SENTIMENT ANALYSIS FOR MARKETING

PHASE:4

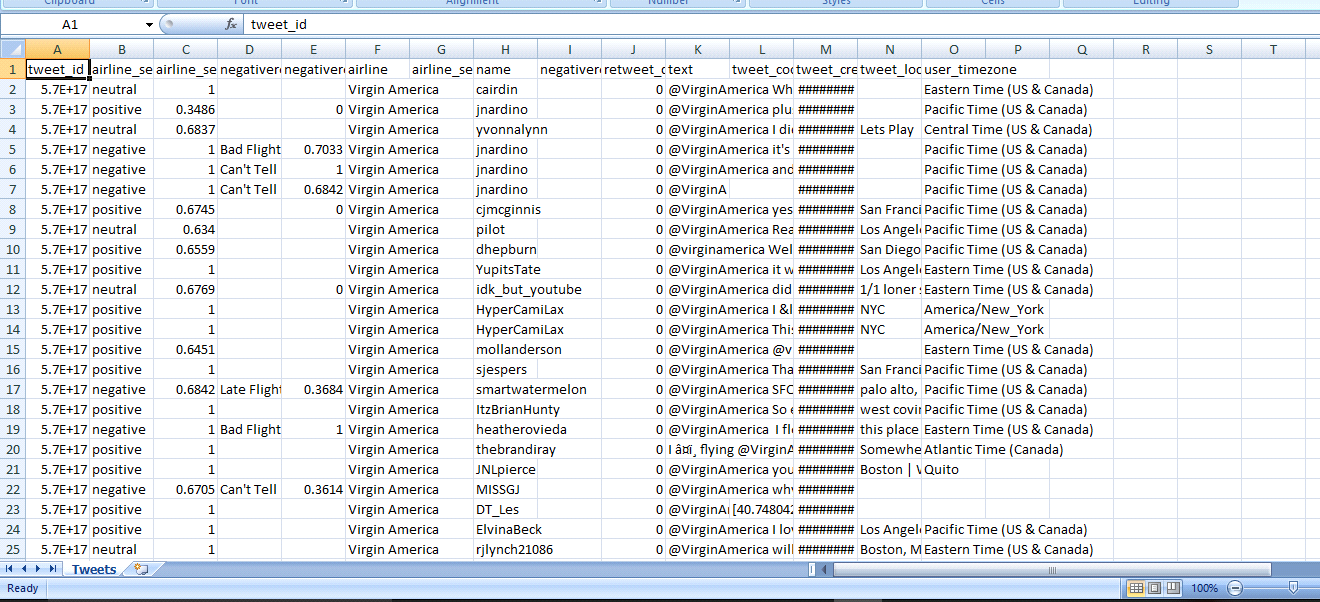
Develpoment :2

**INDROUCTION:**

* Data preprocessing is a crucial step in sentiment analysis, involving transforming raw data into a suitable format for analysis.
* Sentiment analysis is a popular task in natural language processing. The goal of sentiment analysis is to classify the text based on the mood or mentality expressed in the text, which can be positive negative, or neutral.
* Sentiment analysis is the process of classifying whether a block of text is positive, negative, or, neutral. The goal which Sentiment analysis tries to gain is to be analyzed people’s opinions in a way that can help businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.
* Twitter sentiment analysis analyzes the sentiment or emotion of tweets. It uses natural language processing and machine learning algorithms to classify tweets automatically as positive, negative, or neutral based on their content. It can be done for individual tweets or a larger dataset related to a particular topic or event.



**DATA SET:**

****

### Naive Bayes

 this is how Bayes’ theorem works. The probability of A, if B is true, is equal to the probability of B, if A is true, times the probability of A being true, divided by the probability of B being true:

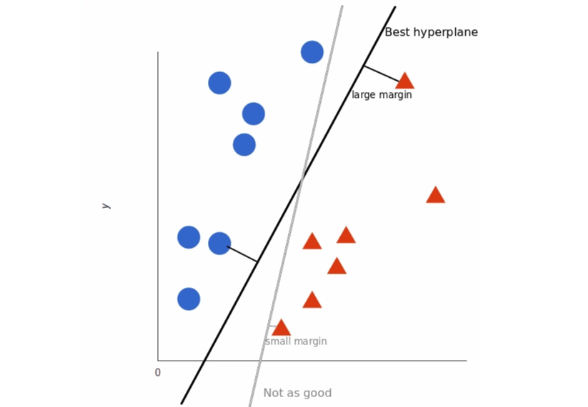
But that’s a lot of math! Basically, Naive Bayes calculates words against each other. So, with machine learning models trained for word polarity, we can calculate the likelihood that a word, phrase, or text is positive or negative.

When techniques like lemmatization, stopword removal, and [TF-IDF](https://monkeylearn.com/blog/what-is-tf-idf/) are implemented, Naive Bayes becomes more and more predictively accurate.

### Linear Regression

### Support Vector Machines (SVM)

A support vector machine is another supervised machine learning model, similar to linear regression but more advanced. SVM uses algorithms to train and classify text within our sentiment polarity model, taking it a step beyond X/Y prediction.



### Deep Learning

Deep learning is a subfield of machine learning that aims to calculate data as the human brain does using “artificial neural networks.”

Deep learning is hierarchical machine learning. In other words, it’s multi-level, and allows a machine to automatically ‘chain’ a number of human-created processes together. By allowing multiple algorithms to be used progressively, while moving from step to step, deep learning is able to solve complex problems in the same way humans do.

**Data Acquisition**

## Get raw data

In the original dataset, I extracted tweets and sentiment only as they are relevant to my semtiment analysis.  
I could have built pandas data frame, but it has no obvious advantage as most of time I will process text and sentiment separately for this specific purpose.

**PROGRAM**

linkcode

raw\_data = pd.read\_csv('/kaggle/input/first-gop-debate-twitter-sentiment/Sentiment.csv', encoding='utf-8')

tweets = raw\_data['text']

labels = raw\_data['sentiment']

print(tweets.head(2))

print(len(tweets),len(labels))

### OUTPUT

0 RT @NancyLeeGrahn: How did everyone feel about...

1 RT @ScottWalker: Didn't catch the full #GOPdeb...

Name: text, dtype: object

13871 13871

## Exploratory Study & Data Visualisation

### Plotting Frequencies of Words

It is common to explore a dataset before starting the analysis, in this section we will find the most common words and plot their frequency.

**PROGRAM**

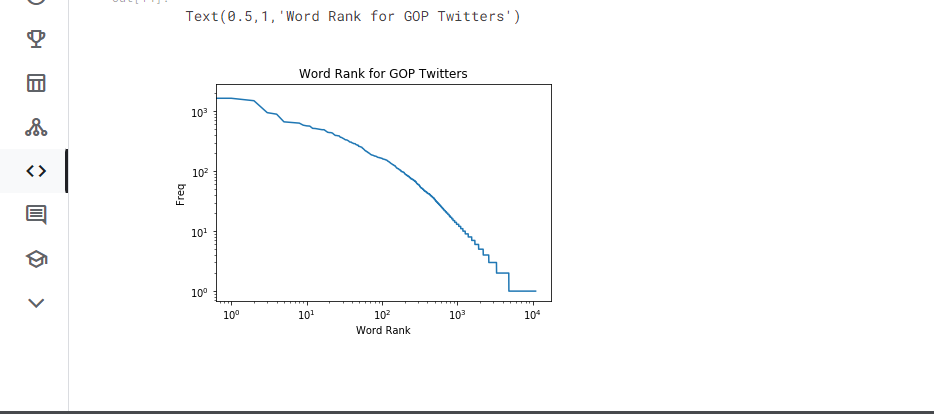
plt.loglog(sorted\_word\_counts)

plt.ylabel("Freq")

plt.xlabel("Word Rank")

plt.title('Word Rank for GOP Twitters')

**OUTPUT**



**PROGRAM**

y\_pos = range(len(label\_count))

*#plt.bar(y\_pos,label\_count,align='center', alpha=.5)*

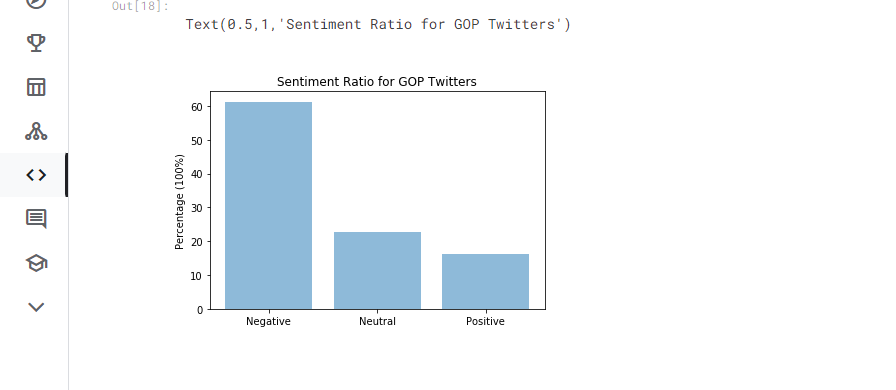
plt.bar(y\_pos,review\_ratio,align='center', alpha=.5)

plt.xticks(y\_pos,label\_count.index)

plt.ylabel('Percentage (100%)')

plt.title('Sentiment Ratio for GOP Twitters')

**OUTPUT**

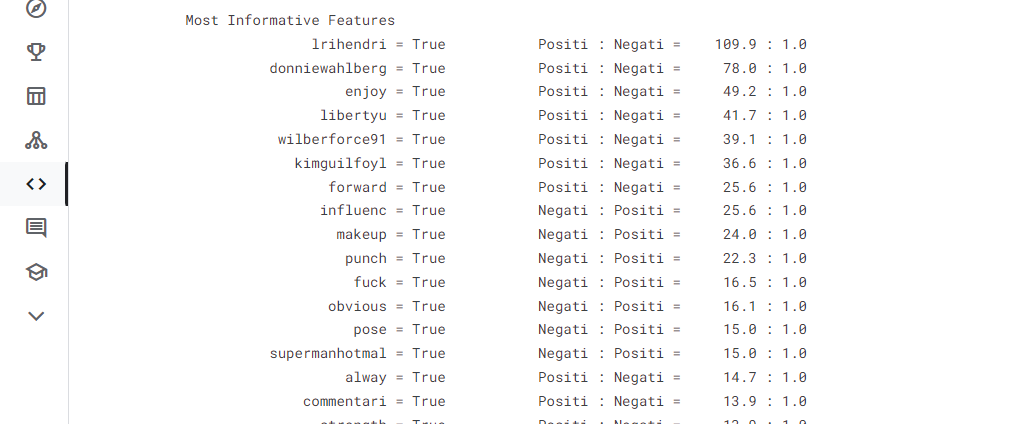
****

Accuracy here is around 81% which is pretty good for such a simple model if we consider that the estimated accuracy for a person is about 80%. We can finally print the most informative features, i.e. the words that mostly identify a positive or a negative review:

**PROGRAM**

sentiment\_classifier.show\_most\_informative\_features(100)

**OUTPUT**

****

### Conclusion

* US Airways, United and Americana are getting the most negative reviews
* Delays and service quality seem to be the biggest concern
* When service has been good, people have gone on twitter to appreciate the same

**We are Implement our AI program by this data set:**

<https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment>