ECE 219 Project5

Evelyn Chen UID: 704332587
Jack Gong UID: 005025415
Jiuru Shao UID:204288539
Haoxiang Zhang UID:104278461

March 5th 2018

1 Introduction

Twitter is a good platform for social network analysis. We want to predict future tweet activity using current tweet activity for a hashtag. The Twitter data used in this project is from 2015 Super Bowl, ranging from 2 weeks before the game to a week after the game. Data from some related hashtags will be used to train a regression model. Then, the model will be used to make predictions for other hashtags. Lastly, we defined several new problems and implemented them.

2 Q1

2.1 Q1.1

#gohawks: Average number of tweets per hour: 324.933
#gohawks: Average number of followers of users posting the tweets: 2203.932
#gohawks: Average number of retweets: 2.015

#sb49: Average number of tweets per hour: 1418.441
#sb49: Average number of followers of users posting the tweets: 10267.317
#sb49: Average number of retweets: 2.511

#gopatriots: Average number of tweets per hour: 45.621
#gopatriots: Average number of followers of users posting the tweets: 1401.896
#gopatriots: Average number of retweets: 1.400

#patriots: Average number of tweets per hour: 834.264
#patriots: Average number of followers of users posting the tweets: 3309.979
#patriots: Average number of retweets: 1.783

#superbowl: Average number of tweets per hour: 2297.729
#superbowl: Average number of followers of users posting the tweets: 8858.975

#superbowl: Average number of retweets: 2.388

#nfl: Average number of tweets per hour: 441.267
#nfl: Average number of followers of users posting the tweets: 4653.252
#nfl: Average number of retweets: 1.539

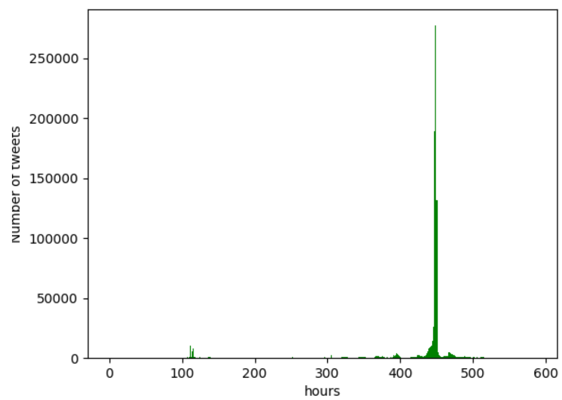


Figure 1: **number of tweets in hour over time for #superbowl**

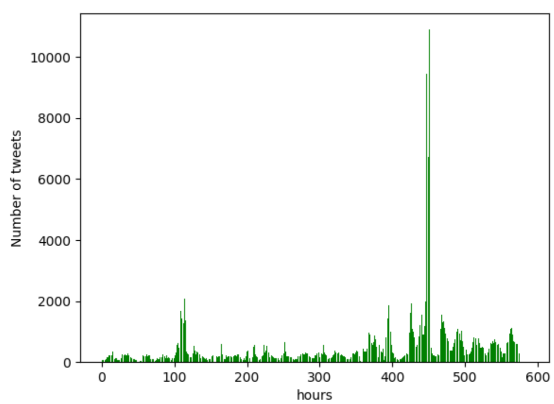


Figure 2: **number of tweets in hour over time for #nfl**

2.2 Q1.2

For each hashtag, fit a linear regression model using 5 features (number of tweets, total number of retweets, sum of number of followers, maximum number of followers, time of the day) to predict the number of tweets in the next hour with features from the previous hour.
For each model (hashtag), we report model's RMSE, R-squared measure. We also analyzed the significance of each feature with t-test and P-value. We used statsmodels.api. Specifically, we use linear regression model OLS for fitting and prediction.
To interpret the OLS output (screenshots), x1 to x5 stands for different features. Specifically, x1=tweet count, x2=retweet count, x3=follower count, x4=max followers, x5=time of the day. For each tag, RMSE and test R-square are shown in the chart. t-value and P-value for each feature for each tag are shown in the screenshots. t-value is under "t" column, and P-value is under "$P >| t |$" column.

	RMSE	Test R-squared
#gohawks	969.32	0.505
#nfl	585.32	0.647
#sb49	4357.89	0.818
#gopatriots	174.47	0.680
#patriots	2517.04	0.684
#superbowl	8330.89	0.789

Table 1: **Metrics of linear regression model**

```
category:  #gohawks
rmse:  969.3208236136504
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.505
Model:                            OLS   Adj. R-squared:                  0.501
Method:                 Least Squares   F-statistic:                     117.1
Date:                Sat, 10 Mar 2018   Prob (F-statistic):           3.33e-85
Time:                        16:35:29   Log-Likelihood:                -4794.8
No. Observations:                 578   AIC:                             9600.
Df Residuals:                     573   BIC:                             9621.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2371      0.127      9.746      0.000       0.988       1.486
x2            -0.1557      0.045     -3.454      0.001      -0.244      -0.067
x3            -0.0005      0.000     -3.151      0.002      -0.001      -0.000
x4             0.0002      0.000      1.026      0.305      -0.000       0.000
x5             6.7740      3.401      1.992      0.047       0.094      13.454
==============================================================================
Omnibus:                      933.725   Durbin-Watson:                   2.248
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           780386.808
Skew:                           9.056   Prob(JB):                         0.00
Kurtosis:                     182.096   Cond. No.                     9.06e+04
==============================================================================
```

(a) #gohawks

```
category:  #nfl
rmse:  585.3194207979475
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     213.2
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          6.76e-129
Time:                        16:35:54   Log-Likelihood:                -4565.6
No. Observations:                 586   AIC:                             9141.
Df Residuals:                     581   BIC:                             9163.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.0581      0.110      9.662      0.000       0.843       1.273
x2            -0.1392      0.063     -2.213      0.027      -0.263      -0.016
x3         -8.332e-05   2.84e-05     -2.936      0.003      -0.000   -2.76e-05
x4          5.736e-05    2.4e-05      2.392      0.017    1.03e-05       0.000
x5             4.6344      2.151      2.154      0.032       0.409       8.860
==============================================================================
Omnibus:                      476.493   Durbin-Watson:                   2.185
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           357611.437
Skew:                           2.283   Prob(JB):                         0.00
Kurtosis:                     123.935   Cond. No.                     1.69e+05
==============================================================================
```

(b) #nfl

```
category:  #sb49
rmse:  4357.88677897403
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.818
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     519.2
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          7.23e-211
Time:                        16:37:06   Log-Likelihood:                -5702.8
No. Observations:                 582   AIC:                         1.142e+04
Df Residuals:                     577   BIC:                         1.144e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.9355      0.058     16.180      0.000       0.822       1.049
x2            -0.1618      0.026     -6.343      0.000      -0.212      -0.112
x3             0.0003    4.9e-05      5.693      0.000       0.000       0.000
x4         -3.613e-05   5.13e-05     -0.705      0.481      -0.000    6.45e-05
x5             2.5998     14.464      0.180      0.857     -25.809      31.009
==============================================================================
Omnibus:                     1240.749   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2871768.756
Skew:                          16.329   Prob(JB):                         0.00
Kurtosis:                     345.574   Cond. No.                     1.01e+06
==============================================================================
```

(c) #sb49

```
category:  #gopatriots
rmse:  174.47409664592223
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.680
Model:                            OLS   Adj. R-squared:                  0.677
Method:                 Least Squares   F-statistic:                     241.7
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          3.80e-138
Time:                        16:37:09   Log-Likelihood:                -3777.3
No. Observations:                 574   AIC:                             7565.
Df Residuals:                     569   BIC:                             7586.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.5720      0.241      2.377      0.018       0.099       1.045
x2            -0.2377      0.211     -1.127      0.260      -0.652       0.176
x3             0.0010      0.000      8.556      0.000       0.001       0.001
x4            -0.0010      0.000     -8.813      0.000      -0.001      -0.001
x5             1.0705      0.566      1.890      0.059      -0.042       2.183
==============================================================================
Omnibus:                      522.450   Durbin-Watson:                   2.036
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           271437.437
Skew:                           2.948   Prob(JB):                         0.00
Kurtosis:                     109.370   Cond. No.                     2.32e+04
==============================================================================
```

(d) #gopatriots

```
category:  #patriots
rmse:  2517.0436225246235
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.684
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     251.0
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          1.51e-142
Time:                        16:37:46   Log-Likelihood:                -5420.4
No. Observations:                 586   AIC:                         1.085e+04
Df Residuals:                     581   BIC:                         1.087e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.9985      0.079     12.624      0.000       0.843       1.154
x2            -0.0458      0.056     -0.824      0.410      -0.155       0.063
x3            -0.0002   9.15e-05     -2.243      0.025      -0.000   -2.55e-05
x4             0.0003      0.000      3.043      0.002       0.000       0.001
x5             1.8266      8.749      0.209      0.835     -15.356      19.009
==============================================================================
Omnibus:                      888.931   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           704259.297
Skew:                           7.952   Prob(JB):                         0.00
Kurtosis:                     172.087   Cond. No.                     3.02e+05
==============================================================================
```

(e) #patriots

```
category:  #superbowl
rmse:  8330.894601699181
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.787
Method:                 Least Squares   F-statistic:                     434.8
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          1.12e-193
Time:                        16:41:22   Log-Likelihood:                -6121.7
No. Observations:                 586   AIC:                         1.225e+04
Df Residuals:                     581   BIC:                         1.228e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.9462      0.082     23.825      0.000       1.786       2.107
x2            -0.4753      0.027    -17.686      0.000      -0.528      -0.423
x3          5.429e-05   5.83e-05      0.931      0.352   -6.02e-05       0.000
x4          9.771e-05      0.000      0.814      0.416      -0.000       0.000
x5            -5.1482     30.563     -0.168      0.866     -65.175      54.879
==============================================================================
Omnibus:                      901.649   Durbin-Watson:                   2.246
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1783902.692
Skew:                           7.797   Prob(JB):                         0.00
Kurtosis:                     272.848   Cond. No.                     1.90e+06
==============================================================================
```

(f) #superbowl

Figure 3: **Statistics for Linear Regression**

2.3 Q1.3

After studying several papers, we designed following new features: mention count, rank score, passivity, co-occurrence of tags, and unique authors.

Mention count. Mention is a directional sharing behavior in Twitter by using @ as the prefix of the user's name. If a user was mentioned in a tweet with a hashtag, he probably took part in the topic. Thus, mention count is one of the new features we tried.

Rank score. Rank score measures the degree of relevance of a tweet to a topic. If there were many tweets related to a specific topic, then these tweets should have high relevance scores to this topic. Thus, mention count is one of the new features we tried.

Passivity. Active users often post or retweet tweets following some hashtags. On the other hand, passive users rarely do so unless the topics are attractive enough. The passivity is defined as following equation:

$$P_{sv}(u_i) = \frac{N_d(u_i)}{1.0 + N_t(u_i)}$$

where $N_d(u_i)$ denotes the number of days since the user account was created, and $N_t(u_i)$ denotes the total number of tweets posted by this user.

Co-occurrence of tags. Sometimes, several hashtags are used together by users if the topic is hot. The co-occurrence of tags is defined as the number of hashtags used in a tweet.

Unique authors. We also consider the unique number of authors who posted tweets containing the hashtag. This feature can help recognize tweets automatically posted by some fake accounts.

We combine these new features, and apply the OLS linear model of stats api to fit our data. We obtain following fitting accuracy metrics:

	RMSE	Test R-squared
#gohawks	904.084	0.570
#nfl	528.773	0.712
#sb49	4224.850	0.829
#gopatriots	136.285	0.806
#patriots	2468.086	0.698
#superbowl	8391.092	0.785

Table 2: **Metrics of linear regression model (New designed features)**

```
#gohawks
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.570
Model:                            OLS   Adj. R-squared:                  0.566
Method:                 Least Squares   F-statistic:                     152.0
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          1.35e-102
Time:                        14:14:31   Log-Likelihood:                -4754.5
No. Observations:                 578   AIC:                             9519.
Df Residuals:                     573   BIC:                             9541.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.8433      0.425      1.985      0.048       0.009       1.678
x2            -0.1611      0.109     -1.483      0.139      -0.374       0.052
x3            -0.1132      0.028     -4.111      0.000      -0.167      -0.059
x4             1.3609      0.201      6.759      0.000       0.965       1.756
x5            74.3725    908.958      0.082      0.935   -1710.924    1859.669
==============================================================================
Omnibus:                     1190.014   Durbin-Watson:                   2.247
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2309786.138
Skew:                          15.106   Prob(JB):                         0.00
Kurtosis:                     311.213   Cond. No.                     4.95e+05
==============================================================================
```

(a) #gohawks

```
#nfl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.712
Model:                            OLS   Adj. R-squared:                  0.710
Method:                 Least Squares   F-statistic:                     287.3
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          1.96e-154
Time:                        14:14:31   Log-Likelihood:                -4506.0
No. Observations:                 586   AIC:                             9022.
Df Residuals:                     581   BIC:                             9044.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.7904      0.559      3.205      0.001       0.693       2.887
x2            -0.3588      0.056     -6.400      0.000      -0.469      -0.249
x3            -0.0719      0.008     -8.615      0.000      -0.088      -0.056
x4             0.9315      0.085     10.963      0.000       0.765       1.098
x5           -37.6525    531.067     -0.071      0.944   -1080.698    1005.393
==============================================================================
Omnibus:                      664.862   Durbin-Watson:                   2.551
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           175209.298
Skew:                           4.784   Prob(JB):                         0.00
Kurtosis:                      87.168   Cond. No.                     3.16e+05
==============================================================================
```

(b) #nfl

```
#sb49
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.829
Model:                            OLS   Adj. R-squared:                  0.828
Method:                 Least Squares   F-statistic:                     559.9
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          1.19e-218
Time:                        14:14:31   Log-Likelihood:                -5684.8
No. Observations:                 582   AIC:                         1.138e+04
Df Residuals:                     577   BIC:                         1.140e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.6792      0.067     10.124      0.000       0.547       0.811
x2            -0.0444      0.093     -0.476      0.634      -0.228       0.139
x3             0.1961      0.027      7.335      0.000       0.144       0.249
x4            -0.8169      0.198     -4.127      0.000      -1.206      -0.428
x5             3.5220   4243.116      0.001      0.999   -8330.313    8337.357
==============================================================================
Omnibus:                     1232.685   Durbin-Watson:                   1.538
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2865975.814
Skew:                          16.071   Prob(JB):                         0.00
Kurtosis:                     345.274   Cond. No.                     2.62e+06
==============================================================================
```

(c) #sb49

```
#gopatriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                     472.7
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          6.69e-200
Time:                        14:14:31   Log-Likelihood:                -3635.5
No. Observations:                 574   AIC:                             7281.
Df Residuals:                     569   BIC:                             7303.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.4558      0.484      5.070      0.000       1.504       3.407
x2            -1.0718      0.127     -8.414      0.000      -1.322      -0.822
x3             0.3039      0.019     15.651      0.000       0.266       0.342
x4             0.7340      0.274      2.677      0.008       0.196       1.272
x5             8.4249    136.883      0.062      0.951    -260.432     277.282
==============================================================================
Omnibus:                      232.846   Durbin-Watson:                   2.221
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           106874.859
Skew:                          -0.107   Prob(JB):                         0.00
Kurtosis:                      69.848   Cond. No.                     9.94e+04
==============================================================================
```

(d) #gopatriots

```
#patriots
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.698
Model:                            OLS   Adj. R-squared:                  0.695
Method:                 Least Squares   F-statistic:                     268.5
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          2.04e-148
Time:                        14:14:31   Log-Likelihood:                -5408.9
No. Observations:                 586   AIC:                         1.083e+04
Df Residuals:                     581   BIC:                         1.085e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1819      0.049      3.712      0.000       0.086       0.278
x2             0.2747      0.116      2.371      0.018       0.047       0.502
x3            -0.0653      0.033     -1.999      0.046      -0.130      -0.001
x4             0.0233      0.192      0.122      0.903      -0.353       0.400
x5           105.9916   2478.995      0.043      0.966   -4762.892    4974.875
==============================================================================
Omnibus:                     1003.460   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           702869.545
Skew:                          10.409   Prob(JB):                         0.00
Kurtosis:                     171.384   Cond. No.                     1.19e+06
==============================================================================
```

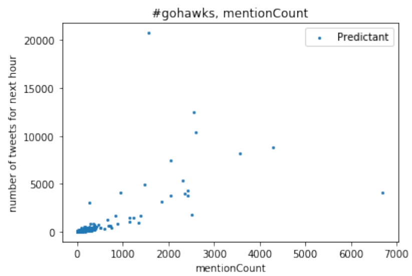
(e) #patriots

```
#superbowl
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.785
Model:                            OLS   Adj. R-squared:                  0.783
Method:                 Least Squares   F-statistic:                     423.4
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          4.81e-191
Time:                        14:14:31   Log-Likelihood:                -6126.0
No. Observations:                 586   AIC:                         1.226e+04
Df Residuals:                     581   BIC:                         1.228e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -5.9103      0.668     -8.848      0.000      -7.222      -4.598
x2             0.6314      0.127      4.981      0.000       0.382       0.880
x3            -0.4438      0.031    -14.262      0.000      -0.505      -0.383
x4             2.6111      0.242     10.770      0.000       2.135       3.087
x5          -104.7545   8427.143     -0.012      0.990   -1.67e+04    1.64e+04
==============================================================================
Omnibus:                      850.822   Durbin-Watson:                   2.515
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           434552.954
Skew:                           7.449   Prob(JB):                         0.00
Kurtosis:                     135.573   Cond. No.                     5.70e+06
==============================================================================
```

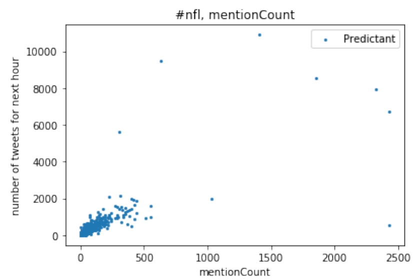
(f) #superbowl

Figure 4: **Statistics for Linear Regression (New designed features)**

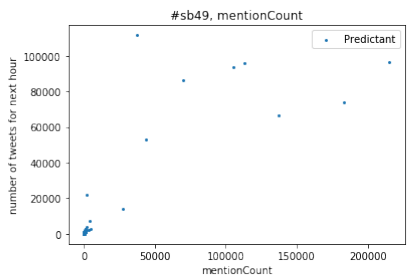
These are statistic results of OLS model for each hashtag. To choose the top 3 features, we simply did majority votes among 5 new features based on the p value of each feature. We found that features x1 (mention count), x2 (passivity), and x4 (co-occurrence of tags) are most significant 3 features. Then, we plot scatter plots (predictant versus the value of feature) for each of top 3 features, and for each hashtag.



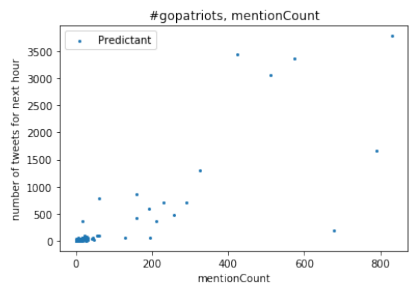
(a) #gohawks



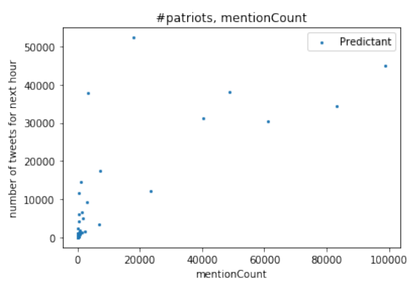
(b) #nfl



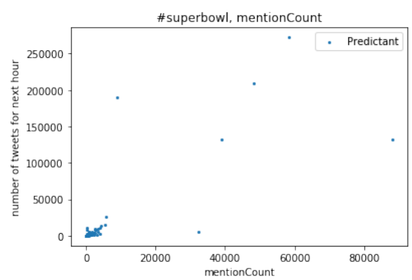
(c) #sb49



(d) #gopatriots



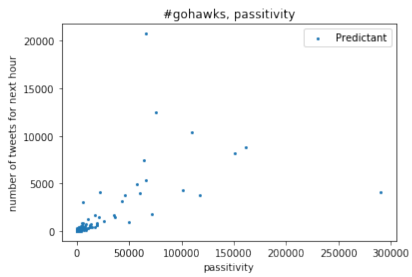
(e) #patriots



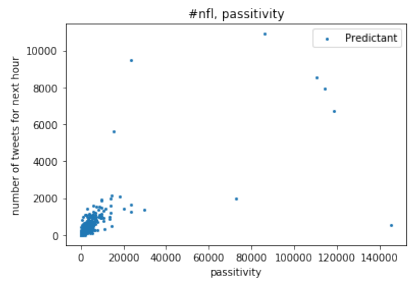
(f) #superbowl

Figure 5: **Scatter plot of predictant versus value of feature (Mention Count)**

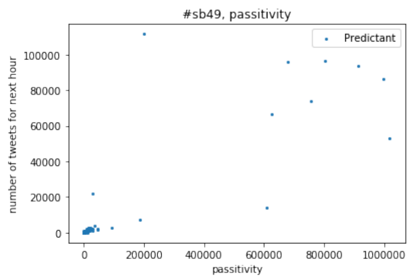
These are scatter plots for predictant versus value of mention count, for each hashtag. We conclude that there is a relatively linear relationship between our mention count feature and the number of tweets for next hours, despite of some extreme points.



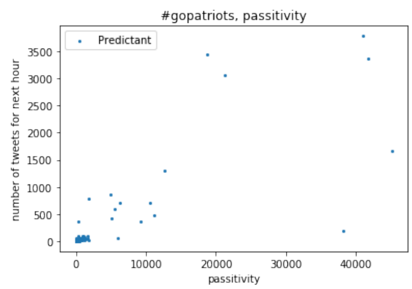
(a) #gohawks



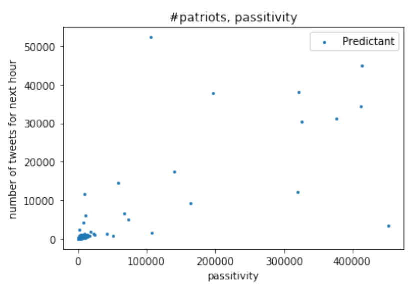
(b) #nfl



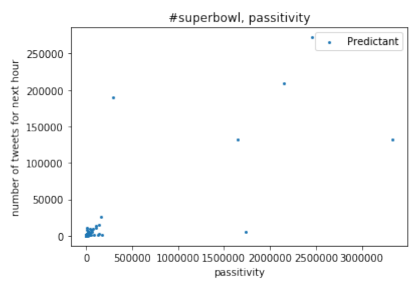
(c) #sb49



(d) #gopatriots



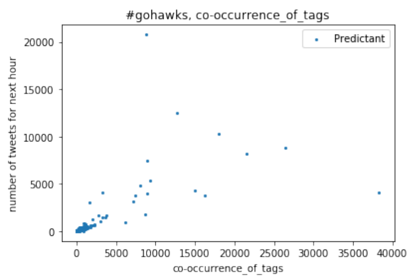
(e) #patriots



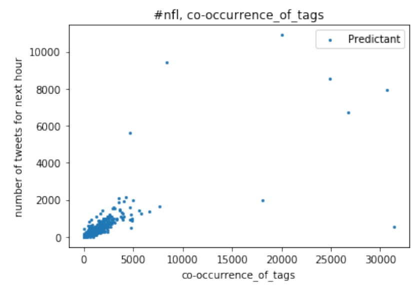
(f) #superbowl

Figure 6: **Scatter plot of predictant versus value of feature (Passivity)**

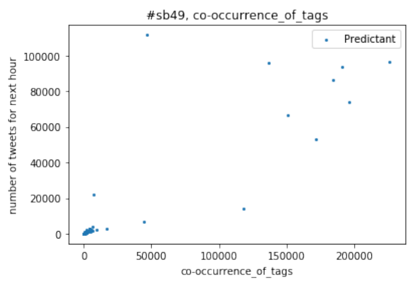
These are scatter plots for predictant versus value of passivity, for each hashtag. We conclude that there is a relatively linear relationship between our passivity feature and the number of tweets for next hours, despite of some extreme points.



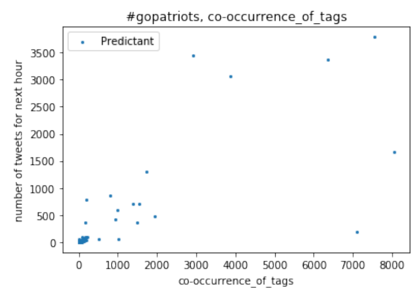
(a) #gohawks



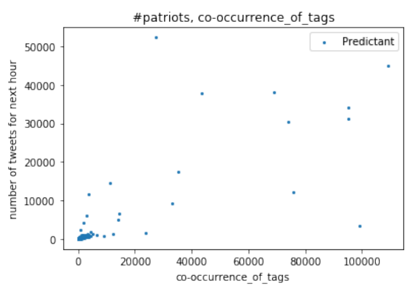
(b) #nfl



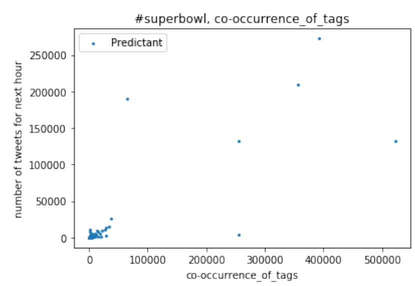
(c) #sb49



(d) #gopatriots



(e) #patriots



(f) #superbowl

Figure 7: **Scatter plot of predictant versus value of feature (Co-occurrence of tags)**

These are scatter plots for predictant versus value of co-occurrence of tags, for each hashtag. We conclude that there is a relatively linear relationship between our co-occurrence of tags feature and the number of tweets for next hours, despite of some extreme points.

2.4 Q1.4

In this part, we use three different models, Random Forest Regressor,Support Vector Regressor, and Linear SVR for cross validation. We use the same features as part 1.3. As we can see, the MAE is particularly large in the second period(the day when the final happens). This is expected since the number of tweets is the largest during that period.

Hashtag	Model	MAE Period 1	MAE Period 2	MAE Period 3
#gohawks	Random Forest Regressor	224.805	2961.000	32.771
#gohawks	Support Vector Regressor	162.873	2057.769	29.308
#gohawks	Linear SVR	230.735	4961.248	19.039
#nfl	Random Forest Regressor	189.435	3360.846	296.178
#nfl	Support Vector Regressor	113.409	1862.838	161.894
#nfl	Linear SVR	131.304	6610.062	184.607
#sb49	Random Forest Regressor	104.355	43000.962	340.302
#sb49	Support Vector Regressor	45.973	24142.238	133.946
#sb49	Linear SVR	50.603	102436.941	85.322
#gopatriots	Random Forest Regressor	12.830	1407.308	4.932
#gopatriots	Support Vector Regressor	11.654	884.046	4.226
#gopatriots	Linear SVR	13.230	1474.449	3.717
#patriots	Random Forest Regressor	265.085	20316.308	143.657
#patriots	Support Vector Regressor	223.789	13326.308	104.829
#patriots	Linear SVR	211.170	59552.666	60.537
#superbowl	Random Forest Regressor	441.546	75255.077	598.250
#superbowl	Support Vector Regressor	248.277	48557.815	256.598
#superbowl	Linear SVR	338.001	69712.643	300.267

Table 3: **MAE of three different models in different periods for 6 hashtags**

Hashtag	Model	MAE Period 1	MAE Period 2	MAE Period 3
total	Random Forest Regressor	1642.051	77828.077	1239.598
total	Support Vector Regressor	1525.975	72535.892	370.987
total	Linear SVR	1386.803	103832.673	408.831

Table 4: **MAE of three different models in different periods using aggregate data**

Hashtag	Model	MAE Period 1	MAE Period 2	MAE Period 3
total	Random Forest Regressor	1238.056	146301.5	1416.091
total	Support Vector Regressor	794.296	95530.546	643.571
total	Linear SVR	818.874	289569.02	603.722

Table 5: **sum of MAE of three different models in different periods using each hashtag**

The first table is generated with the cross-validated MAE of aggregate data, and the second one is simply the sum of errors from all 6 tags. We can observe that the first table is much smaller than the second one. This is expected, since we might over-estimate or under-estimate the result, and aggregating the data will counteract some of this effect.

2.5 Q1.5

The best model we found in Q1.4 is the random forest regressor model and we apply this model here. We set time window of features to 5 hours instead of 1 hour, and predict for the hour after each window. We use the 'first_postdate' instead of 'citation_date', because the test data are collected based on 'first_postdate'. All test samples have 6 hour span except for sample8 which only has 5 hour span. We trained our model by aggregating the data of all hashtags, with time window of 5 hours. Because most each test sample file spans over 6 hours, so we use the first 5 hours data (first 4 hours data for sample8) as the input to the model, and compare the predicted number of tweets in the 6th hour (the 5th hour for sample 8) with the true number of tweets in the 6th hour of each test sample file. Both the true value and the predicted value are listed in the following table.

Hour 6	True value	Predicted Value
Sample1	178	225.5
Sample2	82892	30163.2
Sample3	524	547.9
Sample4	203	426.45
Sample5	211	699.5
Sample6	37279	67951
Sample7	121	224.7
Sample8	12	255.5
Sample9	2791	2349.6
Sample10	62	475

Table 6: **Number of tweets in the next hour (Use 5 hour interval to predict)**

There are some situations where the predicted value (5 hour window) is close to the true value: sample 3 (524 versus 547.9) and sample 9 (2791 versus 2349.6). Also, there are situations where the predicted value based on 5-hour window is much closer to the true value than the predicted value based on 1-hour window. For example, in sample 6, the true value is 37279, the predicted value based on 5-hour window is 67951, and the predicted value based on 1-hour window is 159763.9 which is far from the true value.

In general, predicting number of tweets in the next hour based on 5-hour window is a fair choice. But since this prediction task is hard, it is difficult for training data based on 5-hour window to reach a perfect performance.

3 Q2

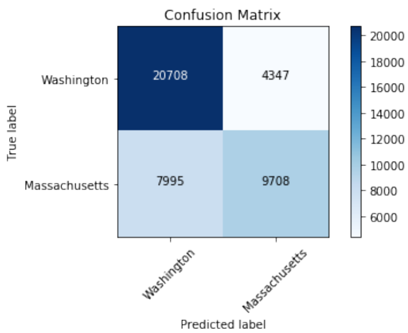
Leveraging the techniques in Project 1, we transform the textual content of the text into matrix with latent semantic information. After preprocessing the data, we tried 6 classification algorithms. These algorithms are **random forest classifier, linear support vector machine classifier, logistric regression classifier, k nearest neighbors, multiple layer perceptron, and decision tree**.

For each classification algorithm, we plot confusion matrix, ROC curve, and calculate accuracy, recall and precision scores. Following table shows the metric scores of 6 classification algorithms. Analysis will be made at the end of this part (Q2).

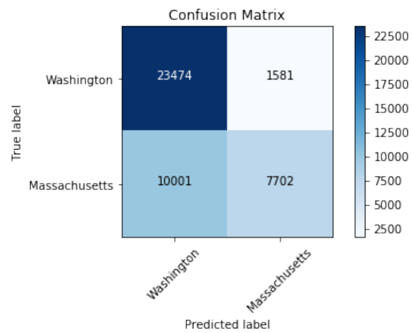
	Accuracy	Recall	Precision
Random Forest	0.7114	0.5484	0.6907
Support Vector	0.7291	0.4351	0.8297
Logistic Regression	0.7300	0.4351	0.8329
K Nearest Neighbors	0.6884	0.5989	0.6302
Multi Layer Perceptron	0.7345	0.5162	0.7664
Decision Tree	0.6710	0.6048	0.6022

Table 7: **Test Metrics of Classification Algorithm (Fan Base Prediction)**

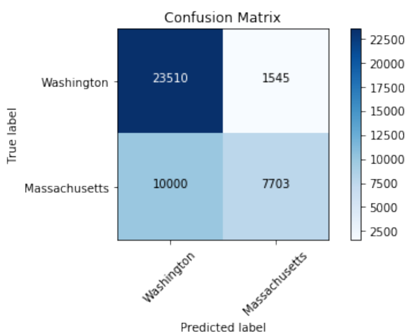
Followings are plots of confusion matrix for each classification algorithm. Analysis will be made at the end of this part (Q2).



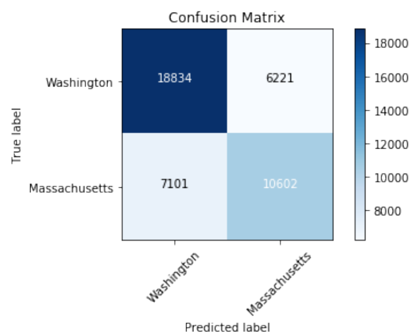
(a) Random Forest Classifier



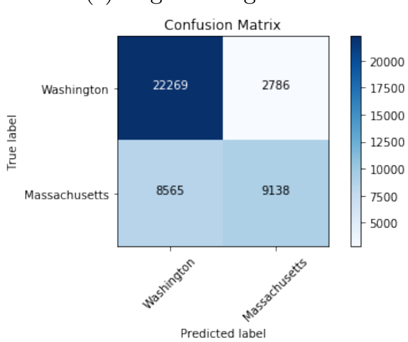
(b) Support Vector Machine Classifier



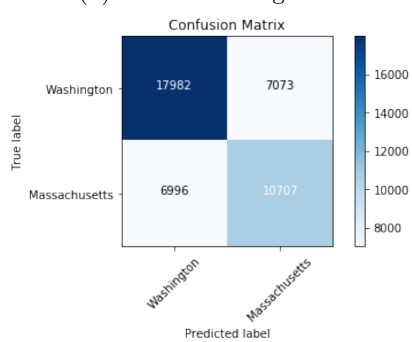
(c) Logistic Regression



(d) K Nearest Neighbors



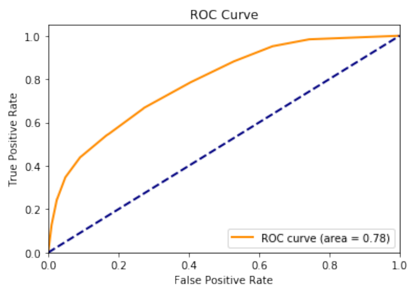
(e) Multi Layer Perceptron



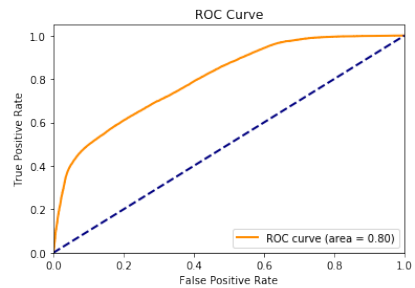
(f) Decision Tree

Figure 8: **Confusion matrix plots for fan base prediction**

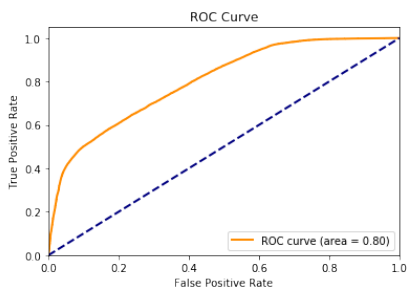
Followings are plots of ROC cureve for each classification algorithm. Analysis will be made at the end of this page.



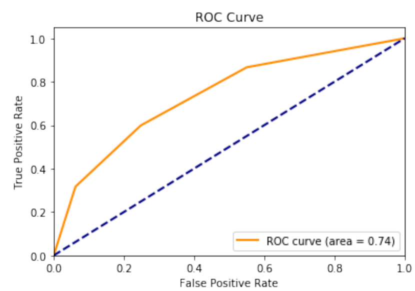
(a) Random Forest Classifier



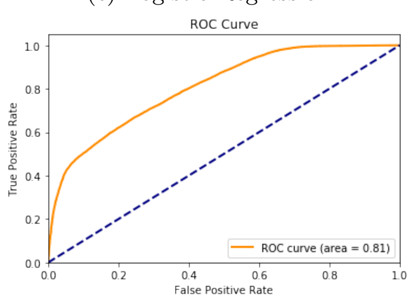
(b) Support Vector Machine Classifier



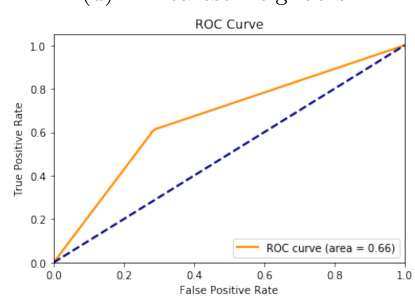
(c) Logistic Regression



(d) K Nearest Neighbors



(e) Multi Layer Perceptron



(f) Decision Tree

Figure 9: **ROC plots for fan base prediction**

All 6 classification algorithms have not bad performances. But based on metric scores (accuracy, recall, precision), ROC curve and confusion matrix, the multiple layer perceptron classification algorithm is the best one among all six classification algorithms.

4 Q3

4.1 Quantitative Sentiment Analysis

For part 3, we are analyzing the change of tweet sentiments for fans of the two teams (hawks and patriots) in the superbowl match. We first plot positive and negative sentiment versus time. The sentiments (y-axis) are obtained from SentimentIntensityAnalyzer from nltk and I use the tweet content (tweet['title']) to analyze the sentiment polarity score.

The time (x-axis) is the number of hours passed from the beginning of the data collection, which is two weeks before the game match. To approximate the time of the game match we can deduct a week from the last data, so the time would be around the middle of 400 500 hours on x-axis.

If we carefully analyze the two sentiment vs time plots, we can see that the negative sentiment polarity score has a peak at around 450 hours on the x-axis for #gohawks and positive sentiment polarity score has a peak around 450 horus on the x-axis for #patriots. This makes sense because The New England Patriots team won Superbowl in 2015, which causes the peak of positive sentiment for the patriots (winner) tag and the peak of negative sentiment for the hawks (loser) tag.

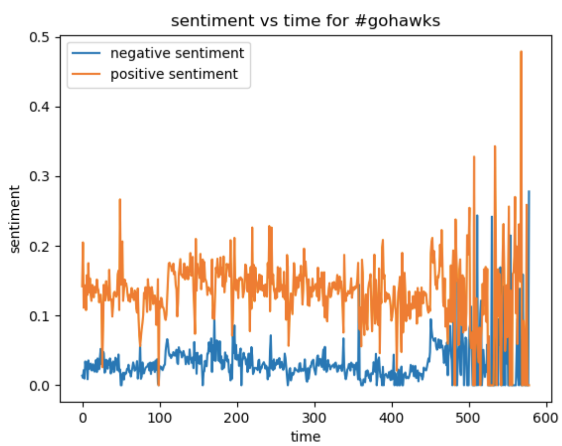


Figure 10: **Sentiment vs time for #gohawks**

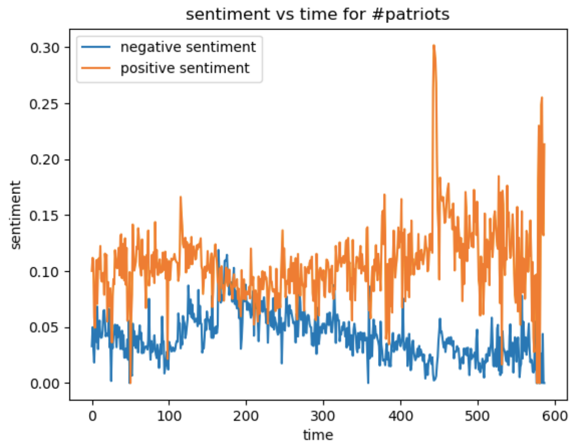


Figure 11: **Sentiment vs time for #patriots**

4.2 Relative Sentiment Analysis

Different from the previous analysis which estimated the exact quantity sum of sentiment coefficient and plotted all the positive and negative sentiments, this time, we took another insight. That is, we consider those who have sentiment quantity > 0 are considered "positive tweets", those who have sentiment quantity < 0 are considered "negative tweets", and those who holds neutral opinion will have the sentiment equal to zero. In other words, there is no "strong sentiment" in this scenario.

In addition, the sentiment quantity is calculated using TextBlob, a python package for Natural Language Processing, in which the sentiment.polarity was invoked, from the sentiment() function. The polarity ranges from -1 to 1, with -1 being most negative texts and 1 being the opposite.
Next, since there are only two teams involved in this data set, we consider those who hold "negative sentiment" against team "hawk" are supporting their opponent team, that is "patriots". Similarly, those who hold "positive sentiment" for "patriots" are essentially "negative sentiment" for the "hawk". One this concept is established, we conclude that, from the hawk's perspective, tweets supporting hawk = positive sentiment for hawk + negative sentiment for patriots; tweets against hawk = positive sentiment towards patriots + negative sentiment for hawk.

Thereafter, four arrays of data are collected, for positive sentiment hawk, negative sentiment hawk, positive sentiment patriots, negative sentiment patriots. Then based in the formula mentioned above, we plot the diagram for relative negative and positive tweets from the hawk's perspective.

As we can see from the diagram below, at around hour 450, there is a huge peak for relative positive response on hawk's perspective, one can easily deduce that hawk was the winner at that moment.

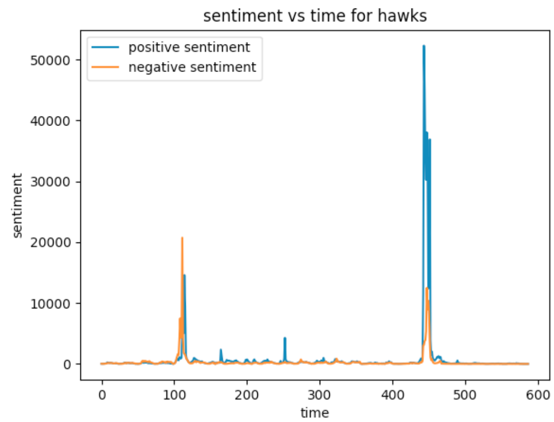


Figure 12: **relative responses from hawk's perspective**

4.3 Comparison between method 4.1 and method 4.2

The sentiment values were calculated differently, due to package differences. In "Quantatative Sentiment Analysis", the sentiment values were computed by the sentiment intensity analyzer, from the ntlk package; whereas in "Relative Sentiment Analysis", the sentiment.polarity from TextBlob was used. The later method took around 3 hours to run through the entire data set, taking 2 hours and 30 min longer than the ntlk package.Both TextBlob and ntlk are part of Natrual Language Processing, but TextBlob was designed to bring more benefits that was not in ntlk. In other words, TextBlob is enhanced super set that is build off of ntlk. This leads to the reason why TextBlob took significantly longer than ntlk, though the data we were interested in wer the same - sentiment quanty. Also, the algorithm was a bit different, in that the "Relative Sentiment Analysis" has to run over the data set for 4 times, whereas "Quantative Sentiment Analysis" only run for twice (see details in implementation).Another difference is that 4.1 is per tweet sentiment. As y-axis in graphs in 4.1 indicates, the sentiments are between 0 and 1. On the other hand, 4.2 is adding up the sentiment scores and so the y-axis is much larger.

4.4 Changes of top words before, during and after the super bowl game

For this newly designed problem, we want find some interesting changes of top words in tweets before, during and after the super bowl game. To perform this experiment, we firstly aggregate tweet contents based on following three time periods (PST times):

1. Before Feb. 1, 8:00 a.m. (i.e. before game)
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m. (i.e. game time)
3. After Feb. 1, 8:00 p.m. (i.e. after game)

Then we perform TfIDF vectorization (the technique we mastered from Project 1) to extract top 20 words.

1. **Before the game**
   ['gohawks', 'http', 'superbowlxlix', 'seattle', 'seahawks', 'superbowl', 'nfl', 'new', 'amp', 'game', 'win', 'football', 'super', 'bowl', 'patriots', 'el', 'sb49', 'colts', 'brady', 'deflategate']
2. **During the game**
   ['gohawks', 'http', 'just', 'super', 'bowl', 'sb49', 'seahawks', 'patriots', 'game', 'superbowl', 'superbowlxlix', 'winning', 'got', 'nfl', 'el', 've', 'halftime', 'katyperry', 'seahawkswin', 'patriotswin']
3. **After the game**
   ['brady', 'http', 'rt', 'seahawks', 'super', 'bowl', 'superbowl', 'sb49', 'superbowlxlix', 'amp', 'win', 'game', 'football', 'nfl', 'patriots', 'https', 'year', 'katyperry', 'new', '2015']

We can find many interesting facts from the change of top words in tweets. For example, 'halftime' and 'katyperry' shows up in the top words from tweets during the game, as Katy Perry was featured in the half time show. Also, 'brady' jumps to the top 1 from tweets after the super bowl game, as Tom Brady was the key person that led New England Patriots win the game!

5 Conclusion

In this project, we used Twitter data for social network analysis. We did popularity prediction using linear regression and other regression models. We designed good features for the regression models. We also used k-fold cross validation for different models. Lastly, we designed our own problems analyzing sentiments and change of top words and implemented them.