CHAPTER 3
RESEARCH METHODOLOGY

Scraping.

Data have to be extracted from twitter on the first time. Twitter may ask API and costumer key to scrap its data. API is available on twitter developer app site, log in into the site with twitter account to go through. Twitter will share the API number key after several terms and condition which is applied before. Start scrapping after API number key is given and put hashtag as front title. At least 100 data is contained in one document.

Pre-Processing.

Before calculate data and algorithm used, data have to be proceed to erase symbols, common words, and number. The process should be:

Tokenized

A sentence will be divided words by words using this method. By putting a tweet inside an array, the work re-do until every tweets on document got into each own array. Afterward, the sentence or tweet will divide word by word.

Stopwords

This used for deleting all symbol, number, and words that classified as common words such as I, You, And, Then, etc.

Stemming

This used for reducing inflection in words, so it became base words. For example for word excitement became excite.

Tf-Idf.

Document which had been pre-proceed, calculated with tf-idf method and also the data testing too. First, calculate the Term frequency on training document. The results will show how often a word came up in one document. Next, every words will be weighting with inverse document frequency. Keep the data training, do the same thing with data testing at least 10 tweets.

k-NN Algorithm.

k-NN stands for k-Nearest Neighbour algorithm which means a simple supervised machine learning algorithm used for solving classification problems. The final step leads document into algorithm calculation. Document added by data testing, overall divided by the amount of document then squared by 2. The results is the document standard. If the result is near to minimal is called negative, otherwise if the results is the same as minimal score called true negative. .

.