SENTIMEN ANALYSIS ON THE NEW VARIANT OF COVID-19 (OMICRON) IN INDONESIA USING BERT TEXT REPRESENTATION

DAMAR LAZUARDI S PUTRA[1] , TUGA MAURITSUS[2]

[1,2] Information System Management Department

BINUS Graduate Program – Master of Information System Management, Bina Nusantara University

Jl. Kebun Jeruk Raya No. 27, Kebun Jeruk, Jakarta, Indonesia, 11530

E-mail: [1] damar.putra@binus.ac.id , [2] tmauritsus@binus.edu

ABSTRACT

The number of opinions that appear can give rise to many perceptions, it is difficult to know the tendency of opinions from the many comments, not only positive perceptions but also negative perceptions including opinions regarding the emergence of a new variant of the Coronavirus. The number of new variants ranging from Alpha to Omicron has resulted in a decrease to an increase in COVID-19 cases, which requires the government to make various policy strategies. All forms of policy changes and these conditions create uncertainty that makes people feel afraid and worried about uncertain conditions. The purpose of conducting a sentiment analysis is to find out public opinion regarding the new variant of Covid-19 more generally and to determine the level of accuracy obtained using the BERT text representation, the CRISP-DM framework and the Naïve Bayes method. The result of this sentiment analysis is that the perception of the Indonesian people towards the new variant of Covid-19 (omicron) tends to be neutral with a percentage of 27,65 (553), followed by a negative percentage of 7,6% (7,6), and a positive percentage of 64,75% (1295) from 2000 tweet data. From the results of testing the accuracy values obtained by BERT and the Naïve Bayes model regarding the perception of the Indonesian people towards the new variant of Covid-19 (omicron) with a comparison of training data and test data in testing using a confusion matrix with both training data and test data comparisons being 80:20 get 77% accuracy for Naïve Bayes, 82% accuracy for Support Vector Machine, and 90% accuracy for Random Forest.

Keywords: *Omnicron, Bert, Naïve Bayes, Support Vector Machine, Random Forest, Sentiment Analysis.*

1. INTRODUCTION

1.1 Introduction

Coronavirus Disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is a new type of coronavirus that has never been previously identified in humans (e.g [1]).

The development of Covid-19 cases in Indonesia has decreased with a total of 9,018 active cases as of November 14 2021. However, WHO announced that a new variant of the corona, Omicron, had been detected in Africa and Indonesia (e.g [2]).

In Indonesia, new subvariants Omicron BA.4 and BA.5 have been detected starting in early June 2022 which have a low morbidity rate in patients who are confirmed positive (e.g [3]). According to Siti Nadia Tarmizi, the number of patients who confirmed the new omicron variant was 3,161 cases (e.g [4]).

According to the Ministry of Health, the general symptoms of Covid-19 are fever 380C, dry cough and shortness of breath. About 80% of cases recover without needing special treatment. About 1 out of every 6 people will develop pneumonia or difficulty breathing, which usually comes on gradually (e.g [5]).

With the advent of Omicron, there have been many pros and cons responses so that this conversation has become a trending topic on Twitter. This opinion gives many perceptions in the form of positive and negative sentiments (e.g [6]).

An example of a tweet from people's anxiety, namely please, if you are still traveling abroad, it's better not to return to Indonesia first, whose economy is still affected by Covid-19, don't add to it a lot, if you have money, you know yourself, if you can help the affected economy there, no problem.

Therefore, the government made various policy strategies such as limiting the movement of people's activities from home or work from home (WFH), transmitting reduced economic growth (e.g [7]). This has an impact on people feeling afraid and worried about Covid-19 (e.g [8]).

From comments on Twitter with the hashtags #coronavirus, #CoronaOmicron, and others, it is difficult to know the trend of opinion with sentiment analysis. Sentiment analysis is a way of gathering public opinion with social networks which contain public services, current issues (e.g [9]).

In this thesis, positive sentiment analysis is awareness of the Covid-19 variant and negative sentiment analysis is an attitude that lowers one's value (e.g [10]). With sentiment analysis one can find out someone's opinion on related issues (e.g [11]). The results of this analysis can be seen the tendency of Indonesian people towards Omicron.

1.2 Identification of Problems

1. How to do a sentiment analysis that can find out public opinion regarding the new variant of Covid-19 using text mining with BERT as the text representation?

1.3 Research Objectives and Benefits

Based on the background of the problem and the formulation of the problem mentioned earlier, here are the objectives of this research:

1. Conduct a sentiment analysis that can find out public opinion regarding the new variant of Covid-19 more generally which will be classified into positive, negative and neutral sentiments using the BERT text representation, the CRISP-DM framework using the Naïve Bayes, Support Vector Machine and Random Forest methods.

Research benefits from this research:

1. Research results in the form of positive, neutral, and negative sentiments can be used for knowledge as a reference in analyzing public perceptions of the new variant of Covid-19.
2. The results of the study can be used as a reference in further research regarding the perception of the Covid-19 variant.

2. LITERATURE REVIEW

This study uses a systematic literature review methodology based on several references from domestic and international publications. Research questions are used to determine the objectives that must be carried out in this study. Bert Text Representation is the latest text mining method with weighted words that can be read not only from left to right but also from right to left. Bert is also a new method invented by Google. With the new word weighting, which method is suitable for Bert's sentiment analysis? The data analyzed is the opinion of the Indonesian public on Twitter, which discusses the new variant of COVID-19, namely Omicron, where data collection is carried out using the web scraping technique on Twitter, which is a developer method on Twitter. After the weighting of the words was carried out, the researcher conducted a sentiment analysis using the Naive Bayes, SVM, and Random Forest classification algorithms. The output of sentiment analysis is positive, neutral, and negative sentiment, depending on the level of accuracy of the method applied.

2.1 Covid-19

Coronavirus Disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is a new type of coronavirus that has never been previously identified in humans. There are at least two types of coronavirus that are known to cause diseases that can cause severe symptoms such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Common signs and symptoms of COVID-19 infection include symptoms of acute respiratory distress such as fever, cough and shortness of breath. The average incubation period is 5-6 days with the longest incubation period being 14 days. In severe cases of COVID-19 it can cause pneumonia, acute respiratory syndrome, kidney failure, and even death (e.g. [12]).

On December 31, 2019, the WHO China Country Office reported a case of pneumonia of unknown etiology in Wuhan City, Hubei Province, China. On January 7, 2020, China identified the case as a new type of coronavirus. On January 30, 2020 WHO declared the incident a Public Health Emergency of International Concern (PHEIC) and on March 11, 2020, WHO had declared COVID-19 a pandemic (e.g. [12]).

2.2 Text Mining

Text mining is the process of exploring and analyzing large amounts of unstructured text data assisted by software that can identify concepts, patterns, topics, keywords, and other attributes in the data. This is also known as text analysis, although some people draw a distinction between the two terms (e.g. [13]).

The stages of Text Mining are as follows:

1. Case Folding
   Converts all letters in the document to lowercase (lowercase). In this stage, characters other than letters are also removed.
2. Tokenizing
   Cut each word in a sentence or parse it by using a space as a delimiter which will generate a token in the form of a word.
3. Filtering
   Filtering words obtained from the tokenizing process that are considered unimportant or have no meaning in the text mining process which is called a stoplist. Each word obtained from tokenizing will be matched in the stopword dictionary in the database, if the word matches one of the words in the stopword then the word will be removed, while those that do not match will be considered suitable and processed to the next stage.
4. Stemming
   Returns the words obtained from the filtering results to their basic form, removes the prefix and the final suffix (suffix) so that the basic word is obtained.
5. Tagging
   Change the word in the past tense (past tense) into the present tense (future tense).
6. Analyzing
   The relationship between words in the document will be determined by calculating the frequency of terms in the document or more commonly known as the weighting stage.

2.2.1 Sentiment Analysis

Sentiment analysis is the extraction of information from text data sources to detect positive or negative views of an object. Usually applied to identify trends in public opinion on a product or company (e.g. [14]). Sentiment analysis is a process to extract, understand, process data in the form of unstructured text automatically to get sentiment information contained in a sentence or opinion (e.g. [15]).

2.3 Text Representation

Text representation is the stage of converting text data into a representation that is easier to process. One approach to text representation is to use a document matrix or what is commonly called a Document Term Matrix. The rows in the matrix represent the documents used, while the columns in the matrix contain words, phrases or other indexing units in a document that are used to determine the context of the document (terms) (e.g. [16]).

2.3.1 Binary Representation

Binary classifications are often required to have fairness in the sense of being less discriminatory with respect to features considered sensitive, e.g. races (e.g. [17]). In the field of information extraction and retrieval, binary classification is the process of classifying a given document/account based on a predefined class. Sockpuppet detection is based on binary, where a given account is detected to be either sockpuppet or non-sockpuppet. Sockpuppets have become a significant problem, where someone can have a false identity for certain purposes or malicious use. Text categorization is also done by binary classification. This study synthesizes binary classification which discusses various approaches to binary classification (e.g. [18]).

2.3.2 TF-IDF

According to R. Mahendrajaya, G. A. Buntoro and M. B. Setyawan (e.g. [19]), Term Frequency Inverse Document Frequency or commonly referred to as TF-IDF is an algorithm used to measure the weight of each word in a document or even a set of documents. document, the weight will represent the importance of a word in the document, the greater the weight value, the more important the role of the word in forming a document. The TF-IDF approach presents text with table spaces where each feature in the text corresponds to a single word. TF (Term Frequency) will calculate the frequency of occurrence of a word and compare the number of all words in the document, following equation 16 is used to calculate TF (Witten & Frank, 2016). The TF-IDF approach presents text with table space in each features in the text correspond to a single word. TF (Term Frequency) will calculate the frequency of occurrence of a word and compare the number of all words in the document, following equation (1) is used to calculate TF (e.g. [19]).

$$t_f(i) = \frac{freq\ (t1)}{\sum freeq\ (t)} \qquad (1)$$

Information:
tf(i) : the Term Frequency value of a word in a document.
Freq (ti) : frequency of occurrence of a word in a document.
$\sum freq\ (t)$: total number of words in the document.

$$tdf(i) = log\frac{|D|}{|\{d:ti\in d\}|} \qquad (2)$$

Information:
idf(i) : the Inverse Document Frequency value of a word (t) throughout the document.
|D| : the total number of documents.
|(d: ti∈ d}| : the number of documents containing the word (t).

2.3.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a large-scale language pre-training model released by Google based on bidirectional transformers. The BERT model aims to train word representation using a two-way converter by adjusting the context on the left and right sides of all layers. Thus, the use of BERT can help prevent ambiguity in a word that results in entity recognition errors (e.g. [20]).

BERT stands for Bidirectional Encoder Representations from Transformers. In Indonesian it means two-way encoder representation. BERT is useful for processing two-way representations in anonymous text by combining the right and left sides of a context in all parts. As a result, by making slight modifications or changes to the processed BERT model, it can produce or solve various existing problems. For example, it provides answers to the questions given and concludes a command/language without adding any other settings. BERT has advantages in terms of practicality (simple) and great observation capabilities. This advantage resulted in BERT being able to understand 11 programming command languages, thereby increasing the GLUE value by 7.7% to 80.5%, MultiNLI increasing by 4.6% to 86.7%, the SQuAD v1.1 value to 93.2 and the SQuAD v2.0 Test F1 value being 83.1 (e.g. [21]).

2.3.4 Word2Vec

Word2vec is a new tool developed by Thomas Mikolov. Word2vec can process words from very large datasets in a relatively short time with better accuracy values compared to previous tools. The way this tool works is by taking the corpustext as input, then producing a vector representation of each word in the text corpus as output. The resulting vector files can be used for research on natural language processing and machine learning applications. The word vector can also be used to measure the proximity between other word vectors. Word2vec has two modeling architectures that can be used to represent word vectors, the architectures are continuous bag-of-word (CBOW) and Skip-gram (e.g. [22]).

2.4 Web Scrapping

Web scraping is the process of retrieving a semi-structured document from the internet, generally in the form of Web pages in a markup language such as HTML (HyperText Markup Language) or TML (Extensible Hyper Text Markup Language), and analyzing the document to retrieve certain data from the page. to be used for other purposes (e.g. [23]).

The purpose of a web scraper is to find certain information and then collect it in a new web. Web scraping focuses on getting data by retrieval and extraction (e.g. [24]).

2.5 Naïve Bayes

Naive Bayes is a probabilistic classification technique based on Bayes' theorem which assumes that there is no relationship between each other between attributes or the presence or absence of certain characteristics in a class that has nothing to do with the characteristics of other classes (e.g. [25]). In addition, Naïve Bayes can also be interpreted as an algorithm that can classify a certain variable using probability and statistical methods. Broadly speaking, the Naïve Bayes algorithm can be explained as equation below (e.g. [26]).

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} \qquad (4)$$

Information:
R: Class unknown data
S: Hypothesis on data R which is a special class
P(R|S): The probability value on the hypothesis R based on the condition S
P(R): The probability value on the hypothesis R
P(S|R): The probability value of S based on the hypothetical condition
R P(S) : S probability value

2.6 Support Vector Machine

Support vector machine (SVM) are a class of linear algorithms that can be used for classification, regression, density estimation, novelty detection, and other applications. In the simplest case of two-class classification, Support Vector Machine has a hyperplane that separates the two data classes by as wide a margin as possible. leads to good generalization accuracy on invisible data, and supports special optimization methods that allow Support Vector Machine to learn from large amounts of data (e.g. [27]).

Classification problems can be translated by trying to find a line (hyperplane) that separates the two classes. In Figure 2.4, it can be seen that the pattern is from two classes, namely +1 and -1. Patterns belonging to class -1 are symbolized by red (squares), while patterns in class +1 are symbolized by yellow (circles). Various alternative lines of separation (discrimination boundaries) are shown in Figure 2.1 (e.g. [28]).

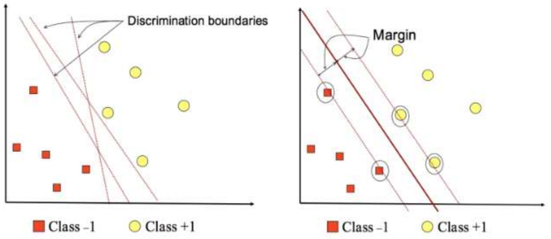


Figure 1. Support Vector Machine (e.g [28])

The Support Vector Machine (SVM) uses a linear model as a decision boundary with the following general form:

$$y\ (x) = w^T \emptyset(x) + b \qquad (5)$$

where x is the input vector, w is the weight parameter, (x) is the basis function, and b is a bias.

2.7 Random Forest

Random forest is a bagging method, which is a method that generates a number of trees from sample data where the creation of one tree during training does not depend on the previous tree then decisions are taken based on the most votes (e.g. [29]).

The RF algorithm has been widely applied in various fields such as predicting the management of social media comments (e.g. [30]). The Random Forest algorithm uses averaging to improve prediction accuracy and control over fitting. The sub-sample size is controlled with the max samples parameter if bootstrap=True(default), otherwise the entire data set is used to construct each tree (e.g. [31]).

When using Random Forest for data classification, the Gini Index formula as shown in the equation is used to decide how the nodes in a decision tree branch (e.g. [32]). This formula uses class and probability to determine the Gini of each branch on a node, determining which branch is more likely to occur.

$$Gini = 1 - \sum_{i=1}^{C}(pi)\ 2 \qquad (6)$$

where pi represents the relative frequency of the observed classes in the data set and C represents the number of classes. In addition to Gini, entropy is also often used in determining how nodes branch in a decision tree. The formula for entropy is in Eq.

$$Entropy = \sum_{i=1}^{C} - pi * \log_2(pi) \qquad (7)$$

Entropy uses outcome probabilities to make decisions about how nodes should branch. Unlike the Gini index, this index is more mathematically intensive because of the logarithmic function used to calculate it.

2.8 Confusion Matrix

Confusion matrix is one method that can be used to measure the performance of a classification method. Basically the confusion matrix contains information that compares the results of the classification carried out by the system with the results of the classification that should be (e.g. [33]). The confusion matrix is depicted by a table that states the number of test data that is correctly classified and the number of test data that is incorrectly classified.

Table 2: Confusion Matrix

		True Value	
		True	*False*
Prediction	*True*	TP *Correct result*	FP *Unexpected Result*
	False	FN *Missing Result*	TN *Corect Absence of result*

Based on the Confusion Matrix table above, it can be explained as follows:

1. True Positives (TP) is the number of positive data records classified as positive values.
2. False Positives (FP) is the number of negative data records classified as positive values.
3. False Negatives (FN) is the number of positive data records classified as positive values.
4. True Negatives (TN) is the number of negative data records classified as negative values.

2.9 CRISP-DM

CRISP-DM (Cross-Industry Starndard Process for Data Mining) which was developed in 1996 by analysis from several industries such as Daimler Chrysler, SPSS and NCR. CRISP-DM provides standardized data mining processes as a general problem solving strategy of a business or research unit. CRISP-DM is a method that can be applied to general problem solving strategies as well as a methodology that provides a standard for data mining (e.g. [34]). The following are the steps in the Cross-Industry Standard Process for Data Mining (CRISP-DM) method:

1. Understanding Business (Business Understanding).
   This stage is the stage of understanding the object of research carried out.
2. Understanding Data (Data Understanding).
   At this stage the researcher aims to collect, identify, and understand the data they have. The data must also be verifiable.
3. Data Collection (Data Collection).
   At this stage the researchers took the data in the form of tweet data.
4. Data Preparation (Data Preparation).
   The data preparation stage can be referred to as the pre-processing stage. This stage is a process to prepare clean data that is ready to be used for research.
5. Modeling (Modeling).
   At this stage a model is made using a classification for the tweet dataset that has been processed through the pre-processing stage.
6. Evaluation.
   At this stage, an evaluation of the classification method will be carried out by measuring performance using a confusion matrix against the algorithm.
7. Conclusion and Suggestions
   At this stage, the conclusions that have been obtained during this research will be given and provide suggestions for further research that will be carried out by other researchers.

3. RESEARCH METHODS

In this study, using the Nave Bayes Method, Support Vector Machine, and Random Forest using BERT Text Representation through the CRISP-DM methodological approach, which is a standard process cross-industry open for data mining, this method is the most used because its application is quite effective and has steps applicable (easy to apply). This research is development research using different word weights, which in this study uses BERT text representation. Through this research, the researcher wants to know which method produces a better value. In addition, researchers wish to show what factors are taken into account when deciding on government policy in light of the Covid incident that was brought on by Twitter social media. The following are the steps of the CRISP-DM methodology.

3.1 Business Understanding

This stage is an understanding stage related to the purpose of conducting research on Identification of Public Perceptions of the New Variant of Covid-19 (Omicron) in Indonesia Using the Naïve Bayes Algorithm and the BERT Method. Following are the steps in Business Understanding.

1. Determination of the objectives of the project and the detailed requirements of the scope of the sentiment analysis system work process.
2. Translating the objectives and scope into a formula for data mining problems.
3. Prepare an initial strategy to achieve the goal.

3.2 Data Understanding

At this stage the researcher aims to collect, identify, and understand the data they have. Following are the steps in Data Understanding.

1. Collecting tweet data contained in social media twitter.
2. Develop data investigation analysis to get to know more about the data and search for the knowledge base.
3. Conducting data evaluation, checking data and cleaning invalid data or data cleansing process.

3.3 Data Collection

At this stage the researchers took tweet data on Twitter social media on Januari 21, 2021 by means of web scraping. The data obtained is a collection of data related to opinions about the new variants in Indonesia obtained using web scraping.

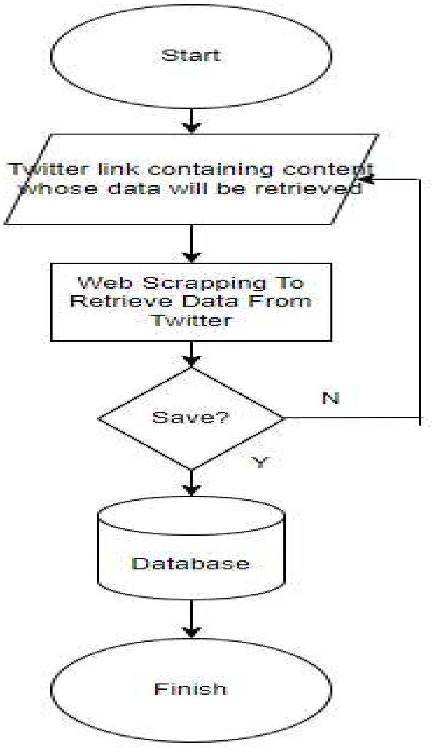

Start
Twitter link containing content whose data will be retrieved
Web Scrapping To Retrieve Data From Twitter
Save?
N
Y
Database
Finish

Figure 2. Flowchart Web Scrapping

3.4 Data Preparation

After inputting the dataset, then the data that has been obtained is given sentiment or labeling, after that the data will enter the data preprocessing stage. In Figure 3 is the flow of the preprocessing process.

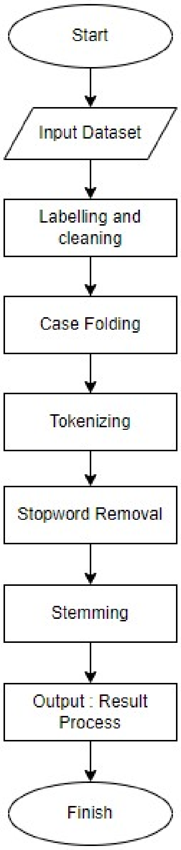


Figure 3. Flowchart Preprocessing

After the data that has been obtained is given a sentiment, after that the data will enter the data preprocessing stage. The following are the steps carried out at the data preprocessing stage as follows:

1. Case Folding, converts all letters in the document to lowercase.
2. Tokenizing, cutting each word in a sentence or parsing it by using a space as a delimiter which will produce a token in the form of a word.
3. Filtering, filtering words obtained from the tokenizing process that are considered unimportant or have no meaning in the text mining process called a stoplist.
4. Stemming, returning the words obtained from the filtering results to their basic form, removing prefix and suffix so that the basic word is obtained.

3.5 Modelling

At this stage a model is made using a classification for the tweet dataset that has been processed through the pre-processing stage. At this stage enter the modeling stage, where the data that has been prepared, cleaned, and labeled will be entered into the model. Following are the steps in Modeling.

1. Selecting a method for sentiment analysis
2. Import library related to sentiment analysis
3. Creating a python program for BERT and sharing datasets. After the dataset goes through the preprocessing process, it is divided into two, namely train data and test data, then it will go through the BERT Tokenizer process using the Transformer & Bert Model. In Figure 4 is the BERT process flow.

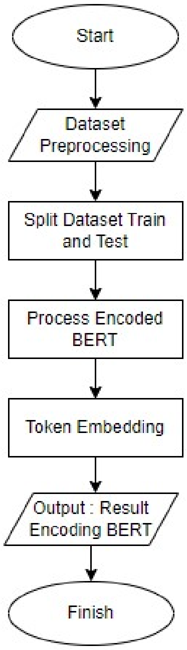


Figure 4. Flowchart BERT

4. Making a python program for nave bayes. After the data goes through the Transformer & Bert Model process which results in BERT encoding. Furthermore, the dataset will be processed with the Multinomial Nave Bayes Algorithm, Support Vection Machine, and Random Forest. The output of this classification results in the form of predictions.

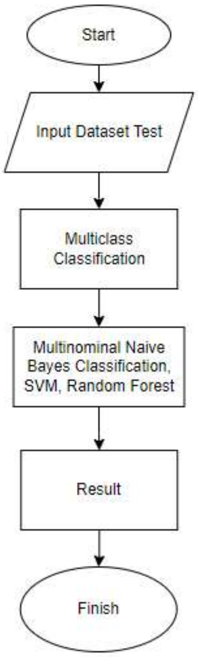

Start
Input Dataset Test
Multiclass Classification
Multinominal Naive Bayes Classification, SVM, Random Forest
Result
Finish

Figure 5. Flowchart Classification

5. Calibrate settings to get maximum results.
6. Can return to the data processing phase if needed.

3.6 Evaluation

At this stage, an evaluation of the classification method will be carried out by measuring performance using a confusion matrix. Following are the steps in Evaluation.

1. Evaluating the results of BERT and all classification that has been carried out will be used as an actual calculation that includes the calculation of the value of accuracy, precision, and F1-Score which is then used as a percentage using a confusion matrix.

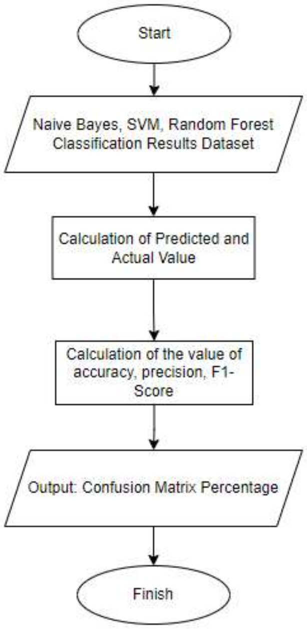

Start
Naive Bayes, SVM, Random Forest Classification Results Dataset
Calculation of Predicted and Actual Value
Calculation of the value of accuracy, precision, F1-Score
Output: Confusion Matrix Percentage
Finish

Figure 6. Flowchart Confusion Matrix

2. Determine whether the model results are in accordance with the initial goal.
3. Determine whether there are important business or research issues that are not being handled properly.
4. If necessary repeat back into the stage of Business Understanding.

3.7 Conclution and Recommendation

At this stage, the conclusions that have been obtained during the research will be carried out. The results of the classification are in the form of sentiment analysis of the classification of the new Covid-19 variant on Twitter, as well as providing suggestions for further research that will be carried out by other researchers.

4. RESULT AND DISCUSSION

4.1 Business Understanding

The Business Understanding stage focuses on understanding the goal, namely conducting experimental research on how to classify sentiments containing the word "omicron" on Twitter social media. Then Twitter data was collected which took 2000 tweets on January 21, 2021 using the Twitter API.

4.2 Data Understanding

The data used in this study were taken from the twitter portal. The data obtained is a collection of data related to the new variant of covid-19 (omicron) obtained by using web scraping.

4.3 Data Collection

The data scraped in this study is tweet data related to the new variant of covid-19 (omicron). The data used are tweets containing the keywords “#omicron”, “omicron di indonesia”. The data obtained is made into an .xlsx file with a total of 2000 tweet scraping data.

4.4 Data Preparation

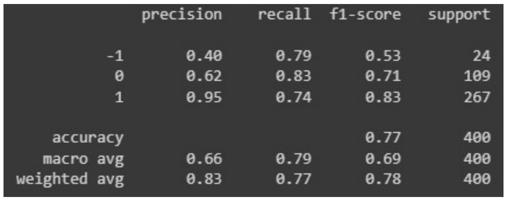
1. Case Folding
   Not all text documents are consistent in using capital letters. Therefore, the role of case folding is needed to convert the entire text in the document into a standard form (lowercase or lowercase). Characters other than letters will be omitted. For example, users who want to get information about “OMNICRON” and type “OmNicRon”, “Omnicron”, or “omnicron” are still given the same retrieval result, namely “omnicron”.
2. Tokenizing
   The tokenizing stage is used to separate the sentences in the string into single word pieces. For example, whitespace characters, such as enter, tabulation, space are considered as word separators. However, for single characters (‘), period (.), semicolon (;), colon (:), or others, it can have quite a lot of role as word separators.
3. Filtering
   At this stage, the disposal of words that are less important or words that often appear (Stopwords), such as connecting words and adverbs that are not unique words, for example "nya", "rt", "pada", and so on.
4. Stemming
   Stemming stage is the process of removing affixes, prefixes, suffixes which aim to change the words according to the basic words.

4.5 Modelling

1. Sentiment Analysis With BERT and Naïve Bayes Classification
   The previously trained data set will perform Bert Tokenizer Using Transformer & Bert Model. The process starts by loading the train data into the BERT model. After that the data is classified with Naive Bayes and produces the following accuracy and confusion matrix.
2. Sentiment Analysis With BERT and Support Vector Machine Classification
   The previously trained data set will perform Bert Tokenizer Using Transformer & Bert Model. The process starts by loading the train data into the BERT model. After that the data is classified with Support Vector Machine and produces the following accuracy and confusion matrix.
3. Sentiment Analysis With BERT and Random Forest Classification
   The previously trained data set will perform Bert Tokenizer Using Transformer & Bert Model. The process starts by loading the train data into the BERT model. After that the data is classified with Random Forest and produces the following accuracy and confusion matrix.

4.6 Evaluation

The Evaluation stage is the stage of the model evaluation report using the Confusion Matrix to measure the performance of the Naïve Bayes classification, Support Vector Machine, Random Forest. Calculations to determine Accuracy, Recall Negative, Recall Positive, Precision Negative, and Precision Positive.



```
              precision    recall  f1-score   support

          -1       0.40      0.79      0.53        24
           0       0.62      0.83      0.71       109
           1       0.95      0.74      0.83       267

    accuracy                           0.77       400
   macro avg       0.66      0.79      0.69       400
weighted avg       0.83      0.77      0.78       400
```

Figure 7. Result Accuracy Naïve Bayes

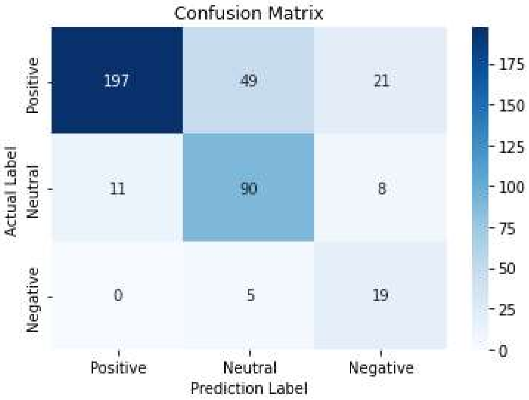
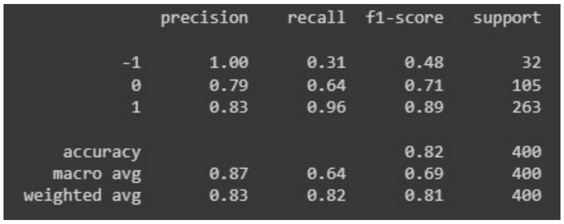


Figure 8. Result Confusion Matrix Naïve Bayes



```
              precision    recall  f1-score   support

          -1       1.00      0.31      0.48        32
           0       0.79      0.64      0.71       105
           1       0.83      0.96      0.89       263

    accuracy                           0.82       400
   macro avg       0.87      0.64      0.69       400
weighted avg       0.83      0.82      0.81       400
```

Figure 9. Result Accuracy Support Vector Machine

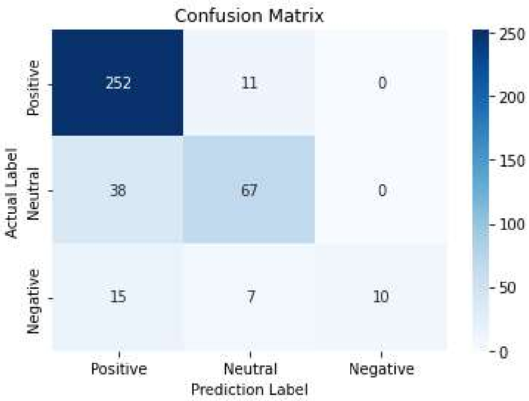
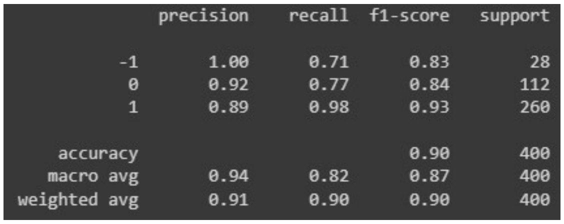


Figure 10. Result Confusion Matrix Support Vector Machine



```
              precision    recall  f1-score   support

          -1       1.00      0.71      0.83        28
           0       0.92      0.77      0.84       112
           1       0.89      0.98      0.93       260

    accuracy                           0.90       400
   macro avg       0.94      0.82      0.87       400
weighted avg       0.91      0.90      0.90       400
```

Figure 11. Result Accuracy Random Forest

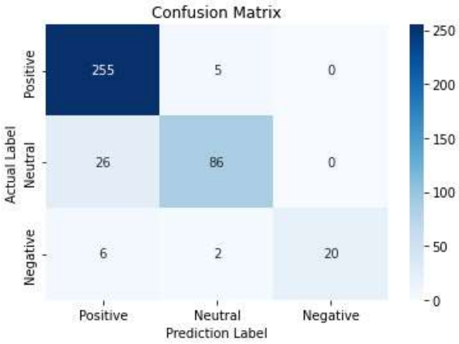


Figure 12. Result Confusion Matrix Random Forest

4.7 Conclution and Recommendation

From the output, the accuracy of various classifications is obtained, namely Naïve Bayes accuracy of 77%, for Support Vector Machine (SVM) accuracy of 82%, and for Random Forest an accuracy of 90%. Of the 2,000 tweet data, 64.75% (1295) received positive sentiment labels, 27.65% (553) neutral sentiment labels, and 7.6% (152) negative sentiment labels from the results of the sentiment analysis. done. From the description, it can be concluded that public sentiment regarding the new variant of Covid-19 (Omicron) tends to be positive.

In the results of the naïve Bayes method accuracy of 77% in sentiment analysis on tweets against the new variant of covid-19 Omicron has 48 tweets of negative sentiment, 19 tweets that are classified correctly and prediction errors on negative reviews 21 tweets, 8 tweets neutral. Meanwhile, there were 208 positive tweets, 197 correctly classified reviews and 11 neutral tweets. This means that tweets on Twitter regarding the new variant of Covid-19 Omicron tend to be positive.

In the results of the Support Vector Machine (SVM) method accuracy of 82% in sentiment analysis on tweets against the new variant of Covid-19 Omicron has 10 tweets of negative sentiment that have been classified correctly. Meanwhile, there are 305 positive tweets, 252 tweets that have been classified correctly and 15 tweets with prediction errors and 38 tweets with neutral reviews. This means that the resulting tweets tend to be positive towards the new variant of Covid-19 Omicron.

In the results of the accuracy of the Random Forest method of 90% in sentiment analysis on tweets against the new variant of Covid-19 Omicron has 20 tweets of negative sentiment that have been classified correctly. Meanwhile, there were 287 positive tweets, 255 correctly classified tweets and 6 tweets of prediction errors and 26 tweets of neutral reviews. This means that the existing tweets tend to be positive about the new variant of Covid-19 Omicron.

5. CONCLUSION

Based on the research that has been done, the following conclusions were obtained:

1. Based on the results of the analysis of public sentiment regarding the perception of the indonesian people towards the new variant of covid-19 (omicron) tends to be positive with a percentage of 64.75%.
2. From the results of testing the accuracy value obtained by bert and naïve bayes, support vector machine, and the random forest model regarding the perceptions of the indonesian people towards the new variant of covid-19 (omicron) with a comparison of training data and test data in the second test comparison of training data and test data is 80:20 get 77% accuracy. For the support vactor machine model, an accuracy of 82%. And for the random forest model, it gets an accuracy of 90%.
3. From the results of the research above, it can be said that sentiment analysis research with text representation is using the random forest method.
4. The data is taken from the Twitter social media, and there are many additional

words, especially those that are popular on Twitter, words or phrases that are often abbreviated, and the word is not standard so that the words can be detected.

5. From the overall results of the sentiment analysis, the sentiment of the indonesian people towards the new variant of covid-19 omicron tends to be positive. From the results of this sentiment, it can be used as input for the government to further educate the public regarding covid-19 in terms of early treatment, symptoms and responses that must be made when experiencing covid-19 symptoms through various advertisements, one of which is social media twitter because many people use it. Those platforms.

REFERENCES:

[1] Republik Indonesia. (2020). Keputusan Menteri Kesehatan Republik Indonesia Nomor Hk.01.07/Menkes/413/2020 Tentang Pedoman Pencegahan Dan Pengendalian Coronavirus Disease 2019 (Covid-19). Jakarta.

[2] Satgas Penanganan Covid-19. (2021, 12 01). Penjelasan Who Tentang Omicron, Varian Baru Covid-19. (Satuan Tugas Penanganan Covid-19) Dipetik 12 21, 2021, Dari Https://Covid19.Go.Id/P/Berita/Penjelasan-Who-Tentang-Omicron-Varian-Baru-Covid-19

[3] Rokom. (2022, Juni 10). Subvarian Baru Omicron Ba.4 Dan Ba.5 Terdeteksi Di Indonesia, Tingkat Kesakitan Rendah. Diambil Kembali Dari Kementerian Kesehatan Republik Indonesia: Https://Sehatnegeriku.Kemkes.Go.Id/Baca/Umum/20220610/2440100/Subvarian-Baru-Omicron-Ba-4-Dan-Ba-5-Terdeteksi-Di-Indonesia-Tingkat-Kesakitan-Rendah/

[4] Sari, H. P. (2022, Februari 3). Kemenkes: 3.161 Kasus Covid-19 Omicron Di Indonesia, 324 Di Antaranya Anak-Anak. Diambil Kembali Dari Kompas.Com: Https://Nasional.Kompas.Com/Read/2022/02/03/08350431/Kemenkes-3161-Kasus-Covid-19-Omicron-Di-Indonesia-324-Di-Antaranya-Anak-Anak

[5] Kemenkes Ri. (2020, Maret). Kementerian Kesehatan Republik Indonesia. Diambil Kembali Dari Pertanyaan Dan Jawaban Terkait Covid-19: Https://Www.Kemkes.Go.Id/Folder/View/Full-Content/Antsmenuheader.Html

[6] Safra, I. A., & Zuliarso, E. (2020). Analisa Sentimen Persepsi Masyarakat Terhadap Pemindahan Ibukota Baru Di Kalimantan Timur Pada Media Sosial Twitter. Proceeding Sendiu 2020, 214-219.

[7] Meirinaldi. (2022). Percepatan Pertumbuhan Ekonomi Sebagai Dampak Teknologi Digital Dan Pandemi Covid 19, Serta Tantangan Dan Peluangnya. Jurnal Ekonomi.

[8] Sholihatunnisa, D., & Desmawati. (2022). Dukungan Sosial Berhubungan Dengan Kesiapan Beradaptasi Dengan Covid-19. Jurnal Keperawatan.

[9] Syarifuddinn, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. Inti Nusa Mandiri 15.1, 23-28.

[10] Ardiani, Sujaini, & Tursina. (2020). Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan Di Kota Pontianak. Justin (Jurnal Sistem Dan Teknologi Informasi), 183-190.

[11] Gunawan, B., Sastypratiwi, H., & Pratama, E. (2018). Sistem Analisis Sentimen Pada Ulasan Produk Menggunakan Metode Naive Bayes. Jepin (Jurnal Edukasi Dan Penelitian Informatika) 4.2, 113-118.

[12] Republik Indonesia. (2020). Keputusan Menteri Kesehatan Republik

Indonesia Nomor Hk.01.07/Menkes/413/2020 Tentang Pedoman Pencegahan Dan Pengendalian Coronavirus Disease 2019 (Covid-19). Jakarta.Ding, W. And Marchionini, G. 1997 A Study On Video Browsing Strategies. Technical Report. University Of Maryland At College Park.

[13] Olhang, M., Achmadi, S., & Wibisono, F. (2020). Analisis Sentimen Pengguna Twitter Terhadap Covid-19 Di Indonesia Menggunakan Metode Naïve Bayes Classifier (Nbc). Jati (Jurnal Mahasiswa Teknik Informatika)Vol. 4 No. 2

[14] Normah. (2019). Naïve Bayes Algorithm For Sentimentanalysis Windows Phonestore Application Reviews. Journal Publications & Informatics Engineering Research, 3(2), 1-19.

[15] Arsi, P., & Waluyo, R. (2021). Analisissentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (Svm). Jurnal Teknologi Informasi Dan Ilmu Komputer (Jtiik), 8(1), 147-156.

[16] Putri, Warsito, & Mustafid. (2019). Implementasi Algoritma Modified Gustafson-Kesseluntuk Clustering Tweets Pada Akun Twitterlazada Indonesia. Jurnal Gaussian, 285 – 295.

[17] Menon, A., & Williamson, R. (2018). The Cost Of Fairness In Binary Classification. *Conference On Fairness, Accountability And Transparency. Pmlr.*

[18] Kumar, R., & Srivastava, S. (2017). Machine Learning: A Review On Binary Classification. *International Journal Of Computer Applications 160.7.*

[19] Witten, I. H., & Frank, E. (2016). Data Mining Practical Machine Learning Tools And Techniques (3rd Ed). Elsevier.

[20] Zahra, A., Hidayatullah, A., & Rani, S. (2021). Kajian Literatur Named Entity Recognition Pada Domain Wisata. Automata 2.1.

[21] Gho, K. (2021). Implementasi Dan Pemodelan Bert Untuk Analisis Sentimen Analisis Aplikasi Gojek Pada Platform Playstore. Diss. Universitas Multimedia Nusantara.

[22] Widyastuti, N., Bijaksana, A., & Sardi, I. (2018). Analisis Word2vec Untuk Perhitungan Kesamaan Semantik Antar Kata. Eproceedings Of Engineering 5.3.

[23] Setiawan, H., Utami, E., & Sudarmawan. (2021). Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine Dan Naive Bayes. Algoritma Support Vector Machine Dan Naive Bayes, 5(1), 43-51.

[24] Ayani, D. D., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi Web Scraping Untuk Pengambilan Data Pada Situs Marketplace. Justin (Jurnal Sistem Dan Teknologi Informasi), 7(4), 257-259.

[25] Kamilah, A. N. (2017). Analisa Sentimen Pelanggan Tokopedia Menggunakan Algoritma Naive Bayes Berdasarkan Review Pelanggan. Simki-Techsain, 01(06), 2-13.

[26] Kurniawan, Y. I. (2018). Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining. Jurnal Teknologi Informasi Dan Ilmu Komputer (Jtiik), 5(4), 455-464.

[27] Sammut, C. And Webb, G. (2010) Encyclopedia Of Machine Learning. Springer

[28] Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Support Vector Machine Dan Aplikasinya Dalam Bioinformatika

[29] Wibowo, A. T., Saikhu, A., & Soelaiman, R. (2016). Implementasi Algoritma Deteksi Spam Yang Tersisipi Informasi Citra Dengan Metode Svm Dan Random Forest.

[30] N. Soonthornphisaj, T. Sira-Aksorn, And P. Suksankawanich, “Social Media Comment Management Using Smote And Random Forest Algorithms,”International Journal Of Networked And Distributed Computing, Vol. 6, No. 4, Pp. 204–209, 2018

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, And B. Thirion, “Scikit-Learn: Machine Learning In Python,”Journal Of Machinelearning Research, Vol. 12, Pp. 2825–2830, 2011

[32] T. P. Pushpavathi, S. Kumari, And N. K. Kubra, “Heart Failure Prediction By Feature Ranking Analysis In Machine Learning,”Proceedingsof The 6th International Conference On Inventive Computation Technologies, Icict 2021, Pp. 915–923, 2021.

[33] Karsito, & Susanti, S. (2019). Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia. Sigma – Jurnal Teknologi Pelita Bangsa, 9(3), 43-48.

[34] Putra, A., & Juanita, S. (2021). Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma Knn. Jatisi (Jurnal Teknik Informatika Dan Sistem Informasi) 8.2, 636-646