**Machine Learning**

**project submission 2**

**Project Group**

Name: peddi sai sravan

Uta id: 1001719086

Name : Elison Tuscano

Uta Id: 1001738728

**Project Title:**

Sentimental Analysis on reviews using naïve bayes classifier.

**Project Description:**

Data Preparation Process:

1. The data sets provided for the movie review are in two folders. One is the positive review and the other one is related negative reviews
2. Each review in both the folders is in separate text file.
3. The create\_csv.py file is used to take each review text file in both positive and negative folders and combine them into train.csv for training the model and test.csv for testing the model.

A screenshot of a cell phone

Description automatically generated

1. DataPrep.py is another file which is used to preprocess before giving it to the naïve Bayes algorithm.

A screenshot of a social media post

Description automatically generated

1. In the DataPrep.py file we will remove the stop words like (the, this….) which do not give crucial information for recognizing whether the review is positive or negative.

Example -> “I really love this movie “will be converted to “really love movie”.

1. Then we remove the unnecessary special characters such as (#, !, @,…) from the reviews .

Example -> “I love this movie….!!!!” Will be converted to “I love this movie “.

1. We convert the string of reviews into lowercase.
2. Then we do stemming on each review where we bring the words to their base form. Example ->” loved, loving” will be converted to “love”.
3. The reason for doing stemming is to reduce the number of unique words in total number of words that we have. This is done to reduce the size of the sparse matrix which helps to increase efficiency and reduce memory requirement.
4. Finally, we convert the review column into vector with help of CountVectorizer library which converts our single column into multiple columns where each column is used as a feature to train our classification model.