CHAPTER 3
RESEARCH METHODOLOGY

3.1 Journal and References

There are 8 journals used in this study. The conclusion of all journals is that the use of Naïve Bayes, NLP, F-Scores, and Hybrid Method were proven to provide accurate results, but for clustering using K-Means algorithm it has not been done for this data. Therefore, the topic in this project is to find out the accuracy level of the K-Means algorithm when compared to the Naïve Bayes which has been proven to have good accuracy.

3.2 Data

The data that used in this project obtained from http://kaggle.com/. The data to be used is game review data on the Steam Store [1]. Data comes in CSV format and will be processed using Pandas Data Frame on Python. There are 8 columns with the total data of 434.891 data.

3.3 Design

The main algorithm that will be used in this study is K-Means with the use of Elbow Method and Euclidean Distance. Elbow Method is a method used for determine the number of clusters. In order to determine the number of clusters, Elbow Method will use the WCSS value. Whereas, Euclidean Distance is a method for determining the distance between each data and cluster's centroid.

First, what will be done in order to clustering the data is to determine the value of k. k is the value that will be use as the number of clusters that will be used. The Elbow Method will be used to find the value of k. After the number of clusters is determined, the next step is to determine the point of centroids, at this state this point will be determined randomly.

Next step is to determine distance each data to the centroid point. Original data will be used for the first iteration. In first iteration, data will be grouped automatically according to their respective clusters. Once done with the first iteration, the new data that already has the cluster from the first iteration will be summed up and divided by the amount of data in the cluster. The result

of this mean operation will be the new centroid point. The iteration will continue until the new centroid point have the same value as the previous centroid point.

3.4 Code Program

The programming language used in this study is Python. This CSV formatted data will be imported and processed using the Pandas Data Frame.

3.5 Implementation and Analysis

The data will be used in this study is game review data on Steam from 2010 to 2019, with a total data of 434,891. Before being processed, the data will enter preprocessing stage, at this state it is possible that the amount of data will be reduced.

3.6 Conclusion

The processed data will produce clusters with various sentiment values. From all existing clusters, the data will be analyzed whether there are deviations or not. After that, the result of the analysis will be compared with the result of data analysis using Naïve Bayes algorithm, whether K-Means can provide a better or even worse level of accuracy.