Naïve Bayes Classification for Sentiment Analysis

In classification tasks, your job is to build a function Y = g(X) that takes in a vector of features X (also called “inputs”) and predicts a label Y (also called the “class” or “output”). Here the words (except stop words) are taken as features. For every word in test data line we calculate its probability of belonging to a particular label. As all the features are independent, we take product of all such features along with probability of the label. As the classification is binary, we compare the two results obtained. Whichever is greater the test data set is said to belong to that class.

Steps

• After reading the input line by line we perform the preprocessing steps listed above.

• We perform 5-fold cross validation.

• After splitting train and test data set, we segregate train data sets according to their classes and also maintain count of words.

• Using Naïve-Bayes we calculate the probability of test data line in each class. Which classify it with the label for which it has maximum value.

• We calculate parameters required to calculate Accuracy and F-Score.

Dataset

https://drive.google.com/file/d/1eFTge09pXFbJ671rafpgkNF0XaG aL01L/view -> a1_d3.txt

Preprocessing

- Converted lines to lower case
- Removed symbols
- Removed stop words

Converting lines to lowercase and removing symbols increased the accuracy but removing the stop words slightly decreased the accuracy. But still it was included in the model as it makes model more meaningful.

Result

Accuracy = 0.814 ± 0.05748043145279966
Precision = 0.8907363135515682 ± 0.08143494964181128
F1 Score = 0.7951490414552304 ± 0.05765598774094459