Project Analysis

For the final project, we are given the Katy Perry Youtube comments document, which contains comments from a video uploaded by singer Katy Perry on Youtube.

Using Data exploration techniques, we loaded the data into a dataframe and did some initial data exploration which gave us the following results:

1. There are 350 rows containing data for each comment such as COMMENT\_ID, AUTHOR, DATE, CONTENT, CLASS.
2. Out of these columns, we only require the CONTENT column as this column contains the actual comment and the other rows have unique values for each row and hence cannot be used for predictions
3. None of the data rows have empty or null data
4. Out of 350 total rows, there are 175 rows with class as 1 and 175 rows with class as 0.

Chart, bar chart

Description automatically generated

1. From the data, rows containing 1 signifies that comment is Spam and rows containing 0 signify that the comment is Ham.
2. We also count the number of stop words in the data.

Chart, bar chart

Description automatically generated

1. We also count the number of words that are most commonly used for spam comments

Chart

Description automatically generated

1. We count the number of words that appear mostly for Ham comments

Chart, funnel chart

Description automatically generated

After getting insights about data using Data exploration techniques, we can use this information to transform the data in order to convert it to appropriate form to give it as an input.

We carry out the following steps:

* Convert all the data to Lower case
* Remove the English stop words
* Replace all punctuations with spaces
* Shuffle the dataset
* Convert the data to vectors using CountVectorizer
* Down scale the data using TF-IDF algorithm

After carrying out the following steps, the data is finally fit to use in order to train a predictor model.

We can split the data into Training and Testing sets in order to first train the data and then test it to get a valid accuracy score. The Split used is 75:25, where 75% of data is used to train the model and 25% data is used to test the model.

After splitting the data, we can now feed the training set data into a predictor model. For this purpose, we have used Naïve Biased algorithm. We have used both Gaussian and Multinomial versions of the algorithm in order to get the better model for this usecase.

We can now feed the training data into the models. We have also used Cross Validation on the training set with 5 folds for both the models. The accuracies for these models are:

**Gaussian Naïve Bias:**

[0.79245283, 0.94339623, 0.90384615, 0.88461538, 0.86538462]

Mean score: 0.8779390420899855

**Multinomial Naïve Bias:**

[0.9245283, 0.94339623, 0.90384615, 0.92307692, 0.94230769]

Mean score: 0.9274310595065313

Using these scores, it is clear that Multinomial Naïve Bias Classifier will give better results for this particular usecase and data.

We can now use the test data set to predict some values which we can compare to actual values and get a realtime accuracy. The confusion matrices for these predictions are:

Square

Description automatically generated

Gausian Naïve Bias Classifier Accuracy on test data: 96.59

Square

Description automatically generated with medium confidence

Multinomial Naïve Bias Classifier Accuracy on test data: 96.59

In order to test on real world data, we have created the following dummy comments, pre-processed the comments and given them to predictor to get predictions.

Comments:

* "Visit my channel for more videos&\*^\*(\*",
* "Like,comment share and subscribe to my channel.",
* "http i love this ong",
* "I LOVE YOOU!!!!",
* "Katy is underrated",
* #"I am a big fan of you and your music!!"
* “Never heard a song as shitty as this one”

For the above comments when we predict the values we get the following predictions:

[1, 1, 0, 0, 0, 0]

The conversion for this data will be:

* Spam
* Spam
* Ham
* Ham
* Ham
* Ham

The predictions match the test comments generated. Hence all predictions are correct.