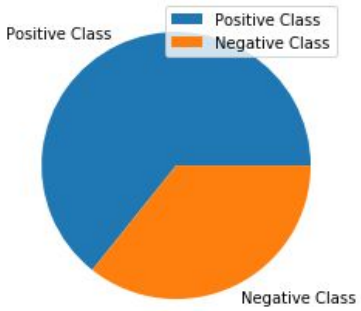
Submitted By: Anish Bhusal
March 18,2020

Naive Bayes using Sentiment Analysis

1. Data Cleaning, Data Analysis and Visualization

The dataset consists of Phrases that are labeled positive or negative based on their sentiment.



Total Phrases	2800
Positive Class Examples	1800
Negative Class Examples	1000

Stopping characters like ‘,’ ‘.’ ‘--’,etc were filtered out from the phrases as part of data cleaning process.

2. Feature Extraction

The dataset was splitted into train and test set in 60:40 ratio. As part of feature extraction step, total of 1842 vocabularies were extracted from whole word bag of 15136 taken by splitting words from all phrases present in training dataset(1680 examples). Labels were encoded as 0 and 1.

Label Name	*Encoded Label*
Positive	1
Negative	0

For feature part, occurence of word and its probability for respective class i.e. P(W | Ci) was calculated with laplacian smoothing given by:

$$P(W \mid Ci) = (n + alpha) / (N + alpha\ |V|\ )$$

Where, n is occurence of W in Ci

N is total number of words in Ci

Alpha is pseudocount and |V| is vocab length

A dictionary of word probabilities w.r.t. Class was created. For eg.:

```
word_probs={
        0:{ “Hello”:”0.0045”,...},
        1:{“Hello”:”0.00034”,...}
}
```

This is used later on to calculate posterior probability of a class.

Submitted By: Anish Bhusal
March 18,2020

3. Model Training and Evaluation

For prediction of posterior probability of a class, following formula was used:

P(Class | Text) = P(Text | Class) * P(Class)
= P(W1|Class)P(W2|Class)P(W3|Class)...P(Wn|Class)* P(Class)

If P(C=0|Text) > P(C=1 | Text) , the model will predict the given phrase as negative, otherwise positive.

Following evaluation metrics show the performance of Naive Bayes model:

Confusion Matrix:

	T	*F*
T	233	173
F	29	685

Model Evaluation Parameters:

Accuracy	0.82
Recall	0.57
Precision	0.88
F1-Score	0.69

For plotting ROC, alpha from Laplacian smoothing was treated as hyperparameter and varied from 1 to 100 and following graph was obtained:

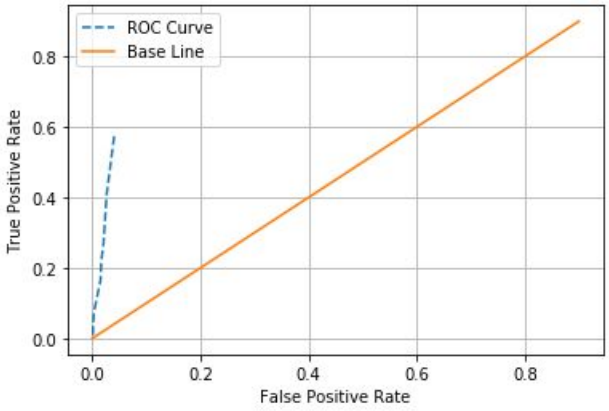


Fig.: ROC Curve

Code is available here