Big Data Analytics – Analysis of Text and Social Media Data

Individual Assignment 1: Sentiment Classifier Modeling Assignment

Sentiment Classifier modelling of Xiaomi Redmi Note 5 Pro reviews from Flipkart

By

Vinay Dalal

Introduction

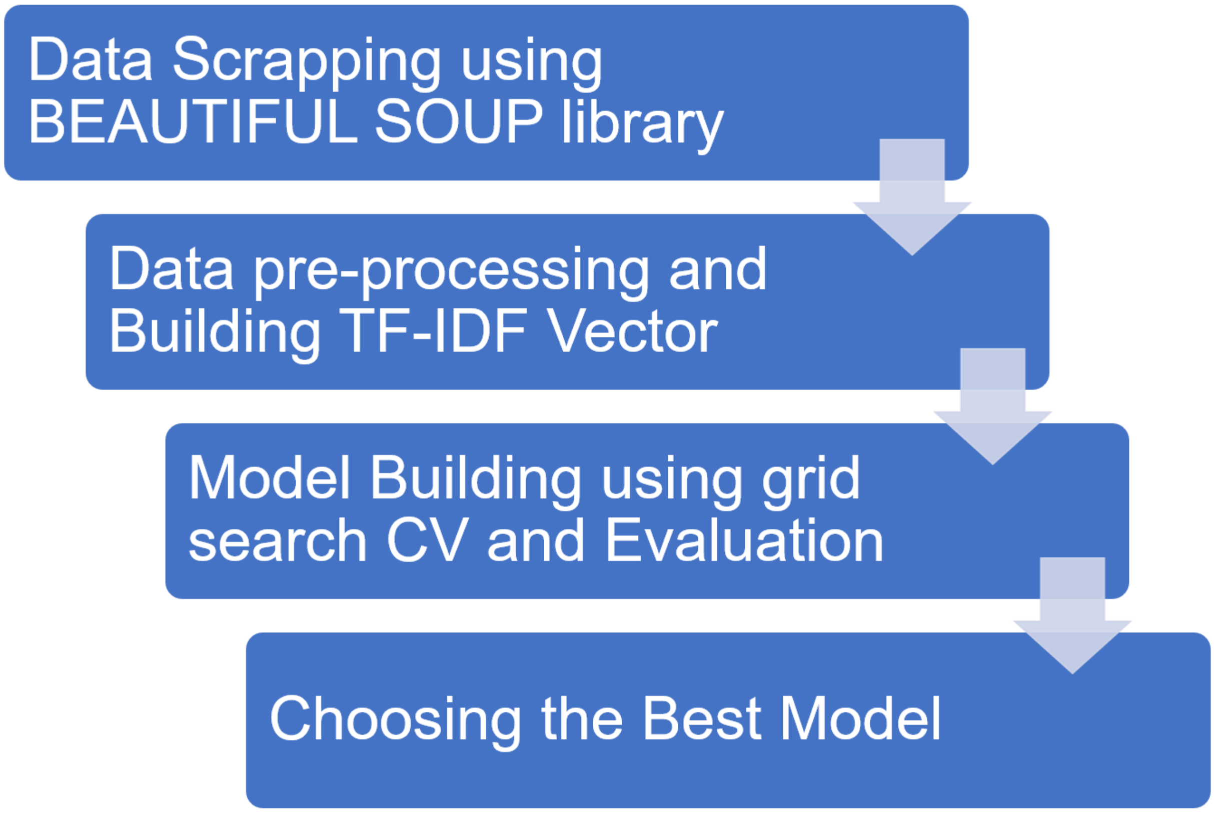
Xiaomi is a smartphone manufacturer based of China, which has also been the largest seller of smartphones in India for some years. The Redmi Note series of phones have been the biggest contributor for Xiaomi in terms of sale and the sale is exclusively done be Flipkart for online mode as an official partner.

Sentiment classifier modelling is generally done, in order to predict the sentiments of the user/customer towards a particular product/service, saving the hassle of going through each review individually and then saying the review is positive, negative etc. In this study I have classified the reviews for Xiaomi Redmi Note 5 pro into 3 classes:

- Positive: 1 (label)
- Negative: 2
- Neutral: 0

While collecting the reviews it was decided by me to sort the reviews in ascending order in order to prevent a future class imbalance and collect only 6000+ reviews at the same time.

Methodology


Data Scrapping using BEAUTIFUL SOUP library
Data pre-processing and Building TF-IDF Vector
Model Building using grid search CV and Evaluation
Choosing the Best Model

Data Insights

The contained the below given number of reviews from each of the classes:

Positive (1)	2185
Negative (2)	2222
Neutral (0)	2183

Web Scraping Using Beautiful Soup

For getting the reviews of user, the Beautiful soup library of python was used. All the reviews were first stored inside a dictionary and then a dataframe. The data frame was also stored in the form of csv file to be shared with faculty. The ratings from user had 3 different classes, so they were fetched by using a try and except loop inside a custom function.

Data Pre-processing and building TF-IDF Vector

Lemmatization was used to pre-process the reviews to create numerical vectors for the same. Post lemmatization using the bag of words a TF-IDF (term frequency-inverse document frequency) was created to be used for modeling. It is created by multiplying two different metrics “term frequency” of a word and “inverse term frequency” of the word. So, if the word is very common and appears in all reviews it will have value closer to 0 otherwise closer to 1. Multiplying these two frequencies the score for all the words is stored in a vector.

Model Building Using Grid search CV

For the purpose of model building the data was divided into 2 parts training (70%) and test (30%) data. For modeling 3 models were used namely

- Decision Tree Classifier
- Random Forest Classifier
- Light Gradient Boosting Machine (LightGBM)

Since this was case of multiclassification the binary classification models were hence not used for modelling. As the dataset in my case had quite the same number of reviews for the classes (a balanced data set), the accuracy measure, AUC ROC and F1 score was used as the main metric for model Evaluation.

Choosing the best Model

After the comparison of

- Accuracy score
- AUC ROC

- F1 SCORE

The best model was the Random Forest classifier model.

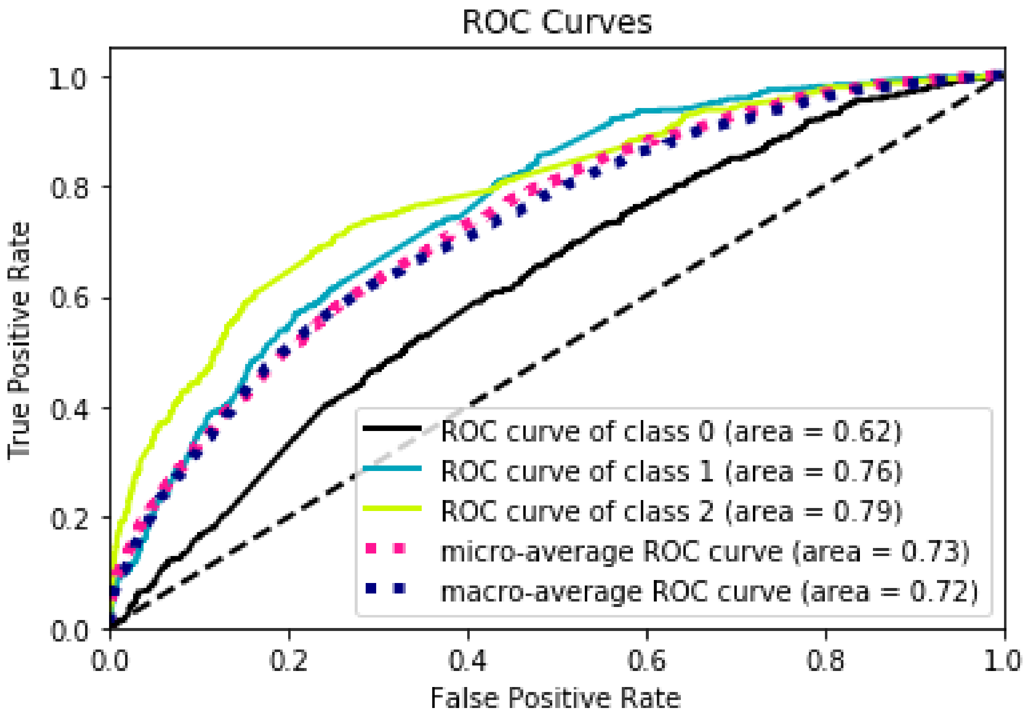


Figure 1: AUC-ROC Curve for the Random Forest model