University of California, Los Angeles

EE 219 Winter 2018

*Project 5: Popularity Prediction on Twitter*



Group Member: Yunhao Ba & Xiaocheng Zhang & Shuangyu Li & Shengrong Wu

Date: March 18, 2018

Table of Contents

[**Introduction**](#_2end0j4mni0t) **3**

[**Part 1: Popularity Prediction**](#_ddum8pqgo2rb) **3**

[Problem 1.1:](#_efb25mwa3dan) 3

[Problem 1.2:](#_2alljuxrrl5z) 4

[Problem 1.3:](#_c2z3scepekka) 9

[Problem 1.4:](#_9rl8kteivlww) 33

[Problem 1.5:](#_4p54fwwaeoq4) 38

[**Part 2: Fan Base Prediction**](#_xo5w5inwr55e) **39**

[**Part 3: Emotion Change of People from Different Cities**](#_ejycz16382b4) **45**

## Introduction

Prediction of future popularity of a subject or event is a useful practice in social network analysis. In this project, the tweets from different hashtags during the Super Bowl 2015 are utilized to train different models for multiple purposes, such as predicting the number of future tweets, and analyzing the location of a user based on the tweet content. Besides, the sentiment analysis for specific regions before and after the Super Bowl event is also investigated by evaluating the positiveness and negativeness of words in tweets.

## Part 1: Popularity Prediction

##### Problem 1.1:

The statistics of hashtags and the plot of number of tweets in hour are posted below.

|  |  |  |  |
| --- | --- | --- | --- |
| Hashtag | Average number of tweets per hour | Average number of followers of users posting the tweets | Average number of retweets |
| #gohawks | 325.37159130433116 | 2203.931767444827 | 2.014617085512608 |
| #gopatriots | 45.69451057356203 | 1401.8955093016164 | 1.4000838670326319 |
| #nfl | 441.3234311373958 | 4653.252285502502 | 1.5385331089011056 |
| #patriots | 834.5555091641886 | 3309.978828415827 | 1.7828156491659402 |
| #sb49 | 1419.8879074871902 | 10267.31684948685 | 2.5111487863247035 |
| #superbowl | 2302.5004018833274 | 8858.974662784603 | 2.3882723999030224 |

Table: Statistics and different hashtags

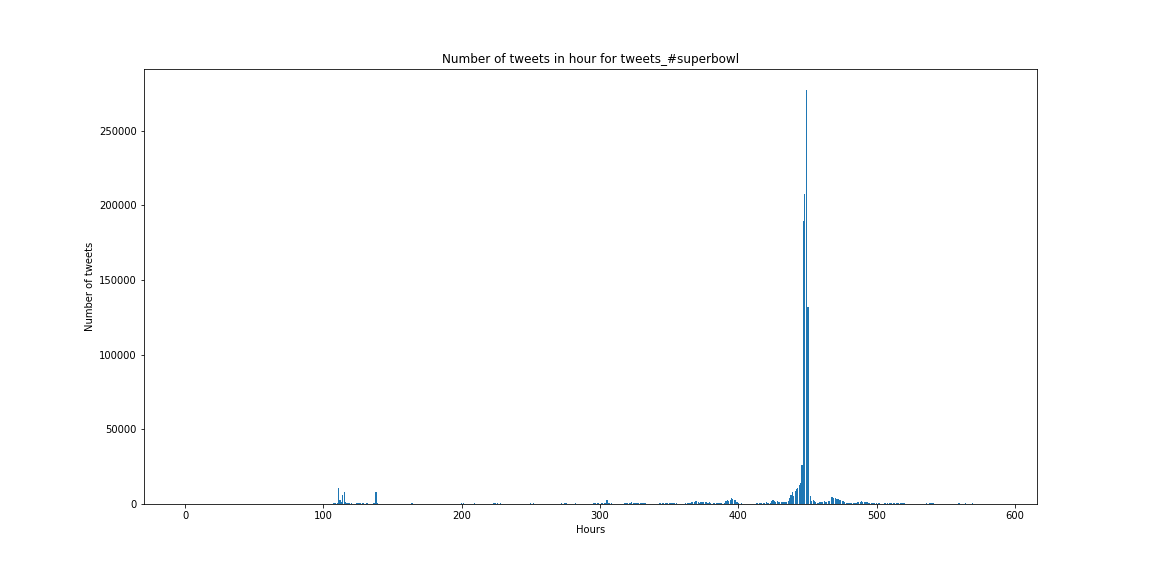


Figure: Number of Tweets per hour during super bowl

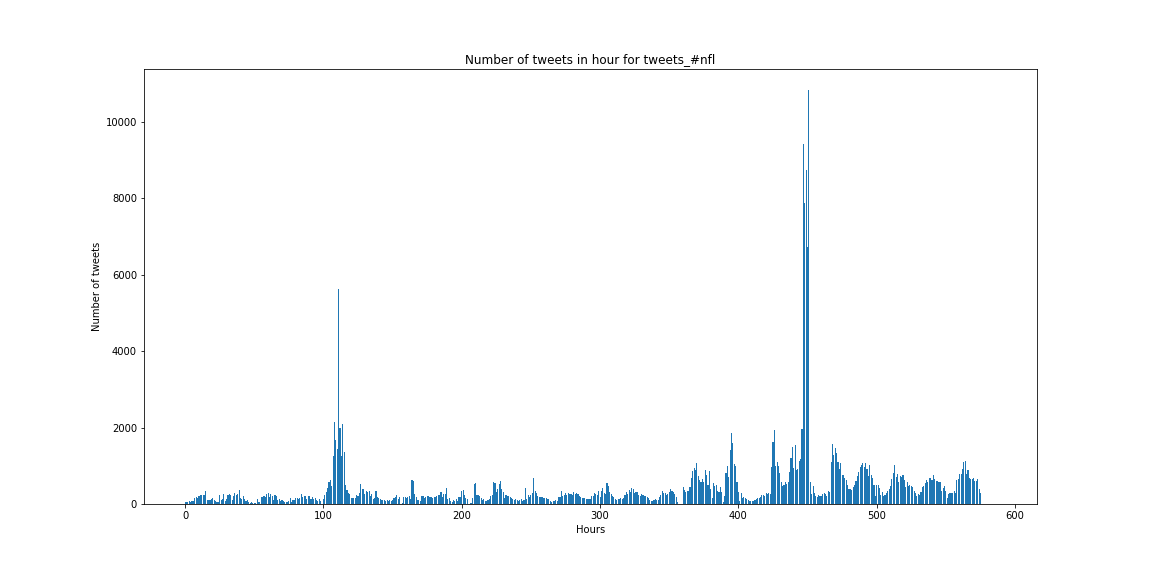


Figure: Number of Tweets per hour during NFL

##### Problem 1.2:

The accuracy (RMSE), R-squared, adjusted R-squared, coefficients, t-test and P-value of different hashtags are presented below:

Note that:

x1 stands for number of tweets;

x2 stands for total number of retweets,

x3 stands for sum of the number of followers of the users posting the hashtag;

x4 stands for maximum of followers of the users posting the hashtag;

x5 stands for time of the day.

**Hashtag: gohawks**

|  |  |
| --- | --- |
| Hashtag | #gohawks |
| RMSE | 974.084617763 |
| R-squared | 0.501 |
| Adjusted R-squared | 0.496 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 1.2319 | 7.253 | 0.000 |
| x2 | -0.1286 | -2.918 | 0.004 |
| x3 | -0.0002 | -2.059 | 0.040 |
| x4 | 2.825e-05 | 0.176 | 0.861 |
| x5 | 8.8110 | 2.660 | 0.008 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, x1, x2, x3, and x5 are considered as significant features, due to that their p-value is below 0.5.

**Hashtag: gopatriots**

|  |  |
| --- | --- |
| Hashtag | #gopatriots |
| RMSE | 185.101783785 |
| R-squared | 0.640 |
| Adj. R-squared | 0.637 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | -0.0808 | -0.316 | 0.752 |
| x2 | 0.5127 | 2.303 | 0.022 |
| x3 | 0.0002 | 1.213 | 0.226 |
| x4 | -0.0004 | -1.881 | 0.060 |
| x5 | 0.5574 | 0.921 | 0.358 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, x2, x3, x4, and x5 are considered as significant features, due to that their p-value is below 0.5.

**Hashtag: nfl**

|  |  |
| --- | --- |
| Hashtag | #nfl |
| RMSE | 585.136733318 |
| R-squared | 0.647 |
| Adj. R-squared | 0.644 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 0.7407 | 5.559 | 0.000 |
| x2 | -0.1785 | -2.789 | 0.005 |
| x3 | 7.885e-05 | 2.998 | 0.003 |
| x4 | -7.251e-05 | -2.017 | 0.044 |
| x5 | 7.5239 | 3.417 | 0.001 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, all five of the features are considered as significant features, due to that their p-value is below 0.5.

**Hashtag: patriots**

|  |  |
| --- | --- |
| Hashtag | #patriots |
| RMSE | 2527.91299191 |
| R-squared | 0.681 |
| Adj. R-squared | 0.678 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 0.9214 | 12.878 | 0.000 |
| x2 | -0.0871 | -1.476 | 0.140 |
| x3 | -1.185e-06 | -0.045 | 0.964 |
| x4 | 0.0002 | 1.770 | 0.077 |
| x5 | 3.9266 | 0.449 | 0.653 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, x1, x2, and x4, are considered as significant features, due to that their p-value is below 0.5.

**Hashtag: sb49**

|  |  |
| --- | --- |
| Hashtag | #sb49 |
| RMSE | 4471.96370687 |
| R-squared | 0.809 |
| Adj. R-squared | 0.809 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 1.1887 | 12.496 | 0.000 |
| x2 | -0.2151 | -2.453 | 0.014 |
| x3 | 1.869e-0 | 1.333 | 0.183 |
| x4 | 0.0001 | 2.117 | 0.035 |
| x5 | -3.7696 | -0.246 | 0.805 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, x1, x2, x3, and x4 are considered as significant features, due to that their p-value is below 0.5.

**Hashtag: superbowl**

|  |  |
| --- | --- |
| Hashtag | #superbowl |
| RMSE | 8004.57533406 |
| R-squared | 0.805 |
| Adj. R-squared | 0.804 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 2.3014 | 28.962 | 0.000 |
| x2 | -0.2899 | -8.059 | 0.000 |
| x3 | -0.0001 | -7.020 | 0.000 |
| x4 | 0.0008 | 5.457 | 0.000 |
| x5 | -38.9335 | -1.308 | 0.191 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this hashtag, all five of the features are considered as significant features, due to that their p-value is below 0.5.

Conclusion:

For different hashtags, the performance varys from a large range in terms of the train RMSE. The best result is achieved in “#gopatriots”, and the worst result is from “#superbowl”. This is because the distribution of “#gopatriots” is more uniform and the distribution of “#superbowl” changes more rapidly. The linear model could handle the uniform distribution, however for the “superbowl”, there is too much nonlinearity near the Super Bowl date, which leads to poor performance.

##### Problem 1.3:

The accuracy (RMSE), R-squared and adjusted R-squared value, coefficient, t-test and P-value of various hashtags are presented below.

Also, the plots of number of tweets vs three most significant features of each hashtag are presented as well. In the plot, the “log1p()” function in numpy is applied to each variable to exhibit their linear relationship more clearly.

Note that:

x1 stands for number of tweets;

x2 stands for total number of retweets,

x3 stands for sum of the number of followers of the users posting the hashtag;

x4 stands for maximum of followers of the users posting the hashtag;

x5 stands for time of the day;

x6 stands for total number of impressions;

x7 stands for total number of momentum;

x8 stands for total number of favorite count;

x9 stands for total number of ranking score;

x10 stands for total number of acceleration;

x11 stands for total number of replies;

x12 stands for total number of unique users;

x13 stands for total number of unique authors;

x14 stands for total number of user mentions.

**Hashtag: gohawks**

|  |  |
| --- | --- |
| Hashtag | #gohawks |
| RMSE | 672.584552561 |
| R-squared | 0.762 |
| Adj. R-squared | 0.756 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | -22.6534 | -4.709 | 0.000 |
| x2 | -0.1168 | -1.996 | 0.046 |
| x3 | 0.0002 | 1.485 | 0.138 |
| x4 | -0.0004 | -2.755 | 0.006 |
| x5 | 1.1086 | 0.424 | 0.671 |
| x6 | 2.846e-05 | 0.341 | 0.733 |
| x7 | -7.1465 | -5.968 | 0.000 |
| x8 | 0.0527 | 2.442 | 0.015 |
| x9 | 4.5469 | 4.954 | 0.000 |
| x10 | -0.7171 | -5.012 | 0.000 |
| x11 | 56.4566 | 5.299 | 0.000 |
| x12 | -63.6179 | -10.070 | 0.000 |
| x13 | 69.4739 | 12.031 | 0.000 |
| x14 | -1.5602 | -3.232 | 0.001 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: number of tweets, total number of ranking score, and total number of unique authors.

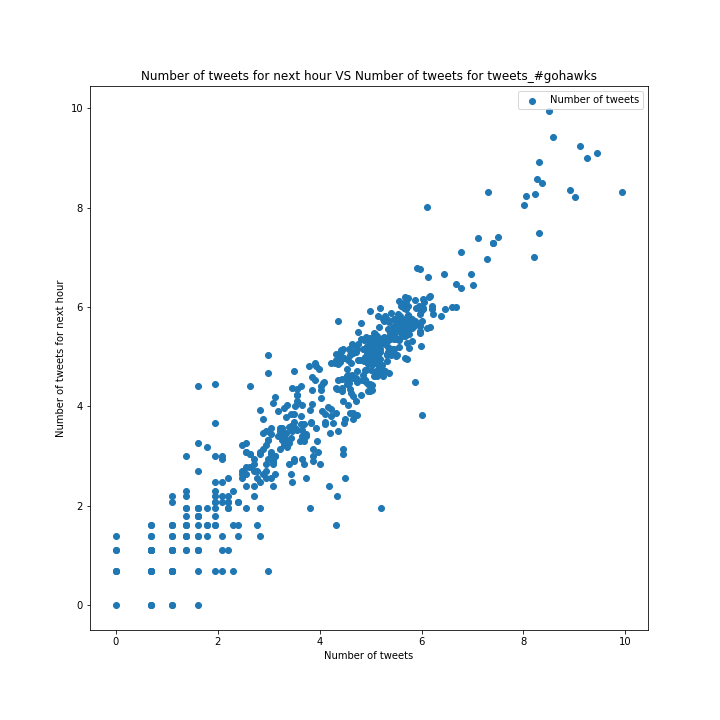


Figure: Number of Tweets of hashtag

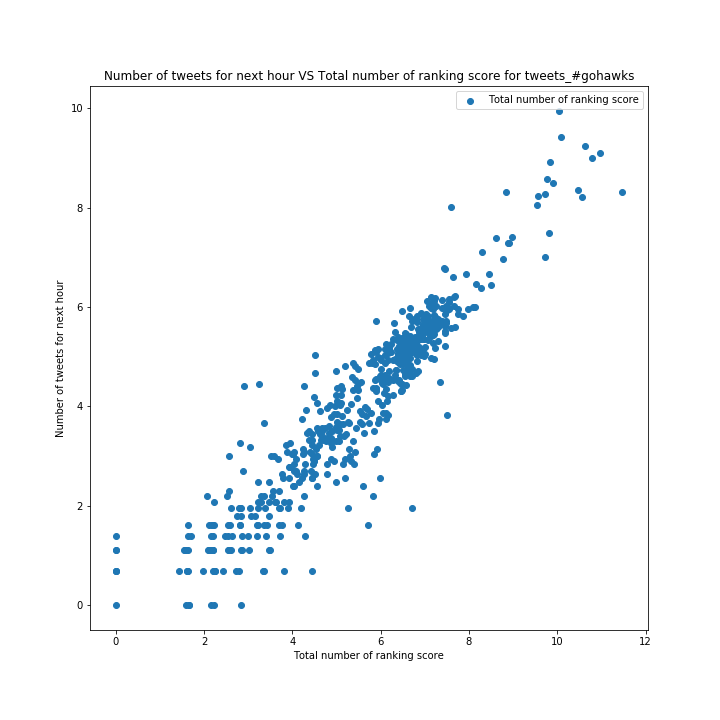


Figure: Total number of ranking score of hashtag

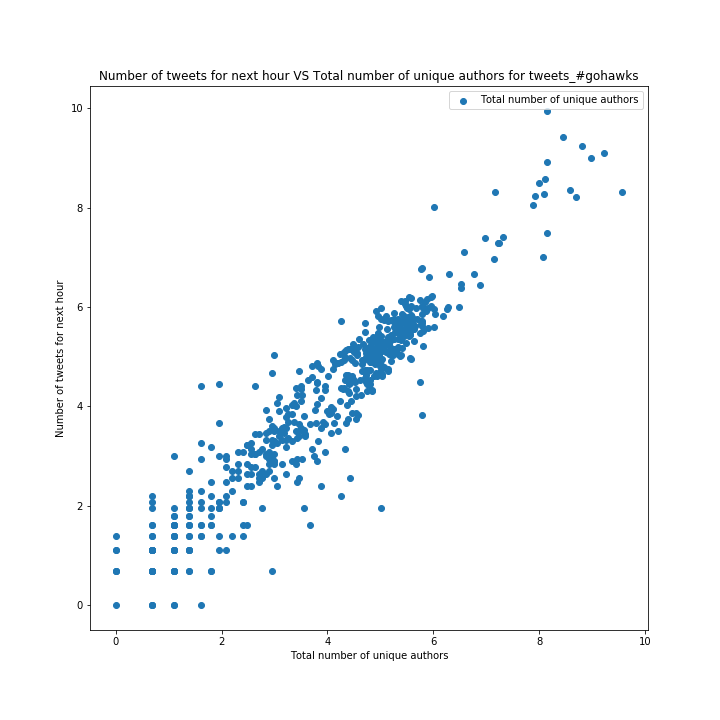


Figure: Total number of unique authors of hashtag

**Hashtag: gopatriots**

|  |  |
| --- | --- |
| Hashtag | #gopatriots |
| RMSE | 121.980426993 |
| R-squared | 0.844 |
| Adj. R-squared | 0.840 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 4.0792 | -1.273 | 0.204 |
| x2 | -1.0640 | -5.112 | 0.000 |
| x3 | 0.0009 | 2.162 | 0.031 |
| x4 | -9.642e-07 | -0.004 | 0.997 |
| x5 | -1.0635 | -2.550 | 0.011 |
| x6 | -0.0009 | -3.712 | 0.000 |
| x7 | -0.8018 | -0.427 | 0.670 |
| x8 | -8.0364 | -4.725 | 0.015 |
| x9 | 1.3826 | 2.432 | 0.000 |
| x10 | 0.4891 | 3.390 | 0.001 |
| x11 | 4.2708 | 0.847 | 0.397 |
| x12 | -12.4246 | -0.744 | 0.457 |
| x13 | 9.8617 | 0.598 | 0.550 |
| x14 | 8.5575 | 13.803 | 0.000 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: total number of ranking score, total number of retweets, and total number of user mentions.

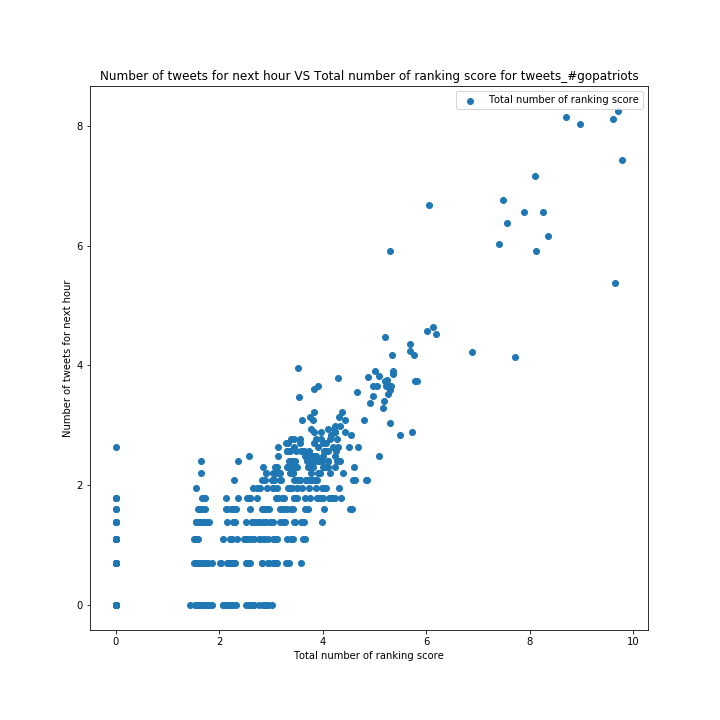


Figure: Total number of ranking score of hashtag

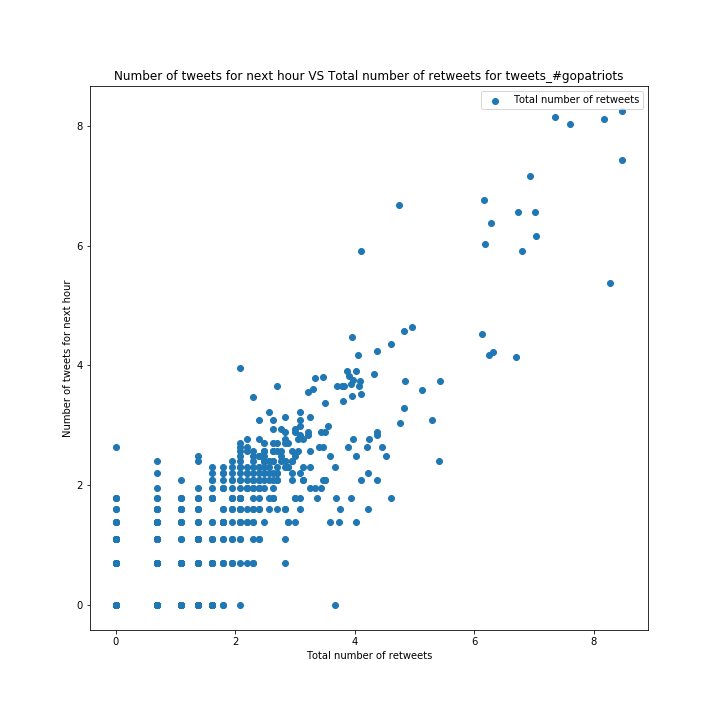


Figure: Total number of retweets of hashtag

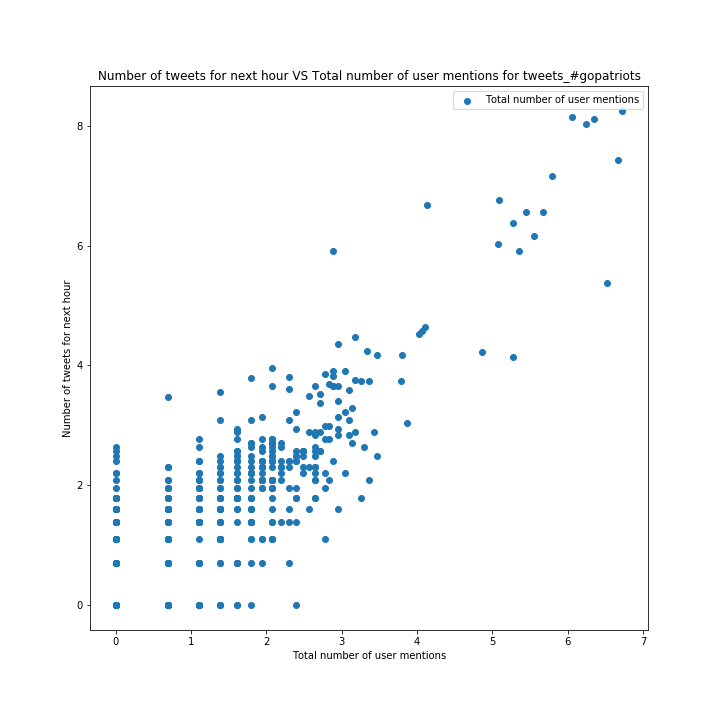


Figure: total number of user mentions of hashtag

**Hashtag: nfl**

|  |  |
| --- | --- |
| Hashtag | #nfl |
| RMSE | 465.152298147 |
| R-squared | 0.777 |
| Adj. R-squared | 0.772 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | 2.9876 | 2.193 | 0.029 |
| x2 | -0.2166 | -3.404 | 0.041 |
| x3 | -7.641e-05 | -2.052 | 0.001 |
| x4 | 0.0001 | 3.822 | 0.000 |
| x5 | 0.2825 | 0.140 | 0.889 |
| x6 | -8.066e-06 | -0.298 | 0.766 |
| x7 | -0.6673 | -0.726 | 0.468 |
| x8 | -1.6581 | -7.585 | 0.001 |
| x9 | -0.2354 | -0.762 | 0.446 |
| x10 | -0.0735 | -0.899 | 0.369 |
| x11 | -1.3699 | -0.366 | 0.715 |
| x12 | -12.3071 | -3.465 | 0.000 |
| x13 | 10.7817 | 3.103 | 0.002 |
| x14 | 4.0500 | 6.578 | 0.000 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: maximum of followers, total number of unique users, and total number of user mentions.

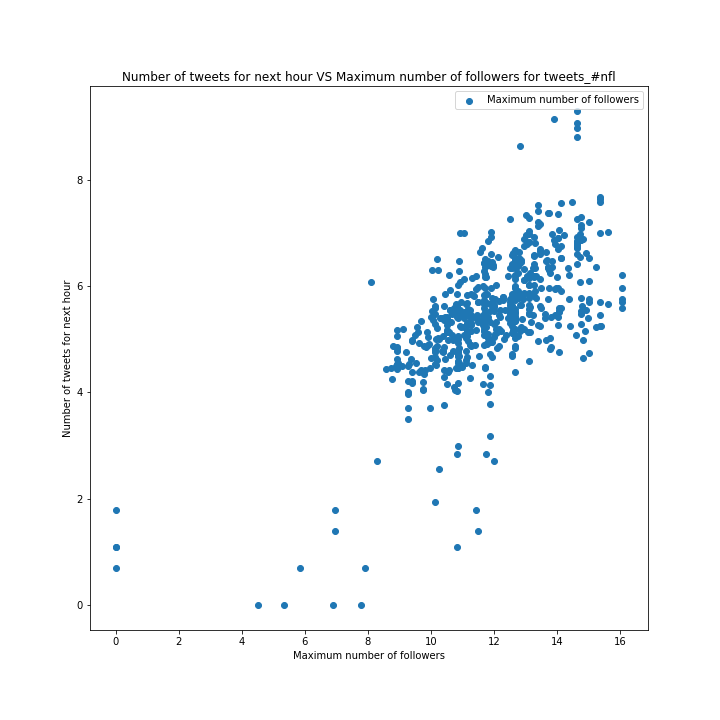


Figure: Maximum number of followers

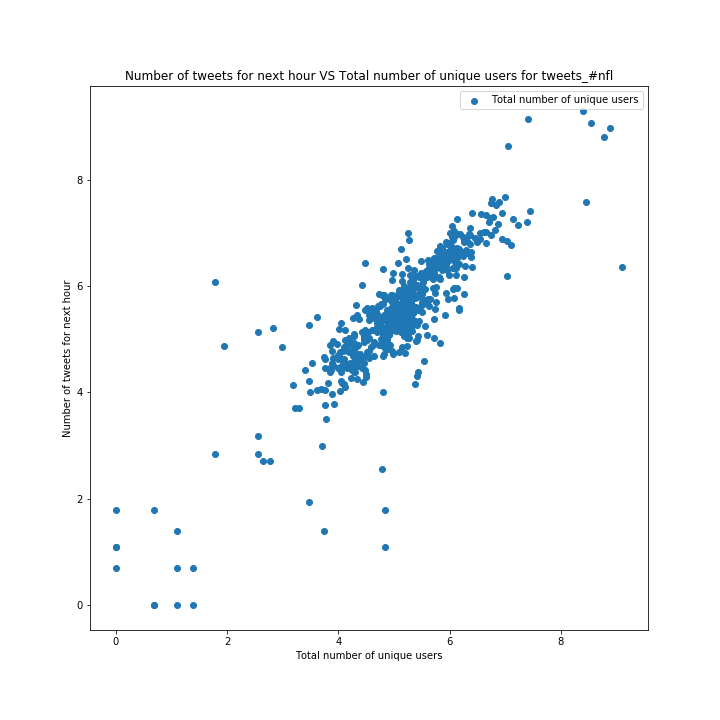


Figure: Total number of unique users

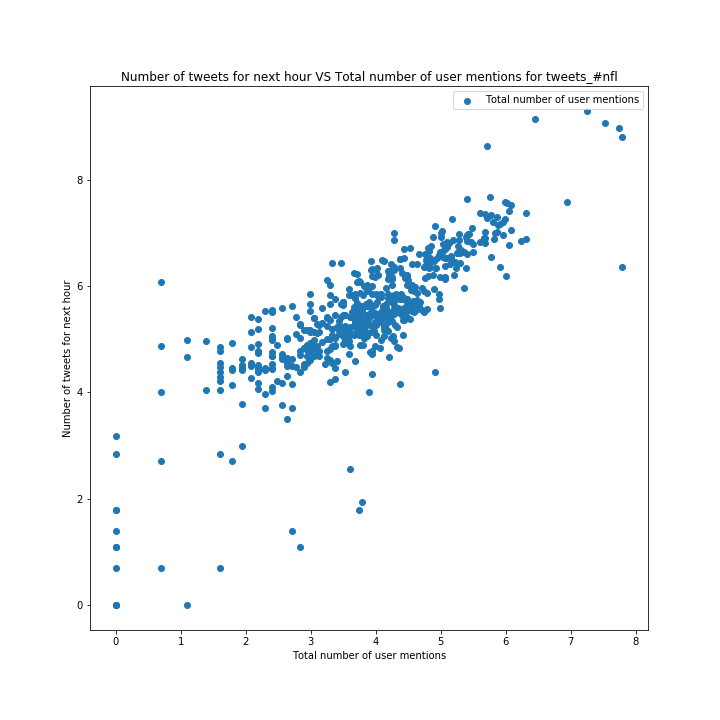


Figure: Total number of user mentions

**Hashtag: patriots**

|  |  |
| --- | --- |
| Hashtag | #patriots |
| RMSE | 1803.55837723 |
| R-squared | 0.838 |
| Adj. R-squared | 0.834 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | -30.7482 | -9.282 | 0.000 |
| x2 | -0.2034 | -2.741 | 0.006 |
| x3 | 0.0005 | 3.139 | 0.002 |
| x4 | -0.0005 | -4.720 | 0.000 |
| x5 | -1.1747 | -0.182 | 0.855 |
| x6 | -0.0002 | -1.355 | 0.176 |
| x7 | -2.5394 | -1.776 | 0.076 |
| x8 | 0.1772 | 1.029 | 0.304 |
| x9 | 6.2363 | 9.750 | 0.000 |
| x10 | 0.2426 | 1.196 | 0.232 |
| x11 | -11.1925 | -2.136 | 0.033 |
| x12 | -21.6183 | -9.218 | 0.000 |
| x13 | 24.8078 | 12.595 | 0.000 |
| x14 | 2.8806 | 17.459 | 0.000 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: number of tweets, total number of ranking score, and total number of unique authors.

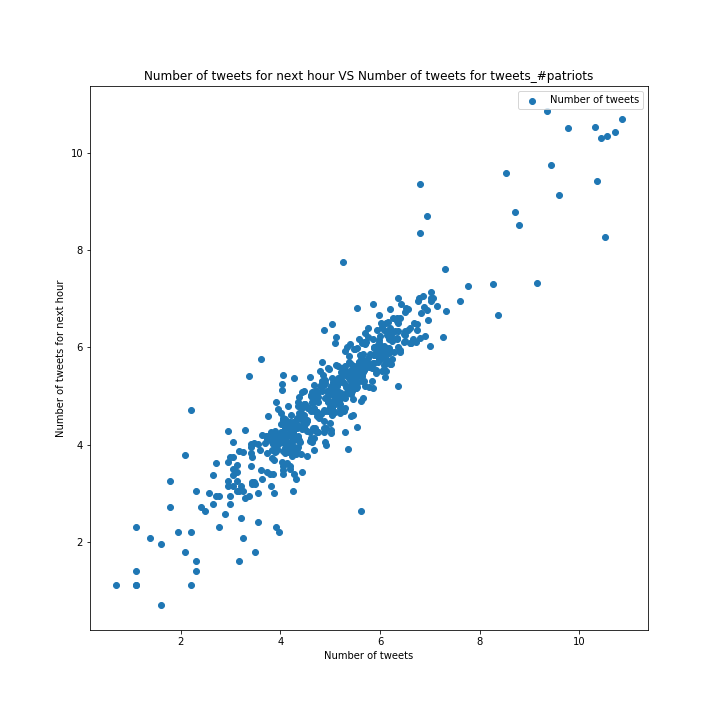


Figure: Number of Tweets of hashtag

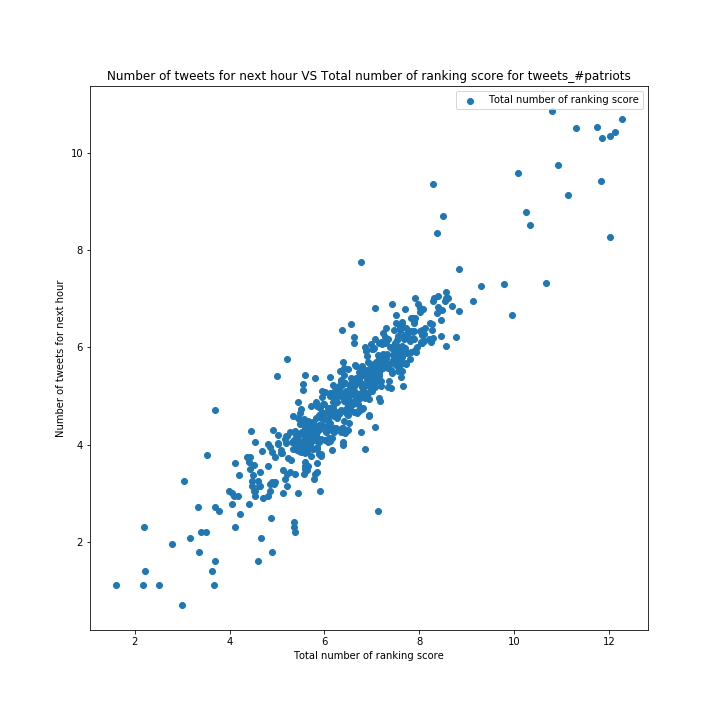


Figure: Total number of ranking score of hashtag

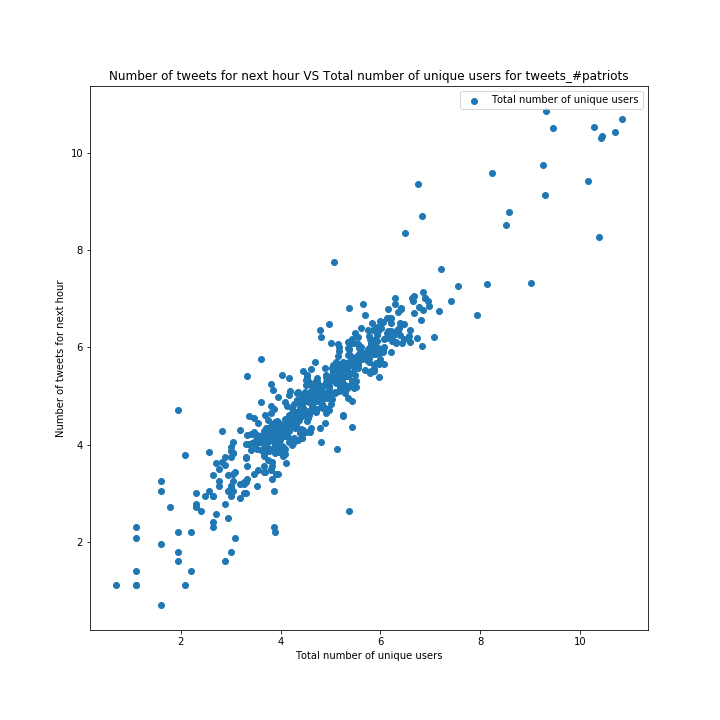


Figure: Total number of unique authors of hashtag

**Hashtag: sb49**

|  |  |
| --- | --- |
| Hashtag | #sb49 |
| RMSE | 3245.01564061 |
| R-squared | 0.899 |
| Adj. R-squared | 0.897 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | -31.5130 | -5.769 | 0.000 |
| x2 | 0.1385 | 1.301 | 0.194 |
| x3 | 8.078e-05 | 1.290 | 0.197 |
| x4 | -0.0003 | -5.438 | 0.000 |
| x5 | 11.3405 | 0.993 | 0.321 |
| x6 | -6.141e-05 | -0.985 | 0.325 |
| x7 | 0.7170 | 1.233 | 0.218 |
| x8 | -0.2500 | -2.832 | 0.005 |
| x9 | 6.1970 | 5.897 | 0.000 |
| x10 | 0.4235 | 2.189 | 0.029 |
| x11 | -11.1917 | -1.470 | 0.142 |
| x12 | -20.9711 | -11.067 | 0.000 |
| x13 | 24.5531 | 16.316 | 0.000 |
| x14 | 3.3118 | 13.842 | 0.000 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: number of tweets, total number of ranking score, and total number of unique authors.

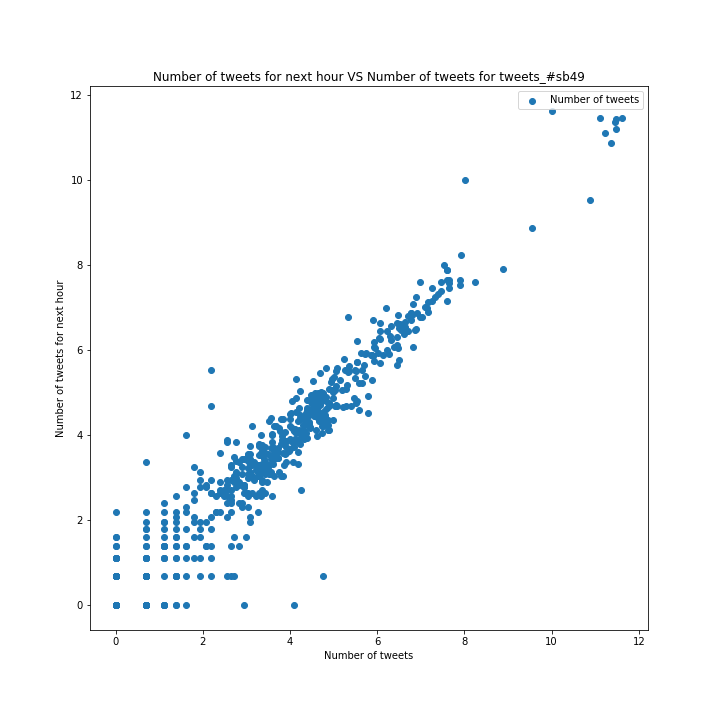


Figure: Number of Tweets of hashtag

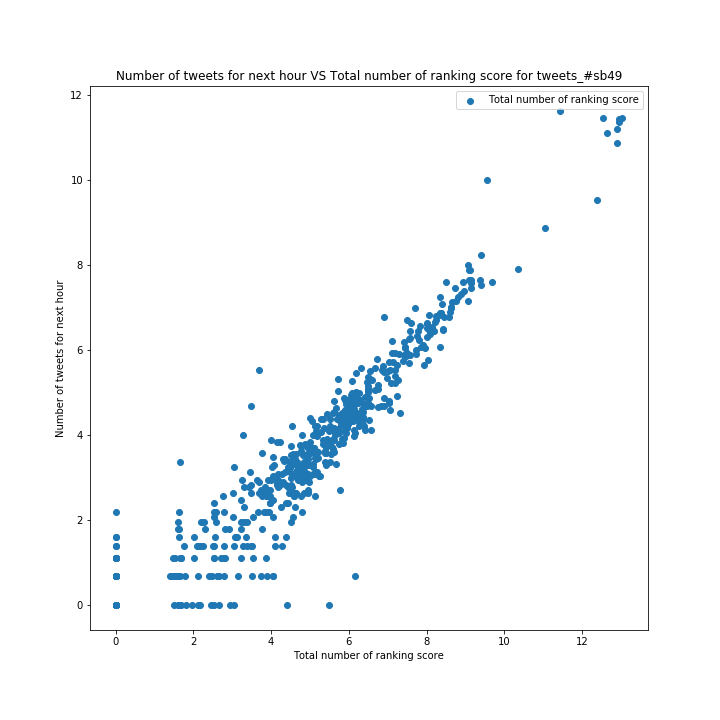


Figure: Total number of ranking score of hashtag

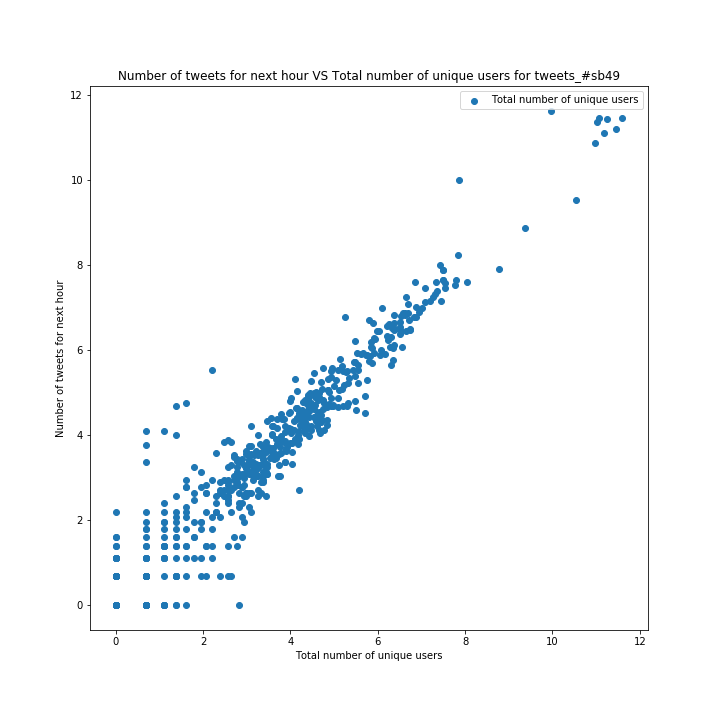


Figure: Total number of unique users of hashtag

**Hashtag: superbowl**

|  |  |
| --- | --- |
| Hashtag | #superbowl |
| RMSE | 5690.04403893 |
| R-squared | 0.902 |
| Adj. R-squared | 0.899 |

Table: RMSE and R-squared Measure of the hashtag

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coefficient | t-test | P-value |
| x1 | -1.8422 | -0.235 | 0.814 |
| x2 | 0.3737 | 2.288 | 0.000 |
| x3 | 6.234e-05 | 0.302 | 0.763 |
| x4 | -0.0001 | -1.079 | 0.281 |
| x5 | -27.4865 | -1.225 | 0.221 |
| x6 | -4.387e-05 | -0.211 | 0.833 |
| x7 | -12.2871 | -5.833 | 0.000 |
| x8 | -0.7528 | -3.173 | 0.002 |
| x9 | 0.7817 | 0.457 | 0.648 |
| x10 | -1.5904 | -4.701 | 0.023 |
| x11 | 4.1731 | 0.175 | 0.861 |
| x12 | -33.6627 | -4.252 | 0.000 |
| x13 | 35.2990 | 4.227 | 0.000 |
| x14 | 2.1936 | 1.184 | 0.237 |

Table: Coefficient, t-test and P-value of the hashtag

Note that for this part, the three most important features are: total number of retweets, total number of unique authors, and total number of unique users.

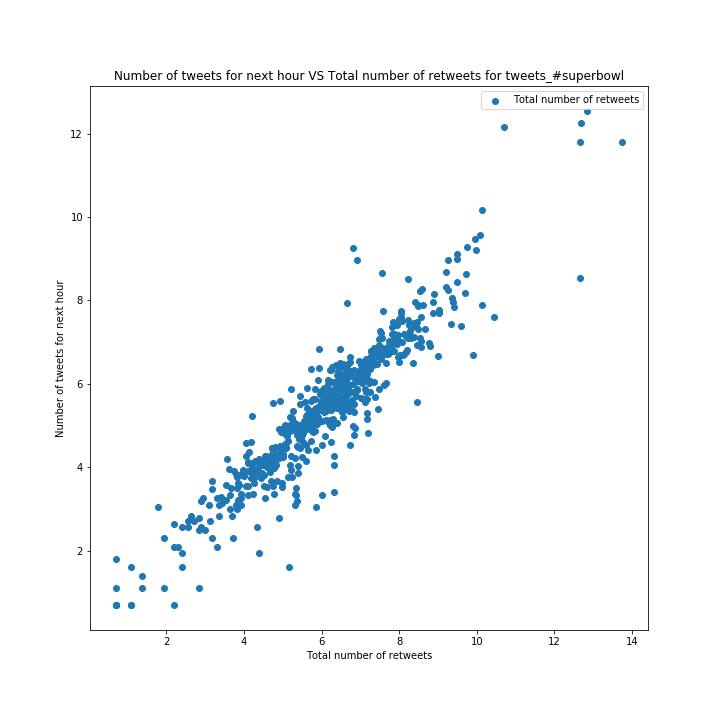


Figure: Total number of retweets of the hashtag

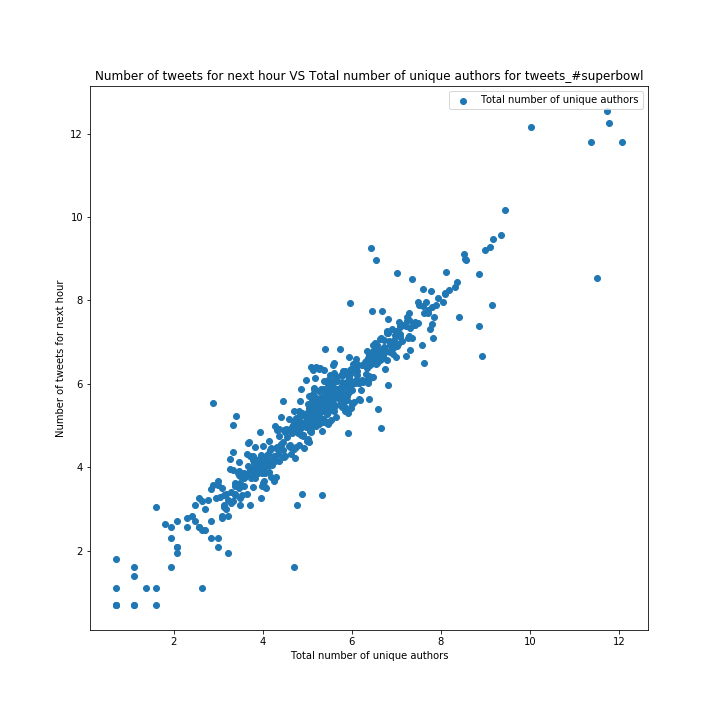


Figure: Total number of unique authors of hashtag

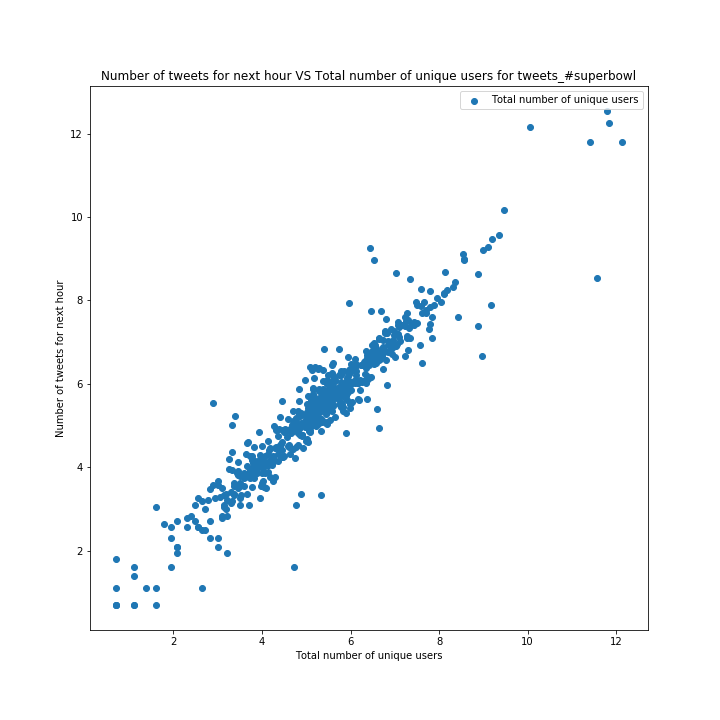


Figure: Total number of unique users of hashtag

Conclusion:

For “#gopatriots” and “#nfl”, the linear relationship is not so clear. The features extracted are not so good for these two hashtags. As to other hashtags, the linear relationship is very clear, which means some relevant features are extracted for these hashtags.

##### Problem 1.4:

**Part a:**

The average 10-fold cross-validation error is presented below:

**Hashtag: gohawks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 440 | Linear Regression | 298.250292112 | 1030.69840366 |
| SVM Regression | 1030.69840366 | 1140.98987637 |
| Random Forest Regression | 159.763159091 | 1098.33721526 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 5084.98598697 | 8918.41019784 |
| SVM Regression | 3158.75 | 4083.77110238 |
| Random Forest Regression | 2157.33 | 3141.50610496 |
| After Feb. 1, 8:00 p.m. | 126 | Linear Regression | 6740.71143658 | 63304.7552305 |
| SVM Regression | 37.5343205237 | 121.575928261 |
| Random Forest Regression | 23.7777042706 | 81.3093638529 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

**Hashtag: gopatriots**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 439 | Linear Regression | 22.0598487327 | 67.6972750246 |
| SVM Regression | 13.2001855072 | 71.9478610947 |
| Random Forest Regression | 11.3983795148 | 69.4109016209 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 8690.18987903 | 13621.6334306 |
| SVM Regression | 1538.08333333 | 1761.37089129 |
| Random Forest Regression | 980.210833333 | 1286.27480734 |
| After Feb. 1, 8:00 p.m. | 123 | Linear Regression | 80.7213810108 | 474.589134017 |
| SVM Regression | 5.55822679077 | 15.3185519403 |
| Random Forest Regression | 4.27702829397 | 11.4140969333 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

**Hashtag: nfl**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 440 | Linear Regression | 128.836245629 | 328.093420821 |
| SVM Regression | 189.434848485 | 425.898115931 |
| Random Forest Regression | 116.546136364 | 288.594964932 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 39709.7504202 | 55347.9044536 |
| SVM Regression | 4284.33333333 | 5099.60981351 |
| Random Forest Regression | 2539.42583333 | 3644.08919497 |
| After Feb. 1, 8:00 p.m. | 134 | Linear Regression | 127.980440101 | 181.07610519 |
| SVM Regression | 294.791044776 | 358.053158732 |
| Random Forest Regression | 151.444104478 | 201.902611064 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

**Hashtag: patriots**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 440 | Linear Regression | 338.884383446 | 1096.04448297 |
| SVM Regression | 265.085 | 998.687747315 |
| Random Forest Regression | 204.608931818 | 905.766484155 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 94902.0112714 | 106332.950136 |
| SVM Regression | 15237.8333333 | 18547.8702102 |
| Random Forest Regression | 18348.645 | 21432.7353688 |
| After Feb. 1, 8:00 p.m. | 134 | Linear Regression | 281.783659442 | 1510.09226455 |
| SVM Regression | 153.241492537 | 355.521696368 |
| Random Forest Regression | 109.471641791 | 270.814047658 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

**Hashtag: sb49**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 436 | Linear Regression | 49.5088492706 | 127.361776957 |
| SVM Regression | 106.364360333 | 283.500776443 |
| Random Forest Regression | 46.7933965706 | 143.550302691 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 96352.9399745 | 122229.459871 |
| SVM Regression | 38629.5833333 | 44724.6151488 |
| Random Forest Regression | 30570.0833333 | 37209.7069289 |
| After Feb. 1, 8:00 p.m. | 134 | Linear Regression | 161.194021903 | 396.936461928 |
| SVM Regression | 358.505970149 | 739.227348544 |
| Random Forest Regression | 131.64261194 | 302.727315668 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

**Hashtag: superbowl**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Time Period | Number of hours in this period | Regression Method | Validation mean absolute error | Validation RMSE |
| Before Feb. 1, 8:00 a.m. | 440 | Linear Regression | 370.599044835 | 1060.59861696 |
| SVM Regression | 441.547272727 | 1111.63057799 |
| Random Forest Regression | 262.827431818 | 782.711914134 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 12 | Linear Regression | 2542185.93525 | 4258684.6098 |
| SVM Regression | 98594.6666667 | 123847.637746 |
| Random Forest Regression | 59965.3483333 | 89870.5818864 |
| After Feb. 1, 8:00 p.m. | 134 | Linear Regression | 406.452285507 | 1460.64574057 |
| SVM Regression | 606.205970149 | 1035.57347489 |
| Random Forest Regression | 284.475597015 | 602.485406493 |

Table: Validation mean absolute error and validation RMSE of different time period of given hashtag

In this part, as required, the cross validation is performed. So, instead of reporting the train RMSE as precious parts, we report the test RMSE. That is the reason why the RMSE of this part is different from other sections. The large RMSE or MAE usually appears in the middle periods. That is because the data tends to change rapidly during the Super Bowl, it is difficult for the model to fit this extreme nonlinearity. Besides, there is only 12 data points within this period, and the insufficient number of training samples may lead to bad performance as well.

**Part b:**

The statistics of combines model is shown below.

The best model is Random Forest Regressor.

|  |  |  |
| --- | --- | --- |
| Time period | Train RMSE | Train mean absolute error |
| Before Feb. 1, 8:00 a.m. | 83.4185978459 | 28.8931528165 |
| Between Feb. 1, 8:00 a.m. and 8:00 p.m. | 82.4672430111 | 29.1761430664 |
| After Feb. 1, 8:00 p.m. | 86.107902935 | 29.5298814152 |

Table: Train RMSE and Train mean absolute error of different time period using random forest regressor

From the value of train RMSE and train MAE, we could tell that both this two values of different time periods have dropped dramatically using combines model. This means that the best model could fit the nonlinear distribution of the data well.

##### Problem 1.5:

In this part, we use “citation date” as our criteria. Since for “sample8\_period1.txt”, there is only 5-hour data, using a 5-hour window is not feasible. So, we use a 4-hour window instead. To make the results consistent, for all the testing samples, we always use the 4-hour window to predict the result for the last hour for all test files. In addition, we only train a universal model across all periods for this prediction as required.

The predicted and true number of tweets of different time periods is presented below.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Period | True number of tweets for the next hour | Predicted number of tweets for the next hour |
| 1 | 1 | 1.0 | 30.39 |
| 2 | 2 | 4.0 | 14.28 |
| 3 | 3 | 523.0 | 679.62 |
| 4 | 1 | 201.0 | 253.2 |
| 5 | 1 | 1.0 | 0.58 |
| 6 | 2 | 14.0 | 24.6 |
| 7 | 3 | 120.0 | 70.67 |
| 8 | 1 | 11.0 | 26.8 |
| 9 | 2 | 1.0 | 0.582802669553 |
| 10 | 3 | 61.0 | 52.77 |

Table: Predicted and True number of tweets of different time periods

The scale of number of can vary largely in different period. Generally speaking, the predicted numbers are on the same scale as the true numbers. The prediction error is within a reasonable range. The predicted number can be more accurate if we can have sufficient data as input (Since the test files are clipped based on the “first post date”, the test set is an incomplete dataset for “citation date” criteria). Besides, the number of tweets are very small in the last hour, which further increases the prediction difficulty. As shown, sometimes the model can handle this nonlinearity, yet sometimes there is a relatively big difference between the prediction and the ground truth.

## Part 2: Fan Base Prediction

In this part, the tweets for two states (Washington and Massachusetts.) are collected for classification. We set Washington as Class 0, and Massachusetts as Class 1. The number of tweets for these two classes are listed below.

The size of class 0 is 41695.

The size of class 1 is 50267.

We first split the data into train set and test set (10% data is used for testing). Then similar to project 1, we perform TFxIDF transformation on the tweet content after stemming. Since the dimension is too high, we also perform LSI for dimension reduction. The dimensions of TFxIDF vectors before and after LSI are shown below.

|  |  |
| --- | --- |
| Train TFxIDF shape | (82765, 21008) |
| Test TFxIDF shape | (9197, 21008) |
| Train shape after LSI | (82765, 60) |
| Test shape after LSI | (9197, 60) |

Table: dimension of TFxIDF vectors

We then use several classifiers to test the performance of classification. Accuracy, precision, recall, confusion matrix and ROC of different models are presented below.

**Model: Naive Bayes**

|  |  |
| --- | --- |
| Accuracy | 0.624442753072 |
| Precision | 0.687998124707 |
| Recall | 0.580268880981 |

Table: Accuracy, Precision and Recall of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 2935 | 2123 |
| True 1 | 1331 | 2808 |

Table: Unnormalized Confusion Matrix of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 0.58026888 | 0.41973112 |
| True 1 | 0.32157526 | 0.67842474 |

Table: Normalized Confusion Matrix of the Model

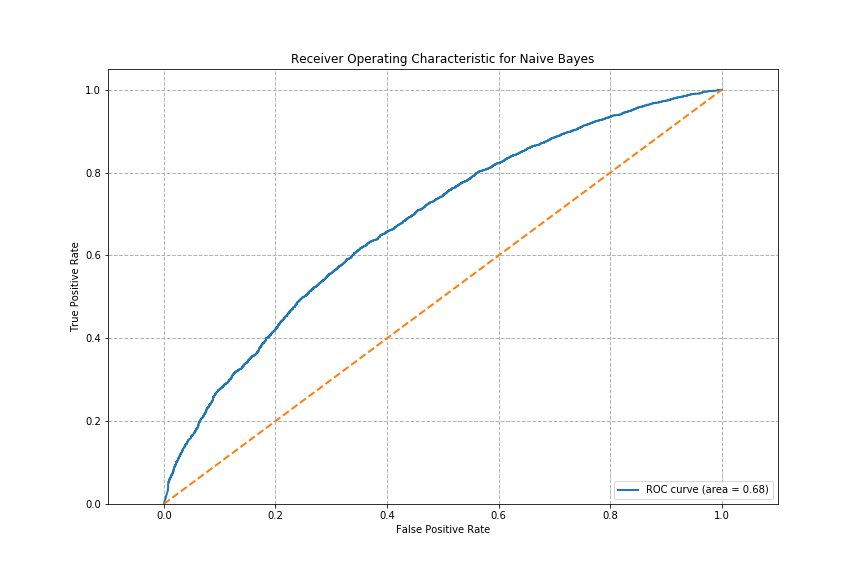


Figure: ROC of the Model

**Model: Random Forest**

|  |  |
| --- | --- |
| Accuracy | 0.647928672393 |
| Precision | 0.654804270463 |
| Recall | 0.689425478768 |

Table: Accuracy, Precision and Recall of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 3312 | 1746 |
| True 1 | 1492 | 2647 |

Table: Unnormalized Confusion Matrix of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 0.65480427 | 0.34519573 |
| True 1 | 0.36047354 | 0.63952646 |

Table: Normalized Confusion Matrix of the Model

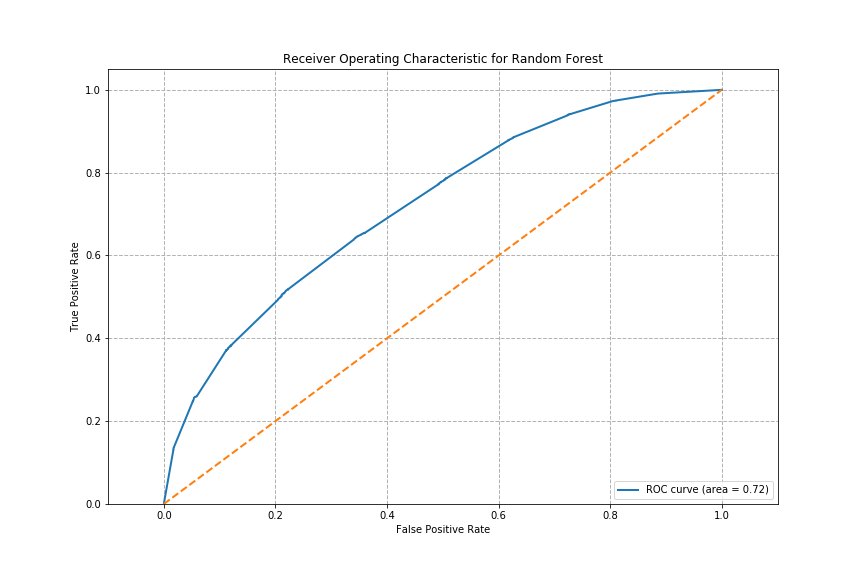


Figure: ROC of the Model

**Model: AdaBoost**

|  |  |
| --- | --- |
| Accuracy | 0.668152658476 |
| Precision | 0.664534120735 |
| Recall | 0.800909450376 |

Table: Accuracy, Precision and Recall of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 4051 | 1007 |
| True 1 | 2045 | 2094 |

Table: Unnormalized Confusion Matrix of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 0.80090945 | 0.19909055 |
| True 1 | 0.4940807 | 0.5059193 |

Table: Normalized Confusion Matrix of the Model

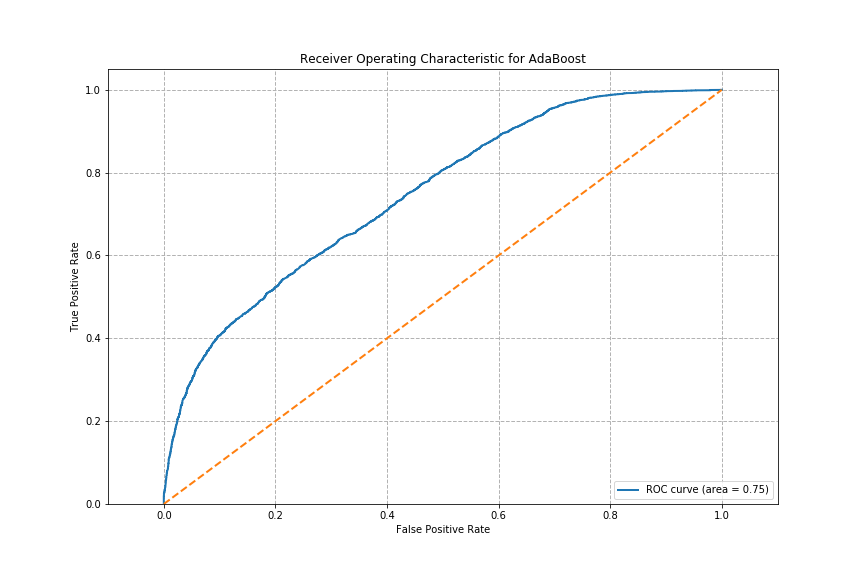


Figure: ROC of the Model

**Model: kNN**

|  |  |
| --- | --- |
| Accuracy | 0.654017614439 |
| Precision | 0.673510913799 |
| Recall | 0.719849742981 |

Table: Accuracy, Precision and Recall of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 3641 | 1417 |
| True 1 | 1765 | 2374 |

Table: Unnormalized Confusion Matrix of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 0.71984974 | 0.28015026 |
| True 1 | 0.42643151 | 0.57356849 |

Table: Normalized Confusion Matrix of the Model

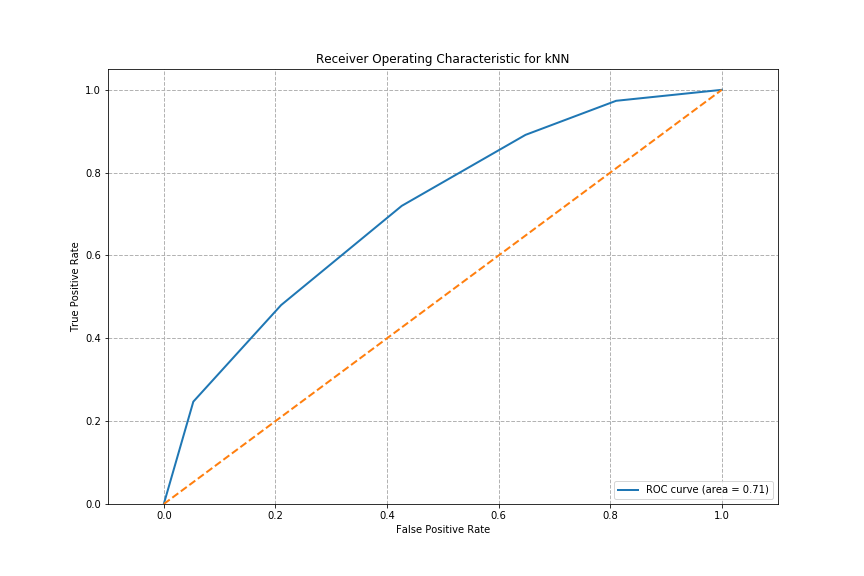


Figure: ROC of the Model

**Model: Neural Network with 3 hidden layers**

|  |  |
| --- | --- |
| Accuracy | 0.658366858758 |
| Precision | 0.66901905434 |
| Recall | 0.749703440095 |

Table: Accuracy, Precision and Recall of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 3792 | 1266 |
| True 1 | 1876 | 2263 |

Table: Unnormalized Confusion Matrix of the Model

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| True 0 | 0.74970344 | 0.25029656 |
| True 1 | 0.45324958 | 0.54675042 |

Table: Normalized Confusion Matrix of the Model

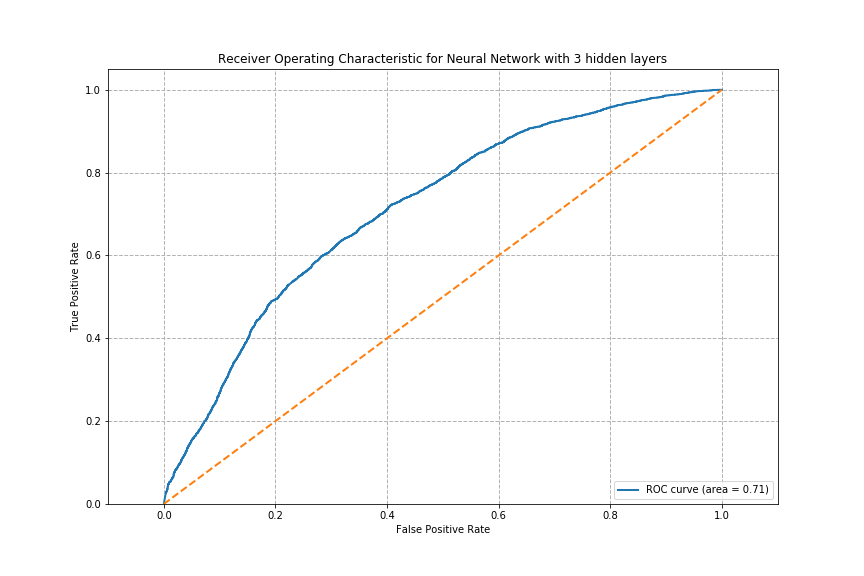


Figure: ROC of the Model

Conclusion:

Of the five models we trained and tested on the fan base prediction data, AdaBoost achieved the best performance with a Recall measurement of over 0.8, it also produced the largest area of 0.75 under ROC curve. KNN and a neural network with 3 hidden layers also performed relatively better than the worst-performing Naive Bayes. KNN achieved a classifying precision of 0.6735. Random Forest also did as well as KNN and neural network with a second highest area under ROC curve.

## Part 3: Emotion Change of People from Different Cities

**Problem Statement:**

We ought to infer the emotional change of twitter users during Super Bowl 2015 by analyzing their tweet data. We used twitter data logged from Seattle and Boston on the game day and the day after Super Bowl 2015.

**Assumption:**

- People’s emotion are oscillating primarily with respect to game results and player’s performance.

- For fans of participating team, their emotion should vary largely with the game results. If they think their team is more likely to winning, they will show more positive emotion through tweets with a hashtag of “#go seahawk” or “#go patriot”.

**Implementation:**

We chose to use the emotion detector ‘textblob’ trained on movie comments. As a demonstration, we use Naive Bayes algorithm to train the classifier to classify the tweet into three categories of emotion, namely positive, negative and neutral. We then did two separate experiments on twitter data set from Boston area and twitter data set from Seattle area. The reason of testing only on dataset from two regions is to reduce the variation of action in different cities. Boston and Seattle are the two most representative and involved cities in this game. In each city, we tested on three subsets, first the set of all the related hashtag tweets, and the sets of containing “#go seahawk” or “#go patriot”. We then compared the variation of emotion inferred from the datasets to how we assumed, as a ground truth, that the group of twitter users would change their emotion.

**Analysis and Evaluation:**

We recorded the number of tweets of each emotional category during the game day and the day after (48 samples for each hour).

In both cities, the number of tweets start climbing up at around noon time, when people are start to feel the vibe of the game when it’s approaching. In both cities, there are a big bump in “positive” tweets at hour 9, which is the noon time in eastern time zone, when the NFL TV coverages began. People are getting excited about the game day. While on the other hand, the number/percentage of “negative” emotional tweets to “positive” tweets has been monotonically increasing right after the hour 9 until the game was over (at hour 20). People tend to show nerves and negative emotions during the game, especially in these cities where there mostly are fans caring about their home team winning.

The most interesting window is the game time of about 4 hours, which corresponds to 16 to 20 in the x-axis. The emotional change in different cities has interesting patterns. In Seattle, the emotional distribution(percentage of each emotion) has been relatively steady from all the tweet data. While the Boston has seen up-and-down , the “negative” emotion tweets kept climbing up throughout the game. This is because the Seattle Seahawks has been leading from start and for most of all the game until the very end. Therefore, Patriots fans in Boston are building up the nerves as the game is progressing, which corresponds to the trend we have seen from the emotion extractor.

Another interesting result we have seen is the “#gopatriots” and “#goseahawks” tweets at the end of the game. “#go patriots” twitter has skyrocketed positive percentage, essentially because they have made a big comeback and won at the end. Although the Seahawk lost at the end, the tweets with “#goseahawks” has been posting a high percentage of “positive” tweets as well the hour right after the game. This could be the case that Seahawk fans were posting encourageous tweets to the team, and appreciating their team as all came to fold.

Overall, we have seen that the emotion extractor has picked up traits of twitter user’s reaction to the game and how their emotion changed. As this is a demonstration of the idea, there are a few future steps to be considered. We could use another model trained from a dataset other than the movie comments, as the movie comments might have bias towards the classification of people’s emotion relative to tweets on sports. We might also want to decrease the time constant to “half hour or 20 mins. The game has changed rapidly, so that we could follow in better resolution how fans react to the game development.

All the plots are presented in following pages.

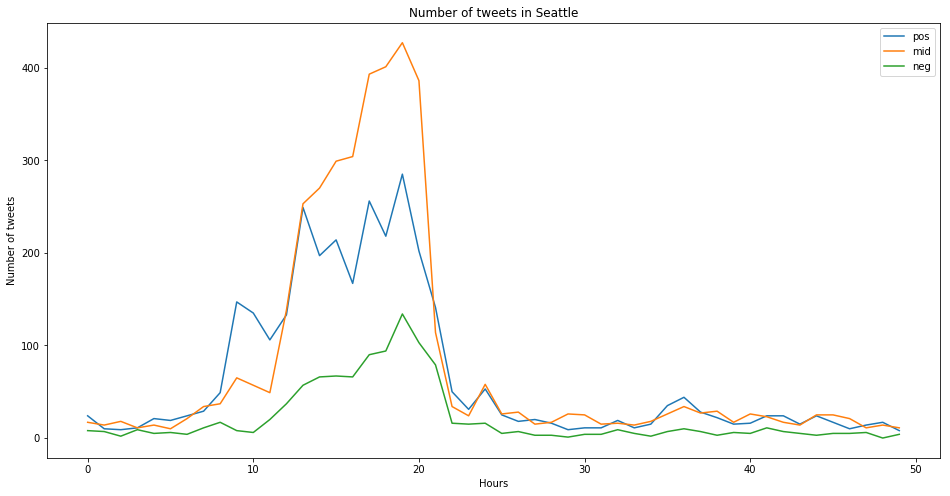


Figure: Numbers of Tweets in Seattle -- All Hashtags

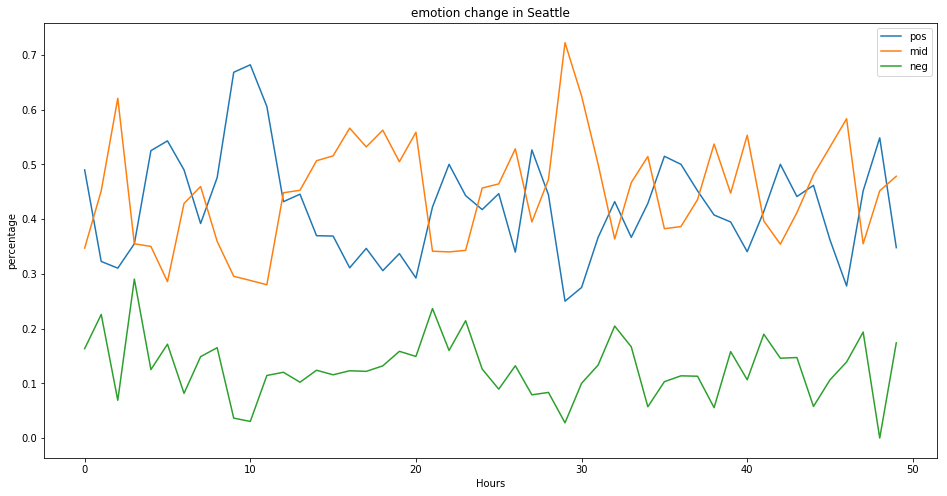


Figure: Percentage of Different Emotion Tweets in Seattle -- All Hashtags

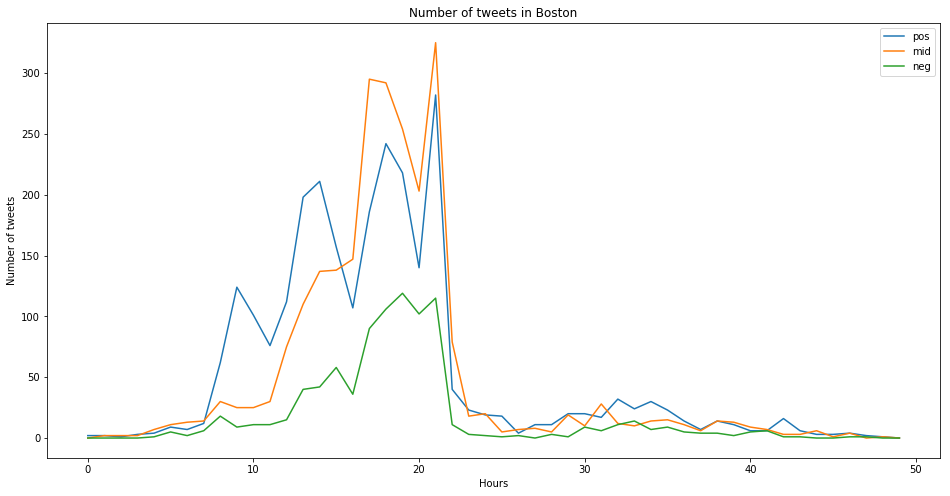


Figure: Numbers of Emotional Tweets in Boston -- All Hashtags

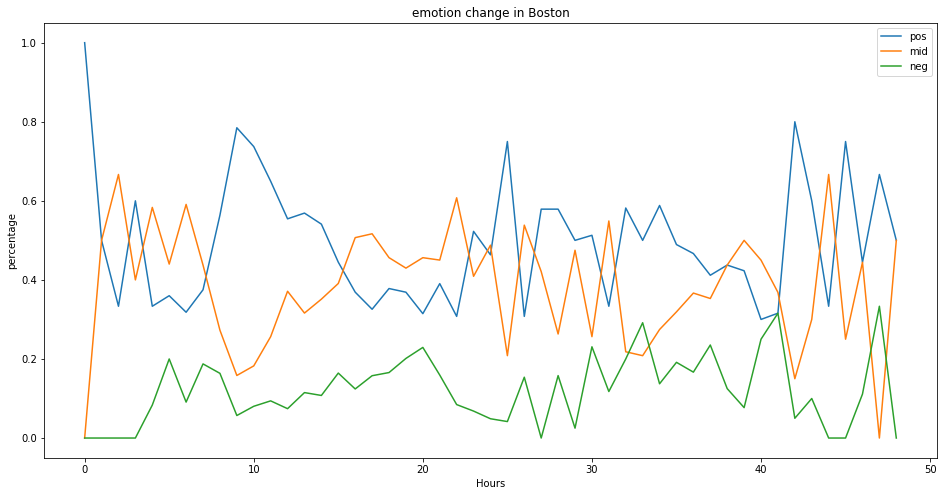


Figure: Percentage of Different Emotion Tweets in Boston -- All Hashtags

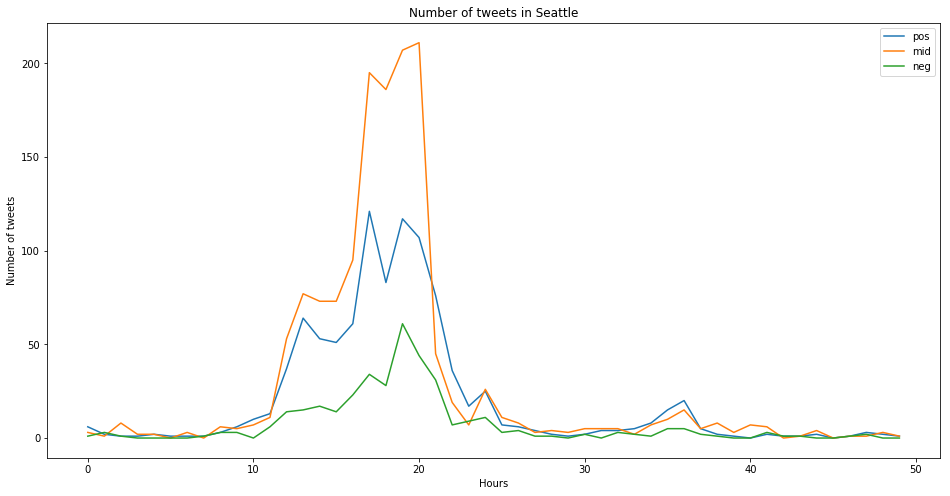


Figure: Numbers of Tweets in Seattle -- #goseahawks

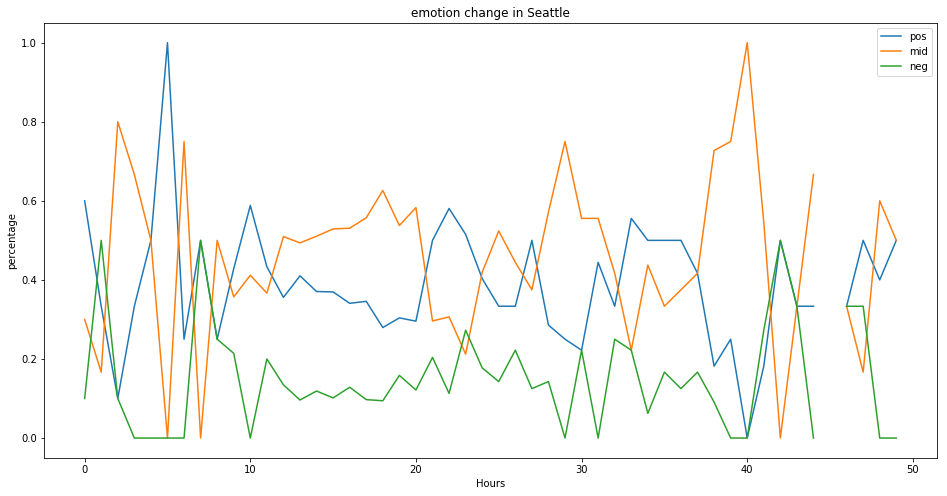


Figure: Percentage of Different Emotion Tweets in Seattle -- #goseahawks

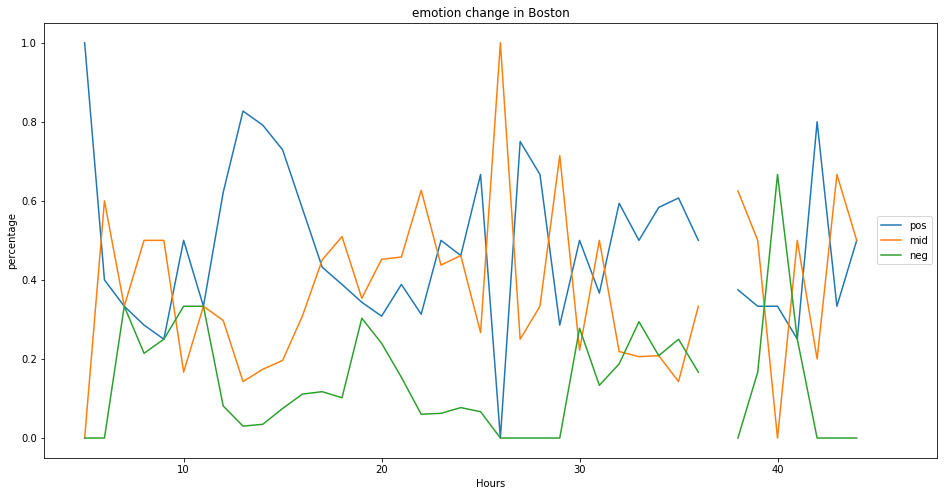


Figure: Numbers of Emotional Tweets in Boston -- #gopatriots

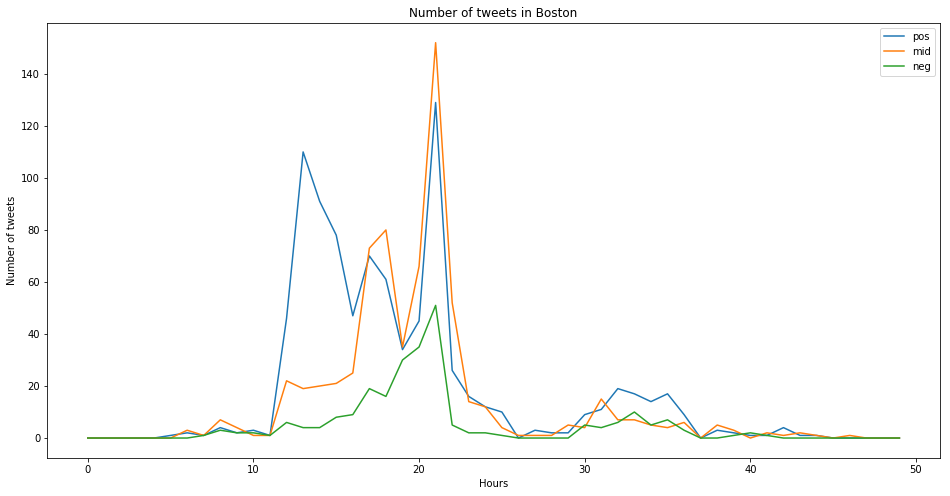


Figure: Percentage of Different Emotion Tweets in Seattle -- #gopatriots