Problem Statement – **Sentiment Analysis of Twitter Data Virgin Airlines**

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. The code given in the repository is implemented in python and is using Naive Bayes Classifier and also it's improved version. The data is not taken directly from Twitter but stored in the CSV format as twitterData.csv

This was built in Python 3.4 and please refer to the pre-requistes required for running it successfully.

Pre-requisites -

1. Python 3.4 (although it shoudn't have problem running on other Python versions apart from some syntax changes like that of print statements.)
2. NLTK for Python
3. Stopwords corpus
4. Wordnet corpus

Algorithm Implemented – Naive-Bayes Algorithm (later used Improved Naive-Bayes Algorithm)

Naive Bayes -

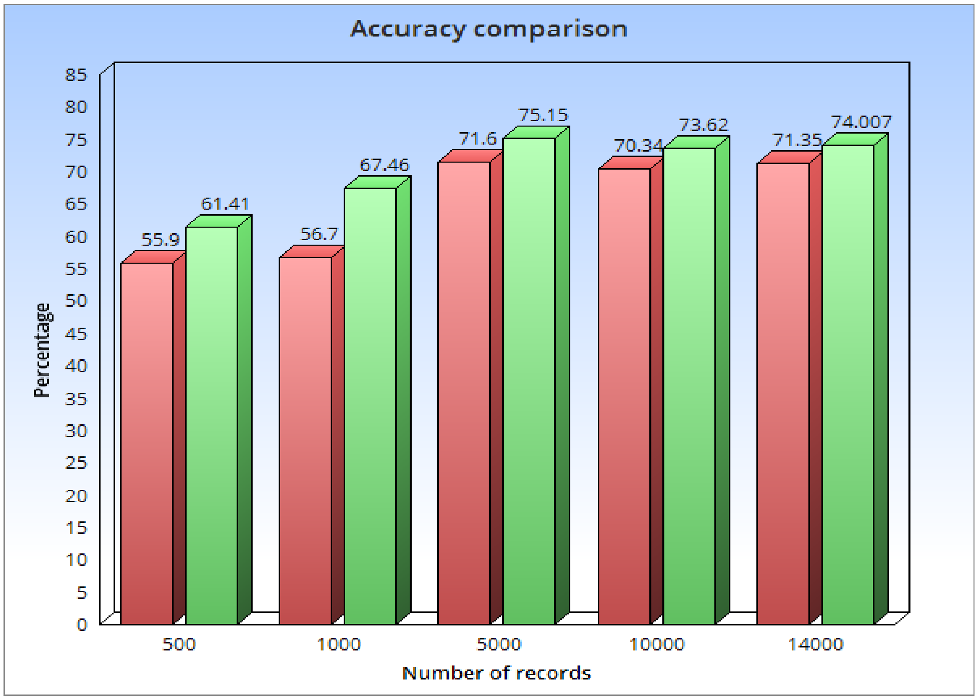
Mathematical formula - $P(cj|d)=P(d|cj)*P(cj)/P(d)$

where, $P(cj|d)$ is the probability of d occuring in class cj

The reason for using an improved version of Naive Bayes because of many false positives and false negatives, i.e., words that are tagged in both positive and negative tags. These words decrease accuracy and also the precision. The improved algorithm takes only the high information words that are unique to that label. The low information words (like: the, and, etc.) which are not unique and occur frequently in both labels are removed. The score of a particular word is found using **BigramAssocMeasures.chi_sq** which takes 4 values -

- **frequency of word for the label**
- **frequency of word for all labels**
- **total frequency of all words for the label**
- **total frequency of all words for all labels**

Accuracy -



Usage -

1. Clone/Download zip of SentAnalysis.
2. Extract content of the zip.
3. Run mainFile.py in IDLE or cmd.