu.
Data Cleaning	@tvonenews selama periode 2024-2029 apakah akan terjadi 2x pelantikan presiden? tidak layak utk ditunggu	selama periode apakah akan terjadi x pelantikan presiden? tidak layak utk ditunggu
Tokenization	selama periode apakah akan terjadi x pelantikan presiden? tidak layak utk ditunggu	['selama', 'periode', 'apakah', 'akan', 'terjadi', 'x', 'pelantikan', 'presiden',’?’, 'tidak', 'layak', 'utk', 'ditunggu']
Negation Handling	['selama', 'periode', 'apakah', 'akan', 'terjadi', 'x', 'pelantikan', 'presiden', ‘?’, 'tidak', 'layak', 'utk', 'ditunggu']	['selama', 'periode', 'apakah', 'akan', 'terjadi', x, 'pelantikan', 'presiden',’?’, 'tidak', 'tidak_layak', utk, 'ditunggu']
Translate	selama periode apakah akan terjadi x pelantikan presiden? tidak_layak utk ditunggu	During what period will there be x presidential inaugurations? not worth the wait

3.3. Sentiment Labeling Result

The sentiment labeling results for each library can be seen in Table 2.

Table 2. Sentiment Results

Library	Positive	Negative	Neutral
TextBlob	2046	937	2300
VADER	2445	1604	1234
Flair	2520	2763	0

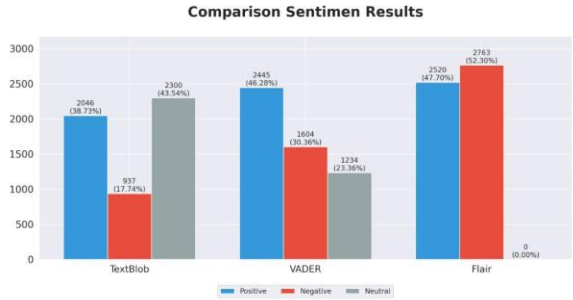


Figure 2. Comparison of Sentiment Results

Table 2 shows different results for each *library*. TextBlob and VADER can classify sentiment into positive, negative, and neutral categories. Flair, using the *pre-trained* model ' *en-sentiment* ', which is a model for English text, cannot detect neutral sentiment. Each labeling has a different approach to analyzing sentiment, so the results vary.

Figure 2 shows the labeling results of each method. TextBlob is a lexicon-based library that uses a simple approach to determine sentiment. The results from TextBlob show that this labelling predominantly classifies neutral sentiment, which is 2300 or 43.51% of the total data. This shows that TextBlob identifies more text that does not have strong emotions so that it is read as neutral sentiment. TextBlob is an easy and fast library, but it needs to be more accurate in handling complex sentences.

Therefore, TextBlob is suitable for fundamental sentiment analysis or on datasets that do not have explicit emotions.

VADER shows dominant results on positive sentiments. 2445 positive sentiments were classified, or 46.28% of the total data. VADER uses a lexicon-based approach designed explicitly for text from social media that tends to be informal. This text can provide a more explicit emotional picture. VADER has the advantage of considering the entire text, including emphasis, emoticons, negation, and amplifying words, to determine sentiment. This ability allows VADER to detect more positive or negative sentiments than neutral, with a neutral classification of only 23.36% or 1234 data. This makes VADER ideal for analyzing texts with a wide variety of emotions, such as product reviews or social media posts that are emotional.

Flair uses a different approach between TextBlob and VADER. Flair is based on a deep learning model. In the research, labelling results using Flair only classify two sentiments, positive and negative, without giving results on neutral sentiments. The labelling results detected 47.70% positive sentiment or 2520 data and 52.10% negative sentiment or 2763 data from the total data. The pre-trained 'en-sentiment' model allows Flair to understand text context better than rule-based libraries. Flair is suitable for analysis that requires in-depth coverage of complex text, but its weakness is that it cannot detect neutral sentiment.

These three libraries offer different advantages and disadvantages when analyzing sentiment. TextBlob is an accessible library that is fast to use and dominant in detecting sentiment neutrality. VADER is a library for reading social media text, so it is better at handling informal language. VADER also considers good emotions through emphasis, emoticons, and reinforcing words in analyzing sentiment to balance positive and negative results. At the same time, Flair, with a deep learning model, is superior in understanding the context of the sentence but cannot detect neutral sentiment.

Table 1. Comparison of Sentiment Analysis on Sentences

Sentence	TextBlob	VADER	Flair
menyatakan oposisi setelah pelantikan presiden baru	Positive	Neutral	Negative
setelah pelantikan presiden bakal bnyk lg tarif yg akan naik	Positive	Neutral	Negative
saya sampai saat ini tidak akan mendukung perubahan apalagi seusai pelantikan presiden dan wapres hasil kecurangan	Neutral	Negative	Negative

Sentiment labelling using three different libraries will produce various results on the analyzed text. Table 3 clearly shows how each library classifies a single sentence with different sentiments. This highlights the different approaches applied by each library in analyzing sentiment. Therefore, Table 3 clearly shows the variation in results produced by each library on the exact text. This shows the importance of choosing the proper labelling for sentiment analysis.

3.4. Split Data Result

The 5,283 data is divided into 80% training data, equivalent to 4225 data; 10% validation data, equivalent to 529 data; and 10% test data, or equivalent to 529 data. The data split process uses stratified sampling to ensure sentiment balance in each data subset. The results of the split data and sentiment distribution in each Library can be seen in Table 4, Table 5, and Table 6.

Table 2. Split Data Results and Sentiment Distribution on TextBlob

Subset	Total Data & Percentage	Sentiment	Count & Percentage
Train	4225 (80.0%)	Negative	749 (17.73%)
		Positive	1636 (38.72%)
		Neutral	1840 (43.55%)
Validation	529 (10.0%)	Negative	94 (17.77%)
		Positive	205 (38.75%)
		Neutral	230 (43.48%)
Test	529 (10.0%)	Negative	94 (17.77%)
		Positive	205 (38.75%)
		Neutral	230 (43.48%)

Table 4 shows the data division and sentiment distribution results on the dataset analyzed using TextBlob. The dataset is divided into three subsets: Train, Validation, and Test, with a proportion of 80% for training data, 10% for validation, and 10% for test data. In the Train subset, there are 4225 data, with a distribution of negative sentiment of 749 data or 17.73%, positive sentiment of 1636 data or 38.72%, and neutral sentiment of 1840 data or 43.55%.

For the Validation subset, from a total of 529 data, 94 data or 17.77%, as negative, 205 data or 38.75% as positive, and 230 data or 43.48% as neutral. Similarly, in the Test subset, which also consists of 529 data, the sentiment distribution shows 94 data or 17.77%, as negative, 205 data or 38.75%, as positive, and 230 data or 43.48%, as neutral.

Table 5. Split Data Results and Sentiment Distribution on VADER

Subset	Total Data & Percentage	Sentiment	Count & Percentage
Train	4225 (80.0%)	Negative	1283 (30.37%)
		Positive	1955 (46.27%)
		Neutral	987 (23.36%)
Validation	529 (10.0%)	Negative	160 (30.25%)
		Positive	245 (46.31%)
		Neutral	124 (23.44%)
Test	529 (10.0%)	Negative	161 (30.43%)
		Positive	245 (46.31 %)
		Neutral	123 (23.25%)

Table 5 presents the data division and sentiment distribution results on the dataset analyzed using VADER. The data is divided into three subsets: Train, Validation, and Test, with equal proportions of 80% for training data, 10% for validation, and 10% for test. In the Train subset, out of 4225 data, 1283 data or 30.37%, were classified as negative, 1955 data or 46.27%, as positive, and 987 data or 23.36%, as neutral.

The Validation subset has 529 data with a negative sentiment distribution of 160 data or 30.25%, 245 positive data or 46.31%, and 124 neutral data or 23.44%. The Test subset with the same amount of data, 529 data, has almost the same distribution as the validation subset, where 161 data or 30.43%, are negative, 245 data or 46.31% are positive, and 123 data or 23.25% are neutral.

Table 6. Split Data Results and Sentiment Distribution on Flair

Subset	Total Data & Percentage	Sentiment	Count & Percentage
Train	4225 (80.0%)	Negative	2209 (52.28%)
		Positive	2016 (47.72%)
		Neutral	0
Validation	529 (10.0%)	Negative	277 (52.36 %)

		Positive	252 (47.64%)
		Neutral	0
Test	529 (10.0%)	Negative	277 (52.36 %)
		Positive	252 (47.64%)
		Neutral	0

Table 6 shows the data division and sentiment distribution results on the dataset analyzed using Flair. As in the previous table, the dataset is divided into three subsets: Train, Validation, and Test, with proportions of 80%, 10%, and 10%, respectively. In the Train subset, there are 4225 data with the following sentiment distribution: 2209 data or 52.28%, are classified as negative and 2016 data or 47.72% as positive, with no neutral sentiment category.

In the Validation subset, out of 529 data, 277 data or 52.36%, are classified as negative, and 252 data or 47.64% as positive. The same distribution is also seen in the Test subset, where 277 data or 52.36%, are negative, and 252 data, or 47.64%, are positive.

The absence of a neutral category in Flair's analysis results confirms that this library only classifies data into two sentiment categories, namely positive and negative.

3.5. Evaluation Performance Comparation

A series of processes, from preprocessing to split data, has been carried out. The results of each library are trained and tested using the IndoBERT model to determine the performance of each labeling. The report classification of TextBlob is shown in Table 7, the report classification of VADER is shown in Table 8, and the report classification of Flair is shown in Table 9.

Table 3. Classification Report TextBlob

	Precision	Recall	F1-score
Negative	0.6552	0.6064	0.6298
Positive	0.7225	0.7366	0.7295
Neutral	0.7725	0.7826	0.7775
Accuracy			0.7335
Macro avg	0.7167	0.7085	0.7123
Weighted avg	0.7323	0.7335	0.7327

Table 7 is TextBlob's Classification Report in sentiment classification based on three main metrics, namely Precision, Recall, and F1-score for Negative, Positive, and Neutral sentiment categories. In the Negative category, Precision reaches a value of 0.6552, which indicates that 65.52% of the negative predictions by TextBlob are correct. Recall in this category was recorded at 0.6064, meaning that TextBlob could identify 60.64% of the negative data. The F1-score in the negative category is 0.6298, which shows a balance between Precision and recall in this category, although the performance is relatively lower than in other categories.

In the Positive category, TextBlob performed better, with a Precision of 0.7225, indicating that 72.25% of the positive predictions were accurate. The recall for this category was 0.7366, indicating that TextBlob could detect 73.66% of the positive data correctly. The F1-score in the positive category reached 0.7295, which shows a stable performance in identifying and classifying positive sentiments.

In the Neutral category, TextBlob showed its best performance with a Precision of 0.7725, which means 77.25% of the neutral predictions by TextBlob were correct. Recall for this category was recorded at 0.7826, which indicates that the model could detect 78.26% of the total data that was truly neutral. The F1-score for the neutral category reached the highest value among the three categories at 0.7775, indicating an excellent performance in identifying neutral sentiments.

Vol. 6, No. 2, April 2025, Page. 803-818
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.2.4015

Overall, TextBlob's accuracy in sentiment classification is 0.7335, indicating that It performs quite well in sentiment analysis. The weighted average values for Precision, recall, and F1-score are 0.7323, 0.7335, and 0.7327. This shows that TextBlob provides balanced results in terms of detecting positive and neutral sentiments, although its performance decreases slightly in negative sentiments. Based on the results obtained, TextBlob is suitable for use in a simple sentiment analysis context.

Table 4. Classification Report VADER

	Precision	Recall	F1-score
Negative	0.7079	0.7826	0.7434
Positive	0.8008	0.7714	0.7859
Neutral	0.7043	0.6585	0.6807
Accuracy			0.7486
Macro avg	0.7377	0.7375	0.7366
Weighted avg	0.7501	0.7486	0.7485

Table 8 is VADER's Classification Report, classifying sentiments into negative, positive, and neutral categories based on precision, Recall, and F1-score metrics. In the Negative category, VADER has a Precision of 0.7079, which means about 70.79% of the negative predictions are correct. Recall in this category reached 0.7826, indicating that VADER could detect 78.26% of the data that was negative. The F1-score for negative sentiment was recorded at 0.7434, a balance between Precision and Recall, indicating a fairly good performance in this category.

In the Positive category, VADER showed a higher performance with a Precision of 0.8008, indicating that 80.08% of the positive predictions were accurate. Recall in this category was 0.7714, indicating that VADER successfully identified 77.14% of the total data that was truly positive. The F1-score for positive sentiment is 0.7859, indicating a fairly good and stable performance in detecting positive sentiment.

Precision was recorded at 0.7043 in the Neutral category, which means 70.43% of the neutral predictions were correct. However, the Recall for neutral sentiment is lower than the other categories, at 0.6585, indicating that VADER can only detect 65.85% of the neutral data. The F1-score for this category is 0.6807, which reflects that VADER is less optimal in classifying neutral sentiment than other categories.

Overall, VADER has an accuracy of 0.7486, with weighted average values for Precision, Recall, and F1-score of 0.7501, 0.7486, and 0.7485, respectively. These results show that VADER generally performs well, especially in detecting positive and negative sentiments, but shows weakness in classifying neutral sentiments.

Table 5. Classification Report Flair

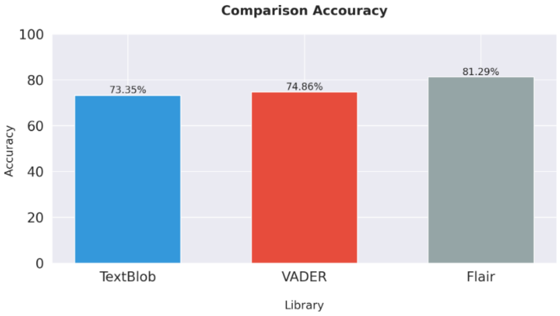
	Precision	Recall	F1-score
Negative	0.8090	0.8412	0.8248
Positive	0.8174	0.7817	0.7992
Accuracy			0.8129
Macro avg	0.8132	0.8115	0.8120
Weighted avg	0.8130	0.8129	0.8126

Table 9 is Flair's Classification Report, which uses a deep learning-based approach in classifying sentiment into Negative and Positive categories. In the Negative category, Flair shows superior performance with a Precision of 0.8090, which means about 80.90% of the negative predictions are correct. Recall in this category reached 0.8412, indicating that 84.12% of the supposedly negative data was correctly identified. The F1-score on the negative category was 0.8248, signifying an excellent performance in detecting and classifying negative sentiments.

In the Positive category, Precision was recorded at 0.8174, indicating that 81.74% of Flair's positive predictions were correct. Recall in the positive category was 0.7817, which means Flair successfully detected 78.17% of the data that was actually positive. The F1 Score for this category is 0.7992, which signifies a balance between Precision and recall, showing that Flair is able to handle positive sentiment with strong performance.

Overall, Flair's accuracy is 0.8129, reflecting its high performance in sentiment analysis. The weighted average values for Precision, recall, and F1-score are 0.8130, 0.8129, and 0.8126, respectively. This shows that Flair, which uses a deep learning model, performs better than other libraries in detecting both negative and positive sentiments. Based on these results, Flair proved to be better at handling complex sentiment classification, with an overall higher accuracy rate..

Based on the classification report's results, each library's accuracy comparison is shown in Figure 3.



Gambar 1. Comparison Accuracy

Figure 3 shows the accuracy of the comparison of each label. TextBlob produces 73.35% accuracy, VADER produces 74.86% accuracy, and Flair produces 81.29% accuracy. Based on the classification report and the accuracy value obtained, Flair performs better than TexBlob and VADER. This is because Flair analyzes by using contextual embedding, thus enabling accurate and better representation.

4. DISCUSSIONS

Sentiment labeling can be done in several ways, such as manual or automatic labeling. In this research, automatic labeling is done using the TextBlob, VADER, and Flair libraries. Research [37] also uses TextBlob in labeling, showing that sentiment labeling using TextBlob can classify sentiment into positive, negative, and neutral. Another study [38] also used TextBlob in labeling. The labeling results using TextBlob dominantly detect neutral sentiments compared to positive and negative. Another study [39] also conducted sentiment labeling using TextBlob with the results of neutral sentiment being more dominant.

In this research, VADER is also used in the labeling process. The results of this study, the labeling process using VADER can classify sentiments into positive, negative, and neutral. This is in accordance with research [40] which conducted sentiment labeling using VADER and classified sentiments into positive, negative, and neutral. Another study [41] also conducted sentiment labeling using VADER and produced dominant positive sentiment. Another study [42] also conducted sentiment labeling using TextBlob and VADER. The results of TextBlob labeling dominantly detect neutral sentiments, while VADER dominantly detects positive sentiments.

In this study, the labeling process using TextBlob and VADER can be classified into three sentiments: positive, negative, and neutral. Meanwhile, Flair labeling only classifies two sentiments, namely positive and negative. This is in line with research [32] prove that TextBlob can classify sentiments into positive, negative, and neutral but dominantly detect neutral and positive sentiments. VADER can classify into positive, negative, and neutral sentiments but dominantly detect positive and negative sentiments. Flair only classifies positive and negative sentiments with a dominant negative sentiment, and neutral sentiments cannot be detected. The same thing was done in research [43], which conducted labelling using TextBlob, VADER, and Flair. TextBlob and VADER can classify positive, negative, and neutral sentiments. Flair only classifies sentiment into positive and negative and cannot detect neutral sentiment.

In research [44] on sentiment analysis of COVID-19 vaccinations, comparing the results of sentiment analysis using TextBlob and VADER, it shows that VADER has better performance than TextBlob, VADER provides an accuracy of 85.22%, while Textblob provides an accuracy of 84.97%. This is in line with research [45] on the use of pre-trained models for sentiment analysis, which results in that VADER has better performance than TextBlob because VADER is specifically designed to predict sentiment from social media data by considering sentiment polarity and emotional intensity that takes into account emphasis, emoticons, slang words, and acronyms.

This research comparing TextBlob, VADER, and Flair labeling shows that Flair has higher accuracy than TextBlob and VADER. Flair labeling is done with deep learning and contextual embedding techniques. Flair is also a pre-trained model that provides better performance and efficiency. The results of research [46] comparing TextBlob, VADER, and Flair to measure sentiment from financial news, where the evaluation results show that Flair has superior performance compared to TextBlob and VADER. This is because Flair uses a deep learning approach that allows the model to understand the context of the sentence better. Research [32] also mentioned that Flair uses contextual embedding so that the model can understand the context of the words in the sentence better.

What has been mentioned is in line with this research because it gives the results of TextBlob, which dominantly detects neutral sentiment by 43.54%, positive sentiment by 38.73%, and negative sentiment by 17.74%. VADER dominantly detects positive sentiment by 46.28%, negative sentiment by 30.36%, and neutral sentiment by 23.36%. Flair only detects positive sentiment by 47.70% and negative sentiment by 52.30%. In addition, from the evaluation results, Flair's performance is better, with an accuracy value of 81.29%, compared to TextBlob, which has an accuracy of 73.36% and VADER, which has an accuracy of 74.86%.

5. CONCLUSION

Based on the research, 5,283 data were obtained from X tweets, YouTube comments, and Instagram comments. Data is preprocessed and translated from Indonesian to English, sentiment labeling using TextBlob, VADER, Flair, IndoBERT implementation, evaluation, and performance comparison. Based on the sentiment labeling results, Textblob dominantly detects neutral sentiments. VADER dominantly detects positive sentiment, and Flair dominantly detects negative sentiment and is unable to detect neutral sentiment. Split data is done with a ratio of 80% train data, 10% validation data, and 10% test data, and a stratified sampling method is done to ensure sentiment balance in each subset. Training and testing were conducted using IndoBERT with an accuracy of 73.35% on TextBlob, 74.86% on VADER, and Flair with an accuracy of 81.29%. Flair gives the best results because the Flair labeling process is done with deep learning and contextual embedding techniques. The accuracy results are more accurate compared to lexicon-based methods because Flair can consider the relationship between words in a sentence. Flair has a pre-trained model that provides better performance, efficiency, and good generalization. This proves that the sentiment labeling process will affect the sentiment label results and the model's performance in sentiment analysis.

Suggestions for future research are to use more data and more diverse sources, different preprocessing techniques, compare labelling techniques that have not been compared, and train and test using different models from this study.

REFERENCES

[1] Zulham, "Communication of Political Identity & Indonesian Presidential Candidacy in the 2024 Election," *Int. J. Humanit. Soc. Stud.*, vol. 11, no. 1, pp. 60–63, 2023, doi: 10.24940/theijhss/2023/v11/i1/hs2301-014.

[2] E. Konovalova, G. Le Mens, and N. Schöll, "Social media feedback and extreme opinion expression," *PLoS One*, vol. 18, no. 11 November, pp. 1–12, 2023, doi: 10.1371/journal.pone.0293805.

[3] A. Elhan, M. K. D. Hardhienata, H. Yeni, S. Wijaya Hartono, and J. Adisantoso, "Analisis Sentimen Pengguna Twitter terhadap Vaksinasi COVID-19 di Indonesia menggunakan Algoritme Random Forest dan BERT Sentiment Analysis of Twitter Users on COVID-19 Vaccines in Indonesia using Random Forest and BERT Algorithms," *J. Ilmu Komput. Agri-informatika*, vol. 9, no. 2, pp. 199–211, 2022, doi: 10.29244/jika.9.2.199-211.

[4] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, no. October 2021, p. 100157, 2022, doi: 10.1016/j.array.2022.100157.

[5] P. K. Rangarjan *et al.*, "The social media sentiment analysis framework: deep learning for sentiment analysis on social media," *Int. J. Electr. Comput. Eng.*, vol. 14, no. 3, pp. 3394–3405, 2024, doi: 10.11591/ijece.v14i3.pp3394-3405.

[6] U. Pathak and E. P. Rai, "Sentiment Analysis : Methods , Applications , and," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. February, 2023, doi: 10.22214/ijraset.2023.49165.

[7] M. Mikula, X. Gao, and M. Mach, "Lexicon-based Sentiment Analysis Using the Particle Swarm Optimization," *J. Electron.*, pp. 1–22, 2020, doi: 10.3390/electronics9081317.

[8] M. Kalaiarasu and C. Ranjeeth Kumar, "Sentiment Analysis using Improved Novel Convolutional Neural Network (SNCNN)," *Int. J. Comput. Commun. Control*, vol. 17, no. 2, pp. 1–15, 2022, doi: 10.15837/ijccc.2022.2.4351.

[9] D. J. Hussein, M. N. Rashad, K. I. Mirza, and D. L. Hussein, "Machine Learning Approach to Sentiment Analysis in Data Mining," *Passer J. Basic Appl. Sci.*, vol. 4, no. 1, pp. 71–77, 2022, doi: 10.24271/psr.2022.312664.1101.

[10] N. G. Ramadhan, M. Wibowo, N. F. L. Mohd Rosely, and C. Quix, "Opinion mining indonesian presidential election on twitter data based on decision tree method," *J. Infotel*, vol. 14, no. 4, pp. 243–248, 2022, doi: 10.20895/infotel.v14i4.832.

[11] B. M. Alenzi, M. B. Khan, M. H. A. Hasanat, A. K. J. Saudagar, M. Alkhathami, and A. Altameem, "Automatic Annotation Performance of TextBlob and VADER on Covid Vaccination Dataset," *Intell. Autom. Soft Comput.*, vol. 34, no. 2, pp. 1311–1331, 2022, doi: 10.32604/iasc.2022.025861.

[12] P. Sivalakshmi, P. U. Kumar, M. Vasanth, R. Srinath, and M. Yokesh, "COVID-19 Vaccine – Public Sentiment Analysis Using Python's Textblob Approach," *Int. J. Curr. Res. Rev.*, vol. 13, no. 11, pp. 166–172, 2021, doi: 10.31782/ijcrr.2021.sp218.

[13] A. Prof and P. Gujjar, "Sentiment Analysis : Textblob For Decision Making Department of Business Analytics," *Int. J. Sci. Res. Eng. Trends*, vol. 7, no. 2, pp. 1097–1099, 2021.

[14] E. Rosenberg *et al.*, "Results in Engineering Sentiment analysis on Twitter data towards climate action," *Results Eng.*, vol. 19, no. June, p. 101287, 2023, doi: 10.1016/j.rineng.2023.101287.

[15] D. A. Darji and S. A. Goswami, "The Comparative study of Python Libraries for Natural Language Processing (NLP)," 2024.

[16] P. Shah, H. Patel, and P. Swaminarayan, "EAI Endorsed Transactions Multitask Sentiment Analysis and Topic Classification," *EAI Endorsed Trans. Scalable Inf. Syst.*, pp. 1–12, doi: 10.4108/eetsis.5287.

[17] A. S. Alammary, “applied sciences BERT Models for Arabic Text Classification : A Systematic Review,” *J. Appl. Sci.*, pp. 1–20, 2022, doi: 10.3390/app12115720.

[18] V. Fitriyana, L. Hakim, D. Candra, R. Novitasari, and A. Hanif, “Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine,” *J. Buana Inform.*, vol. 14, no. April, pp. 40–49, 2023, doi: 10.24002/jbi.v14i01.6909.

[19] D. Rifaldi, A. Fadlil, and Herman, “Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet ‘Mental Health,’” *Decod. J. Pendidik. Teknol. Inf.*, vol. 3, no. 2, pp. 161–171, 2023, doi: 10.51454/decode.v3i2.131.

[20] K. Kusnawi and A. H. Wijaya, “Sentiment Analysis of Pancasila Values in Social Media Life Using the Naive Bayes Algorithm,” in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarangin, Indonesia*, 2021, pp. 96–101. doi: 10.1109/iSemantic52711.2021.9573194.

[21] K. U. Chouhan, R. S. Jha, N. Pradeep, K. Jha, and S. I. Kamaluddin, “Legal Document Analysis,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. IV, 2023, doi: 10.22214/ijraset.2023.50123.

[22] A. Muzaki and A. Witanti, “Sentiment Analysis Of The Community In The Twitter To The 2020 Election In Pandemic Covid-19 By Method Naive Bayes Classifier Sentimen Analisis Masyarakat Di Twitter Terhadap Pilkada 2020 Ditengah Pandemic Covid-19 Dengan Metode Na Ï Ve Bayes Classifier,” *J. Tek. Inform.*, vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.

[23] A. C. Khotimah *et al.*, “Comparison Naïve Bayes Classifier , K-Nearest Neighbor And Support Vector Machine In The Classification Of Individual On Perbandingan Algoritma Naïve Bayes Classifier , K-Nearest Neighbor Dan Support Vector Machine Dalam Klasifikasi,” *J. Tek. Inform.*, vol. 3, no. 3, pp. 673–680, 2022, doi: doi: 10.20884/1.jutif.2022.3.3.254.

[24] R. Puspitasari, Y. Findawati, M. A. Rosid, I. S. Program, and U. M. Sidoarjo, “Sentiment Analysis Of Post-Covid-19 Inflation Based On Twitter Using The K-Nearest Neighbor And Support Vector Machine Analisis Sentimen Terhadap Inflasi Pasca Covid-19 Berdasarkan Twitter Dengan Metode Klasifikasi K-Nearest Neighbor Dan,” *J. Tek. Inform.*, vol. 4, no. 4, pp. 669–679, 2023, doi: 10.52436/1.jutif.2023.4.4.801.

[25] R. Sistem, I. M. S. Putra, P. Jhonarendra, N. Kadek, and D. Rusjayanthi, “Deteksi Kesamaan Teks Jawaban pada Sistem Test Essay Online dengan Pendekatan Neural Network,” *J. RESTI (Rekayasa Sist. Dan Teknol. Inf.*, vol. 5, no. 158, pp. 3–12, 2021, doi: 10.29207/resti.v5i6.3544.

[26] K. Makkar, P. Kumar, M. Poriye, and S. Aggarwal, “Improving Sentiment Analysis using Negation Scope Detection and Negation Handling,” *Int. J. Comput. Digit. Syst.*, vol. 1, no. 1, pp. 239–247, 2024, doi: 10.12785/ijcds/160119.

[27] D. Hazarika, G. Konwar, and S. Deb, “Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing,” *Proc. Int. Conf. Res. Manag. Technovation*, vol. 24, pp. 63–67, 2020, doi: 10.15439/2020KM20.

[28] S. Dewi and D. B. Arianto, “Twitter Sentiment Analysis Towards Qatar As Host Of The 2022 World Cup Using Textblob,” *J. Soc. Res.*, no. 2018, pp. 443–454, 2022, doi: 10.55324/josr.v2i2.615.

[29] M. Arief and N. A. Samsudin, “Hybrid Approach with VADER and Multinomial Logistic Regression for Multiclass Sentiment Analysis in Online Customer Review,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 12, pp. 311–320, 2023, doi: 10.3390/s23010506.

[30] T. Pano and R. Kashef, “A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19,” 2020, doi: 10.3390/bdcc4040033.

[31] G. B. Herwanto, A. M. Ningtyas, I. G. Mujiyatna, and I. Nyoman, “Hate Speech Detection in Indonesian Twitter using Contextual Embedding Approach,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, 2021, doi: 10.22146/ijccs.64916.

[32] M. F. Ali, R. Irfan, and T. A. Lashari, “Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan : a study on governmental preventive measures,” *PeerJ Comput. Sci.*, 2023, doi: 10.7717/peerj-cs.1220.

[33] A. Fadlil, I. Riadi, and F. Andrianto, “Improving Sentiment Analysis in Digital Marketplaces through SVM Kernel Fine-Tuning,” *Int. J. Comput. Digit. Syst.*, vol. 1, no. 1, 2024, doi: 10.12785/ijcds/160113.

[34] E. Yulianti and N. K. Nissa, “ABSA of Indonesian customer reviews using IndoBERT : single-sentence and sentence-pair classification approaches,” vol. 13, no. 5, pp. 3579–3589, 2024, doi: 10.11591/eei.v13i5.8032.

[35] M. F. Cahyadi and T. H. Rochadiani, “Implementasi Ensemble Deep Learning Untuk Analisis Sentimen Terhadap Genre Game Mobile,” vol. 8, pp. 1512–1523, 2025, doi: 10.30865/mib.v8i3.7832.

[36] Kusnawi, M. Rahardi, and V. D. Pandiangan, “Sentiment Analysis of Neobank Digital Banking Using Support Vector Machine Algorithm in Indonesia,” *Int. J. INFORMATICS Vis.*, vol. 7, no. June, pp. 377–383, 2023, doi: 10.30630/joiv.7.2.1652.

[37] M. Özel and Ö. Çetinkaya Bozkurt, “Sentiment Analysis on GPT-4 with Comparative Models Using Twitter Data,” *Acta Infologica*, vol. 0, no. 0, pp. 0–0, 2024, doi: 10.26650/acin.1418834.

[38] E. Miranda, V. Gabriella, S. A. Wahyudi, and J. Chai, “Text Classification untuk Menganalisis Sentimen Pendapat Masyarakat Indonesia terhadap Vaksinasi Covid - 19 Text Classification for Analysing Indonesian People ’ s Opinion Sentiment for,” *J. Sist. Inf.*, vol. 12, pp. 438–451, 2023, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[39] O. Bellar, A. Baina, and M. Bellafkih, “Sentiment Analysis of Tweets on Social Issues Using Machine Learning Approach,” *Proc. - 2023 Int. Conf. Digit. Age Technol. Adv. Sustain. Dev. ICDATA 2023*, vol. 9, no. 4, pp. 126–131, 2023, doi: 10.1109/ICDATA58816.2023.00031.

[40] M. F. Mushtaq, M. M. S. Fareed, M. Almutairi, S. Ullah, G. Ahmed, and K. Munir, “Analyses of Public Attention and Sentiments towards Different COVID-19 Vaccines Using Data Mining Techniques,” *Vaccines*, vol. 10, no. 5, 2022, doi: 10.3390/vaccines10050661.

[41] S. Marrapu, W. Senn, and V. Prybutok, “Sentiment Analysis of Twitter Discourse on Omicron Vaccination in the USA Using VADER and BERT,” *J. Data Sci. Intell. Syst.*, vol. 00, no. January, pp. 1–11, 2024, doi: 10.47852/bonviewjdsis42022441.

[42] F. Illia *et al.*, “Sentiment Analysis on PeduliLindungi Application Using TextBlob and VADER Library,” *Proc. Int. Conf. Data Sci. Off. Stat.*, no. 64, pp. 278–288, 2021, doi: 10.34123/icdsos.v2021i1.236.

[43] P. Rajkhowa *et al.*, “Factors Influencing Monkeypox Vaccination : A Cue to Policy Implementation,” *J. Epidemiol. Glob. Health*, vol. 13, no. 2, pp. 226–238, 2023, doi: 10.1007/s44197-023-00100-9.

[44] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, “Pelabelan Otomatis Lexicon Vader dan KlasifikasAsri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. PETIR, 15(2), 264–275. https://,” *Petir*, vol. 15, no. 2, pp. 264–275, 2022.

[45] A. A. Arifiyanti, D. S. Y. Kartika, and C. J. Prawiro, “Using Pre-Trained Models for Sentiment Analysis in Indonesian Tweets,” *Proc. - Int. Conf. Informatics Comput. Sci.*, vol. 2022-Septe, no. February, pp. 78–83, 2022, doi: 10.1109/ICICoS56336.2022.9930599.

[46] J. Maqbool, P. Aggarwal, R. Kaur, A. Mittal, and I. Ali, “ScienceDirect ScienceDirect Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor : A Machine Learning Approach,” *Elsevier*, vol. 218, pp. 1067–1078, 2023.