Vol. 7, No. 1, February 2025, pp. 73~85, p-ISSN 2685-2381, e-ISSN 2715-2626
http://dx.doi.org/10.28989/avitec.v7i1.2806

The Role of VADER and SentiWordNet Labeling in Naïve Bayes Accuracy for Sentiment Analysis of Rice Price Increases

Ihtiar Nur Furqon, Dewi Soyusiawaty*
Department of Informatics, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:
Submitted January 14, 2025
Accepted February 19, 2025
Published February 25, 2025

Keywords:
Sentiment Analysis;
Naïve Bayes;
VADER;
SentiWordNet;
TF-IDF.

ABSTRACT

The surge in rice prices in Indonesia in 2024 is a critical issue affecting social welfare and national food security, particularly amid rising rice imports. This study evaluates public sentiment on Twitter using the Naïve Bayes method and compares the effectiveness of two automated labeling methods, VADER and SentiWordNet, in improving sentiment analysis accuracy. The research is significant due to the limited literature on automated labeling comparisons, especially in food price crises. The methodology includes data collection, preprocessing, translation, sentiment labeling using VADER and SentiWordNet, TF-IDF feature extraction, Naïve Bayes classification, and performance evaluation across different data split ratios: 60% training and 40% testing, 70% training and 30% testing, 80% training and 20% testing, and 90% training and 10% testing. Results show that VADER excels in detecting positive sentiments, achieving 74.42% accuracy at a 90:10 split but struggles with negative sentiment identification, with a highest F1-score of 56.58%. SentiWordNet performs better for positive sentiment detection, reaching 77.86% accuracy and 96.22% recall at an 80:20 split but yields a low F1-score of 32.15% for negative sentiments. In conclusion, VADER is suitable for balanced sentiment detection, while SentiWordNet is more effective for identifying positive sentiments.

Corresponding Author:
Dewi Soyusiawaty,
Department of Informatics, Universitas Ahmad Dahlan, Indonesia,
Ringroad South Street, Kragilan, Tamanan, Banguntapan District, Bantul Regency,
Special Region of Yogyakarta, 55191, Indonesia.
Email: *dewi.soyusiawaty@tif.uad.ac.id

1. INTRODUCTION

According to information from sp2kp.kemendag.go.id [1], there has been an increase in rice prices in Indonesia, as reported by the Ministry of Trade through the Market Monitoring and Basic Needs System (SP2KP). Data indicates a monthly average increase in national rice prices from March 2023 to March 2024. Domestic production has failed to meet rising demand, forcing Indonesia to import significant quantities of rice, making it one of the world's largest rice importers. It is unsurprising that when rice prices rise, residents in various regions are willing to wait for hours to obtain affordable rice through the government's market operation programs.

Twitter is one of the most popular social media platforms, enabling users to interact with others and share photos and videos. As of July 2023, Indonesia ranked 4th among the top ten countries with the highest number of Twitter users, with 25.25 million users, according to databoks.katadata.co.id [2]. More than 190% of social media users complained about recent increases in the prices of basic necessities, according to a study by the Continuum Institute for Development of Economic and Finance (INDEF) [3]. The study, conducted from February 29 to March 4, 2024, revealed that 67,579 social media users engaged in 74,817 conversations about rising food prices, as reported by Wahyu Tri Utomo, a data analyst at Continuum INDEF. Most of these conversations originated from Twitter.

Sentiment analysis is an automated process for identifying attitudes, opinions, and emotions within textual data. It processes text to classify it into categories of positive or negative emotions. The sentiment analysis process includes defining the dataset domain, preprocessing, feature selection, annotation, classification, and evaluation. Techniques such as Support Vector Machine, K-Nearest Neighbor, and Naïve Bayes are commonly employed in sentiment analysis [4][5][6].

Naïve Bayes is one of the most widely used approaches for understanding public opinion. Despite its simplicity, Naïve Bayes is highly effective and accurate in text classification [7]. A study titled Sentiment Analysis of Election Postponement Issues on Twitter Using Naïve Bayes demonstrated that negative sentiment achieved a precision of 98%, recall of 94%, and an F1-score of 99%. Meanwhile, neutral sentiment attained a precision of 100%, recall of 94%, and an F1-score of 96.9%. Positive sentiment recorded a precision of 96.1%, recall of 100%, and an F1-score of 98% [8].

Previous studies, such as Sentiment Analysis and Topic Modeling of Lombok Tourism using Latent Dirichlet Allocation & Naïve Bayes, achieved an accuracy of 92%, precision of 100%, recall of 83.84%, and specificity of 100% using the Naïve Bayes method [9]. Another study revealed that the Naïve Bayes algorithm achieved an accuracy of 88.24% in Sentiment Analysis of the Relocation of the State Capital, whereas the Support Vector Machine (SVM) algorithm achieved only 78.77% [10]. Additional research highlighted the superiority of Naïve Bayes over SVM in Sentiment Analysis of the Impact of the Coronavirus. Results showed that Naïve Bayes attained an accuracy of 81.07%, compared to SVM's 79.96% [7]. Furthermore, a study conducted by Naraswati et al [11] utilized a dataset comprising 10,000 records related to COVID-19 policy management in Indonesia. Sentiments were classified into two categories: positive and negative. The Naïve Bayes method was employed for analysis, yielding an accuracy of 87.34%, sensitivity of 93.43%, and specificity of 71.76%.

Naïve Bayes is chosen for this research due to its proven effectiveness in sentiment analysis, as demonstrated in previous studies. Despite its simplicity, Naïve Bayes consistently delivers high accuracy, precision, and recall, making it a reliable method for text classification. Studies on various topics, including election postponement, tourism sentiment, and COVID-19 policy analysis, have shown that Naïve Bayes outperforms other algorithms such as SVM in many cases. Its ability to handle large datasets, classify sentiments efficiently, and achieve high sensitivity and specificity further supports its selection for this study.

Data labeling in sentiment analysis is a crucial step that involves assigning labels to text to reflect the sentiment it contains. Traditional methods often rely on manual annotation by humans, which is time-consuming and costly, especially when dealing with large volumes of data. To address these challenges, automating the labeling process can be achieved by leveraging lexical resources. These lexicons are dictionaries or databases of words and phrases pre-labeled with sentiments, enabling systems to automatically identify and assign sentiment to text with greater accuracy and efficiency. Data labeling can be performed using four main lexical resources: VADER, AFINN, SentiWordNet, and the Hu Liu Lexicon [12].

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule-based sentiment analysis tool that uses a sentiment lexicon. VADER combines a list of lexical features labeled based on their semantic orientation, whether positive or negative. This tool is highly effective for analyzing social media texts, movie reviews, and product reviews. The main advantage of VADER is its ability not only to determine a positive or negative score but also to measure the intensity of the sentiment. During the analysis process, VADER scans the text for words found in its lexicon and can determine the polarity index using the polarity_scores() function. This function returns metrics for negative, neutral, positive, and compound values for a given sentence [13].

SentiWordNet is a lexical resource based on the WordNet Lexicon. This resource groups words into synsets such as adjectives, nouns, and verbs, and provides numerical scores based on their objectivity, positivity, and negativity. In the sentiment analysis process using SentiWordNet, a repository of lexical words is used to assign sentiment scores. The sentiment score for each word in a text is calculated by comparing its positive and negative values. The overall sentiment score for the text is then computed by summing these individual scores. The text is broken down into individual words, and preset functions are used to calculate the sentiment score of each word. The sentiment of the text is then evaluated to determine whether it is positive, negative, or neutral based on the cumulative sentiment score [14].

This study aims to collect and evaluate public sentiment regarding the rice price increase in 2024 in Indonesia, particularly through the Twitter platform, using the Naïve Bayes method. The novelty of this research lies in its focus on economic issues, specifically rice price increases, which have received limited attention in sentiment analysis studies using Naïve Bayes. Additionally, unlike previous studies that rely on traditional sentiment labeling, this research compares the role of automated data labeling using VADER and SentiWordNet in improving classification accuracy. This is important because literature comparing automated labeling methods in sentiment analysis remains limited. Based on this background, the research questions formulated are how to use Naïve Bayes to assess public sentiment about the rice price increase and how data labeling influences the accuracy of Naïve Bayes. The objective of this study is to apply Naïve Bayes in sentiment analysis related to the rice price increase on Twitter and evaluate the impact of data labeling on the accuracy of the Naïve Bayes model, providing new insights into both sentiment analysis of economic issues and the effectiveness of automated labeling techniques.

2. RESEARCH METHODS

This research was conducted through organized, structured, and systematic stages to ensure that each step produces valid and reliable data and findings. The stages of this research include data collection, data preprocessing, translation, labeling, data splitting, TF-IDF feature extraction, Naïve Bayes classification, and evaluation. These stages are illustrated in Figure 1.

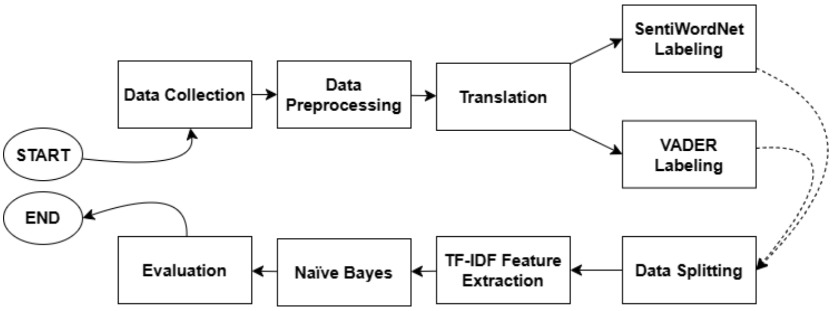


Figure 1. Research stages

2.1 Data Collection

In this study, the data collection method was conducted using data crawling technique. This method was implemented through Python coding and utilizing the tweet-harvest library, which is used to retrieve data from the Twitter API. The purpose of this crawling method is to obtain raw data from relevant tweets, which will later be processed into sentiment analysis products. The data obtained from this process provides insights into public opinion and societal reactions to the rise in rice prices during the specified period [15].

2.2 Data Preprocessing

Data preprocessing aims to improve data quality and facilitate its analysis. The preprocessing stages involve several processes. Data cleaning refers to the removal of irrelevant characters or elements from text data, such as punctuation, numbers, or special symbols. For example, the text “Harga beras naik!!! @123” would be transformed into “Harga beras naik.” Case folding involves converting all letters in the text to lowercase for consistency, such as changing “Harga Beras Naik” to “harga beras naik.” Normalization refers to converting non-standard words or abbreviations into their correct and standard forms according to linguistic rules. For instance, the text “harga beras gk naik yg signifikan” would be normalized to “harga beras tidak naik yang signifikan.” Stopword removal entails eliminating common words that hold little significance in the analysis, such as “dan,” “yang,” and “di.” An example of this process is transforming “Harga beras di pasar naik dengan cepat” into “harga beras pasar naik cepat.” Tokenizing involves breaking the text into smaller units, usually words. For instance, the sentence “Harga beras naik” would be tokenized into [“harga,” “beras,” “naik”]. Lastly, stemming reduces words to their root forms, such as transforming [“menaikkan,” “kenaikan,” “naik”] into [“naik,” “naik,” “naik”].

2.3 Translation

In sentiment analysis based on automated labeling from Indonesian text, translation is utilized to enable access to advanced labeling methods such as VADER and SentiWordNet, which are designed for the English language. By translating the text, the analysis can leverage a more extensive sentiment lexicon, enhancing labeling accuracy and reducing language ambiguity. Translation also ensures consistency in the analysis, especially when models or algorithms are optimized for English [16].

2.4 Labeling

Labeling serves to assign tags or categories to raw data so that it can be utilized in analysis, particularly in machine learning and text analysis. In this study, two types of labeling—negative, positive, and neutral—are used with the automated lexicon labeling methods VADER and SentiWordNet.

VADER Labeling, this is an automation-based labeling method that uses a sentiment lexicon. VADER is implemented using the Sentiment Intensity Analyzer package in Python [17][18][19], where the composite score is divided into three categories: scores ≤ -0.05 are considered negative, scores > -0.05 and < 0.05 are considered neutral, and scores ≥ 0.05 are considered positive [20]. Mathematically, the aggregation calculation for this method is as Equation (1).

$$Compound\ Score = \frac{(sum\ of\ valence\ scores)}{\sqrt{((sum\ of\ valence\ score)^2 + \alpha)}} \tag{1}$$

where α is a scaling constant set for normalization.

SentiWordNet Labeling, this is an automation-based labeling method that relies on lexical resources based on the WordNet Lexicon [21]. SentiWordNet 3.0 uses a semi-supervised learning process involving "seeds" for positive and negative synsets, followed by a classification training process to determine sentiment polarity [22][23]. The polarity results are classified into positive, negative, or objective, with numerical values interpreted into three labels: ≤ -0.05 is considered negative, > -0.05 and < 0.05 is considered neutral, and ≥ 0.05 is considered positive. Mathematically, the algorithm for this method can be written as Equation (2).

$$SC = \frac{1}{N}\sum_{i=1}^{N}(PosScore(i) - NegScore(i)) \quad (2)$$

where SC is the Sentiment Score and N is the number of words in the text.

2.5 Data Splitting

Data splitting divides the dataset into two parts: training data and testing data. By splitting the dataset, model evaluation can be performed using data that has never been seen before (testing data), which tests the model's ability to generalize from the data used in training (training data). An example of this process is splitting the data into 80% for training and 20% for testing.

2.6 TF-IDF Feature Extraction

Feature extraction using TF-IDF (Term Frequency-Inverse Document Frequency) serves as a method to transform text into numerical representations. TF-IDF helps reduce the weight of common words that appear in many documents, while also identifying key and meaningful words unique to each document [24][25]. Mathematically, the calculation of this method is as Equation (3).

$$TFIDF(t, d, D) = TF(t, D) \times IDF(t, D) \quad (3)$$

with t : the word being evaluated
d : the document being evaluated
TF : measures the frequency of a word within a document
IDF : measures the importance of a word

Mathematically, the calculation of the TF and IDF methods is as Equations (4) and (5).

$$TF(t, d) = \frac{n_t}{n_d} \quad (4)$$

with n_t : the number of occurrences of term *t* in document *d*
n_d : the total number of terms in document *d*

$$IDF(t, D) = log\frac{N}{df_t} \quad (5)$$

with df_t : the number of documents containing term *t*

2.7 Naïve Bayes

Naïve Bayes is an algorithm that falls under supervised classification. This algorithm is based on Bayes' Theorem [26], which assumes that the attributes of the data are statistically independent. In sentiment analysis, the combination of TF-IDF and Naïve Bayes allows the system to determine the key words that influence positive, negative, or neutral sentiment. To train the Naïve Bayes model, the TF-IDF representation of the data and training labels are fed into the fit function. Then, the Naïve Bayes model will classify additional data with the help of the training data. Mathematically, the calculation of this method is as Equation (6).

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (6)$$

with: $P(C|X)$: the posterior probability of class *C* given feature *X*
$P(X|C)$: the probability that a document in class *C* will have feature *X*
$P(C)$: the prior probability of class *C*
$P(X)$: the prior probability of feature *X*

2.8 Evaluation

Evaluation serves as a crucial stage after training the model, used to measure and assess the performance of the model in analyzing and predicting data. In this stage, the performance of the VADER and SentiWordNet labeling methods is also compared. During the evaluation phase, various metrics or performance indicators are used to assess how well the model can generate accurate and reliable predictions. Some commonly used evaluation metrics in sentiment analysis and text classification include accuracy score, precision, recall, and F1-Score [27]. These metrics help illustrate the model's success in achieving the analysis objectives, such as identifying positive and negative sentiment from text. Mathematically, these metrics are as follows.

Accuracy, defined as the ratio of correctly classified instances to the total number of predictions made by the model as Equation (7).

$$Accuracy = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (7)$$

Precision, defined as the ratio of correctly classified positive samples to the total number of samples predicted as positive as Equation (8).

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

Recall, defined as the ratio of actual positive occurrences to the total number of actual positive occurrences in the classification as Equation (9).

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

F1-Score, defined as the harmonic mean of Recall and Precision as Equation (10).

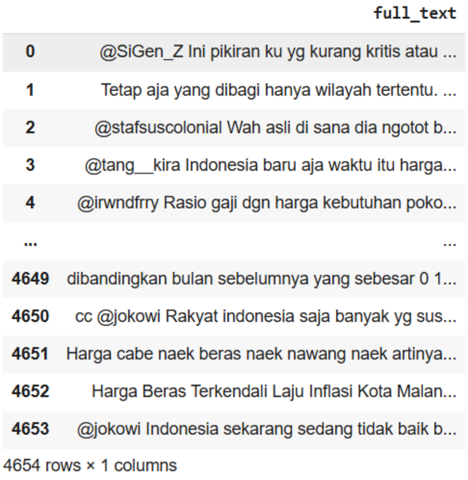
$$F1 - Score = \frac{2(Recall \times Precision)}{Recall + Precision} \tag{10}$$

with TP : True Positive
FP : False Positive
TN : True Negative
FN : False Negative

3. RESULTS AND DISCUSSION

3.1 Data Collection

In this phase, the crawling method is implemented using Python coding and the tweet-harvest library, focusing on collecting data with the Indonesian keyword "kenaikan harga beras 2024" gathered from the period of September 5, 2023, to April 9, 2024. The result of the crawling process is a dataset of 4654 entries, ready for processing in the next phase, as shown in Figure 2.



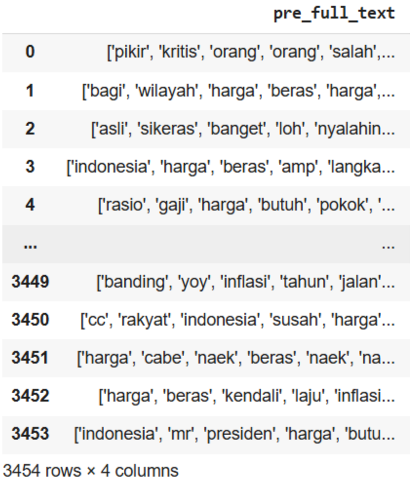
	full_text
0	@SiGen_Z Ini pikiran ku yg kurang kritis atau ...
1	Tetap aja yang dibagi hanya wilayah tertentu. ...
2	@stafsuscolonial Wah asli di sana dia ngotot b...
3	@tang__kira Indonesia baru aja waktu itu harga...
4	@irwndfrry Rasio gaji dgn harga kebutuhan poko...
...	...
4649	dibandingkan bulan sebelumnya yang sebesar 0 1...
4650	cc @jokowi Rakyat indonesia saja banyak yg sus...
4651	Harga cabe naek beras naek nawang naek artinya...
4652	Harga Beras Terkendali Laju Inflasi Kota Malan...
4653	@jokowi Indonesia sekarang sedang tidak baik b...

4654 rows × 1 columns

Figure 2. Data collection

3.2 Data Preprocessing

In this phase, data cleaning is performed on the full_text attribute, which includes the removal of duplicate sentences, normalization of non-standard words, elimination of punctuation and symbols, text consistency by converting it to lowercase, removal of common words (stopwords), conversion of words to their root form (stemming), and text segmentation into tokens (tokenization). As a result, a dataset of 3454 entries is generated, as shown in Figure 3.



	pre_full_text
0	['pikir', 'kritis', 'orang', 'orang', 'salah',...
1	['bagi', 'wilayah', 'harga', 'beras', 'harga',...
2	['asli', 'sikeras', 'banget', 'loh', 'nyalahin...
3	['indonesia', 'harga', 'beras', 'amp', 'langka...
4	['rasio', 'gaji', 'harga', 'butuh', 'pokok', '...
...	...
3449	['banding', 'yoy', 'inflasi', 'tahun', 'jalan'...
3450	['cc', 'rakyat', 'indonesia', 'susah', 'harga'...
3451	['harga', 'cabe', 'naek', 'beras', 'naek', 'na...
3452	['harga', 'beras', 'kendali', 'laju', 'inflasi...
3453	['indonesia', 'mr', 'presiden', 'harga', 'butu...

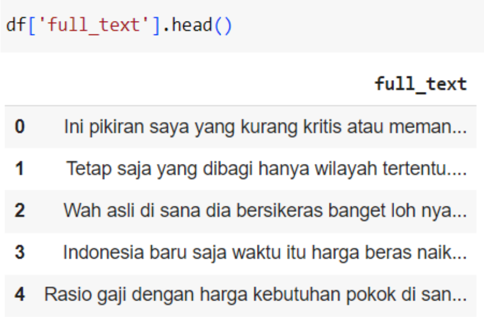
3454 rows × 4 columns

Figure 3. Data preprocessing

3.3 Translation

In this phase, the data is translated from Indonesian to English in the full_text attribute using the Google Translate API. The translated result is stored in a new attribute called text_translate. This process aims to ensure that the data can be further analyzed in the automatic labeling phase.

Figure 4 shows the view of the full_text before translation.

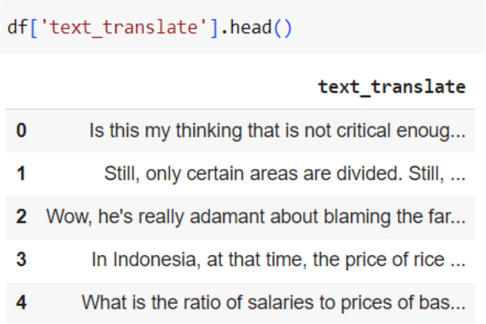


```
df['full_text'].head()
```

	full_text
0	Ini pikiran saya yang kurang kritis atau meman...
1	Tetap saja yang dibagi hanya wilayah tertentu....
2	Wah asli di sana dia bersikeras banget loh nya...
3	Indonesia baru saja waktu itu harga beras naik...
4	Rasio gaji dengan harga kebutuhan pokok di san...

Figure 4. Data full_text before translation

Figure 5 shows the view of the full_text after translation.



```
df['text_translate'].head()
```

	text_translate
0	Is this my thinking that is not critical enoug...
1	Still, only certain areas are divided. Still, ...
2	Wow, he's really adamant about blaming the far...
3	In Indonesia, at that time, the price of rice ...
4	What is the ratio of salaries to prices of bas...

Figure 5. Data full_text after translation

3.4 Labeling

In this phase, automatic labeling is performed using the text_translate attribute by utilizing NLTK (Natural Language Toolkit) with the VADER and SentiWordNet lexicons.

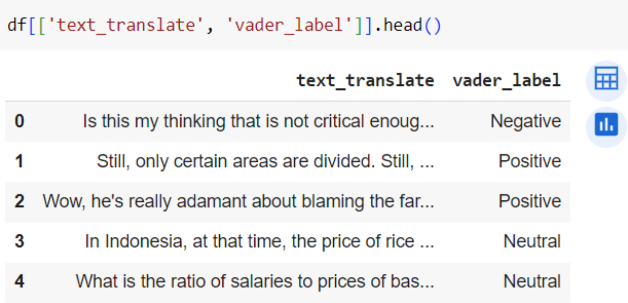
3.4.1 VADER Labeling

This VADER labeling process results in the total sentiment, as shown in Table 1.

Table 1. Total Sentiment for VADER Labeling

Sentiment	Total
Positive	1628
Negative	958
Neutral	868

Figure 6 shows the view of the automatic labeling using the VADER method.



```
df[['text_translate', 'vader_label']].head()
```

	text_translate	vader_label
0	Is this my thinking that is not critical enoug...	Negative
1	Still, only certain areas are divided. Still, ...	Positive
2	Wow, he's really adamant about blaming the far...	Positive
3	In Indonesia, at that time, the price of rice ...	Neutral
4	What is the ratio of salaries to prices of bas...	Neutral

Figure 6. VADER Automatic Labeling

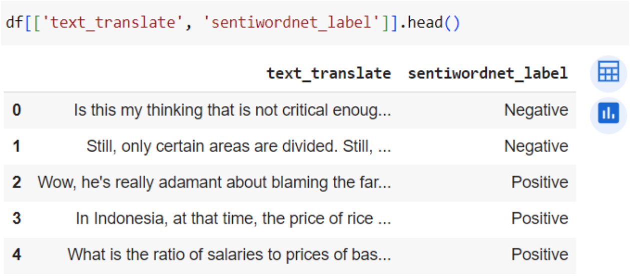
3.4.2 SentiWordNet Labeling

This SentiWordNet labeling process results in the total sentiment, as shown in Table 2.

Table 2. Total sentiment for SentiWordNet labeling

Sentiment	Total
Positive	2464
Negative	919
Neutral	71

Figure 7 shows the view of the automatic labeling using the SentiWordNet method.



```
df[['text_translate', 'sentiwordnet_label']].head()
```

	text_translate	sentiwordnet_label
0	Is this my thinking that is not critical enoug...	Negative
1	Still, only certain areas are divided. Still, ...	Negative
2	Wow, he's really adamant about blaming the far...	Positive
3	In Indonesia, at that time, the price of rice ...	Positive
4	What is the ratio of salaries to prices of bas...	Positive

Figure 7. SentiWordNet Automatic Labeling

3.5 Data Splitting

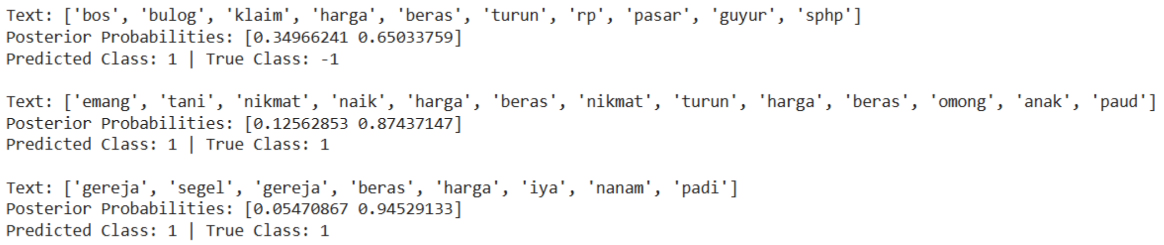
In this phase, the data is divided using negative and positive sentiment. Table 3 shows the comparison of the training and testing data for both the VADER and SentiWordNet methods.

Table 3. Data Split Comparison for VADER and SentiWordNet

No	Data	Split Ratio	VADER	SentiWordNet
1	Train	60%	1552	2030
	Test	40%	1034	1353
2	Train	70%	1810	2368
	Test	30%	776	1015
3	Train	80%	2069	2706
	Test	20%	517	667
4	Train	90%	2327	3045
	Test	10%	259	338

3.6 Naïve Bayes

At this stage, Naive Bayes functions as a probabilistic classification model used to predict text sentiment based on features extracted from the data. The "Text" column contains words or features from the analyzed documents to determine the sentiment class, such as "Positive" or "Negative." The "Posterior Probabilities" column shows the probability of each class calculated by the model, where the model combines the prior class probabilities with the likelihood of words in the text to compute the final probabilities. The "Predicted Class" column represents the prediction results of Naive Bayes, which is the class with the highest probability, while the "True Class" column indicates the original labels of the data used for evaluation. By comparing the Predicted Class with the True Class, we can measure the model's accuracy and understand how the model interprets textual data, as illustrated in Figure 8.



```
Text: ['bos', 'bulog', 'klaim', 'harga', 'beras', 'turun', 'rp', 'pasar', 'guyur', 'sphp']
Posterior Probabilities: [0.34966241 0.65033759]
Predicted Class: 1 | True Class: -1

Text: ['emang', 'tani', 'nikmat', 'naik', 'harga', 'beras', 'nikmat', 'turun', 'harga', 'beras', 'omong', 'anak', 'paud']
Posterior Probabilities: [0.12562853 0.87437147]
Predicted Class: 1 | True Class: 1

Text: ['gereja', 'segel', 'gereja', 'beras', 'harga', 'iya', 'nanam', 'padi']
Posterior Probabilities: [0.05470867 0.94529133]
Predicted Class: 1 | True Class: 1
```

Figure 8. Detailed Naive Bayes computation

3.7 Testing and Evaluation

In this phase, testing is conducted using Feature Extraction, Naïve Bayes, and evaluating the results from various data split comparisons for the VADER and SentiWordNet labeling methods.

3.7.1 VADER Labeling

The accuracy results of the VADER labeling based on the comparison of training and test data splits are shown in Table 4.

Table 4. Results of VADER testing and evaluation

Split Ratio		Precision	Recall	F1-Score	Accuracy
60:40	Negative	0.6360	0.4511	0.5278	0.7117
	Positive	0.7373	0.8565	0.7925	
70:30	Negative	0.6300	0.4406	0.5185	0.6973
	Positive	0.7208	0.8480	0.7792	
80:20	Negative	0.6765	0.4742	0.5576	0.7165
	Positive	0.7309	0.8629	0.7914	
90:10	Negative	0.7414	0.4574	0.5658	0.7442
	Positive	0.7450	0.9085	0.8187	

Based on the VADER labeling results with various data splits (60:40, 70:30, 80:20, and 90:10), there is a noticeable trend of improved model performance on positive sentiment. However, the model's performance on negative sentiment is still unsatisfactory. For negative sentiment, although precision continuously increases with the larger training data portion, ranging from 0.6360 (60:40) to 0.7414 (90:10), recall remains low, between 0.4406 and 0.4742. This indicates that the model is accurate in predicting correctly labeled negative samples, but struggles to capture all the negative samples, resulting in lower F1-Score values.

On the other hand, for positive sentiment, the model shows much better performance. Precision remains high across all splits, ranging from 0.7208 (70:30) to 0.7450 (90:10), while recall is also very high, especially at the 90:10 split with a value of 0.9085. This results in consistently high F1-Scores, indicating that the model is more effective at detecting and predicting positive sentiment compared to negative sentiment. Overall, accuracy also increases from 0.7117 (60:40) to 0.7442 (90:10), suggesting that using more training data helps improve the model's performance.

3.7.2 SentiWordNet Labeling

The accuracy results of the SentiWordNet labeling based on the comparison of training and test data splits are shown in Table 5.

Table 5. Results of SentiWordNet testing and evaluation

Split Ratio		Precision	Recall	F1-Score	Accuracy
60:40	Negative	0.6043	0.2380	0.3215	0.7593
	Positive	0.7771	0.9446	0.8527	
70:30	Negative	0.6526	0.2340	0.3444	0.7663
	Positive	0.7781	0.9557	0.8578	
80:20	Negative	0.6833	0.2398	0.3550	0.7786
	Positive	0.7879	0.9622	0.8664	
90:10	Negative	0.7059	0.2667	0.3871	0.7745
	Positive	0.7822	0.9595	0.8618	

The results of the SentiWordNet labeling with various data splits (60:40, 70:30, 80:20, and 90:10) show significantly different performance between the prediction of negative and positive sentiment. For negative sentiment, although precision gradually increases as the proportion of training data increases, from 0.6043 (60:40) to 0.7059 (90:10), recall remains very low, ranging from 0.2340 to 0.2667. This low recall indicates that the model is unable to capture most of the negative samples, resulting in a low F1-Score for negative sentiment, with the highest value reaching only 0.3871 at the 90:10 split.

On the other hand, for positive sentiment, the model demonstrates very good performance. Precision remains high across all splits, from 0.7771 (60:40) to 0.7879 (80:20), with recall also being very high, ranging from 0.9446 to 0.9622. The F1-Score for positive sentiment is also very good, with the highest value of 0.8664 at the 80:20 split. This indicates that the model is highly effective in detecting and predicting positive sentiment, with very few prediction errors. In terms of accuracy, the model's overall performance shows a slight increase as the amount of training data increases, from 0.7593 (60:40) to 0.7786 (80:20). However, at the 90:10 split, the accuracy slightly drops to 0.7745, which may be due to the imbalance in performance between negative and positive sentiment.

To facilitate the interpretation of the testing and evaluation results, visualizations are presented in the form of charts figure 9 for precision, the precision results obtained from the sentiment analysis comparison reveal significant insights into the performance of VADER and SentiWordNet across different data splits. For VADER, the precision of negative sentiment classification ranged from 0.636 to 0.7414, showing a steady improvement as the training data proportion increased from 60% to 90%. Similarly, VADER's positive sentiment precision varied from 0.7373 to 0.745, indicating relatively consistent performance with minor improvements as the training set grew larger. In contrast, SentiWordNet displayed a more pronounced progression in the precision of negative sentiment classification, increasing from 0.6043 at a 60:40 split to 0.7059 at a 90:10 split. For positive sentiment, SentiWordNet consistently achieved higher precision compared to VADER, with values ranging between 0.7771 and 0.7879, showcasing a stable yet slightly increasing trend across all data splits. These results suggest that while both methods exhibit improvements with larger training data proportions, SentiWordNet consistently outperforms VADER in positive sentiment classification, whereas VADER demonstrates comparable or superior performance in negative sentiment classification depending on the data distribution.

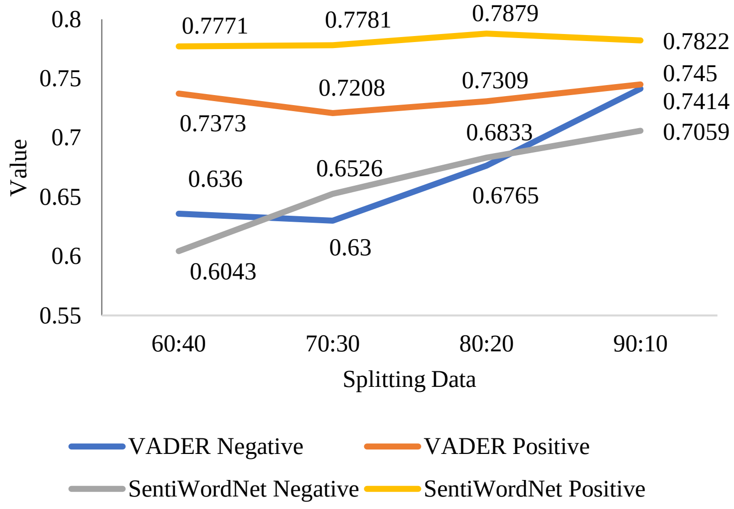


Figure 9. Comparison of precision

Figure 10 for recall, the recall analysis highlights notable differences in the performance of VADER and SentiWordNet across various data splits. For VADER, the recall of negative sentiment classification fluctuated between 0.4406 and 0.4742, peaking at an 80:20 split but slightly declining at the 90:10 split to 0.4574. Conversely, VADER's positive sentiment recall demonstrated a consistent upward trend, increasing from 0.8565 at a 60:40 split to a high of 0.9085 at a 90:10 split, reflecting its strong ability to identify positive sentiment as the training data size grew. SentiWordNet, on the other hand, exhibited considerably lower recall for negative sentiment, ranging from 0.238 at a 60:40 split to 0.2667 at a 90:10 split, indicating limited capability in recognizing negative sentiments regardless of data distribution. However, for positive sentiment, SentiWordNet achieved exceptionally high recall, starting at 0.9446 and reaching up to 0.9622, with only slight variations across the different splits.These findings suggest that VADER is more effective in capturing negative sentiments, albeit with moderate recall values, while SentiWordNet demonstrates superior performance in identifying positive sentiments, achieving consistently high recall across all data splits.

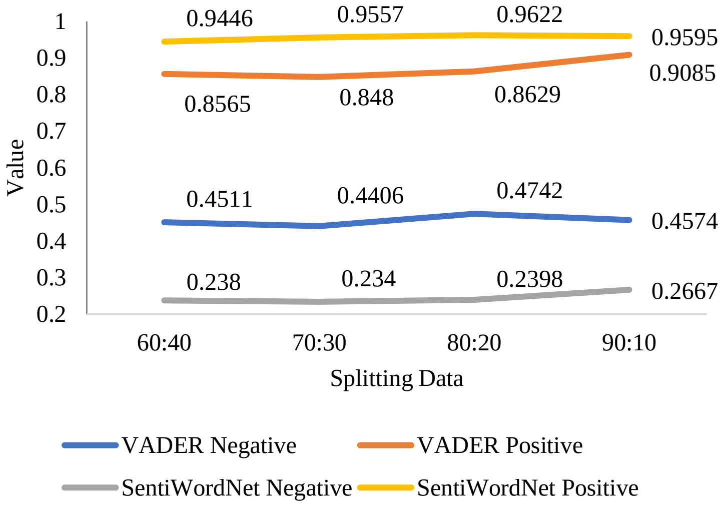


Figure 10. Comparison of recall

Figure 11 for F1-Score, the F1-Score analysis reveals key differences in the performance of VADER and SentiWordNet for sentiment classification across various data splits. For VADER, the F1-Score for negative sentiment classification improved consistently as the training data proportion increased, starting at 0.5278 for a 60:40 split and reaching 0.5658 for a 90:10 split. Similarly, for positive sentiment classification, VADER achieved relatively stable performance, with F1-Scores ranging from 0.7792 to 0.8187, showing a gradual improvement as the training data size grew. In contrast, SentiWordNet demonstrated lower F1-Scores for negative sentiment classification, with values increasing modestly from 0.3215 at a 60:40 split to 0.3871 at a 90:10 split. However, SentiWordNet exhibited consistently high F1-Scores for positive sentiment classification, ranging from 0.8527 to 0.8664, with slight variations across the data splits, indicating its robustness in identifying positive sentiments. These results suggest that VADER outperforms SentiWordNet in negative sentiment classification, while SentiWordNet maintains superior performance in positive sentiment classification. Both methods show improved F1-Scores as the proportion of training data increases, but their strengths differ depending on the sentiment category.

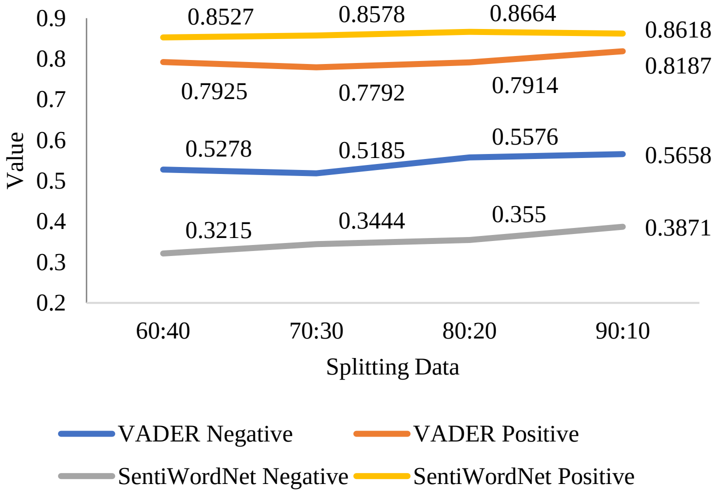


Figure 11. Comparison of F1-Score

Figure 12 for accuracy, the accuracy results highlight the overall effectiveness of VADER and SentiWordNet across different data splits. For VADER, accuracy ranged from 0.6973 at a 70:30 split to 0.7442 at a 90:10 split, indicating steady improvement as the proportion of training data increased. This suggests that VADER's ability to correctly classify sentiments benefits from larger training datasets. SentiWordNet, on the other hand, demonstrated higher accuracy compared to VADER in all data splits. Its accuracy started at 0.7593 for a 60:40 split and peaked at 0.7786 for an 80:20 split, with a slight decline to 0.7745 at a 90:10 split. This reflects SentiWordNet's consistency in sentiment classification performance across various data distributions. Overall, these findings indicate that while both methods improve with larger training datasets, SentiWordNet consistently outperforms VADER in terms of accuracy, particularly in scenarios with balanced or moderately skewed data splits.

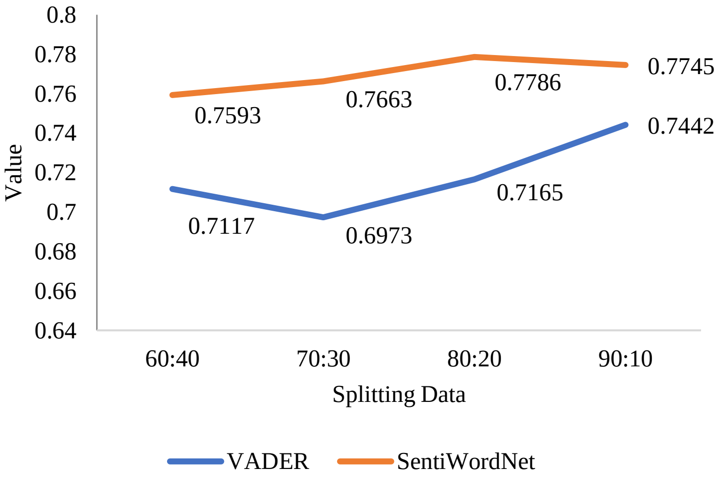


Figure 12. Comparison of Accuracy

The limitations of the methods used in this study are primarily related to the complexity of sentiment classification in Indonesian text and the effectiveness of automated labeling. VADER demonstrates better performance in balanced sentiment detection but struggles to identify negative sentiment accurately. On the other hand, SentiWordNet excels in detecting positive sentiment but has low recall for negative sentiment. These findings highlight the challenge of accurately classifying sentiments in datasets with imbalanced sentiment distributions. Additionally, the translation process from Indonesian to English for sentiment labeling may introduce errors, affecting the overall accuracy of classification. Therefore, while the applied methods are valid for this case, future research should explore improvements in sentiment classification, particularly for negative sentiment detection.

The results of this study align with previous research that highlights the effectiveness of the Naïve Bayes method in sentiment analysis. However, unlike earlier studies that primarily focused on general sentiment classification, this research introduces a comparative evaluation of automated labeling methods VADER and SentiWordNet specifically in the context of economic issues, particularly the rice price surge in Indonesia. Compared to prior research, which demonstrated high accuracy levels of Naïve Bayes in various domains such as 98% precision in election-related sentiment analysis, 92% accuracy in tourism sentiment analysis, and 87.34% accuracy in COVID-19 policy sentiment analysis this study shows that the choice of labeling method significantly impacts classification performance. The findings indicate that VADER achieves a maximum accuracy of 74.42% at a 90:10 data split, excelling in detecting positive sentiments but struggling with negative sentiment classification (F1-score: 56.58%). On the other hand, SentiWordNet attains a higher accuracy of 77.86% at an 80:20 split, with a recall of 96.22% for positive sentiment, but fails to capture negative sentiments effectively (F1-score: 32.15%). These results suggest that while Naïve Bayes remains a strong classifier for sentiment analysis, the effectiveness of sentiment labeling techniques varies depending on the dataset and context. Unlike previous studies where Naïve Bayes alone was evaluated, this research underscores the importance of selecting an appropriate labeling method to optimize classification accuracy, particularly in economic discourse where sentiment expressions may be more nuanced.

4. CONCLUSION

This study compares the performance of VADER and SentiWordNet in sentiment analysis of rice price increases, revealing distinct strengths and limitations. VADER achieves an accuracy of 69.73% to 74.42%, with its highest performance at a 90:10 data split, demonstrating balanced sentiment detection but struggling with negative sentiment classification, where its highest F1-score is only 56.58%. In contrast, SentiWordNet shows higher accuracy (75.93% to 77.86%) and excels in detecting positive sentiment, reaching an F1-score of 86.64% and a recall of 96.22% at an 80:20 ratio, but with a significantly lower recall for negative sentiment (F1-score of 32.15%). These findings highlight the challenges in sentiment classification due to imbalanced sentiment distribution and the limitations of automated labeling. Additionally, translation from Indonesian to English may introduce errors that affect classification accuracy. Future research should focus on improving negative sentiment detection and refining sentiment labeling techniques to enhance overall classification performance.

REFERENCE

[1] Pusat Data dan Sistem Informasi, “Perbandingan Harga Barang Kebutuhan Pokok,” 2025 [Online]. Available: https://sp2kp.kemendag.go.id/

[2] C. M. Annur, “Jumlah Pengguna Twitter Indonesia Duduki Peringkat ke-4 Dunia per Juli 2023,” Katadata Media Network. Accessed: Apr. 27, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/11/01/jumlah-pengguna-twitter-indonesia-duduki-peringkat-ke-4-dunia-per-juli-2023

[3] D. Rachmawati, “Indef: Warganet Mengeluhkan Kenaikan Harga Bahan Pokok di Twitter,” EKONOMI BISNIS.COM. Accessed: Apr. 27, 2024. [Online]. Available: https://ekonomi.bisnis.com/read/20240306/12/1746820/indef-warganet-mengeluhkan-kenaikan-harga-bahan-pokok-di-twitter

[4] A. Rahman *et al.*, “Analisis Perbandingan Algoritma LSTM dan Naive Bayes untuk Analisis Sentimen,” *JEPIN: Jurnal Edukasi dan Penelitian Informatika*, vol. 8, no. 2, 2022. https://doi.org/10.26418/jp.v8i2.54704

[5] F. Sidik, I. Suhada, A. H. Anwar, and F. N. Hasan, “Analisis Sentimen Terhadap Pembelajaran Daring dengan Algoritma Naïve Bayes Classifier,” *Jurnal Linguistik Komputasional* (*JLK*), vol. 5, no. 1, p. 34 2022. http://dx.doi.org/10.26418/jlk.v5i1.79

[6] N. Satya Marga, A. Rahman Isnain, and D. Alita, “Sentimen analisis tentang kebijakan pemerintah terhadap kasus corona menggunakan metode Naive Bayes,” *Jurnal Informatika dan Rekayasa Perangkat Lunak* (*JATIKA*), vol. 2, no. 4, pp. 453–463, 2021. http://dx.doi.org/10.33365/jatika.v2i4.1602

[7] C. F. Hasri and D. Alita, “Penerapan metode naïve bayes classifier dan support vector machine pada analisis sentimen terhadap dampak virus corona di Twitter,” *Jurnal Informatika dan Rekayasa Perangkat Lunak (JATIKA)*, vol. 3, no. 2, pp. 145–160, 2022. https://doi.org/10.33365/jatika.v3i2.2026

[8] A. Perdana, A. Hermawan, and D. Avianto, “Analisis Sentimen Terhadap Isu Penundaan Pemilu di Twitter Menggunakan Naive Bayes Clasifier,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 11, no. 2, pp. 195–200, 2022. https://doi.org/10.32736/sisfokom.v11i2.1412

[9] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, “Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation,” *Jurnal RESTI* (*Rekayasa Sistem dan Teknologi Informasi*), vol. 5, no. 1, pp. 123–131, 2021. https://doi.org/10.29207/resti.v5i1.2587

[10] F. Zamachsari, GV. Saragih, Susafa'ati, W. Gata, “Analisis Sentimen Pemindahan Ibu Kota Negara dengan Feature Selection Algoritma Naive Bayes dan Support Vector Machine,” *JURNAL RESTI* (*Rekayasa Sistem dan Teknologi Informasi*), vol. 1, no. 3, pp. 504–512, 2020. https://doi.org/10.29207/resti.v4i3.1942

[11] N. Putu Gita Naraswati, D. Cindy Rosmilda, D. Desinta, F. Khairi, R. Damaiyanti, and R. Nooraeni, “Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification,” *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 1, 2021. https://doi.org/10.32520/stmsi.v10i1.1179

[12] A. H. Pratama and M. Hayaty, “Performance of Lexical Resource and Manual Labeling on Long Short-Term Memory Model for Text Classification,” *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika* (*JITEKI*), vol. 9, no. 1, pp. 74–84, 2023. https://doi.org/10.26555/jiteki.v9i1.25375

[13] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, “Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile,” *PETIR*, vol. 15, no. 2, pp. 264–275, 2022. https://doi.org/10.33322/petir.v15i2.1733

[14] P. C. Sridevi and T. Velmurugan, “Twitter Sentiment Analysis of COVID-19 Vaccination Integrating SenticNet-7 and SentiWordNet-Adjusted VADER Models,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 16, no. 2, 2024. [Online]. Available: https://cspub-ijcisim.org/index.php/ijcisim/article/view/630

[15] H. F. Rachman and Imamah, “Pendekatan Data Science untuk Mengukur Empati Masyarakat terhadap Pandemi Menggunakan Analisis Sentimen dan Seleksi Fitur,” *JEPIN* (*Jurnal Edukasi dan Penelitian Informatika*), vol. 8, no. 3, 2022. https://doi.org/10.26418/jp.v8i3.56655

[16] Nopan and E. Mailoa, “Perbandingan Beberapa Algoritma Machine Learning Dalam Analisis Sentimen Terkait Pemilihan Presiden RI 2024,” *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 13, no. 2, 2024. http://dx.doi.org/10.35889/jutisi.v13i2.1980

[17] R. Chandrasekaran, R. Desai, H. Shah, V. Kumar, and E. Moustakas, “Examining Public Sentiments and Attitudes Toward COVID-19 Vaccination: Infoveillance Study Using Twitter Posts,” *JMIR Infodemiology*, vol. 2, no. 1, 2022. https://doi.org/10.2196/33909

[18] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, “An analysis of COVID-19 vaccine sentiments and opinions on Twitter,” *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, 2021. https://doi.org/10.1016/J.IJID.2021.05.059

[19] K. Mohamed Ridhwan and C. A. Hargreaves, “Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100021, 2021. https://doi.org/10.1016/J.JJIMEI.2021.100021

[20] 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* (*Basel*), vol. 10, no. 5, 2022. https://doi.org/10.3390/vaccines10050661

[21] V. A. Rao, K. Anuranjana, and R. Mamidi, “A sentiwordnet strategy for curriculum learning in sentiment analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2020, pp. 170–178. https://doi.org/10.1007/978-3-030-51310-8_16

[22] T. A. Tran, J. Duangsuwan, and W. Wettayaprasit, “A new approach for extracting and scoring aspect using SentiWordNet,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1731–1738, 2021. https://doi.org/10.11591/ijeecs.v22i3.pp1731-1738

[23] A. Dadhich and B. Thankachan, “Opinion Classification of Product Reviews Using Naïve Bayes, Logistic Regression and Sentiwordnet: Challenges and Survey,” *IOP Conf Ser Mater Sci Eng*, vol. 1099, no. 1, p. 012071, 2021. https://doi.org/10.1088/1757-899x/1099/1/012071

[24] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa’ah, “Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, 2021. https://doi.org/10.29207/resti.v5i4.3308

[25] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 2, pp. 785–795, 2022. https://doi.org/10.35957/jatisi.v9i2.1835

[26] Y. Astuti, I. R. Wulandarim, A. R. Putra, and N. Khoromadhaona, "Naïve Bayes untuk Prediksi Tingkat Pemahaman Kuliah Online Terhadap Mata Kuliah Algoritma Struktur Data," *JEPIN* (*Jurnal Edukasi dan Penelitian Informatika*), vol. 8, no. 1, 2022. https://doi.org/10.26418/jp.v8i1.48848

[27] A. Rifa'i, Herry. Sujaini, and P. Dian, "Sentiment Analysis Objek Wisata Kalimantan Barat pada Google Maps Menggunakan Metode Naive Bayes," *JEPIN* (*Jurnal Edukasi dan Penelitian Informatika*), vol. 7, no. 3, 2021. https://doi.org/10.26418/jp.v7i3.48132