Assignment Report

Elisabeth Putri - 20306250

Social media has turned into a mainstream media for everybody to share their own opinions. This opportunity is used by companies or media associations to build a web-based social networking collecting each individual consideration about their organization or items.

However, analyzing opinion is not easy because it is a big amount of data. Analyzing people's opinion, such as tweet in Twitter, can be done by classifying the data into two classes. In the tweet case, for instance, classify the data into positive and negative sentiments. After that, that classified data is used for training data to build a training model for further prediction.

There are several basic classifiers to be considered for machine learning usage, such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression[1]. In this assignment, we will use Logistic Regression as the classifier.

Logistic Regression is a discriminative model which learning to distinguish the classes without learning much about each model. For an easy example, logistic regression trains a dog image with blue collar while a cat image doesn't have collar. Once there is another image to test, it will classify the image as a dog whenever the object in that image is wearing collar or classify it as a cat in reverse. Logistic regression is one of a supervised machine learning classifiers (as it trains the data to create a model, then predict it). It extracts the features from the input, creating a weight, then multiplies each input by the weight. It uses probability by sigmoid function. There is also a threshold to be considered as decision.

Logistic regression can classify either two classes or multiple classes. Analyzing part-of-speech can be done using this classifier. In multinomial logistic regression, a well-known title for multiple classes classification by logistic regression, it uses softmax function computing the probabilities of each features. This classifier is one of the most popular analytic tools because it is able to extract and learn the importance of each features transparently.

In this assignment, the given data is classified into two separate files, positive and negative review. The data contains 3,576 negative sentiments and 3,577 positive sentiments. Because the huge amount of data, it is not possible to open each of the data. To tackle this thing, accessing the data is done by using their path directory (i.e.," imdb_datasets"\neg2\"00.txt").

After the data can be accessed, the data is divided into train and test set in ratio 3:1. In this case, each class of data should be mixed in a list either for training data or test data. Each of the document's path should be added an information about its class. Regarding to that, each document is written as [" imdb_datasets"\neg2\"00.txt", 0 or 1]. 0 stands for negative class and 1 stands for positive class. Having this information about the class of each documents makes easier to combine and shuffle the data into further analysis.

Defining the stopwords is helped by nltk library. But, we must avoid the usage of negative stopwords because it can change the sentiment class. The stopwords is benefit to reduce the amount of words to be calculated. It removes all of the less important word in each document. To analyze that, we have to first tokenize each document using the function proc_text. This function is proposed to split the text, get rid all of the stopwords, and return a list of words of each document.

The next function is build_vocab. This function is aiming to create a vocabulary dictionary from all the documents which are used. In this function, we should pay attention of repetition of the word, calculating how many documents contain the word that we are analyzing. After that, the tf-idf function to create the matrix, analyzing the parameter by logistic regression, and checking the metrics.

References

[1] Abhilasha Tyagi, Naresh Sharma, "Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic," *International Journal of Engineering & Technology,* vol. 7, no. 2.24, pp. 20-23, 2018.

[2] Daniel Jurafsky, James H Martin, "Logistic Regresion," in *Speech and Language Processing*, 2020, p. 76.