Aishwarya Gouru

880297521

The given tweetdata.csv file contains of almost 1600000 data entries.

First we load the dataset to the system specifying column names.

Then we preprocess the data by removing all the unnecessary text like removing special characters, https: links and punctuations.

Once the data is cleaned by removing unnecessary data, we are ready to develop a training and a test set.

First, the reviews with 0 polarity are termed as negative and are extracted from the main dataset.

Here, from first half dataset, we’ve chosen random 15000 tweets for the training set containing negative values and from next half dataset we’ve randomly chosen 3000 tweets for negative test set. And similarly for the positive ratings data set.

Once we get positive and negative training and test sets, we concatenate both the positive and negative training sets as training data set and similarly for the test set.

Once the training and the test sets are derived, we obtain features for both the data sets. Once, we train the system by using the training data set, we predict the probability of the text.

Hence, by achieving precision, recall and F1 for the dataset.