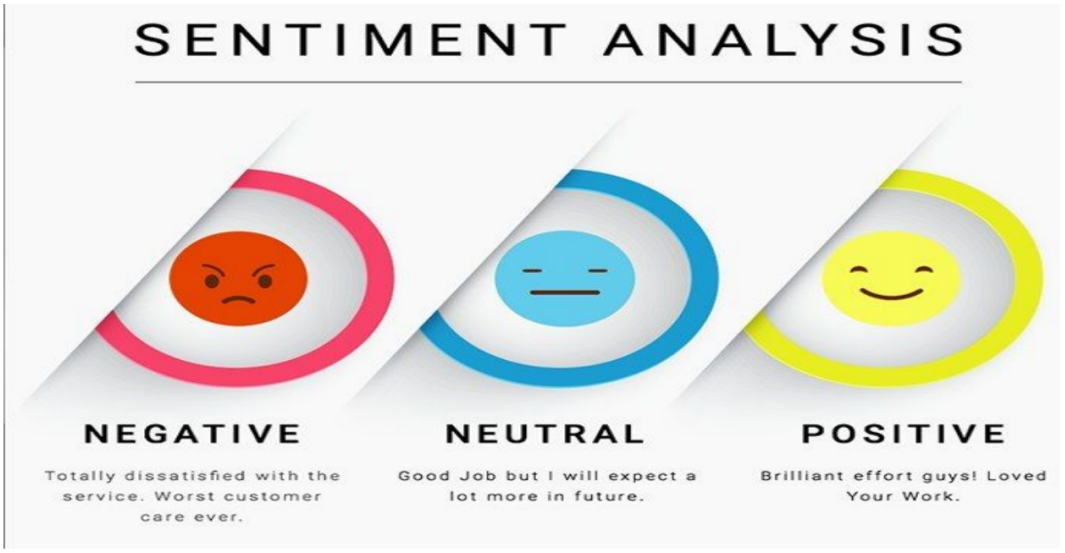
AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

School of Engineering



Submitted by:-
C.Sumanth Kumar Reddy
(AM.EN.U4CSE17019)

1.Problem Definition:

The problem statement which we are dealing with is **Sentiment analysis**. We want to analyze huge volumes of data and want to detect the sentiment of it whether it is positive or negative
TASK(T): To classify if a given text/sentence is positive or negative
EXPERIENCE(E): Corpus files having both positives and negatives
PERFORMANCE(P): Accuracy score. Accuracy is used as a score of performance.

2.Datasets:

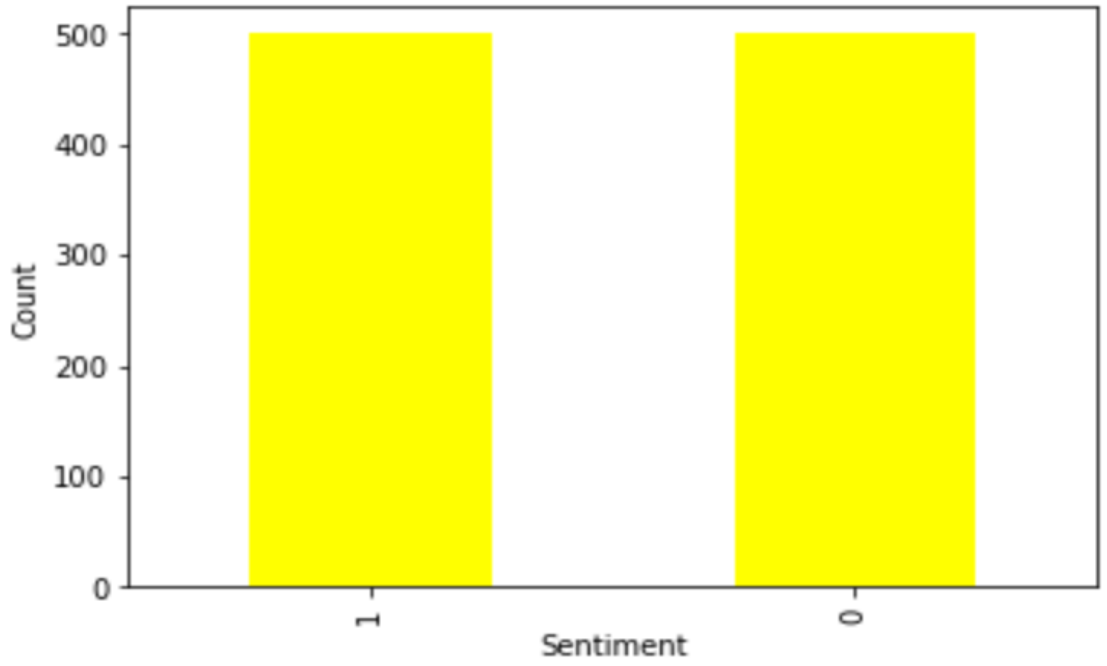
Restaurant Reviews: Dataset having reviews from restaurants. The training set contains reviews as well as their labels, whereas the testing set only reviews.

3.Prepare Data:

- As our input data is text, we used text related preprocessing.
- In the preprocessing step we have done: Removal of stopwords, wild characters, converting uppercase to lowercase letters. Stemming,Tf-IDF/ bag-of-words.
- “Stop words” are commonly used words that are unlikely to have any benefit in natural language processing. So remove them and wild characters.
- TF-IDF is a statistical measure that evaluates how relevant a word is to a document in the collection of documents.
- Removed null values from the dataset.
- Removed duplicates.
- Applied standardization.

Data Visualisation:

Number of positive and negative reviews



Frequency of the reviews