CHAPTER 3
RESEARCH METHODOLOGY

3.1. Literature Study

There are 10 journals used in this study. From all these journals, it is concluded that most of the journals use the nave Bayes algorithm and the rest use the SVM and K-NN algorithms in classifying sentiment analysis. Therefore, in this study, the SVM algorithm is used because it is easy to use for sentiment analysis on Twitter which will be examined.

3.2. Data

The dataset used in this study is a collection of data about vaccines and covid taken from Twitter which was taken using the data crawling method. The data taken is approximately 9000 tweet data. The data taken are datetime, name, and tweet.

3.3. Coding

The programming language used in this research is python. By using the google colab text editor. The dataset used is obtained from the twitter API which is taken using the data crawling method and the results will be used as a CSV file.

3.4. Implementation and Analysis

1. Data Crawling

To be able to retrieve Twitter data, users must have an API Key and a customer key to make it easier for users to retrieve tweet data in real time. The API key can be retrieved on the twitter developer site. To be able to get the API key the user must have a twitter account. Twitter will share the API Key number once certain terms and conditions have been met. After that, the user can start crawling after the API Key is obtained.

2. Pre-processing

Before calculating the data using the algorithm used, the data must be processed first by using the :

Case folding

Converts all letters to lowercase. Characters other than letters a to z will be omitted. Then delete numbers and punctuation that have nothing to do with what will be analyzed.

Tokenization

The process of separating text into chunks of words so that they can then be analyzed.

Filtering Stopwords

the process of taking important words from the tokenization results. Common words that often appear but have no meaning in the data will be deleted and classified as general words.

3. TF-IDF

The data that was previously pre-processed will be calculated using the tf-idf method after which the data will be tested. First, calculate the frequency in the training document. After that the results will show how often the word appears in the document. Then, do this method to test some other data.

4. SVM Algorithm

The classification process will be carried out using the SVM algorithm, which stands for Support Vector Machine. This classification process is carried out using negative and positive data that comes from a review of data that has been taken from Twitter through the crawling method and has been pre-processed, then the data already contains negative and positive comment data that have previously gone through the labeling process, then will be used. SVM algorithm to study data patterns based on the characteristics of the data in each class. After that, the results of this SVM algorithm will be tested using test data, so that after that get the results of the level of accuracy and prediction results.