Project Progress Report

Anant Ashutosh Sharma
anantas2@illinois.edu

Individual Project | Free Topic

Text and Tweet Classification using Machine Learning

1) Which tasks have been completed ?

The following tasks have been successfully completed and corresponding code has been pushed on the GitHub repo:

1. Created a custom dataset from public sources for Political and Non Political Tweets. Total corpus of 6060 tweets, with 4088 labelled as Political and 1972 labelled as Not Political.
2. Cleaned the following datasets for classification
    a. Spam SMS Dataset
    b. Offensive Language Dataset
    c. Political Tweets Dataset
3. Performed the following Pre-Processing on the text data
    a. Removed Stop Words
    b. Removed Non Alphabetic Characters
    c. Performed Stemming
4. Performed the following types of feature extractions:
    a. Count Vectors
    b. Word level TF-IDF
    c. N-Gram level TF-IDF
    d. Character Level TF-IDF
5. Trained the following models:
    a. Naïve Bayes
    b. Linear Classifier
    c. SVM
    d. Random Forest
6. Produced Classification reports for each of the following trained models

Note : The tasks 3 to 6 were performed on all the three datasets. Each model has a separate notebook on GitHub under the Helper Notebooks Directory. In each notebook, the currentDF (i.e. the current Data Frame can be changed to choose one among the three datasets.

2) Which tasks are pending ?

Except CNN, all the models have been trained and tested on the datasets. They are showing good accuracy levels. The following tasks are still pending

1. Train a CNN Model to perform text classification
2. Analyse the accuracies of different model on different datasets with different features by creating a comparison table.
3. Train and Save the most accurate model for classifying political tweets (Indian Context).
4. Fetch Tweets from twitter using the twitter API and classify them as Political or Not Political using the saved model.
5. Create a comprehensive jupyter Notebook to cover all the task performed (for the purposed of presentation). Create readme file on GitHub (documentation with instructions) and Video Presentation for submission.

3) Are you facing any challenges ?

Having no experience with sklean python library, it was initially a bit challenging to get things done. But as I progressed, things became much more clear.

Implementing a CNN to classify texts also seemed a bit challenging at the beginning. However, I am confident that I will be able to accomplish the task. Also, looking forward to use the tweepy python library to extract tweets automatically.