CHAPTER 1
INTRODUCTION

1.1. Background

In the more sophisticated technology era, together with more massive online selling and buying system, there are many widespread informations of item quality and service reviews in all e-commerce platforms. The bulk of various comments can take a long time to identify the sentiment manually. Identification of sentiment based of rating only is not enough to ensure whether the review is positive or negative because of the 3 stars review that not all of them classified as negative or positive which all of them can identify as positive or negative depends on the context that can more likely to. One of the famous e-commerce platforms which used on this project is Amazon. The one of other categories of it focused on cell phones which many people use to buy with.

This problem can be solved by using sentiment analysis program that using polarity values from TextBlob library from Python and mapped each of 1, 2, 3, 4, and 5 overall converted into -1, -0.5, 0, 0.5, and 1 following the polarity value TextBlob range to totaled and averaged which the result determines whether the sentiment of each comments on dataset are positive if it is >0 or negative if it is <=0. If the comment found the only one or more helpful votes, then there are addition of mapped overalls so it can be counted and averaged together with the previous calculation. Then, the words processed into a statistical measure calculation named Term Frequency–Inverse Document Frequency (TF-IDF) which the words on each comments datasets changed into numbers based on the frequency appearance of words. After processed with TF-IDF calculation, it processed to an Support Vector Machine (SVM) algorithm that show cross validation result and measuring the accuracy levels from the processed dataset. Thus there are another prediction dataset used to prove the accuracy based on sentiment identification. Results of this project are compared based of SVM kernels and test and training numbers.

This project shows on accuracy level in a large main dataset using SVM algorithm and analyzing using prediction dataset, the another dataset for sentiment classifying. It used for classify sentiment of each reviews if the sentiment of each comments are positive or negative by detecting the sentiment result of prediction dataset automatically.

1.2. Problem Formulation

The proved questions on this project are :

1. How to get the main dataset and prediction dataset?
2. How to do train and test this datasets using SVM algorithm?
3. How to classify sentiment from prediction dataset after predicting?
4. How to prove the best SVM algorithm kernel and best test size by using main dataset?
5. How to prove the best SVM algorithm kernel and best test size by using prediction dataset?

1.3. Scope

This project will study on how to preprocess the huge Amazon dataset (only phone review), how to labelling the sentiment for each review on it with using TextBlob library, influence on using SVM algorithms kernel and test sizes to evaluate the accuracy on it, and how to predict the review using another dataset (Iphone 11 review from Amazon) automatically in order to detecting sentiment.

1.4. Objective

The purpose on this project is to proving that the result of accuracy levels on SVM algorithm can used as sentiment prediction result on prediction dataset. Besides that, it can used to identify the sentiment from each comments automatically if there are huge review datas that needed to analyze immediately on how bad and good the review of a product.