EE219 Project 5 - Report
Popularity Prediction On Twitter
Winter 2018

Introduction

In this project, we try to solve the problem of predicting the tweet activity in the future, based on the current tweet activity for a hashtag.

Dataset

We use the Super Bowl 2015 tweet dataset which spans from a period starting from 2 weeks before the game to a week after the game. We prepared training data by extracting features and fitting a regression model on the training data. The test data consists of tweets containing a hashtag in a specified window and we used the trained model to predict the number of tweets containing the hashtag posted within one hour immediately following the given time window.

Q1.1:

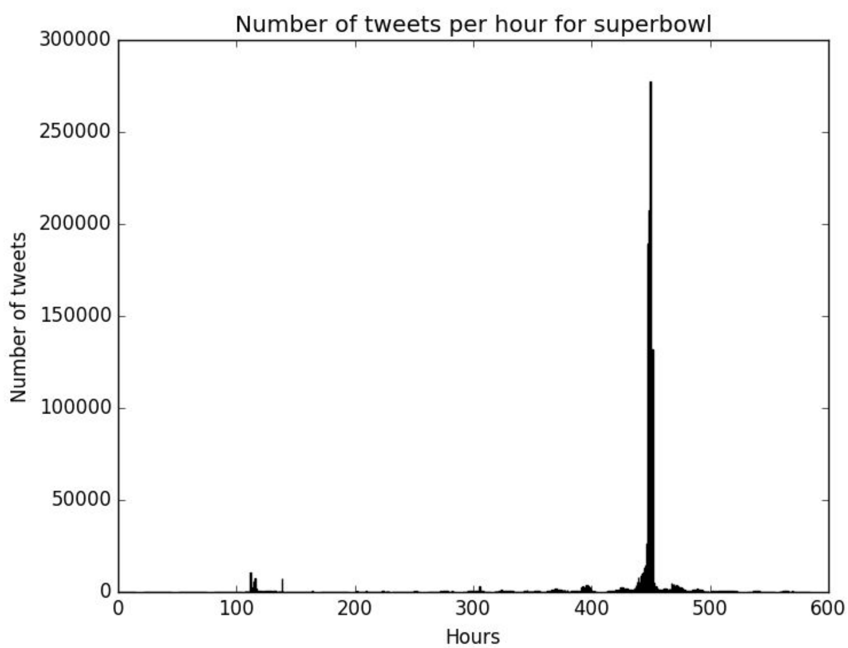
In the original text files, the data is stored in JSON format where each line has a tweet and tweets are sorted with respect to their posting time. We convert these text files into CSV files and use them to load the data into a dataframe for all the questions.

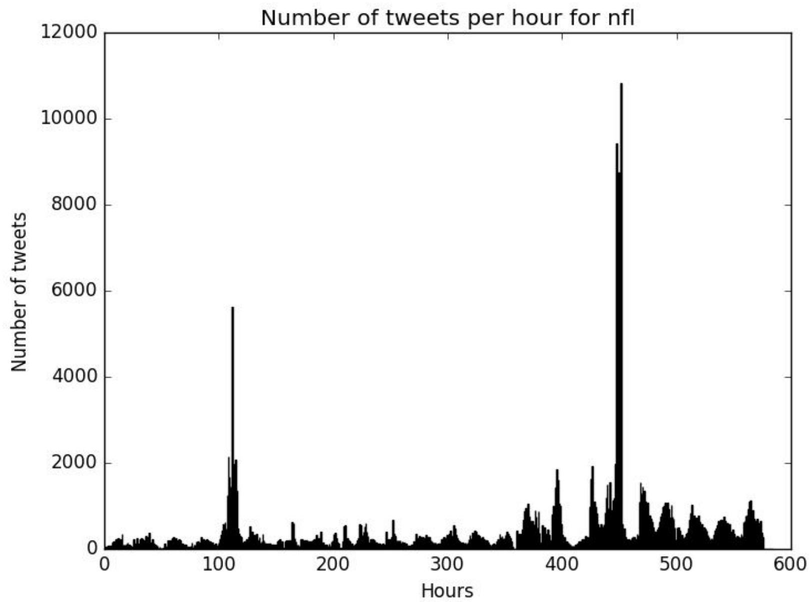
We calculated the following statistics for each hashtag over 1 hour windows:

1. Average number of tweets per hour
2. Average number of followers of users posting the tweets
3. Average number of retweets

Hashtag	Average number of tweets	Average number of followers	Average number of retweets
#superbowl	2297.729	14917.05	1.79
#nfl	441.267	4464.549	1.155
#gohawks	324.93	2486.52	1.6580
#gopatriots	45.62	1554.329	1.179
#patriots	834.264	7202.974	2.01
#sb49	1418.4408	23375.215	2.9536

The plots for number of tweets per hour for #SuperBowl and #NFL are as follows:





Q1.2:

For each hashtag, a model was trained using the following features:

1. Number of Tweets
2. Total number of retweets (hashtag of interest)
3. Sum of the number of followers of the users posting the hashtag
4. Maximum number of followers of the users posting the hashtag
5. Time of the day (which could take 24 values that represent hours of the day with respect to a given time zone)

#Super Bowl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     478.9
Date:                Sat, 17 Mar 2018   Prob (F-statistic):          2.24e-203
Time:                        23:13:31   Log-Likelihood:                -6099.2
No. Observations:                 586   AIC:                         1.221e+04
Df Residuals:                     581   BIC:                         1.223e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             2.2893      0.079     28.836      0.000         2.133     2.445
x2            -0.2921      0.036     -8.118      0.000        -0.363    -0.221
x3            -0.0001   1.86e-05     -6.839      0.000        -0.000 -9.07e-05
x4             0.0007      0.000      4.876      0.000         0.000     0.001
x5             0.5915     29.127      0.020      0.984       -56.615    57.798
==============================================================================
Omnibus:                     1011.558   Durbin-Watson:                   2.312
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1849450.122
Skew:                          10.096   Prob(JB):                         0.00
Kurtosis:                     277.477   Cond. No.                     1.07e+07
==============================================================================
```

The train RMSE was found to be: 8016.349839229025
R-squared value: 0.805

Based on the p-test and t-value the significant features for this model were observed to be: Number of Tweets, maximum number of followers of the users posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

#sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.807
Method:                 Least Squares   F-statistic:                     488.1
Date:                Sat, 17 Mar 2018   Prob (F-statistic):          1.43e-204
Time:                        23:13:58   Log-Likelihood:                -5717.5
No. Observations:                 582   AIC:                         1.144e+04
Df Residuals:                     577   BIC:                         1.147e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             1.1838      0.095     12.438      0.000         0.997     1.371
x2            -0.2091      0.088     -2.379      0.018        -0.382    -0.036
x3          1.786e-05    1.4e-05      1.272      0.204     -9.73e-06  4.54e-05
x4          7.945e-05   4.73e-05      1.679      0.094     -1.35e-05     0.000
x5            14.0622     15.226      0.924      0.356       -15.842    43.967
==============================================================================
Omnibus:                     1186.519   Durbin-Watson:                   1.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2271097.747
Skew:                          14.782   Prob(JB):                         0.00
Kurtosis:                     307.597   Cond. No.                     7.19e+06
==============================================================================
```

The train RMSE was found to be: 4468.896903026785
R-squared value: 0.809

Based on the p-test and t-value the significant features for this model were observed to be: **Number of Tweets, maximum number of followers of the users** posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.506
Model:                            OLS   Adj. R-squared:                  0.501
Method:                 Least Squares   F-statistic:                     117.2
Date:                Sat, 17 Mar 2018   Prob (F-statistic):           3.05e-85
Time:                        23:14:16   Log-Likelihood:                -4794.7
No. Observations:                 578   AIC:                             9599.
Df Residuals:                     573   BIC:                             9621.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             1.2520      0.168      7.444      0.000         0.922     1.582
x2            -0.1230      0.044     -2.813      0.005        -0.209    -0.037
x3            -0.0002   8.38e-05     -2.321      0.021        -0.000 -2.99e-05
x4          4.487e-05      0.000      0.287      0.774        -0.000     0.000
x5            11.4929      3.191      3.602      0.000         5.225    17.760
==============================================================================
Omnibus:                      906.959   Durbin-Watson:                   2.239
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           774023.336
Skew:                           8.495   Prob(JB):                         0.00
Kurtosis:                     181.468   Cond. No.                     2.26e+05
==============================================================================
```

The train RMSE was found to be: 969.1739711655209
R-squared value: 0.506

Based on the p-test and t-value the significant features for this model were observed to be: **Number of Tweets, time of the day** posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.642
Model:                            OLS   Adj. R-squared:                  0.639
Method:                 Least Squares   F-statistic:                     204.0
Date:                Sat, 17 Mar 2018   Prob (F-statistic):          2.32e-124
Time:                        23:14:18   Log-Likelihood:                -3809.4
No. Observations:                 574   AIC:                             7629.
Df Residuals:                     569   BIC:                             7651.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -0.0455      0.255     -0.179      0.858        -0.547     0.455
x2             0.4831      0.220      2.198      0.028         0.051     0.915
x3             0.0002      0.000      1.243      0.214        -0.000     0.001
x4            -0.0004      0.000     -1.945      0.052        -0.001  3.79e-06
x5             1.2488      0.589      2.122      0.034         0.093     2.405
==============================================================================
Omnibus:                      515.025   Durbin-Watson:                   1.954
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           300783.149
Skew:                           2.834   Prob(JB):                         0.00
Kurtosis:                     115.001   Cond. No.                     3.19e+04
==============================================================================
```

The train RMSE was found to be: 184.51110215079322
R-squared value: 0.642

Based on the p-test and t-value the significant features for this model were observed to be: **Total number of retweets**,**Time of the day** posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

#nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.653
Model:                            OLS   Adj. R-squared:                  0.650
Method:                 Least Squares   F-statistic:                     218.4
Date:                Sat, 17 Mar 2018   Prob (F-statistic):          7.55e-131
Time:                        23:13:39   Log-Likelihood:                -4561.0
No. Observations:                 586   AIC:                             9132.
Df Residuals:                     581   BIC:                             9154.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.7199      0.131      5.499      0.000         0.463     0.977
x2            -0.1590      0.064     -2.493      0.013        -0.284    -0.034
x3          7.559e-05   2.55e-05      2.959      0.003      2.54e-05     0.000
x4         -7.566e-05   3.48e-05     -2.171      0.030        -0.000 -7.23e-06
x5             9.4491      2.084      4.535      0.000         5.357    13.542
==============================================================================
Omnibus:                      590.088   Durbin-Watson:                   2.370
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           340862.774
Skew:                           3.559   Prob(JB):                         0.00
Kurtosis:                     120.939   Cond. No.                     4.06e+05
==============================================================================
```

The train RMSE was found to be: 580.7965654820871
R-squared value: 0.653

Based on the p-test and t-value the significant features for this model were observed to be: **Number of Tweets, Sum of the number of followers of the users, Time of the day** posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

#patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     249.0
Date:                Sat, 17 Mar 2018   Prob (F-statistic):          7.04e-142
Time:                        23:14:10   Log-Likelihood:                -5421.9
No. Observations:                 586   AIC:                         1.085e+04
Df Residuals:                     581   BIC:                         1.088e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.9189      0.071     12.939      0.000         0.779     1.058
x2            -0.0895      0.058     -1.531      0.126        -0.204     0.025
x3          1.486e-06   2.63e-05      0.057      0.955     -5.01e-05   5.3e-05
x4             0.0001      0.000      1.382      0.167        -6e-05     0.000
x5            12.5167      8.592      1.457      0.146        -4.358    29.391
==============================================================================
Omnibus:                      884.023   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           695058.464
Skew:                           7.863   Prob(JB):                         0.00
Kurtosis:                     170.986   Cond. No.                     7.55e+05
==============================================================================
```

The train RMSE was found to be: 2523.7470774584417
R-squared value: 0.682

Based on the p-test and t-value the significant features for this model were observed to be: **Number of Tweets, Time of the day** posting the hashtag. The other features have either a negative t-value or a high p-value indicating that they are not significant.

Observation: We can observe that the larger the size of the training dataset, the better the accuracy. For instance, we can see that the R-squared value for #superbowl which contains 1348767 tweets is 0.805 and the R-squared value for #gopatriots which contains 26232 tweets is 0.642.

Q1.3:

The model designed in Part 1.2 was augmented with the following additional features:

1. Number of distinct users tweeting in the given hour
2. Sum of the Impression counts in the given hour
3. Sum of the ranking scores of the tweets in the hour
4. Number of tweets of length > 200
5. Sum of the number of favorites

The RMSE was found to improve for each hashtag.
For each hashtag the 3 best features were determined.

#Superbowl

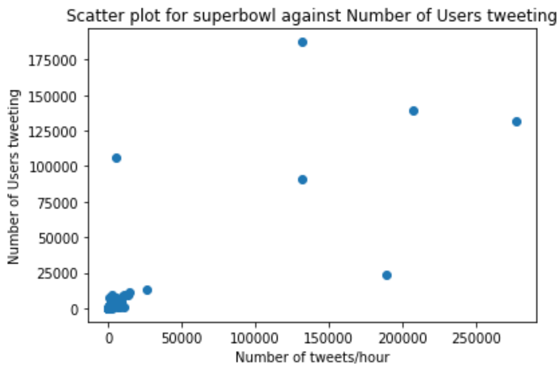
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.820
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     292.2
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          2.28e-208
Time:                        00:03:13   Log-Likelihood:                -6075.3
No. Observations:                 586   AIC:                         1.217e+04
Df Residuals:                     577   BIC:                         1.221e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             1.9738      0.133     14.837      0.000         1.713     2.235
x2            -0.3918      0.039    -10.107      0.000        -0.468    -0.316
x3         -7.903e-05   2.04e-05     -3.880      0.000        -0.000  -3.9e-05
x4             0.0007      0.000      4.460      0.000         0.000     0.001
x5           101.6508     47.770      2.128      0.034         7.826   195.476
x6            -0.0359      0.021     -1.722      0.086        -0.077     0.005
x7            29.0681    216.139      0.134      0.893      -395.446   453.583
x8          -214.6649    157.848     -1.360      0.174      -524.691    95.362
x9            -2.4834      0.385     -6.453      0.000        -3.239    -1.728
x10            1.9738      0.133     14.837      0.000         1.713     2.235
==============================================================================
Omnibus:                     1262.722   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3692736.122
Skew:                          16.629   Prob(JB):                         0.00
Kurtosis:                     390.469   Cond. No.                     4.51e+19
==============================================================================
```

The RMSE value for the model was found to be: 7695.53

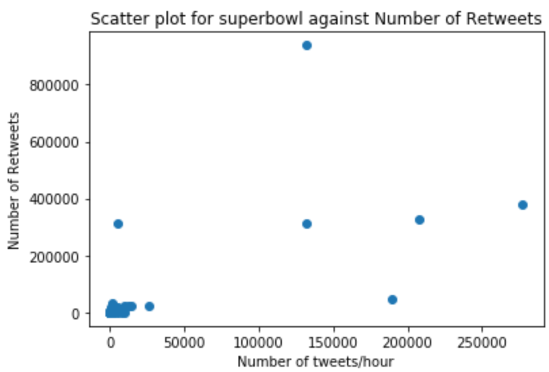
The best features were found to be: Number of tweets, Number of retweets, Number of tweets of length greater than 200 characters

Scatter plots of the best features vs number of tweets:

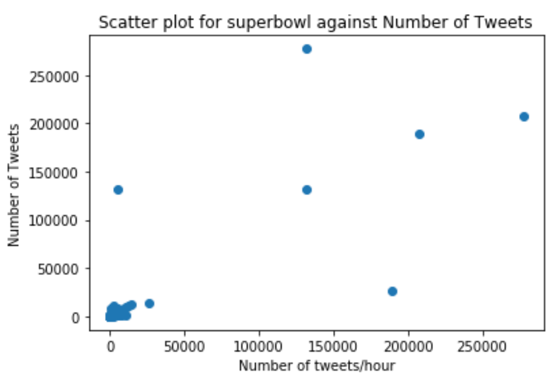
Number of Users tweeting



Number of Retweets



Number of Tweets



#nfl

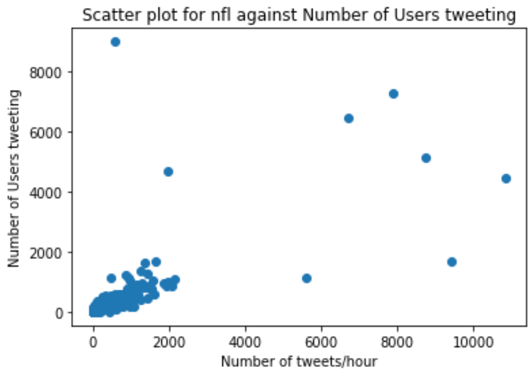
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.693
Model:                            OLS   Adj. R-squared:                  0.689
Method:                 Least Squares   F-statistic:                     145.1
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          6.91e-142
Time:                        00:03:20   Log-Likelihood:                -4524.4
No. Observations:                 586   AIC:                             9067.
Df Residuals:                     577   BIC:                             9106.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             1.1396      0.113     10.068      0.000         0.917     1.362
x2            -0.0360      0.062     -0.582      0.561        -0.157     0.085
x3          4.214e-05   2.56e-05      1.648      0.100     -8.09e-06  9.24e-05
x4         -5.702e-05   4.05e-05     -1.407      0.160        -0.000  2.26e-05
x5             8.1310      3.339      2.435      0.015         1.573    14.689
x6             0.0029      0.008      0.372      0.710        -0.012     0.018
x7            -7.0138     35.081     -0.200      0.842       -75.916    61.888
x8           -12.7756     12.398     -1.030      0.303       -37.126    11.575
x9            -2.2023      0.256     -8.599      0.000        -2.705    -1.699
x10            1.1396      0.113     10.068      0.000         0.917     1.362
==============================================================================
Omnibus:                      827.532   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           383998.353
Skew:                           7.077   Prob(JB):                         0.00
Kurtosis:                     127.606   Cond. No.                     3.21e+19
==============================================================================
```

The RMSE value for the model was found to be: 545

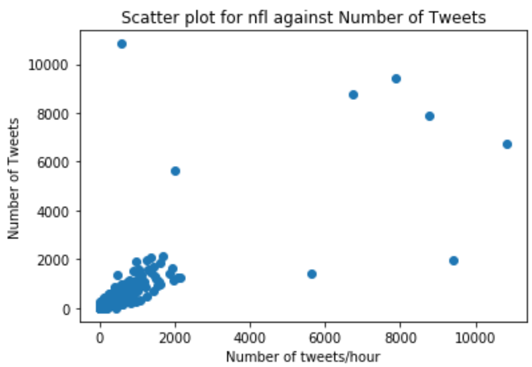
The best features were found to be: Number of tweets of length > 200, Number of unique users tweeting, number of tweets

Scatter plots of the best features vs number of tweets:

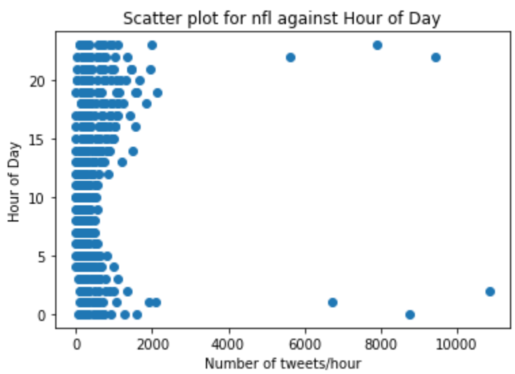
Number of Users tweeting



Number of Tweets



Hour of Day



#sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.813
Method:	Least Squares	F-statistic:	281.3
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	9.77e-204
Time:	00:03:35	Log-Likelihood:	-5707.2
No. Observations:	582	AIC:	1.143e+04
Df Residuals:	573	BIC:	1.147e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	-0.4479	0.253	-1.767	0.078	-0.946	0.050
x2	0.0154	0.104	0.148	0.882	-0.188	0.219
x3	3.164e-05	1.43e-05	2.209	0.028	3.51e-06	5.98e-05
x4	8.468e-06	5.88e-05	0.144	0.885	-0.000	0.000
x5	31.3446	25.365	1.236	0.217	-18.474	81.164
x6	-0.0033	0.004	-0.838	0.402	-0.011	0.004
x7	-4.4762	26.283	-0.170	0.865	-56.099	47.147
x8	-65.2675	79.324	-0.823	0.411	-221.068	90.534
x9	1.7914	0.431	4.160	0.000	0.946	2.637
x10	-0.4479	0.253	-1.767	0.078	-0.946	0.050

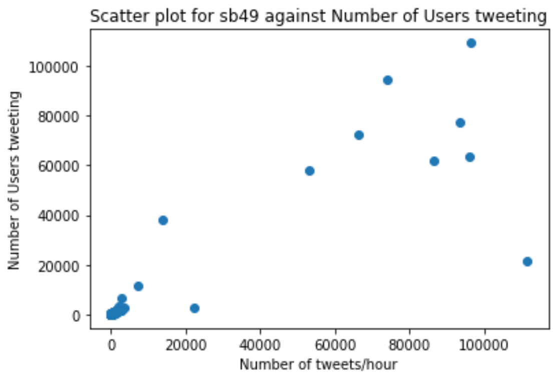
Omnibus:	1156.302	Durbin-Watson:	1.750
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2084470.163
Skew:	13.942	Prob(JB):	0.00
Kurtosis:	294.856	Cond. No.	6.63e+19

The RMSE value for the model was found to be: 4390

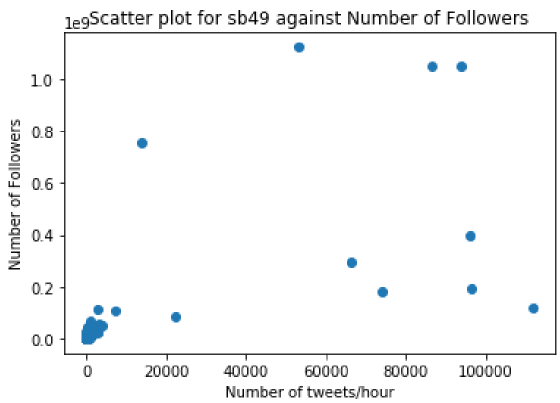
The best features were found to be:Number of tweets of length > 200, Sum of the number of followers of the authors , Sum of the number of followers

Scatter plots of the best features vs number of tweets:

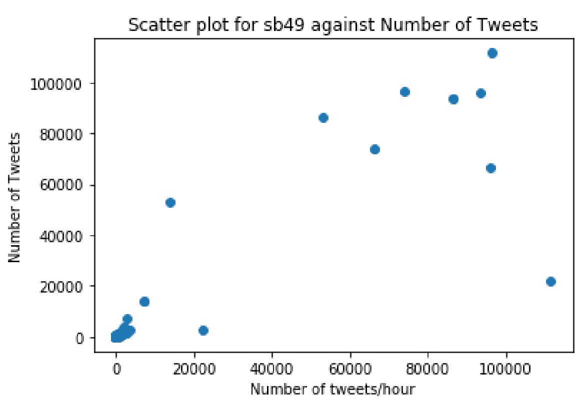
Number of Users tweeting



Number of Followers



Number of Tweets



#patriots

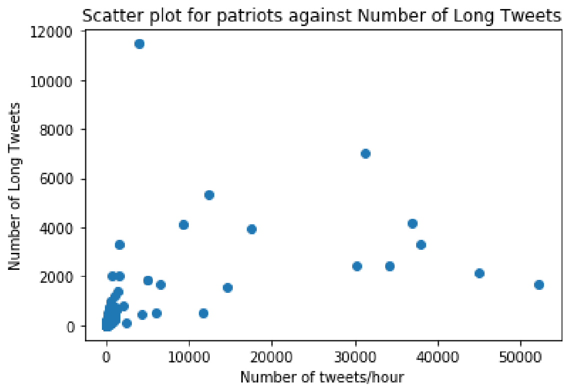
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     151.2
Date:                Mon, 19 Mar 2018   Prob (F-statistic):         1.65e-145
Time:                        00:03:45   Log-Likelihood:                -5402.5
No. Observations:                 586   AIC:                         1.082e+04
Df Residuals:                     577   BIC:                         1.086e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -1.9719      0.407     -4.849      0.000        -2.771    -1.173
x2            -0.1088      0.058     -1.892      0.059        -0.222     0.004
x3             0.0003   5.57e-05      5.288      0.000         0.000     0.000
x4            -0.0002      0.000     -1.320      0.188        -0.000   8.9e-05
x5             8.5798     15.430      0.556      0.578       -21.727    38.886
x6            -0.0176      0.009     -2.047      0.041        -0.035    -0.001
x7             5.7453     30.358      0.189      0.850       -53.880    65.371
x8            24.4065     46.415      0.526      0.599       -66.756   115.569
x9             4.7760      0.796      5.999      0.000         3.212     6.340
x10           -1.9719      0.407     -4.849      0.000        -2.771    -1.173
==============================================================================
Omnibus:                      909.510   Durbin-Watson:                   2.008
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           687411.510
Skew:                           8.373   Prob(JB):                         0.00
Kurtosis:                     169.952   Cond. No.                     3.76e+19
==============================================================================
```

The RMSE value for the model was found to be: 2441

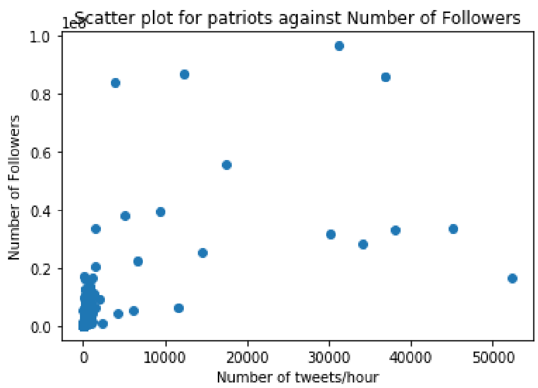
The best features were found to be: Number of unique users tweeting, sum of the number of followers, number of tweets of length >200

Scatter plots of the best features vs number of tweets:

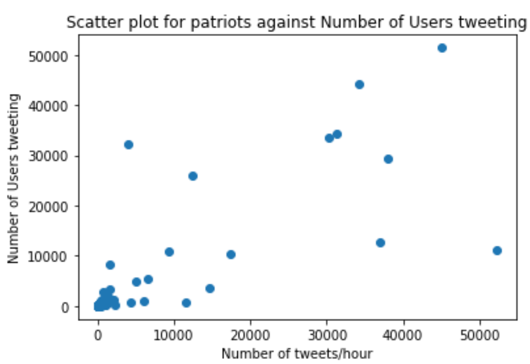
Number of Long Tweets



Number of Followers



Number of Users tweeting



#gohawks

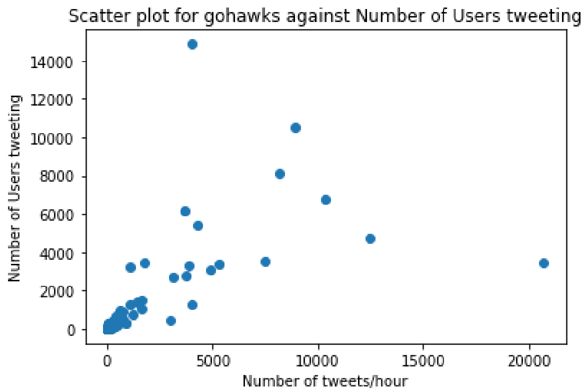
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.576
Model:                            OLS   Adj. R-squared:                  0.569
Method:                 Least Squares   F-statistic:                     85.83
Date:                Mon, 19 Mar 2018   Prob (F-statistic):         5.44e-100
Time:                        00:03:49   Log-Likelihood:                -4750.4
No. Observations:                 578   AIC:                             9519.
Df Residuals:                     569   BIC:                             9558.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -0.9526      0.194     -4.919      0.000        -1.333    -0.572
x2            -0.1232      0.050     -2.454      0.014        -0.222    -0.025
x3            -0.0003   8.03e-05     -4.217      0.000        -0.000    -0.000
x4          4.441e-05      0.000      0.284      0.777        -0.000     0.000
x5            11.5626      5.536      2.089      0.037         0.689    22.437
x6             0.0070      0.007      1.024      0.306        -0.006     0.020
x7            21.2451     14.373      1.478      0.140        -6.985    49.475
x8           -20.8693     17.523     -1.191      0.234       -55.287    13.548
x9             4.4983      0.480      9.370      0.000         3.555     5.441
x10           -0.9526      0.194     -4.919      0.000        -1.333    -0.572
==============================================================================
Omnibus:                     1113.184   Durbin-Watson:                   2.147
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1671842.211
Skew:                          13.053   Prob(JB):                         0.00
Kurtosis:                     265.178   Cond. No.                     6.86e+18
==============================================================================
```

The RMSE value for the model was found to be: 897

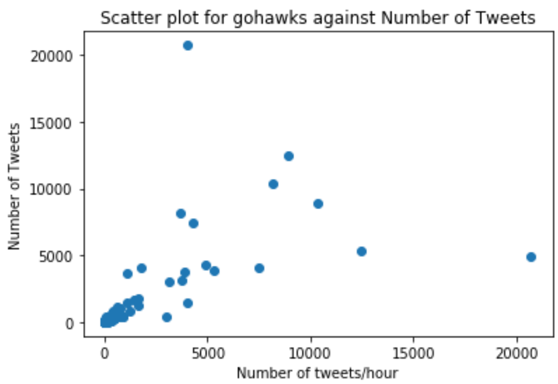
The best features were found to be: Number of tweets, Number of unique users tweeting, Number of tweets of length > 200

Scatter plots of the best features vs number of tweets:

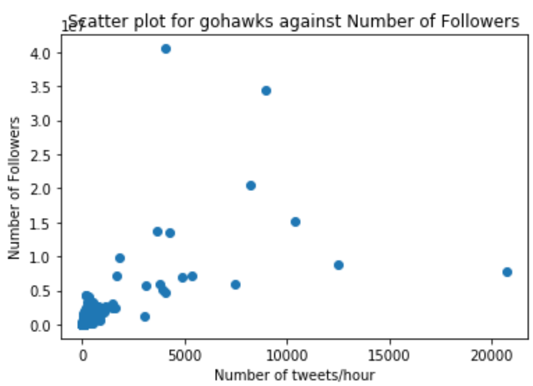
Number of Users tweeting



Number of Tweets



Number of Followers



#gopatriots

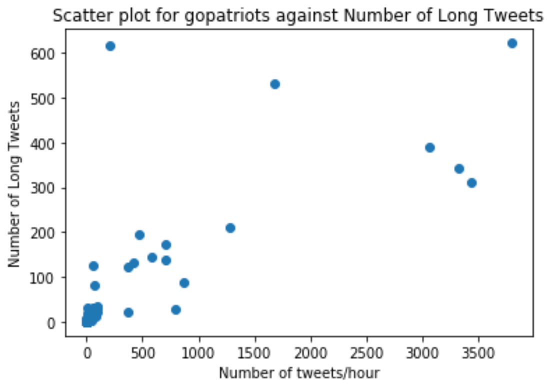
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     116.4
Date:                Mon, 19 Mar 2018   Prob (F-statistic):         1.66e-122
Time:                        00:03:50   Log-Likelihood:                -3803.3
No. Observations:                 574   AIC:                             7625.
Df Residuals:                     565   BIC:                             7664.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.5001      0.198      2.532      0.012         0.112     0.888
x2             0.5953      0.223      2.670      0.008         0.157     1.033
x3             0.0004      0.000      1.733      0.084     -4.71e-05     0.001
x4            -0.0005      0.000     -2.420      0.016        -0.001 -9.16e-05
x5             1.7376      0.907      1.916      0.056        -0.044     3.519
x6             0.0002      0.001      0.211      0.833        -0.001     0.002
x7           -11.9151     28.330     -0.421      0.674       -67.560    43.730
x8            -1.3467      3.187     -0.423      0.673        -7.607     4.914
x9            -1.5113      0.438     -3.447      0.001        -2.372    -0.650
x10            0.5001      0.198      2.532      0.012         0.112     0.888
==============================================================================
Omnibus:                      683.563   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           322802.058
Skew:                           5.001   Prob(JB):                         0.00
Kurtosis:                     118.745   Cond. No.                     7.17e+18
==============================================================================
```

The RMSE value for the model was found to be: 182

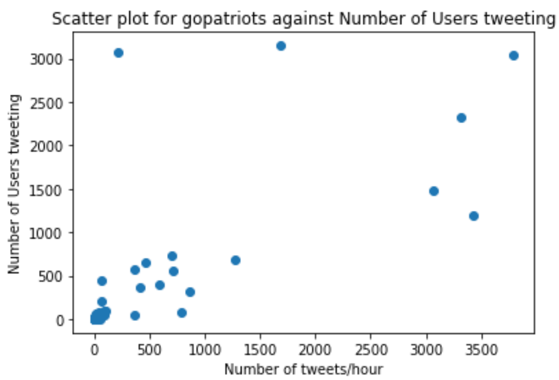
The best features were found to be: Number of tweets, Number of retweets, number of distinct users tweeting

Scatter plots of the best features vs number of tweets:

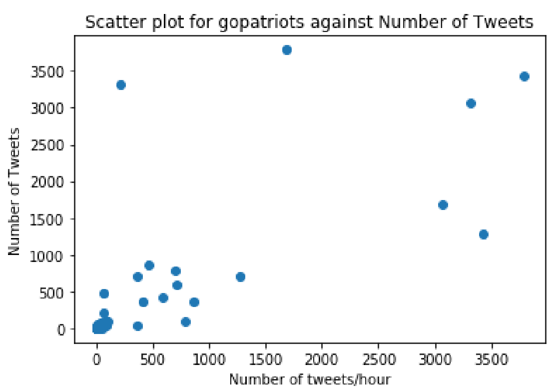
Number of Long Tweets



Number of Users tweeting



Number of Tweets



Q1.4:

In this section, 10 fold cross validation is performed on the chosen model. The dataset is split into 10 parts, where 9 parts are used to train the model and the last part is used to test it. The process is repeated ten times and the value of the root mean squared error is calculated in each run and the average is determined at the end of the process to determine the effectiveness of the model. The tweets are grouped into hourly intervals and the model is trained on hour 'n' using 10 features to predict the number of tweets in the subsequent hour 'n+1'.

For this task, the dataset is split into three periods: the period of the super bowl, before and after.

Three different models, one linear and two non-linear are trained on each of these periods, for each of the hashtags and a total of 54 models are obtained.
The linear model used is OLS while the two non linear models are Random Forest Regressor and K Nearest Neighbour Regressor.
In the below table, the period mappings are as follows:
Period 1: Before Super Bowl, Feb 1 8.00 a.m
Period 2: During Super Bowl, Feb 1 8.00 a.m to 8.00 p.m
Period 3: After Superbowl, Feb 1 8.00 p.m
Period 4: Over all the time, two weeks before the game to one week after

The following table contains RMSE values for different models over different time intervals:

Hashtag	**OLS Model**				**KNN Regressor**				**Random Forest Regressor**			
Period	1	2	3	4	1	2	3	4	1	2	3	4
gohawks	406.431151116	9280.34284477	1987.78006521	432.364382008	311.895489853	648.375772142	594.221610402	329.659475719	291.09584753	496.668778126	615.218150932	295.862196319
nfl	138.147323714	2650.69394566	695.152627351	286.098965059	171.835789636	358.857200002	741.206690644	360.511712604	126.124280328	258.185525493	585.863734872	359.938256436
sb49	51.67284643	56295.9922901	2770.56024376	1887.47959896	77.353585948	11863.728455	5191.21578466	3316.13168873	49.1988821848	12329.2849801	5059.20713384	3340.75045705
gopatriots	30.7884732182	116.346449281	162.002986782	121.996377185	32.0853710262	100.050457837	193.013870858	118.974803306	26.9163442114	85.089091895	195.628146752	99.3443152621
patriots	337.528333651	17531.6878557	1605.05266139	1565.17484762	311.078252235	6397.26232428	2084.806349	1607.33362833	305.946759473	7115.49362383	2256.09300486	1346.54920589
superbowl	350.442736534	3715.60109272	12624.4784532	4049.27051754	391.662815757	1788.35721044	12873.103	4764.7614765	326.570170631	1730.96289499	12070.3741014	4704.28666981

It is clear that the Random Forest Regressor performs the best across most hashtags and time periods.
The linear model does not perform well because of bursts in tweets within small time intervals.
The higher error values in the second period can be attributed to the relatively lesser number of tweets within this shorter time span, preventing the fitting of a very good model.

The data from all the hashtags are the combined together and 10 fold cross validation is performed. The data is again grouped by the hour and trained on the previous hours data to predict the popularity of the hashtag in the next hour. The Random Forest Model is used for prediction.

period	1	2	3	4
RMSE	861.819942694	25138.1972434	20156.9864592	9015.42504924
MAE	349.188986012	25069.0391667	13023.4550406	2668.51380121

The maximum amount of data is found in the first time period, followed by the third and finally the second. This explains the lower error values for the first period and the highest for the second period.

On comparing the obtained RMSE values with the sum of values obtained over each hashtag for the different models, it is found that the error value of the aggregate for all the periods is lower than the sum of the errors for each hashtag, in the Random Forest Model that is used to train the aggregate dataset as well. A similar trend is observed for the K Nearest Neighbour and Linear models. This can be attributed to the larger dataset and the training of the model over a five hour period. 0

Q1.5:

In this part, we used the best model found in part 1.4(ii), which is random forest, to predict the number of tweets in the next hour on the provided test data files. Each file in the test data contains a hashtag's tweets for a 6 hour window for different periods during the game. We used a window-size of 5 hour instead of one hour to predict the number of tweets in the next hour.

We aggregated the data from all the six files in the training set, divided the data into three parts for three different time periods - before Feb 1, 8:00 am, between Feb 1, 8:00 am and 8:00 pm and after Feb 1, 8:00 pm, and trained three models for these three time periods. Each file from the test set was tested on one of these three models according to the period.

Since we are using the window of size 5, number of features become 50 from 10 and these 50 features are used to make the predictions for the next one-hour window or the 6th hour.

Our predictions are as shown in the table below:

Test File	**Prediction for the next hour**
sample1_period1.txt	451.50716638
sample2_period2.txt	121629.1
sample3_period3.txt	1448.06160057
sample4_period1.txt	666.882538
sample5_period1.txt	556.41099568
sample6_period2.txt	119720.4
sample7_period3.txt	78676.06511113
sample8_period1.txt	11.48449541
sample9_period2.txt	56449.0
sample10_period3.txt	59088.21910277

Q2:

Question

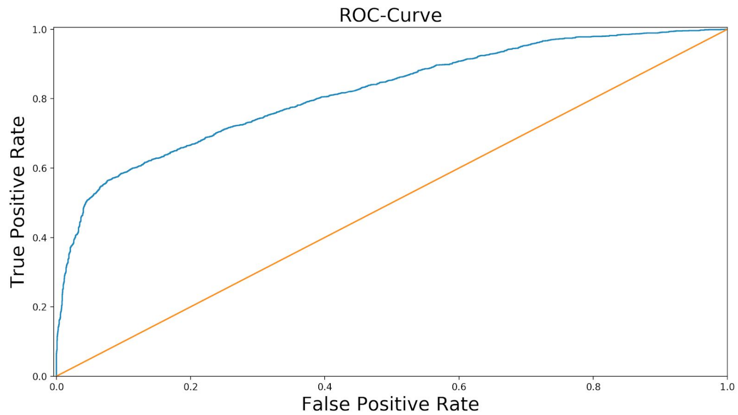
The problems asks us to use different classification algorithms to train a classifier to predict the location of the author of a tweet given only the textual content of the tweet. We consider all the tweets including #superbowl, posted by the users whose specified location is either in the state of Washington or Massachusetts. To evaluate our classifiers, we plot the ROC curve, report the confusion matrix and calculate the accuracy, recall and precision of the classifiers. The classifiers used are Binary Classifier, Logistic Regression Classifier, Naive Bayes Classifier, Multi-layer Perceptron Classifier and Random Forest Classifier.

Preprocessing

In this problem, we use tweets_#superbowl.txt as the dataset. The size of the dataset is around 5.8GB, which is so big that makes it inconvenient for us to manipulate the data. The data set contains many attributes but we only care about two of them, which are the title of tweet and the location of the user. Therefore, we decide to first retrieve the necessary data from the original dataset to speed up our program. The <location> element in the object is not necessarily semantic and hence we need to come up with a keyword-set to do the string matching between the <location> text with our keyword-set. We use the main city names as well as two state names for this purpose.

Once we have our minimised dataset, we use Term Frequency-Inverse Document Frequency (TFxIDF) metric to capture the importance of a word with respect to a document. We also tokenize the documents and exclude the stop words, punctuations, and different stems of a word. After the TF-IDF matrix has been constructed, it is seen that it is a highly sparse matrix and performing classification on a highly sparse matrix does not yield good results. So we perform dimensionality reduction using SVD and use the output of that to train our classifiers. We split the data into train and test set and perform classification.

SVM Classifier



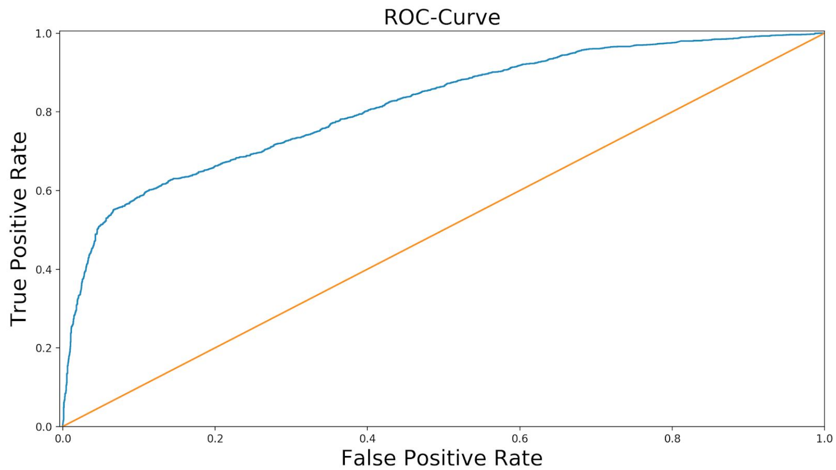
Accuracy: 0.778404952658

Precision: 0.80

Recall: 0.78

```
              Classification report:
=======================================================
Classifier: SVM
=======================================================
               precision    recall  f1-score   support

   Washington       0.75      0.95      0.84      3336
Massachusetts       0.86      0.52      0.65      2156

  avg / total       0.80      0.78      0.76      5492
=======================================================

Confusion Matrix:
==============
[[3154  182]
 [1035 1121]]
==============

Total accuracy:
0.778404952658
```

Logistic Regression Classifier



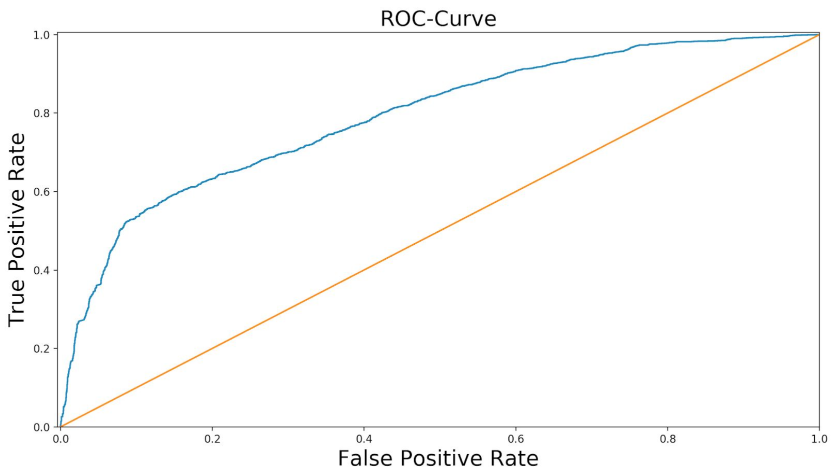
Accuracy: 0.780589949017

Precision: 0.80

Recall: 0.78

```
              Classification report:
=======================================================
Classifier: Logistic Regression
=======================================================
              precision    recall  f1-score   support

   Washington       0.76      0.94      0.84      3336
Massachusetts       0.86      0.53      0.65      2156

  avg / total       0.80      0.78      0.77      5492
=======================================================

Confusion Matrix:
==============
[[3144  192]
 [1013 1143]]
==============

Total accuracy:
0.780589949017
```

Naive Bayes Classifier



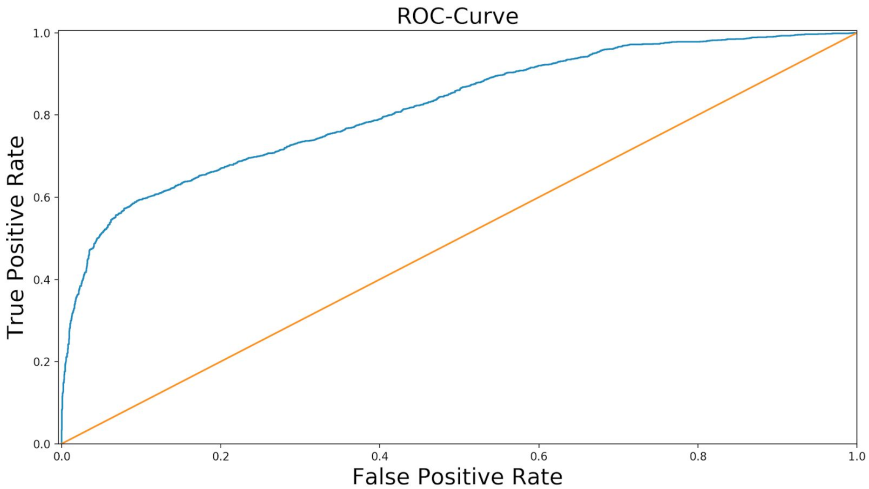
Accuracy: 0.752549162418

Precision: 0.76

Recall: 0.75

```
               Classification report:
=======================================================
Classifier: Naive Bayes
=======================================================
              precision    recall  f1-score   support

   Washington      0.74      0.92      0.82      3336
Massachusetts      0.80      0.49      0.61      2156

  avg / total      0.76      0.75      0.74      5492
=======================================================

Confusion Matrix:
==============
[[3079  257]
 [1102 1054]]
==============

Total accuracy:
0.752549162418
```

Multi-layer Perceptron



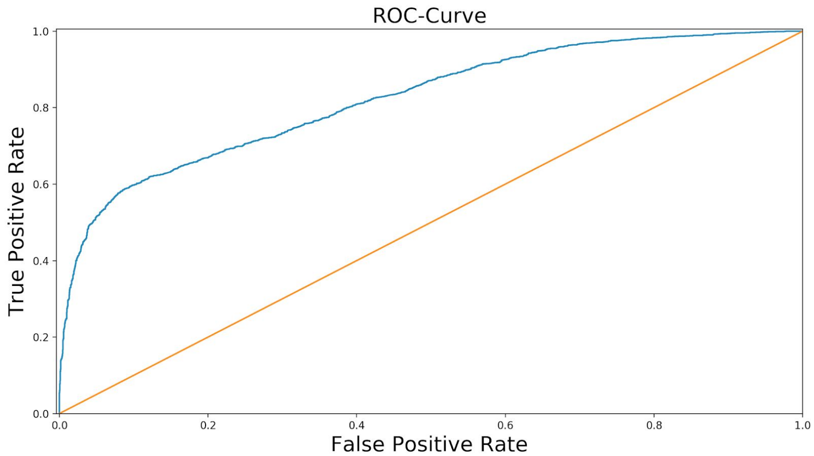
Accuracy: 0.782774945375

Precision: 0.79

Recall: 0.78

```
                Classification report:
==========================================================
Classifier: Multi-layer Perceptron
==========================================================
                precision    recall  f1-score   support

   Washington       0.76      0.93      0.84      3336
Massachusetts       0.84      0.55      0.67      2156

  avg / total       0.79      0.78      0.77      5492
==========================================================

Confusion Matrix:
==============
[[3111  225]
 [ 968 1188]]
==============

Total accuracy:
0.782774945375
```

Random Forest Classifier



Accuracy: 0.783685360524

Precision: 0.79

Recall: 0.78

```
              Classification report:
=======================================================
Classifier: Random Forest
=======================================================
             precision    recall  f1-score   support

   Washington       0.76      0.93      0.84      3336
Massachusetts       0.84      0.56      0.67      2156

  avg / total       0.79      0.78      0.77      5492
=======================================================

Confusion Matrix:
==============
[[3103  233]
 [ 955 1201]]
==============

Total accuracy:
0.783685360524
```

Outcome

Out of the the five classifiers used it can be seen that the best accuracy is obtained by Random Forest Classifier. The best precision is obtained by SVM and Logistic Regression Classifier. The best recall is obtained by all algorithms except Naive Bayes. Random Forest Classifier provides the best overall performance.

Q3: Sentiment Analysis

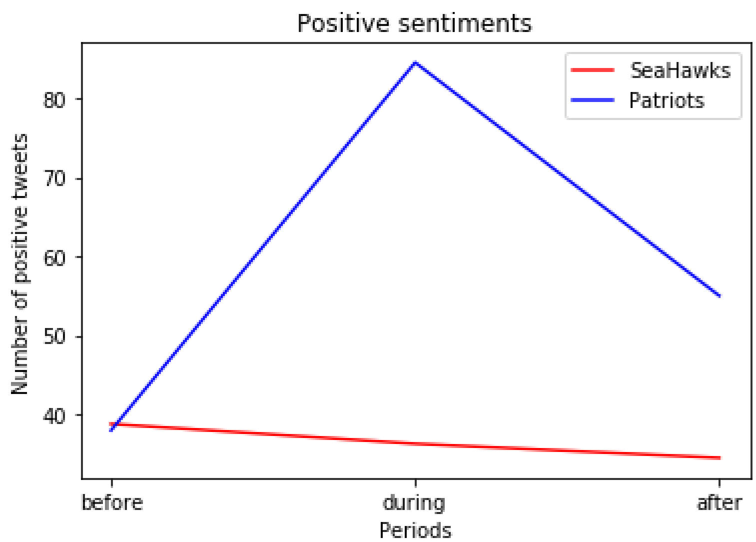
In this part, our task is to define our own project. We decide to analyze the sentiments of the tweets based on locations. On February 1, 2015, the game was played between New England Patriots and Seattle SeaHawks, with Seattle SeaHawks winning the game (28-24). We calculated the percentage of positive, negative and neutral tweets from the Seattle and Massachusetts area and analyzed the drop and rise in the percentage of positive and negative tweets. We observed that there was a huge rise in the percentage of positive tweets from Massachusetts during the game. This means that the team supported by Massachusetts' users i.e. the Patriots would win the game, which is true. This analysis can also be used to predict the results of presidential elections. We can predict which party would win in which state.

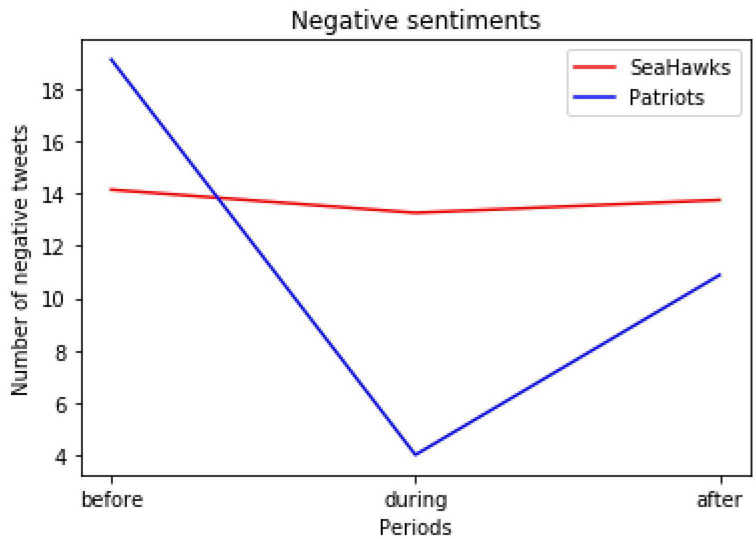
We used the Python TextBlob API to do the sentiment analysis. TextBlob is a Python library for processing textual data and is used for common natural language processing tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, etc.

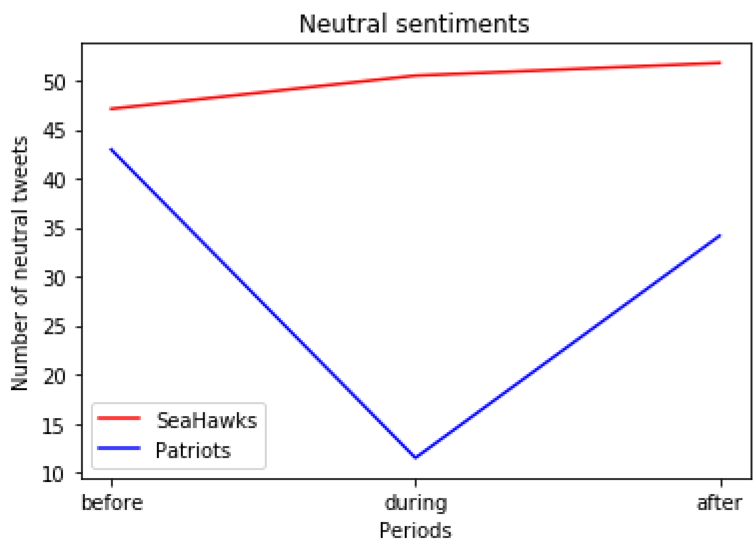
We first clean the tweet's data and then use TextBlob to predict the sentiment of the tweet. The polarity is 0 if the tweet's sentiment is neutral, 1 is it's positive and -1 is it's negative. We perform the sentiment analysis on two datasets: #gohawks and #gopatriots. We divide the data according to the time interval i.e. before the game, during the game and after the game and calculate the number of positive, negative and neutral tweets during each time interval.

Our observations and plots are as follows:

	Sentiment	Before	During	After
#gohawks (Seattle)	**Positive**	38.7605850654	36.2537764350	34.4689993861
	Negative	14.1369429475	13.2552870090	13.7446286065
	Neutral	47.1024719869	50.4909365558	51.7863720073
#gopatriots (Massachusetts)	**Positive**	37.9425937565	84.4979919678	54.9729641160
	Negative	19.1074795725	4.01606425702	10.87989513354
	Neutral	42.9499266708	11.4859437751	34.1471407504







It can be observed from the table and the graph that there's a sudden increase in the percentage of positive tweets during the game and a significant decrease in the percentage of negative and neutral from the users in Massachusetts whereas the number of positive tweets from Seattle users decrease during the game whereas not a big change is observed in the number of neutral and negative comments. From these observations, we can infer that the patriots must have won the game, which is true.

This analysis can further be improved to predict which team scored a goal. We can observe the rise and fall in the number of positive and negative tweets from the Massachusetts and Seattle users to predict which team scored a goal. A sudden rise in the positive tweets from Massachusetts can mean that the patriots scored a goal around that time. This type of sentiment analysis can be very useful for making predictions during the presidential elections.