CHAPTER 3
RESEARCH METHODOLOGY

3.1 Journal

The amount of journal that being used in this research is total of 6 journals. In all 6 of the cited journal can be concluded that almost all the journal is discussing about sentiment analysis using Natural Language Processing, and classifying between positive, negative, and neutral sentiment in their datasets. Therefore this research main topic of discussion is the usage of K-Means in clustering and classifying of Hate Speech.

The amount of journal that being used in this research is total of 6 journals. In all 6 of the cited journal can be concluded that almost all the journal is discussing about sentiment analysis using Natural Language Processing, and classifying between positive, negative, and neutral sentiment in their datasets. Therefore this research main topic of discussion is the usage of K-Means in clustering and classifying of Hate Speech.

3.2 Data

The data that being used here is hate speech dataset which being downloaded from www.kaggle.com/datasets. This dataset contains twitter post, and hate speech indication. This dataset is all in Indonesian Language.

3.3 Program Design

This research is begin from collecting or searching for the dataset. Then the step is to clean the dataset with text pre-processing method.

1. Stopwords

   The first process of text pre-processing is to clean the stopwords inside the dataset. This step is useful because stopwords has almost high frequency compare to the important words inside the data. So it is better to remove the stopwords out from the data. This is to avoid the sudden change topic because of the frequencies of stopwords occurs very often in each sentences.

2. Case Folding

   This process is to convert the dataset into all lowercase. Usually sentences begin with a capital alphabet, but computer will read it different than the lower counterpart. So case folding is best to avoid this.

3. Stemming

The next step is Stemming. Stemming is to simplified words. For example we take the word “Compaaaarree” which the normal version is “Compare”. In the example, there are several “a” and two other alphabet that does not needed. So after stemming, the base word is “Compare”. This is very helping because computer will read “Compare” and “Compaaaarree” as different words when starting the next step. So to avoid miss prediction, the data needed to stem before.

4. N-Gram

This phase is to help preprocess the data to create a better prediction later. N-Gram is to divide the data into set of strings. For example the data in the document is ‘This is the data’, then by using Bi-Gram method of N-Gram, the data will become ‘This is’, ‘is the’. ‘the data’. By using Tri-Gram, the data will be ‘This is the’, ‘is the data’. After this method, TF-IDF will count the N-Gram instead.

5. TF-IDF

After passing the stemming phase and all of the datas are already cleaned, the next step is to use TF-IDF to calculate the frequencies and weight of every documents. First thing to do is to count the TF. Then after that, count the DF. Then adter got the DF, move on to count the IDF. Finally TF-IDF value can be got from TF * IDF. The weight is sum of TF-IDF in a single document. The weight of TF-IDF will be used later for the algorithm.

6. Clustering

After the TF-IDF process done, proceed to the clustering phase. The usage of this step is to cluster whether is a Hate Speech or not a Hate Speech. In this step, the K-Means Algorithm is being used. Before getting the results, the first thing to do is to define the k value. After the k value is being defined, then continue to decide which centroids to use. After the centroids already defined and k has been defined, next step is to calculate each euclidian distance between each documents to each centroid. The centroids are manually defined to differentiate between Hate Speech and Not Hate Speech clusters. This is to make easier observation for later. After the centroids are defined then can continue to the next step for calculating each euclidian distance. For each data, take the lowest value (closest) to one of the centroids to cluster into that centroid. Then the next step is to calculate the new Centroids value. Repeat the methods until the centroids new value is not changing from the previous iteration.

3.4 Coding

The programming language that will be used in this research is python. With the help of Jupyter Notebook text editor. The dataset that will be used in this research is a CSV file.

3.5 Implementation and Analysis

The data that will be used in this research is downloaded from www.kaggle.com/datasets. The dataset is in Indonesian Language. It Contains 12 rows and data attributes to be predict.

3.6 Conclusion and Report Writing

After the data has been processed and calculated, there will be displayed how much is the hate speech posts from overall twitter posts in the dataset in the form of a percentage of accuracy from the grouping dataset.