EE219 Project 5

Popularity Prediction on Twitter
Winter 2017

Shubham Mittal - 104774903
Anshita Mehrotra - 904743371
Swati Arora - 404758379

Introduction:

Twitter, today, is a good platform to predict future popularity of a topic or event. We can perform social network analysis, by knowing current and previous tweet activity for a hash-tag (#), we can predict if it became more prominent and trendy in the future and if yes by how much.

In the project, the twitter data used is collected by querying popular hash-tags related to the 2015 Super Bowl. The data is collected starting from 2 weeks before the game to a week after the game. This data is then used to train a regression model and then the model is used for making predictions for other hash-tags. The test data consists of tweets containing a hash-tag in a specified time window, and we have then used our model to predict number of tweets containing the hash-tag posted within one hour immediately following the given time window.

Part 1: Tweet Data Analysis and Statistics

In this problem, we load calculate some statistics for each hashtag such as average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets.
Every hashtag information is loaded from the corresponding hashtag file into python by reading JSON objects. Each line in the file representing a tweet was parsed to extract information regarding the corresponding hashtag.

We calculate the metrics using following formulas:

$$\text{Average number of tweets per hour} = \frac{\text{Total number of tweets}}{\text{Total number of hours}}$$

$$\text{Average number of retweets per hour} = \frac{\text{Total number of retweets}}{\text{Total number of tweets}}$$

$$\text{Average number of followers per user} = \frac{\text{Total number of followers}}{\text{Total number of unique users}}$$

Metrics for each tag is as follows:

Hashtag	Average number of tweets per hour	Average number of retweets per hour	Average number of followers per user
#gohawks	193.5555	2.01461	1544.9697
#gopatriots	38.4070	1.4000	1298.8242
#nfl	279.4217	1.5385	4289.7466
#patriots	499.1977	1.7828	1650.3219
#sb49	1420.8780	2.5111	2235.1636
#superbowl	1400.5887	2.3882	3591.6044

Bar graph for each hashtag is as follows:

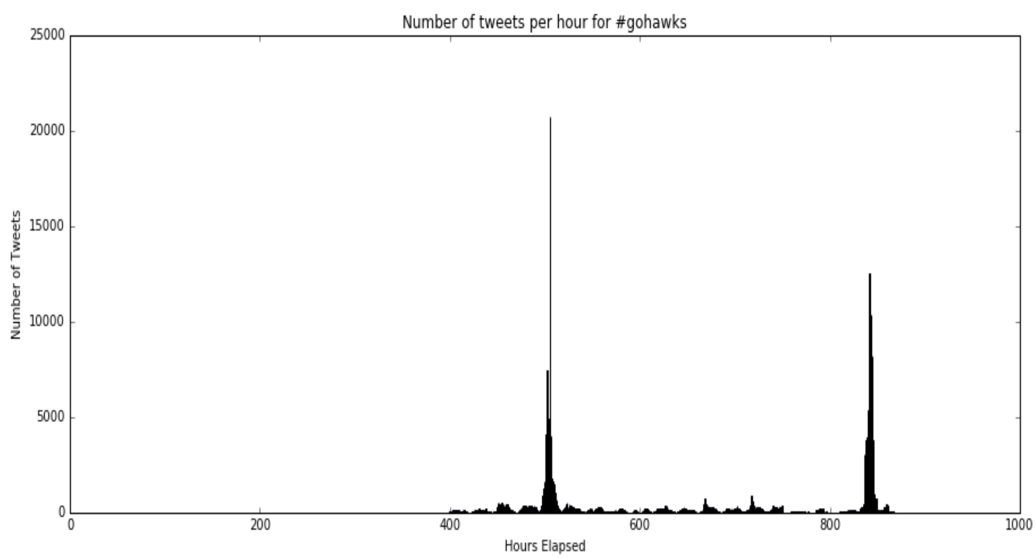


Figure 1 : Figure illustrating Number of tweets per hour for gohawks hashtag over the duration of time

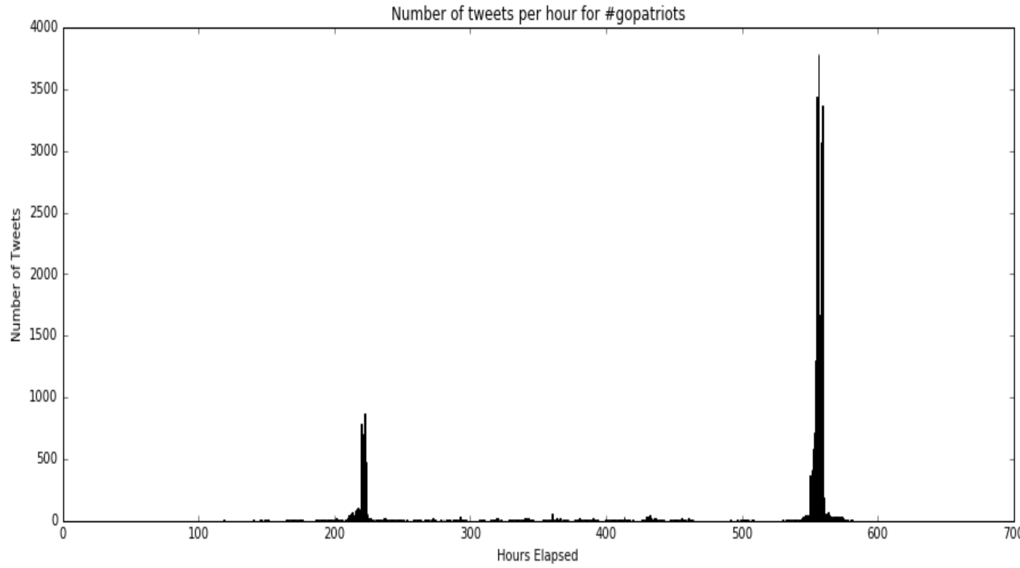


Figure 2: Figure illustrating Number of tweets per hour for gopatriots hashtag over the duration of time

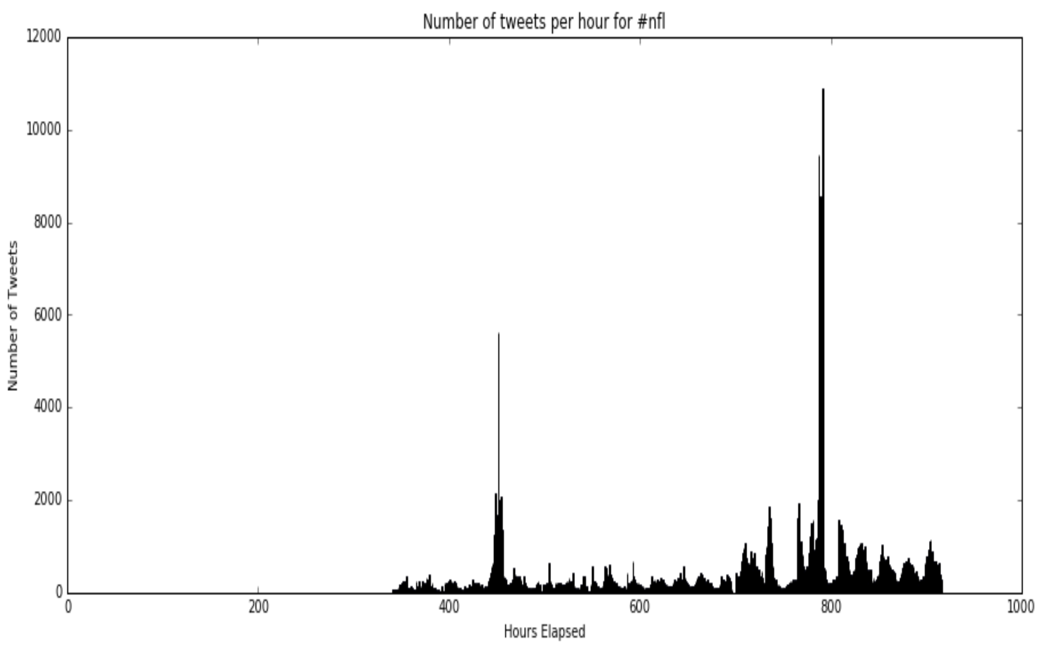


Figure 3: Figure illustrating Number of tweets per hour for nfl hashtag over the duration of time

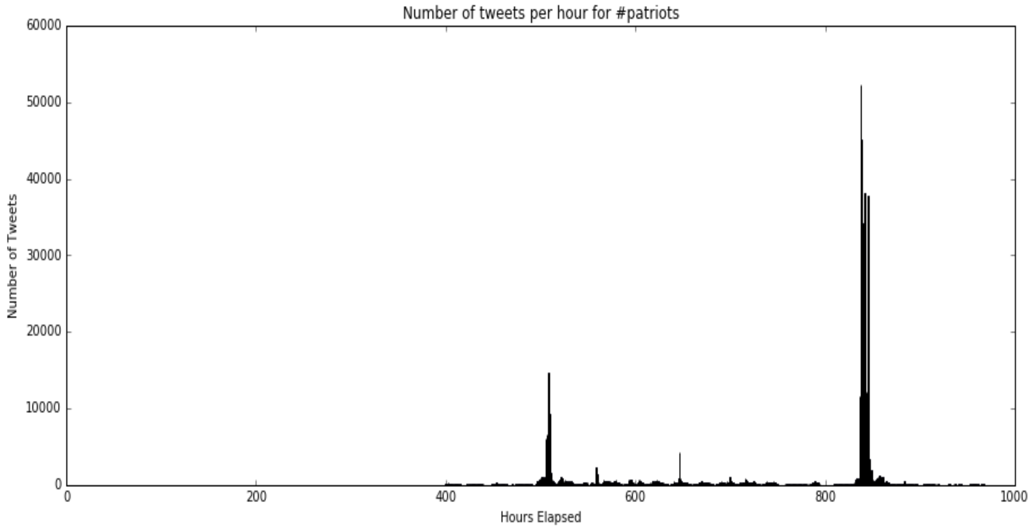


Figure 4 : Figure illustrating number of tweets per hour for patriot's hashtag over the duration of time

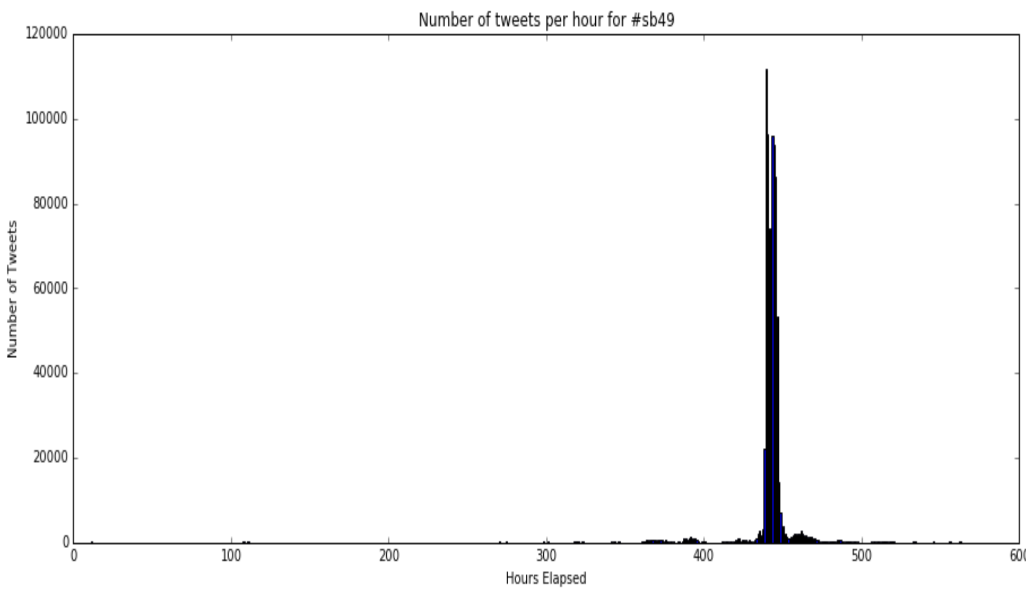


Figure 5 : Figure illustrating number of tweets per hour for sb49 hashtag over the duration of time

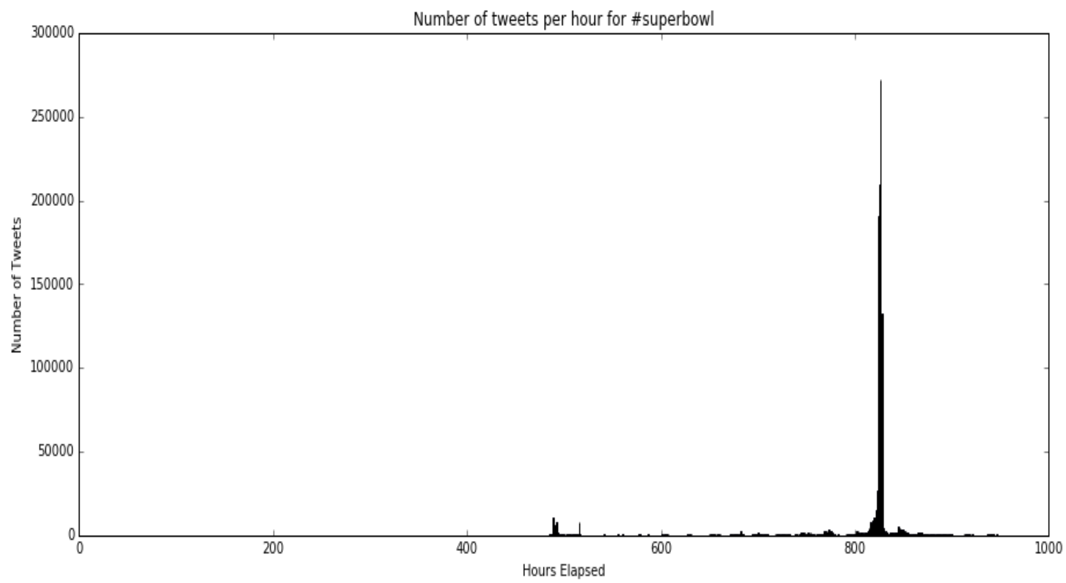


Figure 6 : Figure illustrating number of tweets per hour for superbowl hashtag over the duration of time

Analysis of Statistics:

Most retweeted Hashtag on an average	#superbowl and #sb49
Most tweeted Hashtag on an average	#superbowl and #sb49
Most followers to users tweeting hashtag	#nfl and #superbowl

This analysis can also be visualized from Figure 6 (superbowl) , Figure 5 (sb49) and Figure 3(nfl). Figure 5 and Figure 6 show steep rise which is indicative of high activity for that hashtag.

Part 2: Linear Regression

In this problem, we make predictions for tweets in the next hour based on the features collected from previous hour tweet. The features on which this linear regression model is build are as follows:

1. Number of tweets
2. Total number of retweets
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 (indicating an hour out of 24 hours)

An hour window approach same as problem I is employed to calculate the features information for any given hour to make predictions for next hour.

We obtained the summarized results after fitting the linear regression model for each hashtag. Summarized reports for each hashtag are as follows:

1) #gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.488
Model:                            OLS   Adj. R-squared:                  0.486
Method:                 Least Squares   F-statistic:                     184.6
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          5.44e-138
Time:                        03:03:45   Log-Likelihood:                -7818.8
No. Observations:                 973   AIC:                         1.565e+04
Df Residuals:                     967   BIC:                         1.568e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         66.4568     46.714      1.423      0.155       -25.216   158.129
x1             0.0004   8.16e-05      4.480      0.000         0.000     0.001
x2            -0.1657      0.043     -3.825      0.000        -0.251    -0.081
x3             1.9230      3.485      0.552      0.581        -4.915     8.761
x4             0.5770      0.121      4.750      0.000         0.339     0.815
x5            -0.0006      0.000     -4.711      0.000        -0.001    -0.000
==============================================================================
Omnibus:                     1843.210   Durbin-Watson:                   2.337
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4367808.162
Skew:                          13.186   Prob(JB):                         0.00
Kurtosis:                     330.171   Cond. No.                     3.17e+06
==============================================================================
```

Report 1 : OLS Regression Results for gohawks hashtag. R-squared accuracy can be observed as 0.488 indicating not a good fir to the model for this hashtag.

2) #superbowl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     552.4
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          3.18e-279
Time:                        03:11:20   Log-Likelihood:                -9919.1
No. Observations:                 964   AIC:                         1.985e+04
Df Residuals:                     958   BIC:                         1.988e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        274.0961    456.701      0.600      0.549      -622.155  1170.347
x1            -0.0004   2.58e-05    -13.831      0.000        -0.000    -0.000
x2             0.0247      0.126      0.196      0.845        -0.222     0.271
x3           -12.0668     33.373     -0.362      0.718       -77.559    53.425
x4             1.6753      0.258      6.493      0.000         1.169     2.182
x5             0.0013      0.000      9.867      0.000         0.001     0.002
==============================================================================
Omnibus:                     1889.772   Durbin-Watson:                   1.699
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5798609.766
Skew:                          14.133   Prob(JB):                         0.00
Kurtosis:                     381.899   Cond. No.                     9.08e+07
==============================================================================
```

Report 2 : OLS Regression Results for superbowl hashtag. R-squared accuracy can be observed as 0.742 indicating a good fit to the model for this hashtag.

3) #gopatriots

```
|                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                     268.0
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          6.85e-158
Time:                        03:03:50   Log-Likelihood:                -4453.7
No. Observations:                 684   AIC:                             8919.
Df Residuals:                     678   BIC:                             8947.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          8.0759     12.238      0.660      0.510       -15.952    32.104
x1             0.0011      0.000      5.358      0.000         0.001     0.002
x2             0.4126      0.260      1.588      0.113        -0.098     0.923
x3             0.1524      0.907      0.168      0.867        -1.629     1.934
x4            -0.5873      0.239     -2.455      0.014        -1.057    -0.118
x5            -0.0012      0.000     -6.290      0.000        -0.002    -0.001
==============================================================================
Omnibus:                      794.712   Durbin-Watson:                   2.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           452279.898
Skew:                           4.816   Prob(JB):                         0.00
Kurtosis:                     128.605   Cond. No.                     6.45e+05
==============================================================================
```

Report 3 : OLS Regression Results for gopatriots hashtag. R-squared accuracy can be observed as 0.664 indicating a decent fit to the model for this hashtag

4) #sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     528.7
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          9.98e-213
Time:                        03:08:03   Log-Likelihood:                -5702.2
No. Observations:                 583   AIC:                         1.142e+04
Df Residuals:                     577   BIC:                         1.144e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        162.1198    349.674      0.464      0.643      -524.670   848.909
x1             0.0002   2.96e-05      7.417      0.000         0.000     0.000
x2            -0.3676      0.043     -8.472      0.000        -0.453    -0.282
x3           -16.3934     25.989     -0.631      0.528       -67.438    34.652
x4             1.1411      0.052     21.904      0.000         1.039     1.243
x5            -0.0003   6.91e-05     -4.106      0.000        -0.000    -0.000
==============================================================================
Omnibus:                     1163.209   Durbin-Watson:                   1.726
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2251571.836
Skew:                          14.043   Prob(JB):                         0.00
Kurtosis:                     306.150   Cond. No.                     6.19e+07
==============================================================================
```

Report 4 : OLS Regression Results for sb49 hashtag. R-squared accuracy can be observed as 0.821 indicating a very good fit to the model for this hashtag.

5) #nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.604
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                     281.3
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          1.39e-182
Time:                        03:04:36   Log-Likelihood:                -6999.8
No. Observations:                 927   AIC:                         1.401e+04
Df Residuals:                     921   BIC:                         1.404e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         61.6215     29.813      2.067      0.039         3.111   120.132
x1            -0.0001    2.5e-05     -5.701      0.000        -0.000 -9.34e-05
x2            -0.1778      0.065     -2.718      0.007        -0.306    -0.049
x3            -1.2123      2.197     -0.552      0.581        -5.524     3.100
x4             1.3406      0.110     12.223      0.000         1.125     1.556
x5             0.0002   3.38e-05      5.815      0.000         0.000     0.000
==============================================================================
Omnibus:                     1046.976   Durbin-Watson:                   2.159
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1267037.679
Skew:                           4.467   Prob(JB):                         0.00
Kurtosis:                     183.897   Cond. No.                     5.42e+06
==============================================================================
```

Report 5 : OLS Regression Results for nfl hashtag. R-squared accuracy can be observed as 0.604 indicating a decent fit to the model for this hashtag.

6) #patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     491.6
Date:                Wed, 22 Mar 2017   Prob (F-statistic):         1.51e-263
Time:                        03:05:54   Log-Likelihood:                -8761.5
No. Observations:                 981   AIC:                         1.754e+04
Df Residuals:                     975   BIC:                         1.756e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        136.2579    114.513      1.190      0.234       -88.462   360.978
x1             0.0003   4.28e-05      7.783      0.000         0.000     0.000
x2            -0.9485      0.073    -13.027      0.000        -1.091    -0.806
x3            -1.4698      8.488     -0.173      0.863       -18.126    15.186
x4             1.7832      0.079     22.500      0.000         1.628     1.939
x5            -0.0002   8.94e-05     -2.751      0.006        -0.000 -7.05e-05
==============================================================================
Omnibus:                     1875.685   Durbin-Watson:                   1.696
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4060230.796
Skew:                          13.536   Prob(JB):                         0.00
Kurtosis:                     317.006   Cond. No.                     9.74e+06
==============================================================================
```

Report 6 : OLS Regression Results for patriots hashtag. R-squared accuracy can be observed as 0.716 indicating a good fit to the model for this hashtag.

The results obtained for each hashtag shown above are summarized below:

Hashtag	Accuracy value
#gohawks	48.8329
#superbowl	74.2456
#gopatriots	66.4003
#sb49	82.0842
#nfl	60.4277
#patriots	71.6007

Table 1 : Model accuracy for each hashtag

Hashtag	Maximum followers count	Retweet count	Time	Tweet Count	Follower Count
#gohawks	8.345 * 10-6	1.392 * 10-4	5.811 * 10-1	2.347 * 10-6	2.826 * 10-6
#superbowl	8.458 * 10-40	8.445 * 10-1	7.177 * 10-1	1.351 * 10-10	6.144 * 10-22
#gopatriots	1.155 * 10-7	1.127 * 10-1	8.666 * 10-1	1.433 *10-2	5.690 * 10-10
#sb49	4.308 * 10-13	2.014 * 10-16	5.284 * 10-1	7.413 *10-78	4.600 * 10-5

#nfl	1.606 * 10-8	6.691 * 10-3	5.812 * 10-1	6.032 * 10-32	8.350 * 10-9
#patriots	1.806 * 10-14	7.023 * 10-36	8.625 * 10-1	1.257 * 10-90	6.057 * 10-3

Table 2 : p-values for model parameters

Hashtag	Maximum followers count	Retweet count	Time	Tweet Count	Follower Count
#gohawks	4.478	-3.824	0.551	4.747	-4.710
#superbowl	-13.830	0.196	-0.361	6.492	9.867
#gopatriots	5.357	1.587	0.167	-2.455	-6.290
#sb49	7.416	-8.471	-0.630	21.904	-4.106
#nfl	-5.700	-2.717	-0.551	12.222	5.815
#patriots	7.782	-13.026	-0.173	22.499	-2.750

Table 3 : t-values for model parameters

Analysis of results:

1. **Accuracy:**
   a. We have used R-squared accuracy as a parameter to analyze the fit to the data. Higher R-squared accuracy indicates a better fit of model to the data.
   b. It can be observed that model fits the data better for #sb49 and #superbowl with accuracies as 82% and 74.2%

2. **P-value and t-value**:
   a. T-value is used to statistically determine significance of a variable in the model. It is important to pick the significant features while designing the regression model. The larger the absolute value of t, less likely the value of parameter could be zero. Thus we would want to select features with higher t-value.
   b. Lesser the p-value, higher the probability of being important feature for the model. Thus we want to select features with lower p-value.
   c. Based on analysis of p-value and t-value, significant features for each hashtag are as follows:
      i. #gohawks : Tweet count, Maximum followers count
      ii. #superbowl : Tweet count, Follower count
      iii. #gopatriots : Maximum followers count, Retweet count
      iv. #sb49 : Tweet count, Maximum followers count
      v. #nfl : Tweet count, Follower count
      vi. #patriots : Tweet count, Maximum followers count
   d. It can be observed that Tweet count and Maximum followers count are important features across all hashtags

Part 3: Regression Model With Extra Features

In this problem, we had to add additional features to regression model and analyze the significant features which can be used for the model and accuracy obtained for the model. We have used 11 features which are as follows:

1) Total Favorite Count
2) Ranking Score
3) Impression Count
4) Number of tweets
5) Individual listed frequency
6) Individual mention frequency
7) Total follower count
8) Number of retweets
9) Number of long tweets
10) Time
11) Number of maximum followers

#gohawks

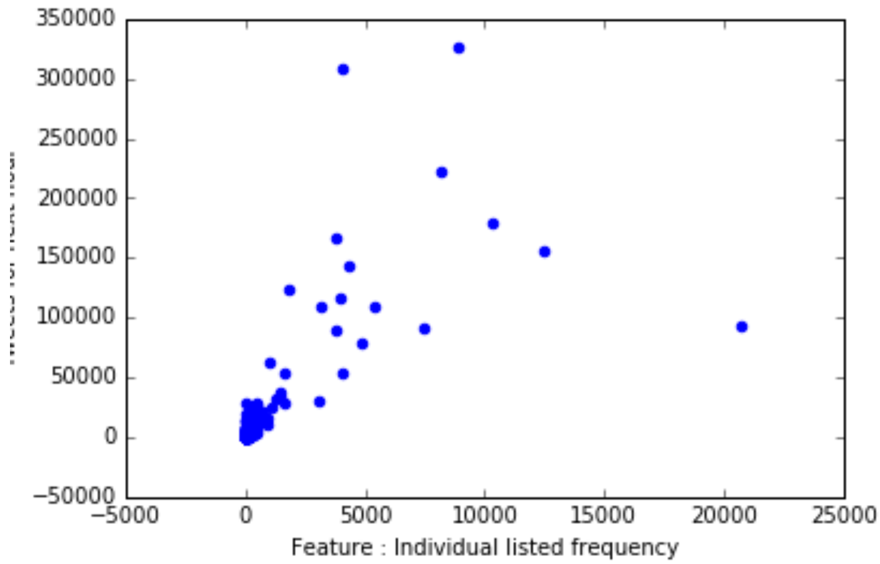
```
|                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.638
Model:                            OLS   Adj. R-squared:                  0.634
Method:                 Least Squares   F-statistic:                     154.0
Date:                Wed, 22 Mar 2017   Prob (F-statistic):         2.60e-203
Time:                        22:49:00   Log-Likelihood:                -7650.4
No. Observations:                 973   AIC:                         1.532e+04
Df Residuals:                     961   BIC:                         1.538e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        -20.2428     39.920     -0.507      0.612       -98.583    58.098
x1             0.1059      0.022      4.745      0.000         0.062     0.150
x2             3.9687      0.660      6.013      0.000         2.674     5.264
x3           1.94e-05   5.79e-05      0.335      0.737     -9.42e-05     0.000
x4           -18.5137      3.165     -5.849      0.000       -24.725   -12.302
x5             0.0707      0.005     13.273      0.000         0.060     0.081
x6             3.5554      0.518      6.858      0.000         2.538     4.573
x7            -0.0005      0.000     -4.277      0.000        -0.001    -0.000
x8            -0.3409      0.061     -5.547      0.000        -0.461    -0.220
x9            -2.9283      0.869     -3.369      0.001        -4.634    -1.222
x10           -0.4335      2.957     -0.147      0.884        -6.237     5.370
x11          6.102e-05     0.000      0.463      0.643        -0.000     0.000
==============================================================================
Omnibus:                     1825.420   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3619146.183
Skew:                          12.987   Prob(JB):                         0.00
Kurtosis:                     300.649   Cond. No.                     5.45e+06
==============================================================================
```
Report 7 : OLS Regression Results for gohawks hashtag with addition of new features

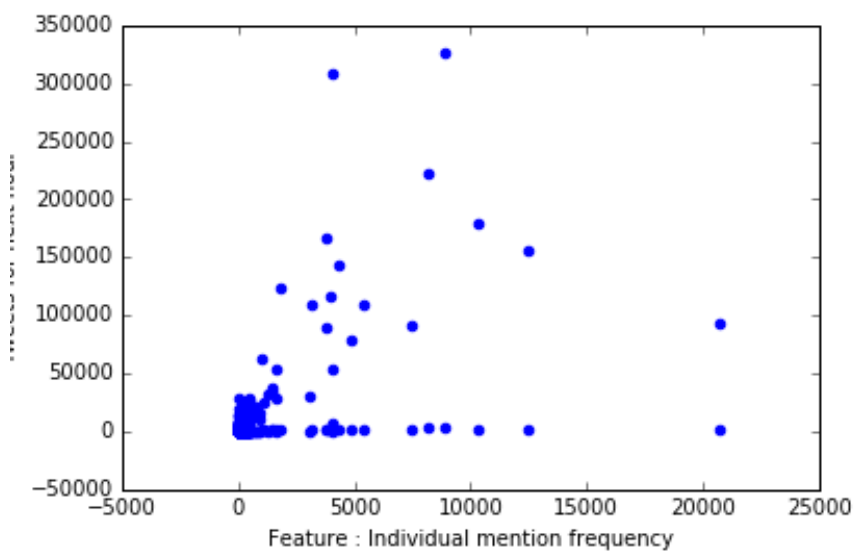
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

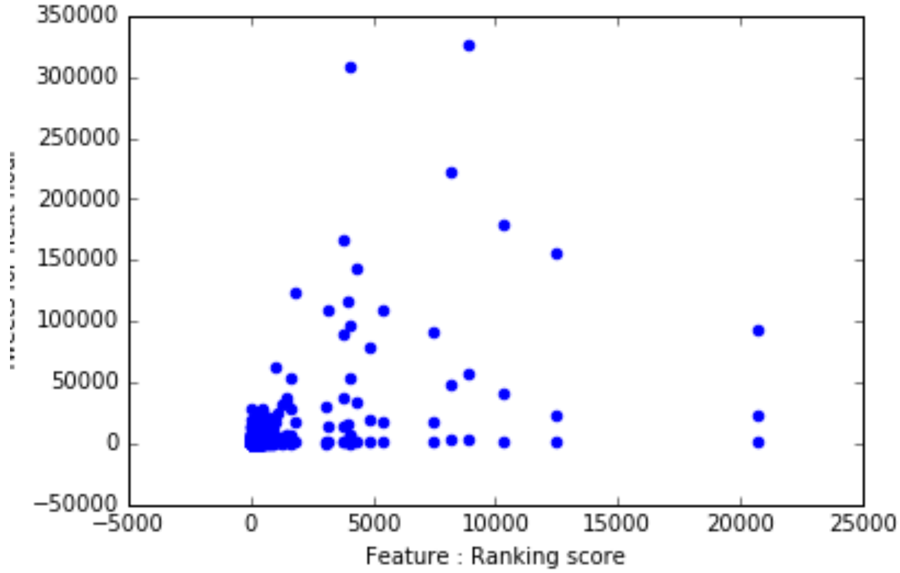
Top 3 Features for #gohawks :

1) Individual listed frequency
2) Individual mention frequency
3) Ranking Score

Scatter plot for these features are as follows:


350000
300000
250000
200000
150000
100000
50000
0
−50000
−5000
0
5000
10000
15000
20000
25000
Feature : Individual listed frequency


350000
300000
250000
200000
150000
100000
50000
0
−50000
−5000
0
5000
10000
15000
20000
25000
Feature : Individual mention frequency



350000
300000
250000
200000
150000
100000
50000
0
−50000
−5000
0
5000
10000
15000
20000
25000
Feature : Ranking score

#gopatriots

```
|                           OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.845
Model:                            OLS   Adj. R-squared:                  0.843
Method:                 Least Squares   F-statistic:                     333.7
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          1.08e-263
Time:                        22:49:05   Log-Likelihood:                -4188.5
No. Observations:                 684   AIC:                             8401.
Df Residuals:                     672   BIC:                             8455.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -0.9249      8.364     -0.111      0.912       -17.347    15.498
x1           -12.5521      1.246    -10.071      0.000       -14.999   -10.105
x2             4.1187      0.371     11.096      0.000         3.390     4.848
x3            -0.0016      0.000     -7.770      0.000        -0.002    -0.001
x4           -19.1470      1.764    -10.853      0.000       -22.611   -15.683
x5            -0.0397      0.011     -3.454      0.001        -0.062    -0.017
x6             3.2204      0.651      4.944      0.000         1.941     4.499
x7             0.0027      0.000      7.281      0.000         0.002     0.003
x8            -1.2168      0.209     -5.810      0.000        -1.628    -0.806
x9             6.7937      0.967      7.024      0.000         4.895     8.693
x10           -0.2536      0.621     -0.408      0.683        -1.474     0.967
x11           -0.0008      0.000     -4.907      0.000        -0.001    -0.000
==============================================================================
Omnibus:                      745.187   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           351919.307
Skew:                           4.309   Prob(JB):                         0.00
Kurtosis:                     113.787   Cond. No.                     9.84e+05
==============================================================================
```

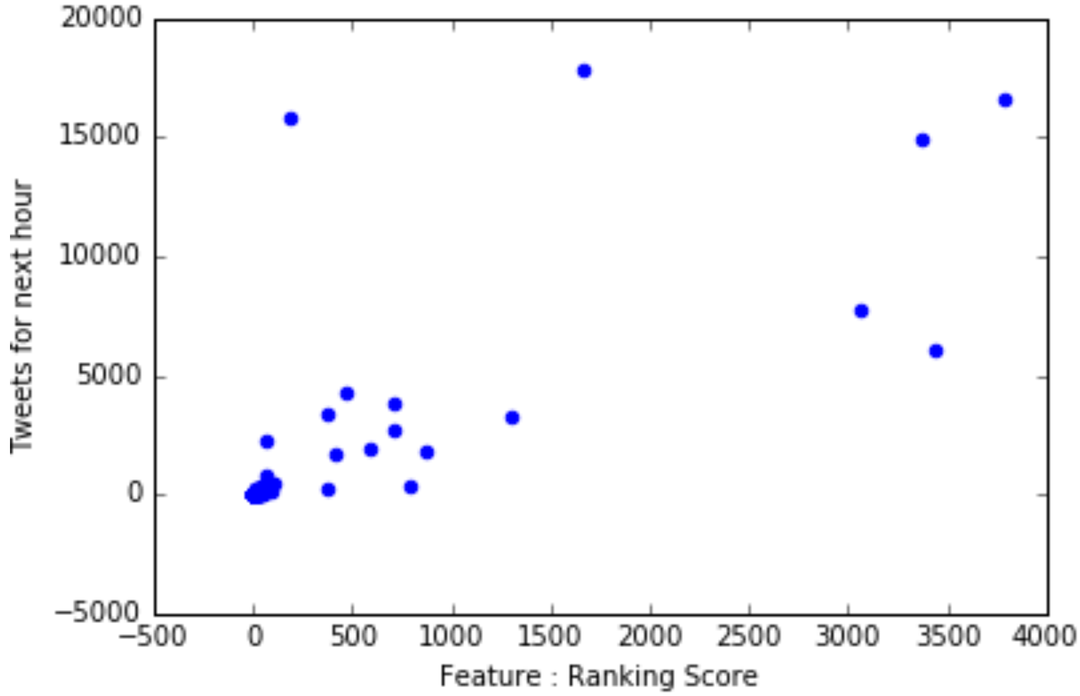
Report 8 : OLS Regression Results for gopatriots hashtag with addition of new features

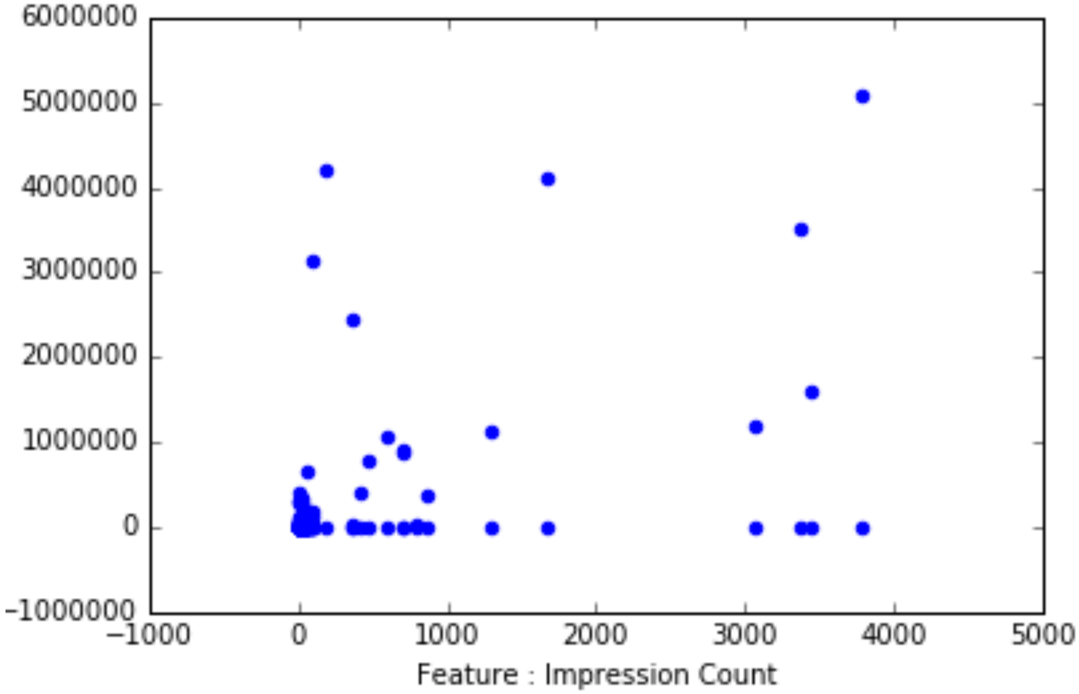
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

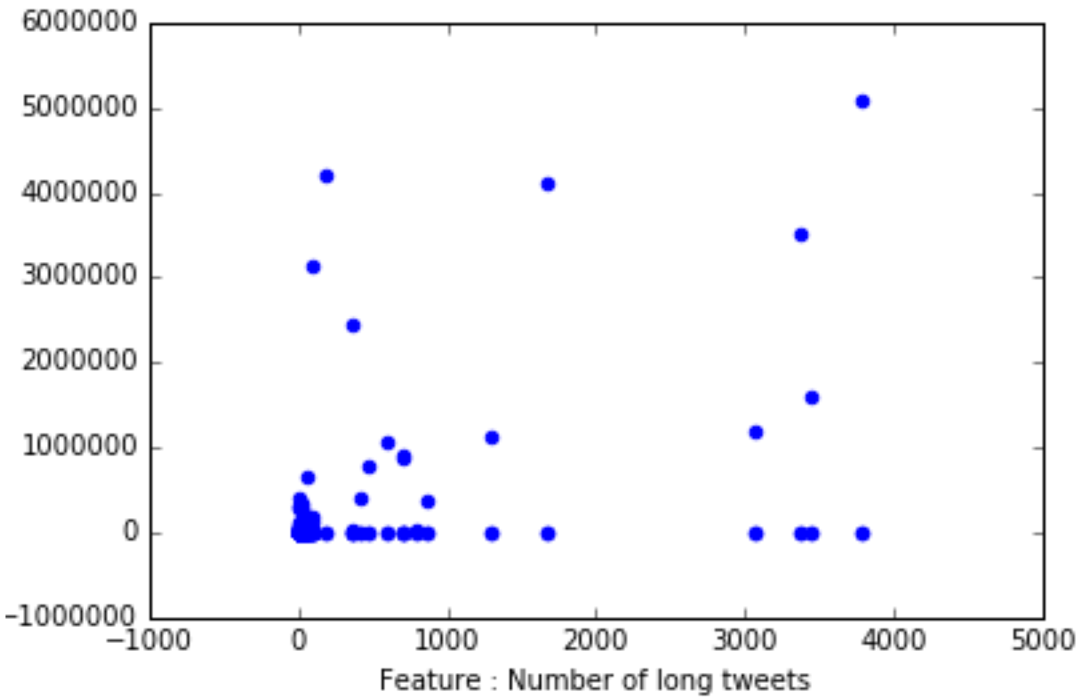
Top 3 Features for #gopatriots

1) Ranking Score
2) Number of long tweets
3) Impression count

Scatter plot for these features are as follows:




6000000
5000000
4000000
3000000
2000000
1000000
0
−1000000
−1000
0
1000
2000
3000
4000
5000
Feature : Impression Count


6000000
5000000
4000000
3000000
2000000
1000000
0
−1000000
−1000
0
1000
2000
3000
4000
5000
Feature : Number of long tweets

#nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.755
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     256.7
Date:                Wed, 22 Mar 2017   Prob (F-statistic):         1.13e-270
Time:                        22:49:52   Log-Likelihood:                -6777.1
No. Observations:                 927   AIC:                         1.358e+04
Df Residuals:                     915   BIC:                         1.364e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         38.9574     24.585      1.585      0.113        -9.292    87.207
x1            -2.2327      0.129    -17.359      0.000        -2.485    -1.980
x2            -1.0340      0.246     -4.195      0.000        -1.518    -0.550
x3         -3.699e-05   1.65e-05     -2.247      0.025     -6.93e-05 -4.68e-06
x4             5.7848      1.137      5.088      0.000         3.553     8.016
x5             0.0279      0.003     10.115      0.000         0.022     0.033
x6             1.7898      0.442      4.052      0.000         0.923     2.657
x7            -0.0003   3.44e-05     -7.963      0.000        -0.000    -0.000
x8            -0.0758      0.056     -1.366      0.172        -0.185     0.033
x9            -1.0138      0.213     -4.758      0.000        -1.432    -0.596
x10           -1.7927      1.739     -1.031      0.303        -5.207     1.621
x11            0.0003   3.15e-05      8.198      0.000         0.000     0.000
==============================================================================
Omnibus:                     1408.854   Durbin-Watson:                   2.288
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           656232.179
Skew:                           8.718   Prob(JB):                         0.00
Kurtosis:                     132.174   Cond. No.                     9.48e+06
==============================================================================
```

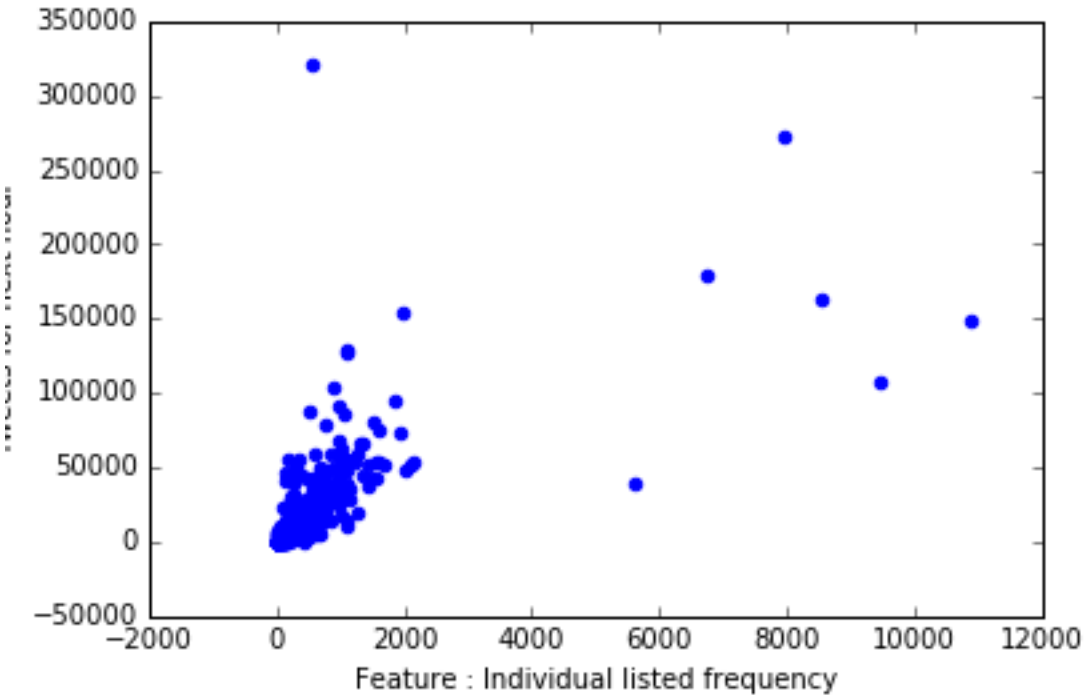
Report 9 : OLS Regression Results for nfl hashtag with addition of new features

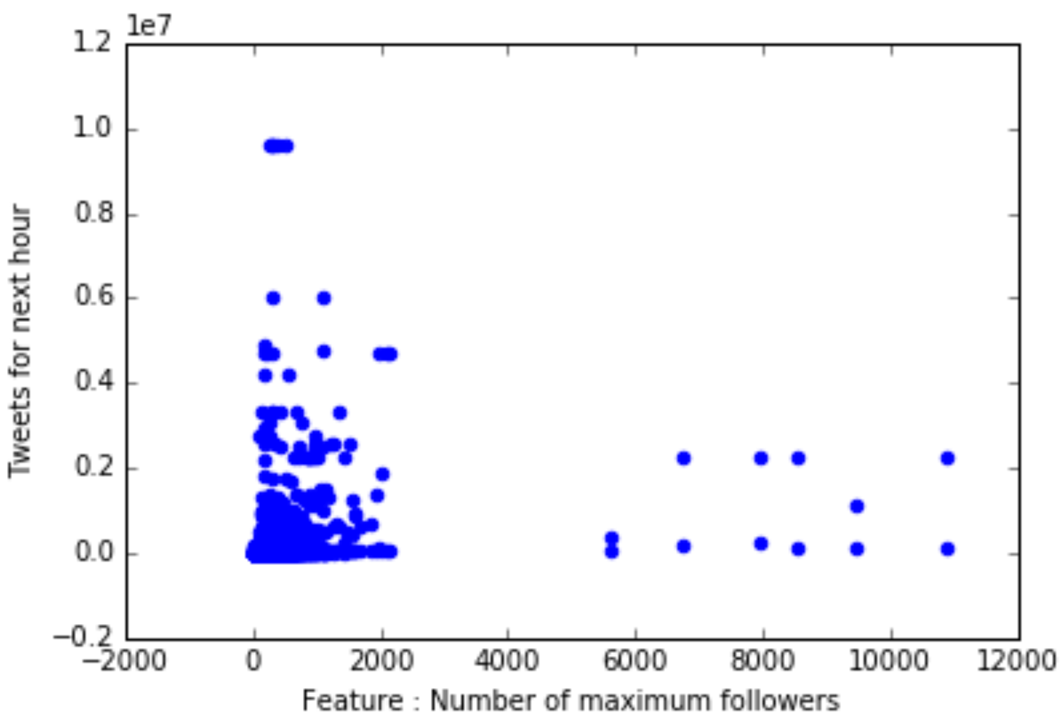
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

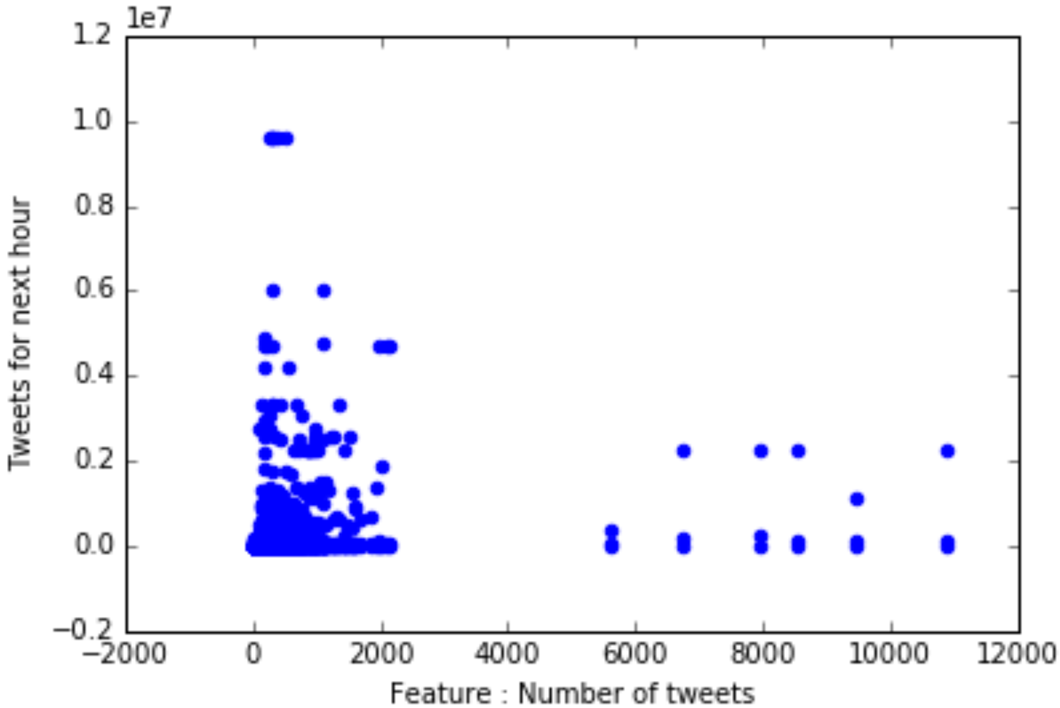
Top 3 Features for #nfl

1) Individual listed frequency
2) Number of maximum followers
3) Number of tweets

Scatter plot for these features are as follows:







1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Tweets for next hour
−2000
0
2000
4000
6000
8000
10000
12000
Feature : Number of tweets

#patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.775
Method:                 Least Squares   F-statistic:                     308.7
Date:                Wed, 22 Mar 2017   Prob (F-statistic):         1.67e-307
Time:                        22:51:20   Log-Likelihood:                -8640.8
No. Observations:                 981   AIC:                         1.731e+04
Df Residuals:                     969   BIC:                         1.736e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        -25.2626    104.137     -0.243      0.808      -229.622   179.097
x1             0.0086      0.180      0.048      0.962        -0.344     0.361
x2             4.4899      0.481      9.338      0.000         3.546     5.433
x3         -5.619e-05    6.9e-05     -0.814      0.416        -0.000  7.93e-05
x4           -20.9754      2.105     -9.962      0.000       -25.107   -16.844
x5            -0.0051      0.007     -0.757      0.449        -0.018     0.008
x6             1.8656      0.138     13.539      0.000         1.595     2.136
x7             0.0007      0.000      5.295      0.000         0.000     0.001
x8            -0.3758      0.133     -2.822      0.005        -0.637    -0.115
x9             1.1963      0.610      1.963      0.050         0.000     2.393
x10           -0.7473      7.548     -0.099      0.921       -15.559    14.064
x11           -0.0008      0.000     -7.433      0.000        -0.001    -0.001
==============================================================================
Omnibus:                     1997.270   Durbin-Watson:                   1.708
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5732115.574
Skew:                          15.535   Prob(JB):                         0.00
Kurtosis:                     376.189   Cond. No.                     1.72e+07
==============================================================================
```

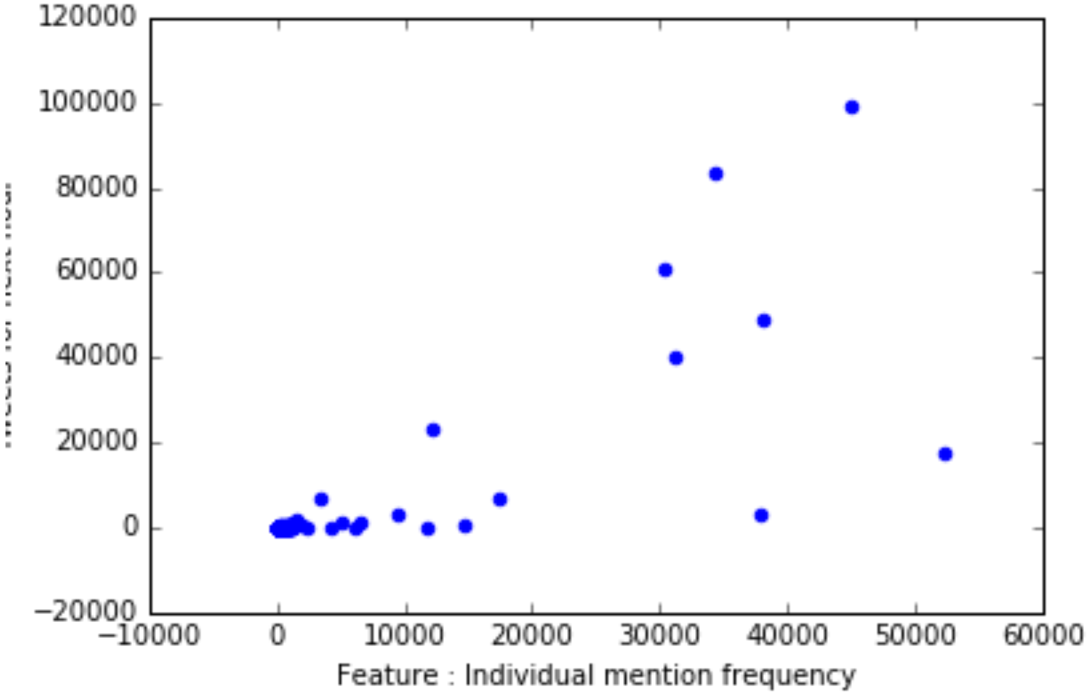
Report 10 : OLS Regression Results for patriots hashtag with addition of new features

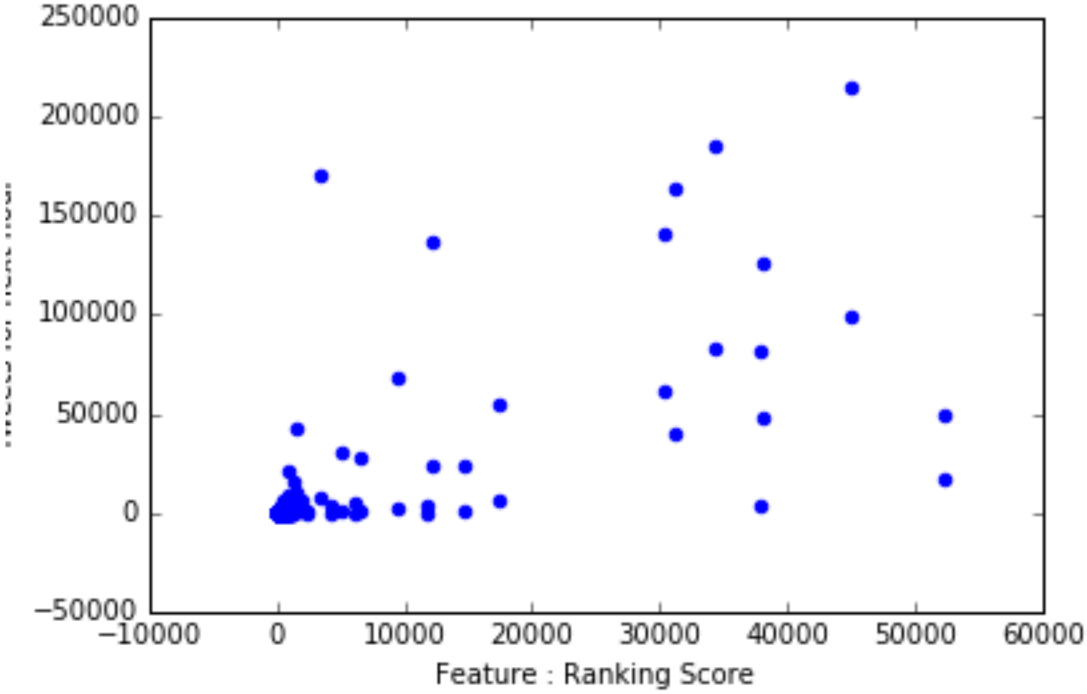
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

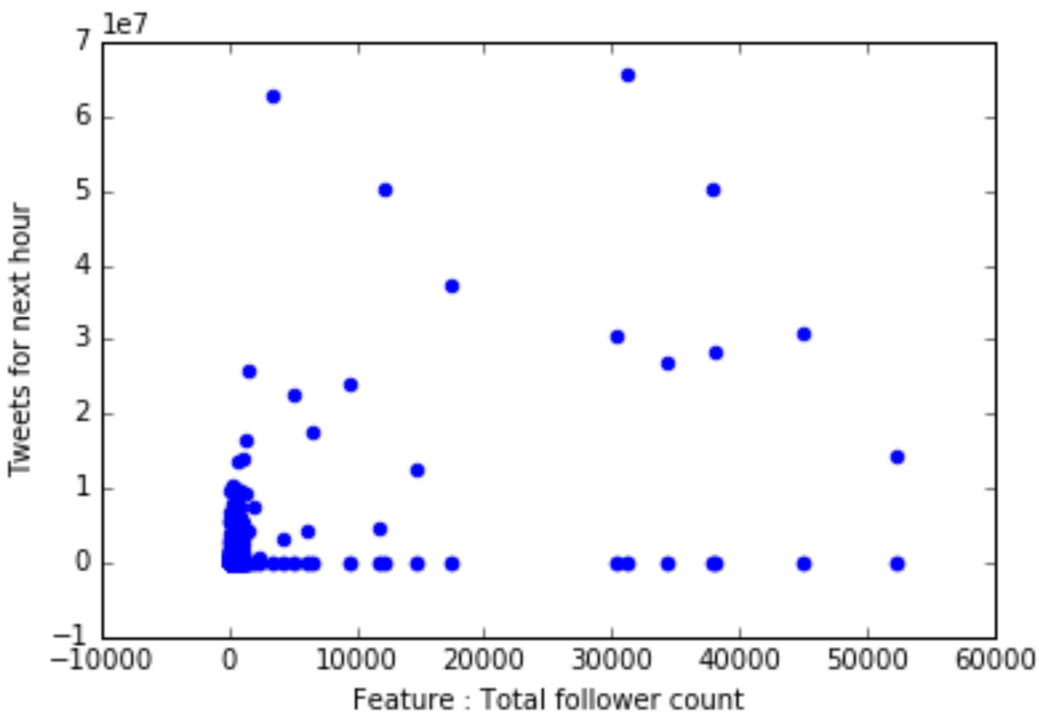
Top 3 Features for #patriots

1) Individual mention frequency
2) Ranking Score
3) Total follower count

Scatter plot for these features are as follows:




250000
200000
150000
100000
50000
0
−50000
−10000
0
10000
20000
30000
40000
50000
60000
Feature : Ranking Score


1e7
7
6
5
4
3
2
1
0
−1
Tweets for next hour
−10000
0
10000
20000
30000
40000
50000
60000
Feature : Total follower count

#sb49

```
|                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.873
Model:                            OLS   Adj. R-squared:                  0.870
Method:                 Least Squares   F-statistic:                     355.9
Date:                Wed, 22 Mar 2017   Prob (F-statistic):          3.19e-247
Time:                        22:53:53   Log-Likelihood:                -5602.5
No. Observations:                 583   AIC:                         1.123e+04
Df Residuals:                     571   BIC:                         1.128e+04
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const       -177.3666    302.685     -0.586      0.558      -771.878   417.144
x1            -0.5351      0.088     -6.089      0.000        -0.708    -0.362
x2             6.2547      0.820      7.630      0.000         4.645     7.865
x3            -0.0001    2.7e-05     -5.133      0.000        -0.000 -8.55e-05
x4           -31.1991      3.929     -7.941      0.000       -38.916   -23.482
x5            -0.0029      0.008     -0.360      0.719        -0.019     0.013
x6             2.7048      0.409      6.608      0.000         1.901     3.509
x7             0.0005    8.7e-05      5.843      0.000         0.000     0.001
x8             0.6434      0.108      5.955      0.000         0.431     0.856
x9             1.1311      1.967      0.575      0.565        -2.732     4.994
x10            0.8389     22.189      0.038      0.970       -42.743    44.421
x11           -0.0007   7.81e-05     -8.688      0.000        -0.001    -0.001
==============================================================================
Omnibus:                     1115.173   Durbin-Watson:                   1.848
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1686259.107
Skew:                          12.862   Prob(JB):                         0.00
Kurtosis:                     265.213   Cond. No.                     1.84e+08
==============================================================================
```

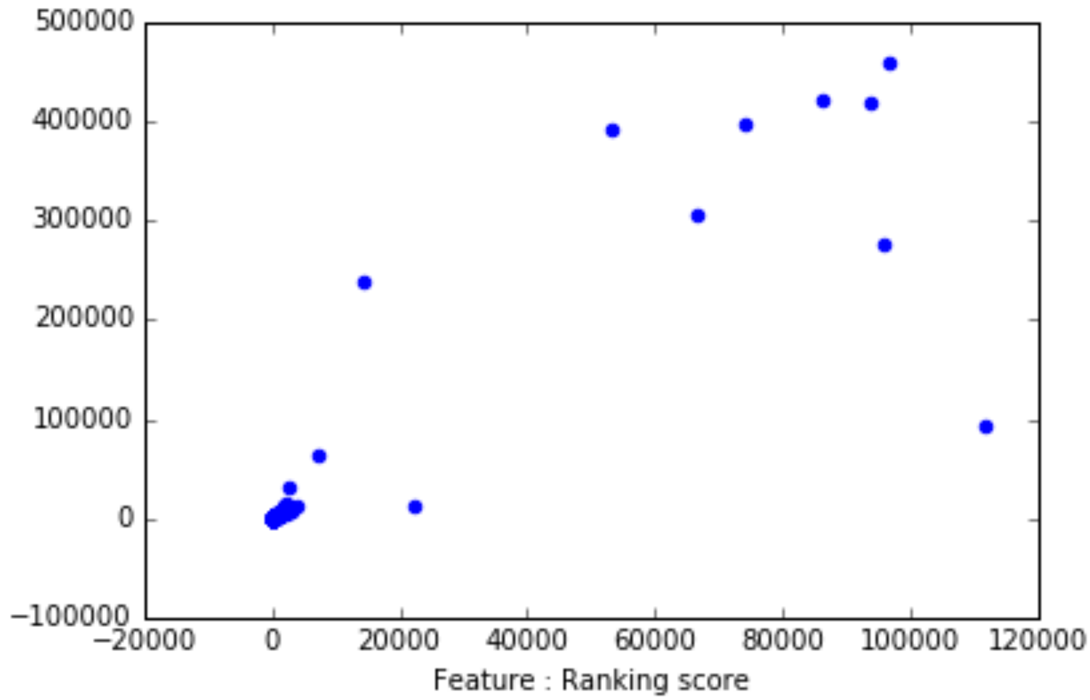
Report 11 : OLS Regression Results for sb49 hashtag with addition of new features

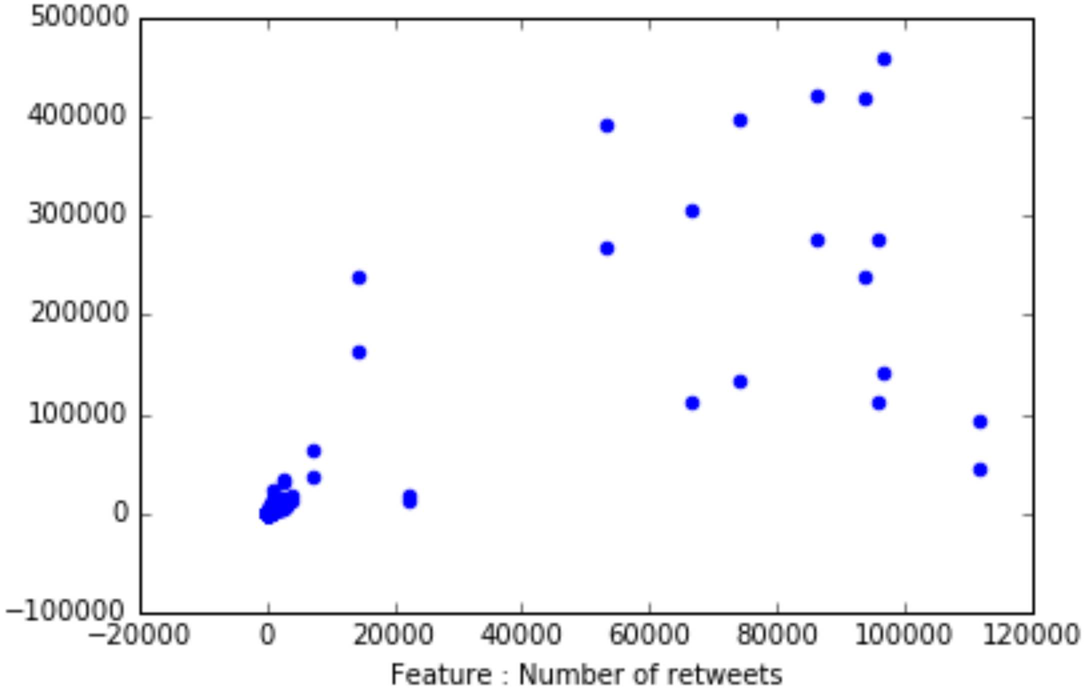
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

Top 3 Features for #sb49

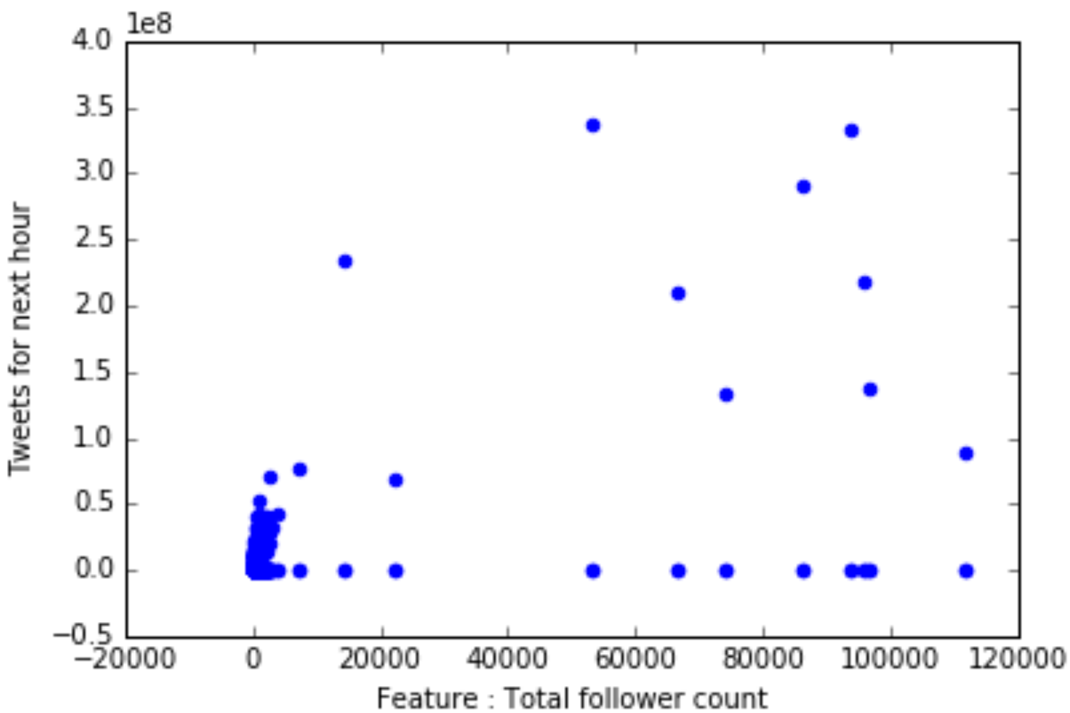
1) Ranking score
2) Number of retweets
3) Total follower count

Scatter plot for these features are as follows:





500000
400000
300000
200000
100000
0
−100000
−20000
0
20000
40000
60000
80000
100000
120000
Feature : Number of retweets



1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Tweets for next hour
−20000
0
20000
40000
60000
80000
100000
120000
Feature : Total follower count

#superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	755.5
Date:	Wed, 22 Mar 2017	Prob (F-statistic):	0.00
Time:	23:02:28	Log-Likelihood:	-9476.4
No. Observations:	964	AIC:	1.898e+04
Df Residuals:	952	BIC:	1.904e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-130.2744	292.067	-0.446	0.656	-703.444	442.896
x1	-2.4725	0.177	-13.955	0.000	-2.820	-2.125
x2	-5.7952	1.028	-5.637	0.000	-7.813	-3.778
x3	6.8e-05	3.37e-05	2.015	0.044	1.78e-06	0.000
x4	25.7548	4.770	5.400	0.000	16.394	35.115
x5	0.0664	0.005	12.344	0.000	0.056	0.077
x6	4.2988	0.881	4.881	0.000	2.570	6.027
x7	-0.0006	8.22e-05	-6.946	0.000	-0.001	-0.000
x8	0.7864	0.135	5.829	0.000	0.522	1.051
x9	-6.7998	1.332	-5.104	0.000	-9.414	-4.185
x10	5.1253	21.245	0.241	0.809	-36.566	46.817
x11	0.0001	9.44e-05	1.500	0.134	-4.37e-05	0.000

Omnibus:	1911.689	Durbin-Watson:	1.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5445448.028
Skew:	14.567	Prob(JB):	0.00
Kurtosis:	370.046	Cond. No.	2.08e+08

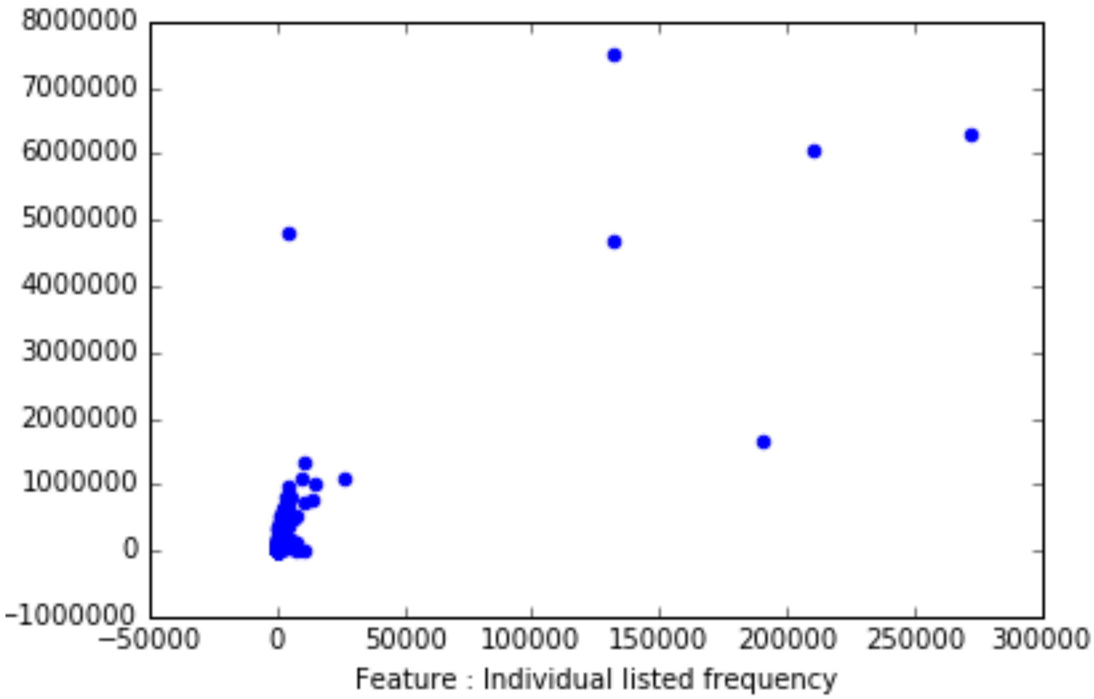
Report 12 : OLS Regression Results for superbowl hashtag with addition of new features

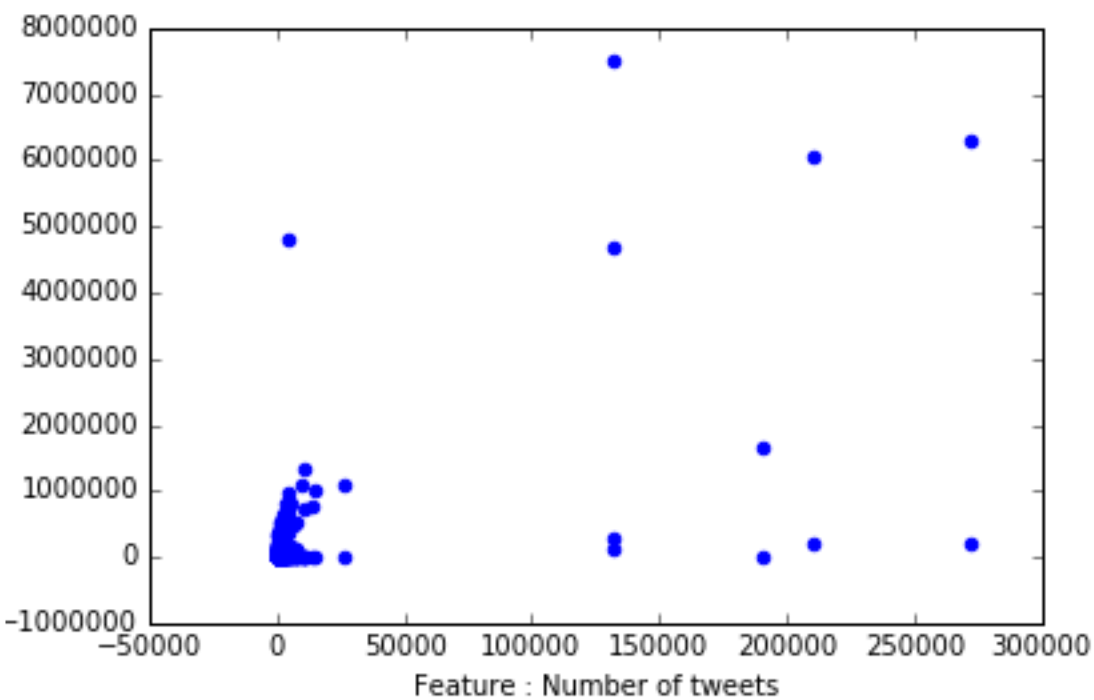
Based on p-value and t-value, we analyze the top features to be used. Based on this we train the model again. R-squared value give us the accuracy obtained with the freshly model trained with significant features analyzed in this step.

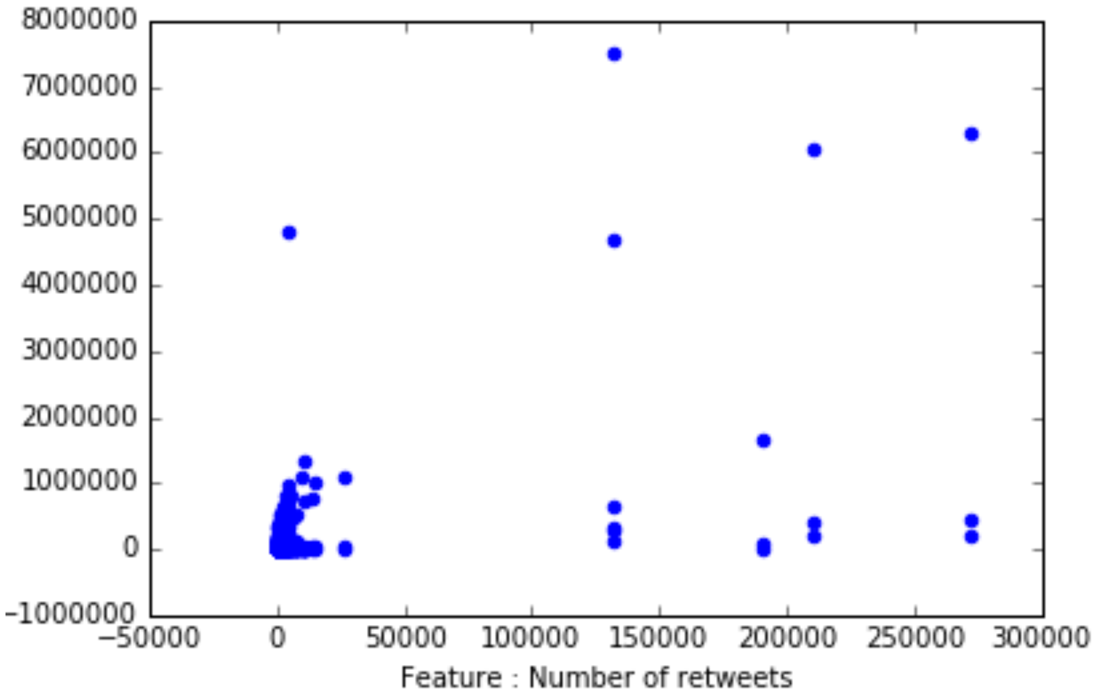
Top 3 Features for #superbowl

1) Individual listed frequency
2) Number of tweets
3) Number of retweets

Scatter plot for these features are as follows:







After obtaining the most significant features, model is trained again for each hashtag and following accuracies are obtained.

Hashtag	Accuracy
#gohawks	63.8
#superbowl	89.7
#gopatriots	84.5
#sb49	87.3
#nfl	75.5
#patriots	77.8

Table 4 : Table showing accuracies corresponding to each hashtag after evaluating significant features

Part 4 (1) – Cross Validation

In Part 4, we are required to use the 11 features obtained in Question-3 and to perform 10-fold Cross Validation across data.
The features are organized in the form of (features, predicant) pairs for each window. The feature data is split into 10 parts, such that 90% of our data will be used for fitting our model and 10% of the data will be used for testing the model.
The process mentioned above, is performed 10 times on the feature data for each of our hastags. To evaluate the performance of the model, we use Prediction error for every fold.

Prediction error is calculated as = |Npredicted – Nreal|

The accuracy results obtained across various hash-tags and over every fold given below:

Fold No	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
(1)	5.497	3.624	5.8562	20.235	55.370	26.962
(2)	5.811	3.649	6.581	20.355	48.427	27.006
(3)	8.896	5.692	6.732	21.208	97.462	26.949
(4)	90.584	9.758	41.497	32.991	53.527	30.424
(5)	16.709	195.137	275.896	137.657	151.170	63.362
(6)	15.890	669.995	161.242	1033.888	300.923	1004.981
(7)	13.719	143.188	159.437	416.451	1726.575	231.682
(8)	12.524	151.169	349.7592	362.720	9300.742	904.430
(9)	289.840	828.921	780.542	3553.646	938.063	11322.880
(10)	5.756	8.919	300.279	108.647	278.735	708.964
Average Error	46.523	202.005	208.782	570.780	1295.099	1434.764

Table 5 : Average Error of 10 Fold Cross Validation

Observation:

- We can see that there is a relationship between the number of tweets for a hash-tag and the average error of cross validation. Greater the number of tweets leads to a higher absolute average error for the hash-tag.

- In particular, it is seen that for each hash-tag the error of one of the cross-validation fold is too high due to the uneven distribution of the data-set. A fold might consider a split wherein the test-data has all high values for the class (tweets during the time of the SuperBowl) and training-data has all low values for the class (tweets before and after the SuperBowl), hence producing a high error value for that fold.

Part 4 (2) - Cross Validation with Time Periods

The second part of Question-4 deals with analysis of regression models created for different time-periods during the SuperBowl. Three different time-periods were considered to create the regression models,

1. Before Feb. 1, 8:00 a.m. [when the hashtags haven't become very active]
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m. [active period]
3. After Feb. 1, 8:00 p.m. [after they pass their high-activity time]

Each tweet was segregated based on the time it was posted and split into windows of one-hour. The models were tested using 10-fold Cross Validation and the average errors for all folds obtained were as follows:

HashTag	Period 1	Period 2	Period 3
#gohawks	200.038	5391.083	3619.449
#gopatriots	15.037	5511.565	3.407
#nfl	129.083	6274.101	320.641
#patriots	193.210	35029.398	119.486
#sb49	99.697	89845.155	233.074
#superbowl	242.084	894816.135	456.501

Table 6 : Average Error of 10 Fold Cross Validation for each Time Period

Observation:

- The error seems to be extremely high for period 2. The reason can be that it is extremely difficult to achieve high accuracy using 12 training points. To deal with this problem we could use sliding windows to increase the number of data points

Part 5 - Testing Data

In this part, we test the models trained by us in part 4 and try to predict the values for the next hour.

The testing data was downloaded and for each file in the testing data features were collected using methods employed in the previous questions. There were 10 files in all, each of them corresponding to one of the three time periods. However, unlike before, the files had a mixture of all hashtags. But the models we had trained earlier were specific to a specific hashtag. So, we found the most dominant hashtag in each of the ten files. The dominant hashtags were:

Test File	Model Used	Dominant HashTag
Sample1_period1	Superbowl model for period1	#superbowl
Sample2_period2	Superbowl model for period2	#superbowl

Sample3_period3	Superbowl model for period3	#superbowl
Sample4_period1	Nfl model for period 1	#nfl
Sample5_period1	Nfl model for period 1	#nfl
Sample6_period2	Superbowl model for period 2	#superbowl
Sample7_period3	Nfl model for period 3	#nfl
Sample8_period1	Nfl model for period 1	#nfl
Sample9_period2	Superbowl model for period2	#superbowl
Sample10_period3	Nfl model for period 3	#nfl

Table 7 : Dominant hashtag for the 10 testing files

For all tags the data for 6 hours had been provided. We had to predict the value for next hour. So given the data from hour 1 to hour 6, we had to predict from hour 2 to hour 7.

Test File	Hour 2	Hour 3	Hour 4	Hour 5	Hour 6	Hour 7	Error
Sample1_period1	115.21	50.32	176.80	265.35	461.99	652.02	213.771
Sample2_period2	614779.9	68409.37	503125	412958	3331221	1806319	1124174.596
Sample3_period3	510.03	723.36	715.78	628.06	643.21	651.30	197.783
Sample4_period1	1375.94	562.02	221.95	342.30	134.77	86.02	332.014
Sample5_period1	491.76	542.83	397.72	308.70	448.62	263.73	253.39
Sample6_period2	11855.12	10885539(	66174686	5643991.7	4233358.1	347051.3	35124214.656
Sample7_period3	86.61	69.31	60.58	51.63	54.21	68.96	31.343
Sample8_period1	NA	57647.17	47250.27	58692.12	72259.96	101448.2	67423.561
Sample9_period2	907629	936522	790894	750649	1019	895972	715378.320
Sample10_period3	43.57	41.00	38.55	36.31	35.28	32.25	25.278

Table 8: Predicted Value for 7th Hour using Regression Model

Error = Actual – Predicted Vale.
Note : hour 7 is skipped over here as the data for hour 7 was not available

Part 6 – Fan Base Prediction

In this part, we train a classifier to predict the location of the author, given only the textual context of the tweet. Because often the textual context reveals some information about the author. Recognizing that supporting a sport team has a lot to do with the user location, so, we try to use the textual content of the tweet posted by a user to predict her location.

For this part, we consider all the tweets including #superbowl, by users whose location has been specified as either Washington state or Massachusetts state. [we consider the tweets that include the following substrings in the location field: Seattle, Washington Washington WA Seattle, WA Kirkland, Washington]
We train different classifiers and evaluate their performance. The classifiers used are:

1) Multinomial Naive Bayes
2) Logistic Regression
3) Linear SVM

The following steps are followed:

1) Collect tweets from superbowl
2) Filter out tweets by appropriate location data
3) Create target labels (0: MA 1: WA) and Balance the datasets
4) Vectorize the tweets:
   Since there are lot of common words, the data needs to be preprocessed. For this, first the punctuations are removed, followed by the common stop words. We then find which words share the same stem so that they can be counted together while finding their TF-IDF vectors. To do the latter, we used a SnowBall stemmer (nlkt) to achieve this. Once the data has been pre-processed, we move on to finding the TF-IDF vector for each term. For this we convert the document into a set of numerical features. This is done using CountVectorizer
5) Truncate twitter data to 50 features using Truncated SVD
6) Perform Feature Scaling for Certain Algorithms Require Nonnegative Values
7) Perform 5-Fold CV to fit different models. The results are given below. The best accuracy was obtained for Linear SVM as 0.8117. The performance for Logistic Regression was comparable.

Multinomial Naive Bayes

Parameter	Value
Average CV- Accuracy	0.7370

	Precision	Recall	F1-Score	Support
0 (MA)	0.70	0.85	0.77	3357
1 (WA)	0.81	0.63	0.71	3351
Avg/Total	0.75	0.74	0.74	6708

Confusion Matrix	
2862	495
1251	2100

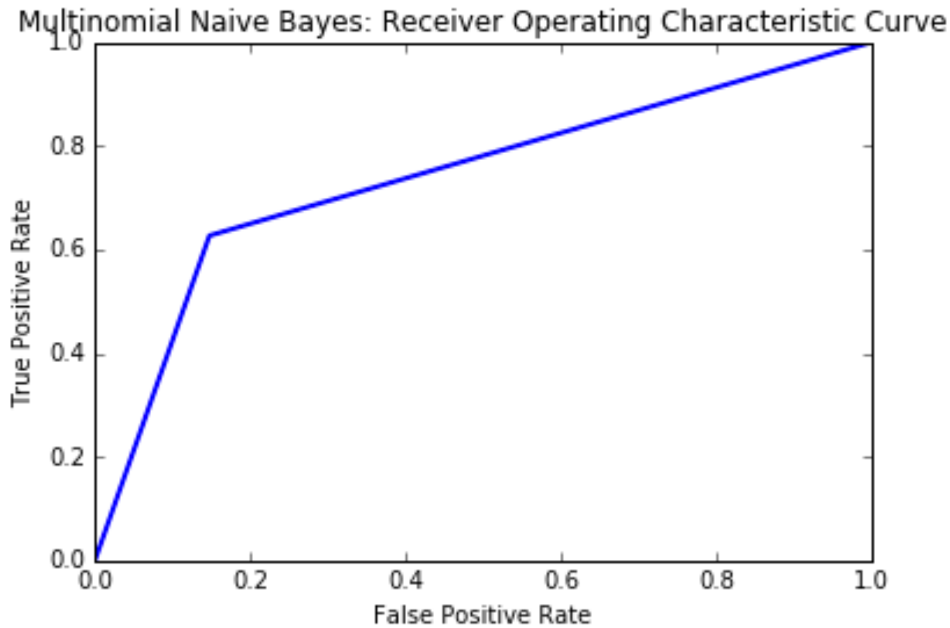


Figure 7: Performance evaluation for Multinomial Naïve Bayes

Logistic Regression:

Parameter	Value
Average CV- Accuracy	0.8092

	Precision	Recall	F1-Score	Support
0 (MA)	0.75	0.94	0.84	3357
1 (WA)	0.92	0.69	0.79	3351
Avg/Total	0.84	0.81	0.81	6708

Confusion Matrix	
3161	196
1052	2299

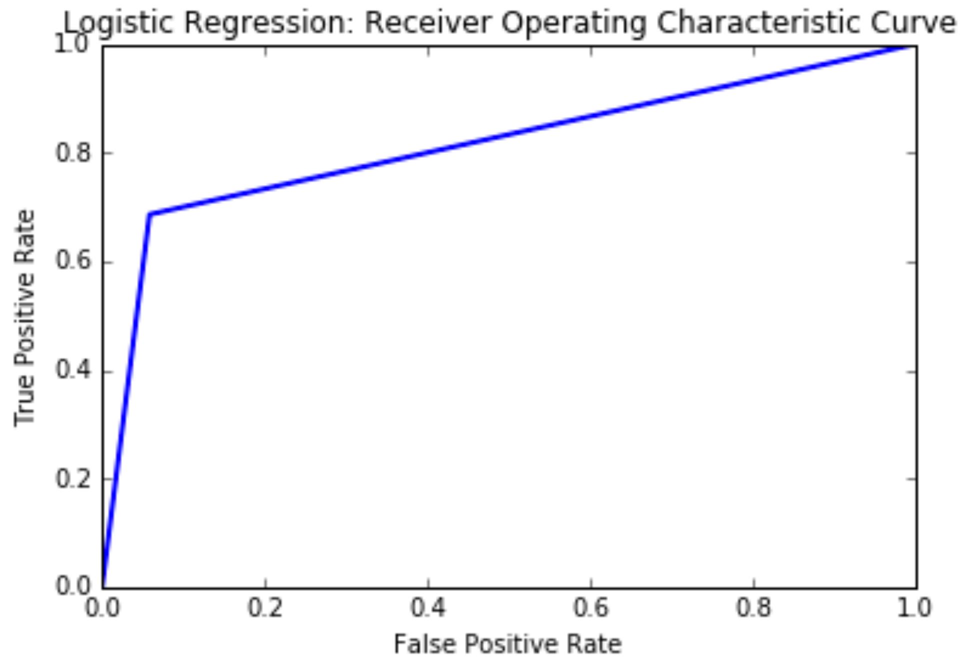


Figure 8: Performance evaluation for Logistic Regression

Linear SVM:

Parameter	Value
Average CV- Accuracy	0. 8117

	Precision	Recall	F1-Score	Support
0 (MA)	0.75	0.95	0.84	3357
1 (WA)	0.93	0.68	0.79	3351
Avg/Total	0.84	0.82	0.81	6708

Confusion Matrix	
3179	178
1059	2292

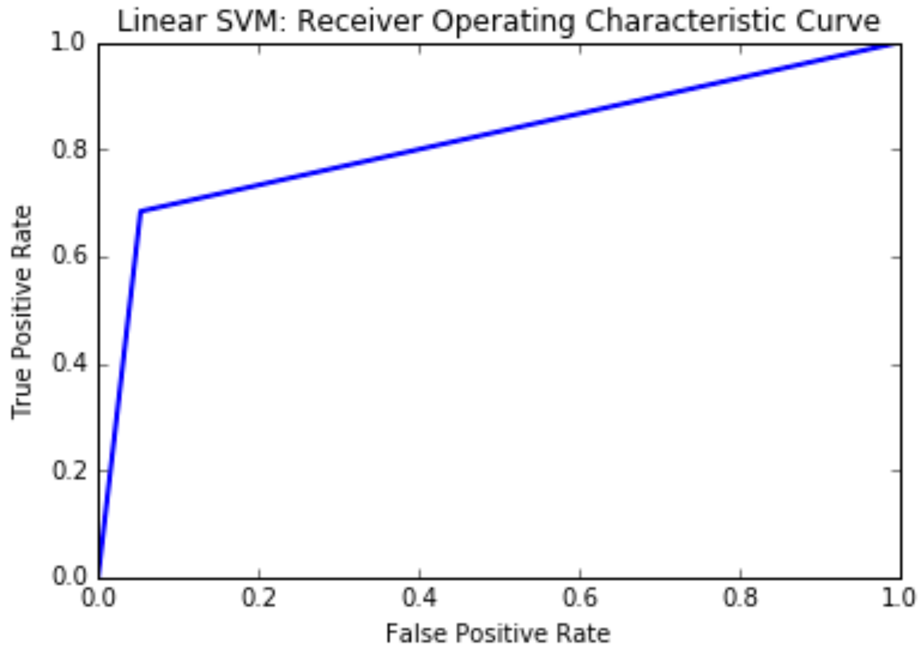


Figure 9: Performance evaluation for Linear SVM

Part 7 - Twitter Ad Week (Event Sequencing)

The SuperBowl event has a huge fan following. It is one of the most popular and widely watched events in the United States. Which has an impact on the activities on Twitter. It is an extremely important platform for some very high profile advertisements given its huge fan base and outreach. We propose an **event sequencing** and analytics problem. We aim to recreate the flow of events that happened at the Superbowl using the tweets. At the same time **analyze how the popularity of different advertisements varied during the timeline of the superbowl**. Through that we aim to project a **brand comparison**, which shows which brands gathered the most attention and at which point of time in the superbowl. Which depicts how the event impacted these brands.

The procedure for the same has been documented below:
We first get all the tweet text for a hash tag using the concept of one hour windows. The data is then preprocessed to get rid of the special characters and the stop words. The keywords are then tokenized into two groups. Which is hash tags and non-hash tags data. The data is classified into different advertisement groups. Some brands that we will be considering in our analysis are :

1) T-mobile
2) Budweiser
3) Snicker
4) McDonald's
5) Coca Cola
6) Dove's Mencare

For brands which have two words in their name. We count them using bigrams where the commonly occurring pairs are put into a counter which puts together key word pairs for every hour. The result of this classification is the hourly count of occurance of every advertisement. We then show the even flow for SuperBowl by dividing the flow into four topics of

1) Team Chatters
2) Touchdowns or goals
3) Advertisements
4) Celebrities

Each topic is analyzed for every hour. Which means it's based on one-hour window.

The graph below covers the timeline – 19th January, 2015 to 7th February,2015.

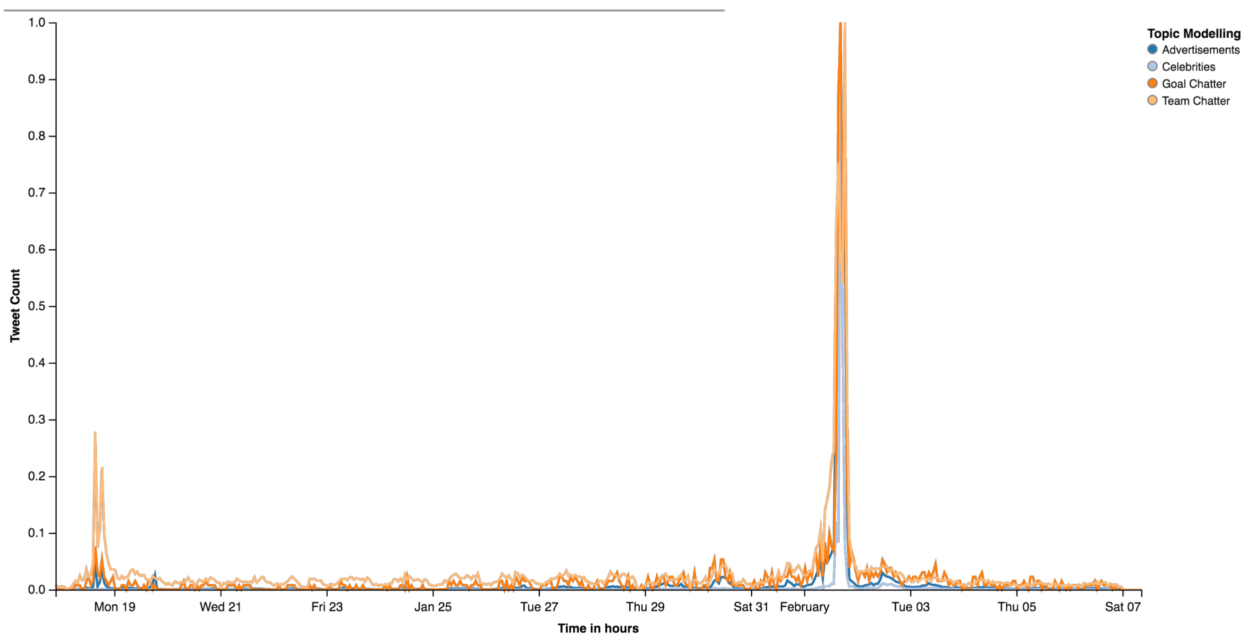
Observation

- Advertisements are present through the event timeline.
- Advertisments have a huge peak during the SuperBowls finals in February. The cause of

this can be that these ads are broadcasted during the half time of the Superbowl finals.

- The orange in the graph represents goal chatter which represents the time when there have been goals or touchdowns. Thus the sudden peaks there represent the time during games. Note, we could change the range of time as the duration of a particular game and get more insights for the number of goals etc. during the game.
- The overall peaks during the finals shows how popular SuperBowl is. The peak in Advertisements show the extent to which advertising agencies are willing to be part of the SuperBowl and how much impact these Ads can have.

Figure 9: Time Series for Event Flow



The graph below shows the popularity of the different brands during different points of time during the superbowl. From the graph we can infer that the popular brands are T mobile, Toyota and Snickers. Their main peaks is between February 1-3.

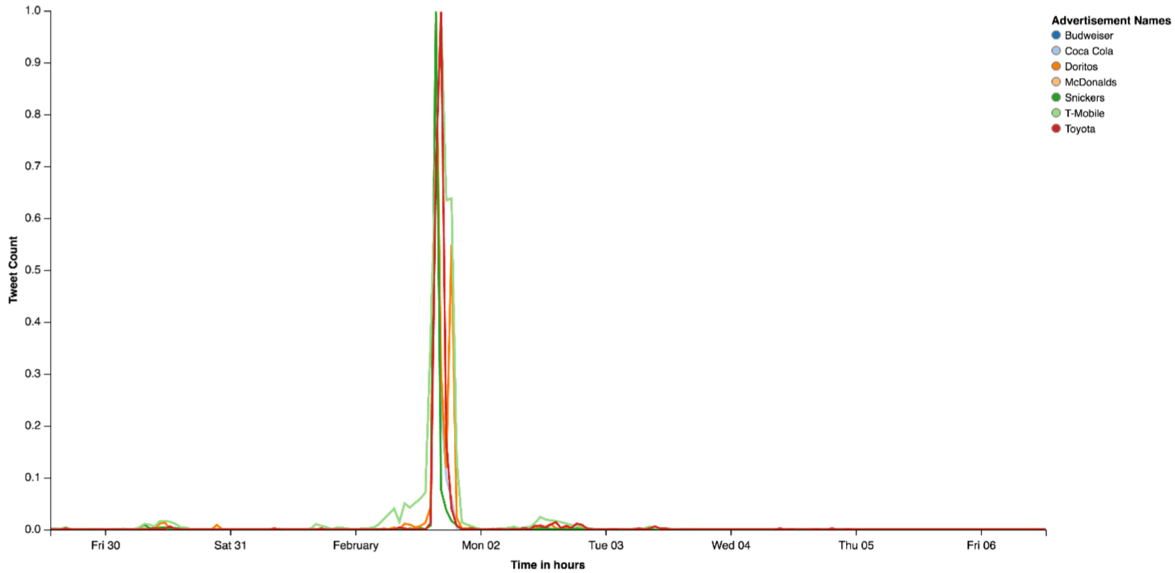


Figure 10 : Time series for brand comparison

Future Work:

Sentiment analysis of the tweets collected can further represent the feelings of an advertisement or a celebrities performance during the SuperBowl.