HOMEWORK 3 REPORT

Feature Extraction:

I analyzed the dataset collected by me, which were targeted tweets of 11 user profiles, which included the tweet text, number of followers, retweets for each of the tweet that user put on his timeline along with the likes each of the collected tweet got.

In addition to this I have identified the pronoun count using the Stanford Core NLP tool. The pronoun count for users in each dataset was then compared and analyzed.

The sentiment of the tweets was also extracted by using just the tweet of the text, human annotators were used to annotate 110 tweets which were used as a reference point to again annotate about 450 more tweets manually. Thus making the Human annotated twitter dataset of about 550 tweets. Apart from sentiment that was assigned to each tweet, the following features were also used namely unigrams, bigrams, trigrams, POS, Line Length, Stemming of N-grams, Stopwords removal was also done.

Description of Classifier: I used Support Vector Machine for the classification of the data. I chose the SVM because given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. Thus helping in classify the tweets of the users whose data was collected in positive, negative and neutral format.

As mentioned above the training data was a set of approximately 550 tweets that were used to train the classifier and then used to predict the sentiment of the users unclassified tweets.

Implementation:

(A) Preprocessing Data:

1. **For the Classifier:** The Data collected from twitter using the tweepy library was very targeted, yet it had several inconsistencies which had to be removed before using it to train and predict the data. I used the Code attached with this pdf in the file to clean the data. I removed all the unwanted symbols and made it in a specific format using python code which was attached to the code used to collect the targeted data.

   Please refer to the Clean.py as well the data extraction code in the file included with this report to see the exact code used.

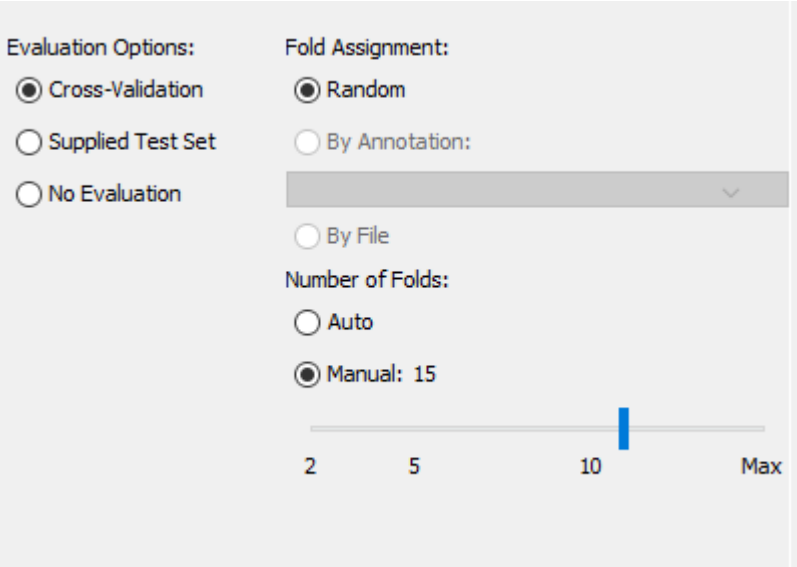
2. **For the Pronoun Count:** Again Python code was used to bring the file in a specific format. Which was then used as an input to the Stanford core nlp software where the command was used in order to tokenize it and gather features which we needed.
3. **For the Information Derived from Twitter Account and Tweets:** I wrote the python code which was used to get the number of likes, followers count and retweet count for each user and save it with the file where tweets were stored.

(B) The features I selected as mentioned above were extracted using the open source tool (Lightside). As mentioned above apart from using sentiment assigned to each tweet, I also

used Uni, Bi, Tri grams along with Stopword removal, POS tagger. The data was already very targeted and did not need any kind of normalizationor standardization after it was cleaned using the python code which I have included in the folder.

(C) As stated I used the Support Vector machine to train and predict my data, this SVM implementation was present in the weka toolkit and was utilized.

(D) A k-fold random validation was also done using lightside. I kept the number of k-folds as 15 because after trying the K-fold validation with both k=10 and 15, I got a better accuracy of my classifier with 15 fold validation. Thus the training and cross validation was performed on the dataset that we had. The training set was the set that I had retrieved from the twitter and had manually annoted using human annotators, the Twitter Dataset which was divided in 15 folds for cross validation and each fold was validated by keeping the other 9 sets as training data. This was repeated 15 times for 15 sets by the tool.



(E) The tool provides the performance metrics by giving us the accuracy values along with the kappa values. In case of feature extraction both the precision and recall values are given. When the dataset is trained both accuracy and kappa values along with the confusion matrix are provided. The precision and recall values were calculated from the confusion matrix of our training data and has been discussed below.

Model Evaluation Metrics:

Metric	Value
Accuracy	0.6234
Kappa	0.3567

Analysis of Results:

Analysis of Classifier: The Accuracy and the Kappa Values for the model have been included in the file as shown above which is the measure of overall accuracy of the classifier.
The Model Confusion Matrix that contains the number of irrelevant, negative, positive, neutral instances has been included in the folder in .csv format. The two identical column and row names are the correctly identified instances from the data after training and the other instances show the mislabeled data that is 'actual/prediction' matrix in case of sentiment analysis while training the classifier. Here we have calculated the precision and recall of our sentiments by the classifier we built using the confusion matrix that we generated after the training of our classifier.

Sentiment	Precision	Recall
Positive	74.917	72.99
Negative	47.979	49.686
Neutral	45.977	47.059

Explanation of the above Results: It was noticed during the Data Collection that most of the tweets by the Leaders from one data set were positive and a majority of those from the self made ones were also positive. This resulted in the accuracy of the classifier when it came to predicting the positive tweets while being almost half when it came to the negative or neutral tweets.

Analysis of the pronoun Count: The Stanford Core NLP tool was used to tokenize and extract all the necessary features of the tweets by a particular user in both the datasets that I collected.

I further wrote a python code to count all the instances of pronouns in the output file in each of the twitter user data that we had and display it in a single screen and store the result in a file so that it becomes easier to understand.

Analysis of Other Data extracted from the twitter: Here I analyzed the data like number of followers of the users in each data set, number of likes to the tweets collected of the user, number of retweets of a users tweet in each dataset. These numbers were then compared for each dataset so as to analyze the results and differences that exist between these two kind of Twitter users.

RESULTS:

Applying the Classifier to our Test Data (Tweets of the users in each Dataset): The trained classifier was then used on the Conversational data provided to us, to find out the sentiment of the text involved. The .csv file that was obtained as a result has been included in the folder.

For DataSet 1: This Dataset included the people who are already established leaders and have twitter verification, two of these were in the Times most Influential list, while the other two are internationally recognized people with one being the editor in chief of indias most famous newspaper.

After applying the classifier that was trained on all the combined tweets of 11 people who have been studies in this homework, to the tweets of the above 5 leaders following results were obtained.

Total Positive Identified	450
Total Negative Identified	154
Total Neutral Identified	114
Total Tweets	717

For Dataset 2: This Dataset included the twitter members that I call self made leaders. I collected their names based on an article published in India Today/Outlook Magazine that featured twitter users that became famous on twitter and have a huge following.

Total Positive Identified	361
Total Negative Identified	306
Total Neutral Identified	234
Total Tweets	900

Explanation: As we can see that the number of Positive tweets is way higher in ratio in Dataset 1 when compared to Dataset 2. This means that the Already established leaders write overwhemingly positive tweets and this is not the case for the users that I have classified as the self made leaders. This corresponds to the ground truth that we collected.

(For 150 tweets calculated predicted user)	Dataset 1 (5 users)	Dataset 2 (6 users)
Average Positive tweets per user	90	60
Average Negative tweets per user	30.8	51
Average Neutral tweets per user	22.8	39

The above results do correspond to the ground truth collected in general sense, as in the majority of tweets by users in dataset one is positive while negative and neutral are very few. Although it still is less than what was expected according to the ground truth. But in overall sense the data agrees with the ground truth.

Results of the pronoun Count corresponding to both the Datasets: I wrote the code to read the xml file format generated by the Stanford Core NLP tool and count the pronoun for each file and list them. The following is the output that was produced.

```
Python 2.7.9 (default, Dec 10 2014, 12:24:55) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ========== RESTART ==========
>>>
input_GabbarSingh.txt.out 91
input_Greatbong.txt.out 151
input_ikaveri.txt.out 123
input_Jaihind.txt.out 216
input_jhunjhunwala.txt.out 114
input_narendramodi.txt.out 184
input_Obama.txt.out 142
input_praveenswami.txt.out 147
input_roflindian.txt.out 121
input_sachinrt.txt.out 191
input_srbachchan.txt.out 92
>>>
```

Dataset 1	Pronoun Count	Dataset 2	Pronoun Count
Obama	142	GreatBong	151
PraveenSwami	147	ikaveri	123
Narendra Modi	184	Jaihind	216
SrBachchan	92	jhunjhunwala	114
Sachin	191	roflindian	121

On comparing the Pronoun Count for both the Dataset users:

Dataset 1 (Established Leaders)	816
Dataset 2 (Self Made Leaders)	756

The Average count for users in both dataset comes out to be:

Average Pronoun count in dataset 1	163.2
Average Pronoun count in dataset 2	126

As we can see from the above results that on an average the established leaders use more pronouns when compared to the other twitter users, who interact more with users with their initials rather than pronouns.

Results of Other Data extracted from the Twitter:

Likes And Retweets: Following are the number of Likes and Retweets by the individual twitter users

```
...
JayHind
Number of retweets: 3584
Number of likes: 207
GabbbarSingh
Number of retweets: 3809
Number of likes: 4222
greatbong
Number of retweets: 802
Number of likes: 494
jhunjhunwala
Number of retweets: 3814
Number of likes: 1243
Roflindian
Number of retweets: 10140
Number of likes: 3255
BarackObama
Number of retweets: 370566
Number of likes: 834303
narendramodi
Number of retweets: 382397
Number of likes: 995250
ikaveri
Number of retweets: 6682
Number of likes: 407
SrBachchan
Number of retweets: 26080
Number of likes: 149817
sachin_rt
Number of retweets: 325093
Number of likes: 863369
praveenswami
Number of retweets: 1694
Number of likes: 746
>>>
```

Thus the Number of Likes and retweets for both the datasets are:

	Dataset 1	Dataset 2
Total Likes	2843745	9830
Total Retweets	1105902	28837

Thus average for each dataset comes out to be:

	Dataset 1	Dataset 2
Average Likes	568,749	1683
Average Retweets	221,180	4806

Explanation: As we can see from the above data that the gap is huge with respect to the number of likes and number of retweets between the twitter users in dataset one and dataset two. This can be attributed to the fact that already established leaders on twitter cater to a larger segment of twitter population whereas the self established ones cater to the niche followers that they have created in their domain.

Number Of Followers and Following: I have also analyzed the number of followers that each member has, and that of the people in the particular Dataset. The number of people followed by these people has also been included in this file.

	Dataset 1	Dataset 2
Total No. of Followers	124950509	585744
Total Number of Following	639119	4339

The average number of followers and those being followed by the users is:

	Dataset 1	Dataset 2
Average no. of followers/user	24990101.8	97624
Average no. of following/user	127823	723

Explanation: The Average no. of followers might not be as an accurate indicator in this case as I hoped as in both the datasets there were outlier profiles that increased this values dramatically when the others did not show shuch a huge numbers as they had a lot of followers of their own.

To see this Let us take a look at the following table with individual values listed:

Dataset 1:

Name of user	Followers	Following
Obama	73988163	636613
Narendra Modi	19619806	1372
Amitabh Bachchan	20643658	999
Sachin	10642322	15
Praveen Swami	56560	120

Dataset 2:

Name of User	Followers	Following
Jaihind	2210	143
Gabbar Singh	277135	1230
GreatBong	23151	140
JhunJhunwala	95667	705
RoflIndian	125883	469
ikaveri	41095	1652

As we can see from the Individual dataset table, the number of people followed by the users in our dataset is relatively less when compared to the number of followers following them. This can be attributed to the fact that these people are the influencers in the Twitter World and are seen as leaders in their area.

USER INTERACTION:

I then also calculated the total number of interactions made by each member of the dataset. The code for doing so has been included in the folder 10. This was calculated for each user based on the 200 tweets collected for each user from their profile using Tweepy.

Following is the output I get aftter running my code:

```
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
usertweets-GabbbarSingh_2.txt.out 129
usertweets-greatbong_2.txt.out 114
usertweets-ikaveri_2.txt.out 115
usertweets-JayHind_2.txt.out 148
usertweets-jhunjhunwala_2.txt.out 153
usertweets-Roflindian_2.txt.out 122
usertweets-BarackObama_2.txt.out 51
usertweets-narendramodi_2.txt.out 36
usertweets-praveenswami_2.txt.out 92
usertweets-sachin_rt_2.txt.out 115
usertweets-srbachchan_2.txt.out 103
>>>
```

Now looking for individual results in the datasets:

Dataset 1:

Number of references in usertweets-BarackObama_2.txt.out: 51

Number of references in usertweets-narendramodi_2.txt.out: 36

Number of references in usertweets-praveenswami_2.txt.out: 92

Number of references in usertweets-sachin_rt_2.txt.out: 115

Number of references in usertweets-srbachchan_2.txt.out: 103

Total number of references: 397

Average Replies: 79.4 tweets per user

Dataset 2:

Number of references in usertweets-GabbbarSingh_2.txt.out: 129

Number of references in usertweets-greatbong_2.txt.out: 114

Number of references in usertweets-ikaveri_2.txt.out: 115

Number of references in usertweets-JayHind_2.txt.out: 148

Number of references in usertweets-jhunjhunwala_2.txt.out: 153

Number of references in usertweets-Roflindian_2.txt.out: 122

Total number of references: 781

Average Replies: 130.1 tweets per user

Evaluation: We can say after looking at this data that Established leaders interact less with public and tweet the things they think are important.

Whereas the self made leaders interact way more with their followers. This can be due to the fact that they need to interact with people in order to remain an influencer on twitter world. They have to be more involved in the twitter universe if they want to maintain their status as a twitter leader and influencer.

Conclusion: There is significant difference between the already established leaders and the Self Made leaders on the Twitter Platform. As we have seen from different measures and criteria above how they differ and have discussed the possible reason for such Behavior.