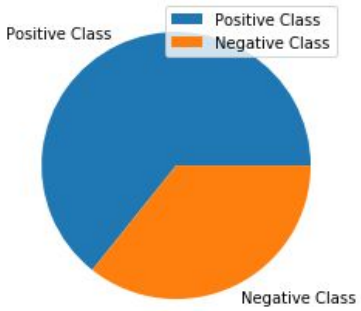
Submitted By: Anish Bhusal
March 18,2020

Logistic Regression 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, validation and test set in 60:20:20 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. Labels were encoded as 0 and 1.

Label Name	*Encoded Label*
Positive	1
Negative	0

For feature part, binary encoding of the form [1 0 0 1] was created for a vocab indicating where a word is present in that phrase or not. 1 indicates the presence of word while 0 doesn’t. For a phrase:

The weather is good.

If the vocab is: *[It The weather good bad is]*, the encoded feature will look like:

[0 1 1 1 0 1]

Submitted By: Anish Bhusal
March 18,2020

3. Model Training, Evaluation and Grid Search

Model was trained, evaluated and grid search was carried out with following params:

num_of_iterations	[1000,1500,2000]
learning_rate	[0.001,0.1,0.01]
Loss Function	Binary Cross Entropy Loss
Activation Function	Sigmoid

The best model with least loss was selected with following hyperparameters:

num_of_iterations	1000
learning_rate	0.1

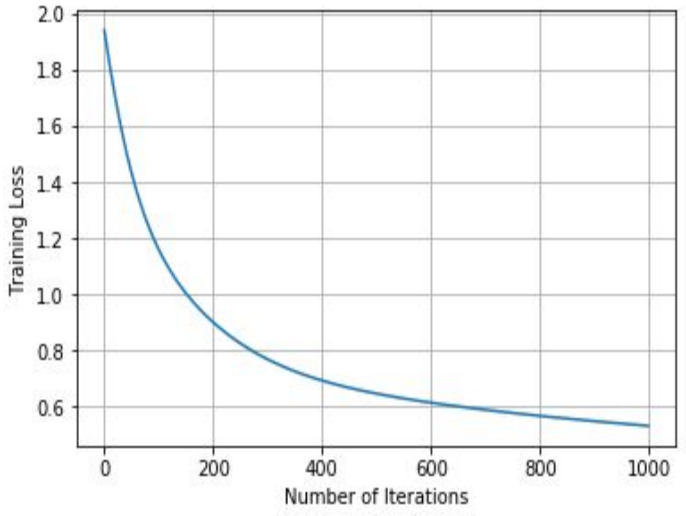


Fig.: Training Loss vs Number of Iterations Plot

Confusion Matrix:

	T	*F*
T	81	132
F	64	283

Model Evaluation Parameters:

Accuracy	0.65
Recall	0.38
Precision	0.55
F1-Score	0.45

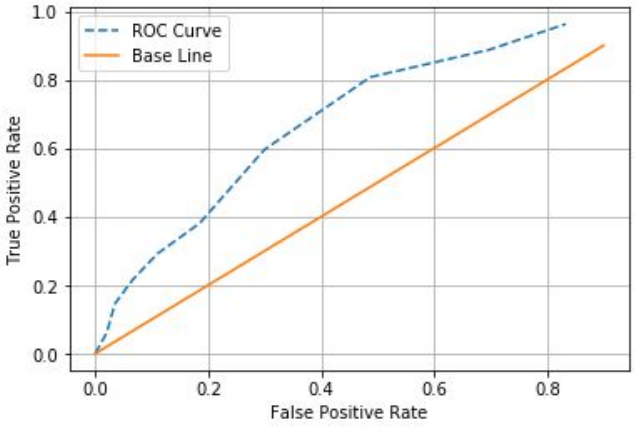


Fig.: ROC Curve

Code is available here