
Asilah Salma¹, Widya Silfianti²
¹Master of Information Systems Management, Business Information System, Gunadarma University, Indonesia
²Faculty of Technology and Engineering, Gunadarma University, Depok, West Java, Indonesia-16424
Email address: ¹asilah.isil(at)gmail.com, ²wsilfi(at)staff.gunadarma.ac.id

Abstract—Sentiment analysis is an application of natural language processing to analyze opinion mining or emotion extraction in text mining by identifying the polarity of the text and categorize it into positive, negative, and neutral sentiments. During the COVID-19 pandemic, the government has carried out various programs to break the chain of spread, one of which is by creating an application for providing information and services regarding COVID-19, called Peduli Lindungi. This study aims to analyze the public’s response to the Peduli Lindungi application using the Naive Bayes Classifier, Support Vector Machine and K-Nearest Neighbors. Results for the classification of sentiments based of SVM obtained 76.5%, the accuracy of the NBC is 72.3%, and the accuracy of the KNN is 59.1% and the average sentiments result is Neutral.

Keywords—Algorithm, Classification, COVID-19, K-Nearest Neighbor, Naive Bayes, Sentimental Analysis, Support Vector Machine.

I. INTRODUCTION

Sentiment analysis is an application of natural language processing (NLP) or also known as opinion mining or emotion extraction and is part of text mining. The main purpose of sentiment analysis is to identify the polarity of the data in the form of text and generate information by displaying three categories, namely positive, negative and neutral. Internet users provide sentiment analysis on the internet, one of which is by providing feedback or user ratings of the application in the form of an assessment or evaluation, emotional at the time of writing, so that it can be used as a reference in making decisions for other readers in real time. Sentiment analysis can be done in different ways because in addition to text, it can also analyze sound [1]. The sentiment analysis process requires a better and more sophisticated level of accuracy simultaneously in order to display more accurate results, because sentiment can not only be categorized as positive, negative, and neutral [2].

Corona Virus Disease 19 (COVID-19) is a highly contagious virus that has hit the world, including Indonesia, for almost 2 years. Various efforts have been made by the government to contain the spread of the coronavirus by imposing programs such as large-scale social restrictions (PSBB), the implementation of community activity restrictions (PPKM) levels 1-4. In addition to the program, the government has also created an application in the field of Health that provides information and services for the community during the pandemic, namely the Peduli Lindungi application which can be downloaded on the Playstore and Appstore. This application provides information such as statistics on the spread of the corona virus and the risk of transmission, information and quotas for COVID-19 referral hospitals, and is currently being developed with additional features such as registering for vaccines, QR codes to facilitate screening administration, digital passports to view status or vaccination certificates and COVID-19 test results, travel history, as well as teledoctors to check their own health and also allow direct consultation with general practitioners and specialists, besides that the protective care application has been integrated with a travel tracer application for flights called e-HAC.

Several studies have been conducted on sentiment analysis on reviews taken online, such as in research [3], the classification results using the NBC algorithm with an accuracy of 88.2% and SVM 85.5%. Research [4] resulted in a classification with a comparison of the SVM algorithm having an accuracy of 80.83% and KNN 73.06% by combining particle swarm optimization (PSO). Research [5] resulted in a classification with an accuracy of 67.9% SVM and 60.3% KNN. So it can be concluded that the most used classifier is SVM.

In this study, the data was taken by scraping from the Playstore on user reviews of the Peduli Lindungi application using a Colab notebook integrated with Google Drive with the Python programming language. This study aims to classify user sentiment based on review data on the Peduli Lindungi application into positive, negative, and neutral categories using the NBC, SVM, and KNN classification algorithms. The data collected are 6000 reviews with data collection in September 2021. The results of this study are expected to provide information and find the ideal algorithm used for classification in sentiment analysis.

II. LITERATURE REVIEW

A. Naive Bayes Classifier (NBC)

Naive Bayes Classifier (NBC) is a machine learning algorithm that can predict and calculate the posterior
probability in a class based on the distribution of words in a document. According to [7] NBC is a popular classification algorithm in data mining by combining prior probabilities and conditional probabilities in a formula that can be used to calculate the probability of each possible classification in the training data.

**B. Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a method in supervised learning that is used in the classification process in machine learning. The first SVM was introduced by Vapnik and coworkers in 1992 as a harmonious series of leading concepts in the field of pattern recognition. Currently SVM is growing rapidly because it is included in a machine learning method that can work on the principle of structural risk minimization (SRM) with the main goal of finding the best hyper lane that separates two classes in the input space [6].

**C. K-Nearest Neighbor (KNN)**

K-Nearest Neighbor (KNN) is algorithm supervised learning non-parametric to classify objects or datasets based on learning data that are closer to the object (similarity) and data that have been previously classified or their similarity or proximity to other data. KNN works by taking a number of nearest K data as a reference in determining the class of new data.

### III. RESEARCH METHODS

**A. Research Flow**

This research is an experimental study of sentiment analysis on the review of information and service provider application services during the COVID-19 pandemic using the Naïve Bayes Classifier, Support Vector Machine, and k-Nearest Neighbor classification models. Research flow describes the sequence in research in general in each process of analysis carried out. The stages in this sentiment analysis research start from collecting the dataset by retrieving the data by web scraping and then processing the dataset with text preprocessing. Then the dataset processed with a predetermined classification model to find out the final result so that it is known that the sentiment is divided into three categories, namely positive, negative, and neutral sentiment, as well as knowing the level of accuracy of each classification model. The flow chart in general can be seen in Figure 1.

Figure 1 shows the sequence of research methodologies carried out starting with collecting data by web scraping, reviewing applications from Google Play, preprocessing, feature extraction using CountVectorizer, classification using NBC, SVM and KNN, ending with analysis and evaluation of the results.

**B. Collect Data**

This study using web scraping techniques for data collection by taking data in the form of user reviews of the Peduli Lindungi application on Google Playstore using an interactive Google Collaboratory notebook with the Python programming language. At the initial stage, a scrape review was carried out by installing the google-play-scraper package and literary module, then importing the libraries to be used for data get purposes, namely json to convert data scripts, pandas to create data frames, NumPy for scientific computing on multidimensional arrays and matrices. Then for formatting the results, import highlights, import the natural language toolkit (NLTK) library for data preprocessing purposes in the form of split data, data cleaning, tokenize and filtering, import seaborn to create plots, and import machine learning and evaluation libraries. Scraping review data is taken based on sort, reviews and app, then the data is converted to a data frame and saved to Google Drive. Sampling of user review data is taken based on the most relevant and newest data, to avoid data unbalance, then ratings with a score of 1, 2, 4, 5 will be taken each of 300 and a rating with a score of 3 will be taken 600 each, so the data collected 6000 reviews. The results of scraping data for each review consist of 12 data frame attributes.

**C. Preprocessing**

The data that has been collected is converted into a data frame and saves the data with the name of the review result to Google Drive.

a. Cleaning Data

At the initial preprocessing stage, fill in the blank words (na/na/na) with "No Reply", then check if there are still empty data or not.

b. Tokenizing

Tokenizing is done to separate and break sentences into words in order to distinguish between word separators or not. In this case, the process is punctuation to remove unnecessary symbols and punctuation marks. At this stage, literature is also used as NLTK which has supported the Indonesian language.

c. Filtering

Filtering is done to take important words needed from the tokenizing results, then eliminate stopwords, namely connecting words that often appear and have no meaning in order to reduce index, noise level, and time during data processing. Stopwords are taken from NLTK stopwords in Indonesian by adding a stopword list extend.

At this stage the data frame takes only 2 of the 12 attributes, namely content and score, where content is a review in the form of text and score is a rating in the form of numbers given by the user on a scale of 1-5.

d. Feature Selection

At this stage, data analysis is carried out by determining the attributes that will be used for analysis so that the data is easier to process and understand. In this study, the review data is taken based on sort, reviews and app, then the data is converted (convert) to a data frame. In the data frame, only 2 of the 12 attributes are taken, namely content and score, where content is a review in the form of text, while score is a rating in the form of numbers given by users on a scale of 1-5.
review data samples were taken based on the most recent and most relevant review data.

D. Feature Extraction

After the preprocessing and testing stages are complete, feature extraction is carried out using the CountVectorizer (CV) method. CountVectorizer functions to convert a collection of text documents into a token matrix for each token that appears in the document before generating a vector, so the CV module is very flexible for feature representation in text form.

E. Classification

The classification process begins with preparing the model to be used in the analysis, namely NBC, SVM, and KNN, then the model is run to train sentiment data and make model predictions. After that, divide the training and testing data with the proportion of 80% training data and 20% testing data.

F. Analysis of Results

The analysis phase of the model used is to evaluate the model based on the confusion matrix to display and compare the actual value with the predicted value of each model used, resulting in evaluation metrics in the form of f1-score or f-measure, precision, recall, and accuracy. At this stage, visualizations are also made to display graphs and diagrams of data that have been processed in sentiment analysis which can be described in various types such as pie charts, bar charts, scatter plots and others. In this case, what is displayed is a bar chart for classification as a visual comparison of results and also word cloud.

IV. RESULTS

The data collected from the scraping of the Playstore is as many as 6000 public opinion reviews of the Peduli Lindungi application. At this stage, the results of the analysis that have been carried out will be displayed.

A. Collect Data

Collection and processing is done entirely on Colab notebooks using Python programming language. The results of the collection of Peduli Lindungi application reviews are stored in javascript form with json format and converted to a data frame and then stored on Google Drive.

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>reviewId</td>
<td>String/Object</td>
<td>User Review ID</td>
</tr>
<tr>
<td>2</td>
<td>userName</td>
<td>String/Object</td>
<td>User Name</td>
</tr>
<tr>
<td>3</td>
<td>userImage</td>
<td>Object/Image</td>
<td>User Photo Profile</td>
</tr>
<tr>
<td>4</td>
<td>content</td>
<td>String/Object</td>
<td>Review</td>
</tr>
<tr>
<td>5</td>
<td>score</td>
<td>Integer64</td>
<td>Rating (scale 1-5)</td>
</tr>
<tr>
<td>6</td>
<td>thumbsUpCount</td>
<td>Integer64</td>
<td>Total Likes</td>
</tr>
<tr>
<td>7</td>
<td>reviewCreatedVersion</td>
<td>String/Object</td>
<td>Review Version</td>
</tr>
<tr>
<td>8</td>
<td>at</td>
<td>DateTime64</td>
<td>Review Date</td>
</tr>
<tr>
<td>9</td>
<td>replyContent</td>
<td>String/Object</td>
<td>Reply Content</td>
</tr>
<tr>
<td>10</td>
<td>repliedAt</td>
<td>String/Object</td>
<td>Time Replied</td>
</tr>
<tr>
<td>11</td>
<td>sortOrder</td>
<td>String/Object</td>
<td>Sort by Order</td>
</tr>
<tr>
<td>12</td>
<td>appId</td>
<td>String/Object</td>
<td>Application ID</td>
</tr>
</tbody>
</table>

In table 1 it can be seen that the data frame attributes obtained from scraping are 12 review attributes, but the attributes that will be taken in this study are only two attributes, namely content and score.

B. Preprocessing

Initial data processing used is cleaning, tokenizing, filtering and feature selection as follows.

a. Cleaning data

At this stage, it checks whether there is still empty data or not, after scanning there is no empty data as shown in Figure 2.

b. Tokenizing

The tokenizing process is carried out to break or separate data or sentences into stand-alone words as shown in Figure 3.

c. Filtering

This stage removes the stopwords list in the form of conjunctions/personal pronouns that do not contain important meanings, as shown in table 2.

<table>
<thead>
<tr>
<th>Bahasa Indonesia</th>
<th>Stopwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahasa Indonesia</td>
<td>yg, dg, rt, dgn, ny, d, dpt, klo, kalo, arap, biar, bikin, bilang, kkn, nva, nih, sib, sis, taw, tuk, utk.</td>
</tr>
</tbody>
</table>

The preprocessing stage also uses sastrawi processing library to perform data cleaning, stopword removal and stemming that already supports text in Bahasa Indonesia.

d. Feature selection

At this stage, 2 out of 12 review attribute classes are selected, namely content and score with the results as shown in the following Figure 4.
Visualization plot of the results of image classification in Figure 7 is a comparison of the results of such testing data classification count and in the form of percentage with the proportion of sampling test data as much as 20% in Figure 8.

Fig. 6. Visualization of Comparison Data Training Results (%)

Fig. 7. Visualization of Comparison Data Testing Results (Count)

Fig. 8. Visualization of Comparison Data Testing Results (%)

Visualization of the plot using Word Cloud to see the highlight/trend of words that often appear in reviews given by users can be seen in Figure 9.

Fig. 9. Visualization Total Reviews in Word Cloud

Based on the total reviews in the word cloud above, it can be seen that the words that often appear include: applications, certificates, vaccines, and date of birth and is good as a representation of the reviews given by users of the Peduli Lindungi application.

V. CONCLUSION

There’s a plenty of scope for algorithms in sentiment analysis method. This paper shows the concepts of various classifiers Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor. This concludes that NBC and SVM are the foremost well-known algorithms for sentiment analysis. The sentiment classification obtained on average is a neutral sentiment with a thin comparison in each category. The results of SVM has 76.5% accuracy, followed by NBC 72.3%, and KNN has 59.1% accuracy. However deeper analysis with better accuracy can be expected with KNN using different techniques. There is a need for other methods in conducting sentiment analysis and further classification to improve accuracy results, such as using emoticon analysis and other tools.

REFERENCES