Building Text classifier using IMDB sentiment Classification Dataset.

Ronak Karmur

Reg No: 2202285

University of Essex

I. Introduction

The sentimental analysis is a computational operation of analyzing the sentiment in the text(information). It is a powerful machine learning application of the classification of text data into different classes. Sentimental analysis is a technique for evaluating reviews, much as how Twitter reviews can be used to analyses public sentiment or how Netflix reviews can be used to choose the best movies and web series to watch. By analyzing user comments, we can assess the internet food. We can use Subjectivity as well as polarity for analysis Whereas, Subjectivity tells us that how much actual information is there in the review and polarity show us the positive & negative of the review.

II. Pre-Processing

Pre-processing is the very important task in EDA (Exploratory Data Analysis). The data we used here is raw data. In raw data we have some special symbols & character, URL, hashtags. Elimination of punctuation, URLs, and numbers. then change all of the text to lowercase. Stopwords will be eliminated from the text when the word "the" is changed to lower case. After basic pre-processing, tokenization will be performed. Moreover, we have use wordcloud. The magnitude of each word in a word cloud, a data visualization approach for expressing text data, shows its frequency or relevance. Using a word cloud, significant textual data points may be displayed. Word clouds are frequently employed for social network data analysis. For instance, using the raw data, we should first identify the frequent terms. After that, we used a word cloud to graphically represent the pleasant and unpleasant evaluations.

Fig 1. Word Cloud Image of Most common words in the datasets.

Fig 2. Word Cloud Image of Most Negative words in the datasets.

Fig 3. Word Cloud Image of Most Positive words in the datasets.

Subsequently, I have added two new columns into the datasets i.e., Subjectivity and Polarity. Subjectivity in Text blob refers to Subjectivity measures how much factual information and value judgement are present in the text. Because of the text's greater subjectivity, personal opinion has taken the place of objective knowledge. Subjectivity lies between [0,1]. The polarity scale is [-1,1], where -1 represents a negative emotion and 1 represents a good emotion. Negative words turn the polarity around.

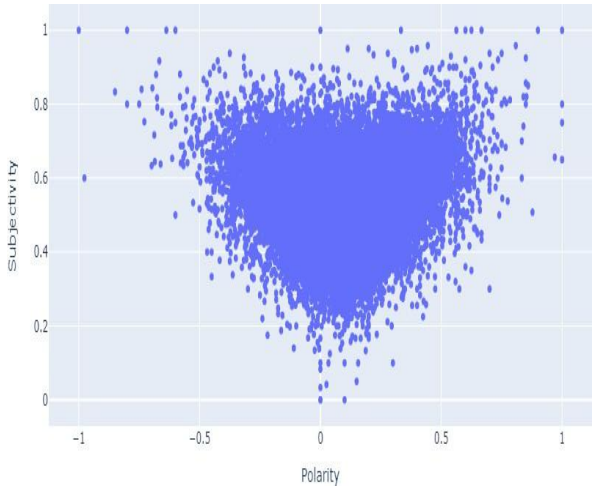

Fig 4. Show's the major sentiments w.r.t. Subjectivity & Polarity.

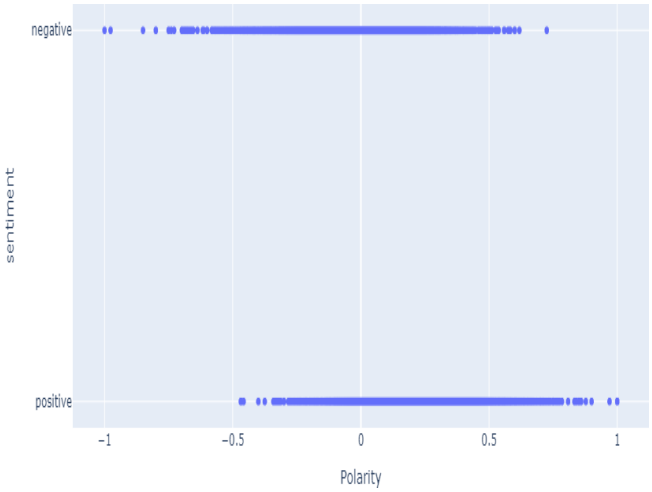

Fig 5. This graph shows the positivity & negativity of Polarity.

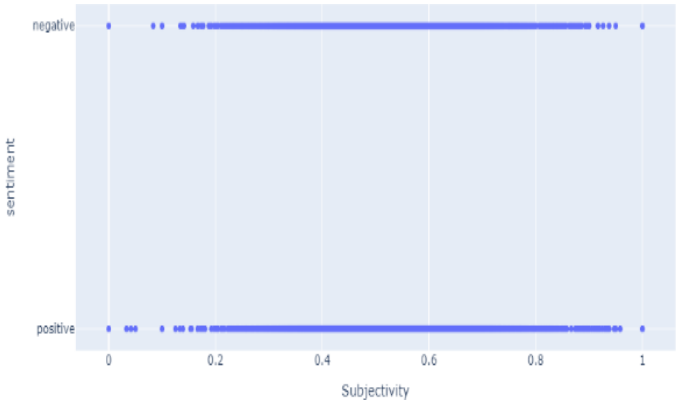

Fig 6. This graph shows the positivity & negativity of Subjectivity.

The raw datasets have now been divided into Training & Testing. I have 40000 rows and 4 columns in my training set, but only 10,000 rows and 4 columns in my testing set. I divided the training datasets into train and test using train-test-split, with a test size of 30%. Following this, we used the Term Frequency-Inverse Document Frequency model to translate the reviews

of the train datasets into numerical form (TFIDF).

III. Algorithms

Text categorisation models come in a wide variety of types, but some of the most typical ones are as follows:

1. Gradient Boosting classifiers: Each prediction in gradient boosting aims to outperform the one before it by lowering the errors. Gradient Boosting's intriguing concept, however, is that it really fits a new predictor to the residual errors created by the preceding predictor, rather than fitting a prediction on the data at each iteration.
2. *Random Forest Classifiers:* The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

3. Logistic regression classifier: A statistical model for binary classification is shown here. To determine the values of the model parameters that minimise the error between the predicted probabilities and the actual class labels of the training data, it is trained using an optimisation procedure.
4. Passive Aggressive classifier: Passive-Aggressive algorithms are somewhat similar to a Perceptron model, in the sense that they do not require a learning rate. However, they do include a regularization parameter.
5. Decision Tree classifiers: Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

IV. Reference

[1] R Raja Subramanian, Nukala Akshith, Gogula Narasimha Murthy et al. " A Survey on Sentiment Analysis".

[2] https://medium.com/@pyashpq56/sentiment-analysis-on-imdb-movie-review-d004f3e470bd.

[3] https://monkeylearn.com/sentiment-analysis/.

[4] https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

MLA Name	Training Accuracy	Testing Accuracy	Testing Precision	Testing Recall	Testing F1-Score
PassiveAggressiveClassifier	0.9979	0.7575	0.764569	0.754271	0.757548
DecisionTreeClassifier	0.9979	0.5177	0.637868	0.113995	0.523675
RandomForestClassifier	0.9979	0.4965	0.979167	0.00772	0.503775
GradientBoostingClassifier	0.5141	0.4928	0.555556	0.001643	0.500145
LogisticRegressionCV	0.5034	0.4927	0	0	0.5