EE219 Project5
Popularity Prediction on Twitter

Qidi Sang 705028670
Hui Wang 205036597
Zhonglin Zhang 005030520

Introduction

Twitter, with its public discussion model, is a good platform to perform social network analysis and to predict future popularity of a subject or event. With Twitter's topic structure in mind, we can predict its tweet activity in the future, or predict if it will become more popular and if so by how much, knowing current (and previous) tweet activity for a hashtag.

In this project, the Twitter dataset is collected by querying popular hashtags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. We will use data from some of the related hashtags to train a regression model and then use the model to make predictions for other hashtags. To train the model, we need to prepare training sets out of the data, extract features for them, and then fit a regression model on it. The regression model will try to fit a curve through observed values of features and outcomes to create a predictor for new samples.

We will use the given training data to create the model, and test data to make predictions. The test data consists of tweets containing a hashtag in a specified time window, and will be predicted number of tweets containing the hashtag posted within one hour immediately following the given time window, using the model we created.

Part 1: Popularity Prediction

Problem 1.1

For hashtag #gohawks:
Average number of tweets per hour is 325.37159130433116
Average number of followers of users posting the tweets per hour is 2203.931767444827
Average number of retweets per hour is 2.014617085512608

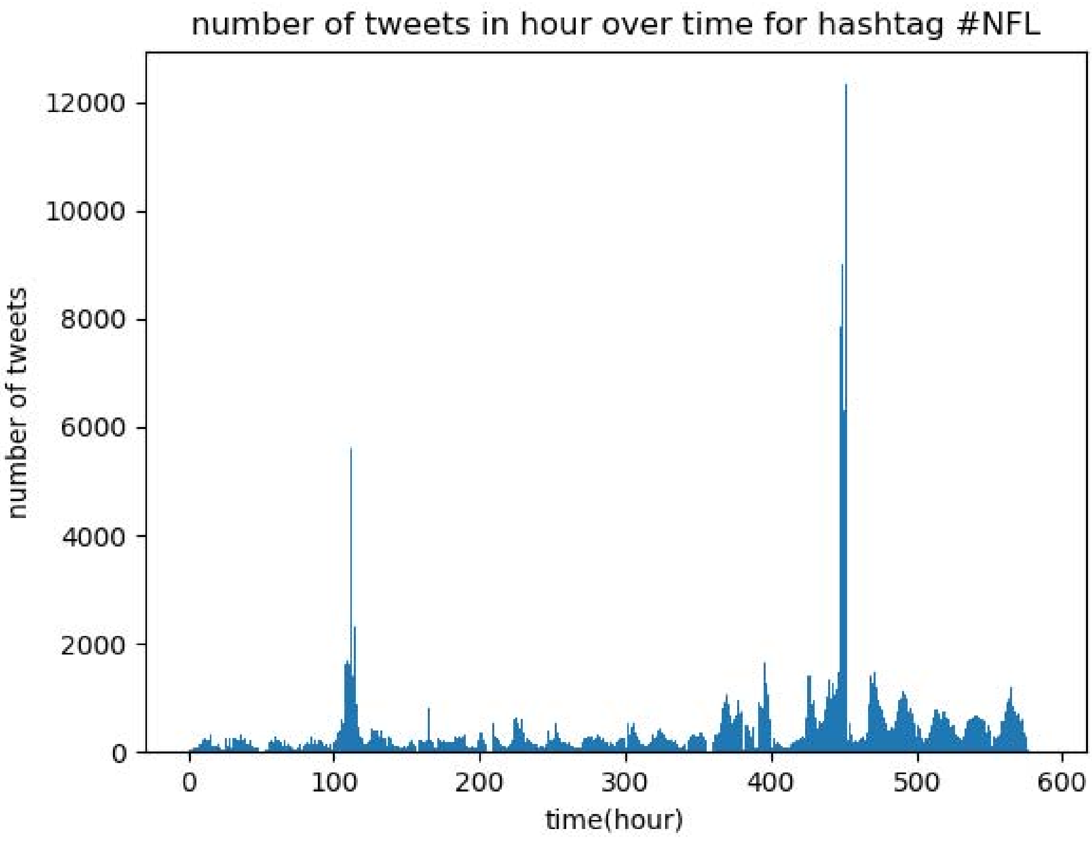
For hashtag #gopatriots:
Average number of tweets per hour is 45.69451057356203
Average number of followers of users posting the tweets per hour is 1401.8955093016164
Average number of retweets per hour is 1.4000838670326319

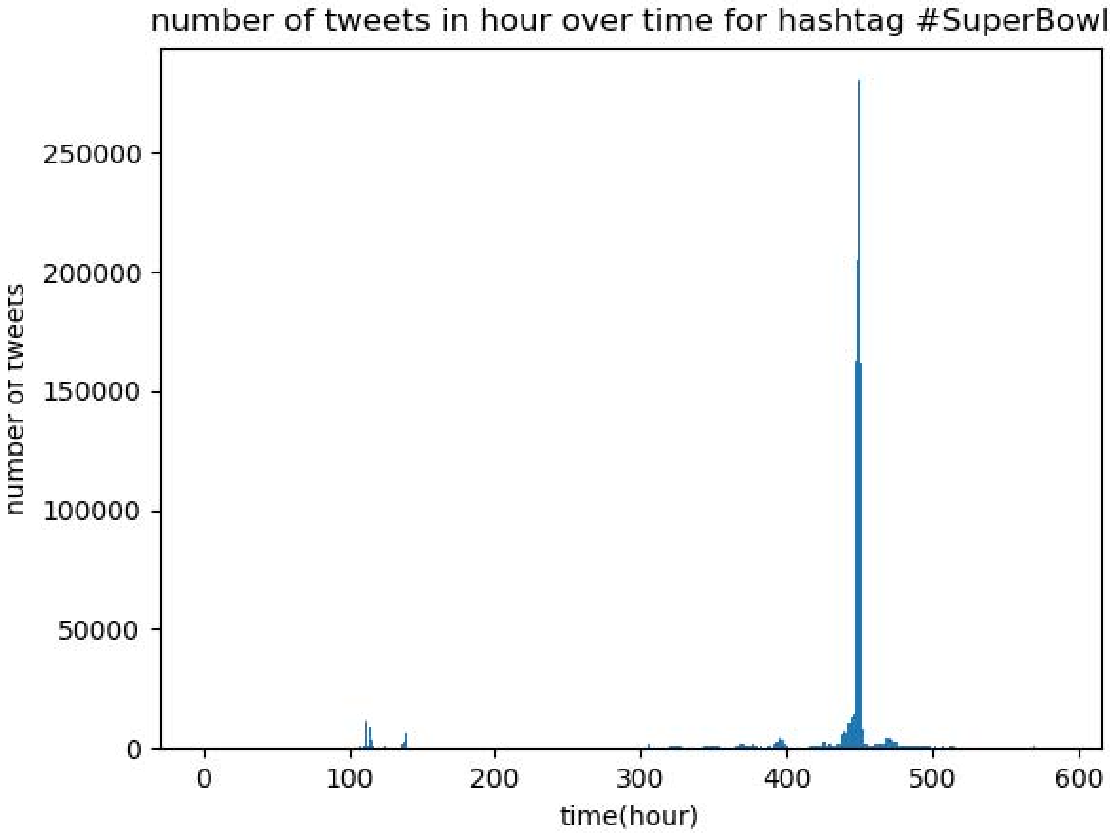
For hashtag #nfl:
Average number of tweets per hour is 441.3234311373958
Average number of followers of users posting the tweets per hour is 4653.252285502502
Average number of retweets per hour is 1.5385331089011056

For hashtag #patriots:
Average number of tweets per hour is 834.5555091641886
Average number of followers of users posting the tweets per hour is 3309.978828415827
Average number of retweets per hour is 1.7828156491659402

For hashtag #sb49:
Average number of tweets per hour is 1419.8879074871902
Average number of followers of users posting the tweets per hour is 10267.31684948685
Average number of retweets per hour is 2.5111487863247035

For hashtag #superbowl:
Average number of tweets per hour is 2302.5004018833274
Average number of followers of users posting the tweets per hour is 8858.974662784603
Average number of retweets per hour is 2.3882723999030224





We can find that there are 2 spikes in #NFL and 1 spike in #SuperBowl. And they both have a spike when hour is about 450, which indicates there may be a major event at that time.

Problem 1.2

Now we are going to use 5 features to fit a linear regression model to predict number of tweets next hour. To extract features, we created time windows of an hour and divide the whole data into roughly 500 hours. Then using the timestamp of each tweet we can decide which hour it should belong to and calculate features of this hour.

R-squared is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

And p-value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary (such as the sample mean difference between two compared groups) would be the same as or of greater magnitude than the actual observed results.

When calculating error, we use mean-absolute-error.

Here are the results.

For hashtag #gohawks:
R-squared = 0.520

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	1.3775	0.165	8.342	0.000	1.053	1.702
Total number of retweets	-0.1466	0.039	-3.779	0.000	-0.223	-0.070
Sum of followers	-0.0002	8.36e-05	-2.909	0.004	-0.000	-7.9e-05
Maximum number of followers	0.0002	0.000	1.387	0.166	-9.84e-05	0.001
Time of the day	7.8263	3.291	2.378	0.018	1.362	14.290

The training error(MAE) is 407.2872228213191

For hashtag #gopatriots:
R-squared = 0.611

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	-0.4214	0.264	-1.597	0.111	-0.940	0.097
Total number of retweets	0.4604	0.230	2.001	0.046	0.009	0.912
Sum of followers	0.0006	0.000	3.163	0.002	0.000	0.001
Maximum number of followers	-0.0007	0.000	-3.737	0.000	-0.001	-0.000
Time of the day	0.7731	0.633	1.222	0.222	-0.470	2.016

The training error(MAE) is 75.93780508984794

For hashtag #nfl:
R-squared = 0.648

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	0.7504	0.135	5.555	0.000	0.485	1.016
Total number of retweets	-0.1751	0.066	-2.662	0.008	-0.304	-0.046
Sum of followers	7.398e-05	2.62e-05	2.827	0.005	2.26e-05	0.000
Maximum number of followers	-7.32e-05	3.6e-05	-2.034	0.042	-0.000	-2.53e-06
Time of the day	8.2286	2.231	3.687	0.000	3.846	12.611

The training error(MAE) is 365.5353919928616

For hashtag #patriots:
R-squared = 0.716

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	1.2170	0.079	15.399	0.000	1.062	1.372
Total number of retweets	-0.3396	0.069	-4.957	0.000	-0.474	-0.205
Sum of followers	3.504e-05	2.63e-05	1.335	0.182	-1.65e-05	8.66e-05
Maximum number of followers	0.0002	9.54e-05	1.599	0.110	-3.48e-05	0.000
Time of the day	8.6652	8.304	1.044	0.297	-7.644	24.974

The training error(MAE) is 1211.6200408456325

For hashtag #sb49:
R-squared = 0.844

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	1.2898	0.095	13.535	0.000	1.103	1.477
Total number of retweets	-0.2961	0.087	-3.387	0.001	-0.468	-0.124
Sum of followers	2.883e-05	1.38e-05	2.083	0.038	1.65e-06	5.6e-05
Maximum number of followers	0.0002	4.25e-05	4.216	0.000	9.57e-05	0.000
Time of the day	-15.8124	13.762	-1.149	0.251	-42.843	11.218

The training error(MAE) is 2486.962548546062

For hashtag #superbowl:
R-squared = 0.869

	coef	std err	t	P>|t|	[0.025	0.975]
Number of tweets	2.5447	0.107	23.708	0.000	2.334	2.756
Total number of retweets	-0.1548	0.035	-4.380	0.000	-0.224	-0.085
Sum of followers	-0.0002	1.08e-05	-20.176	0.000	-0.000	-0.000
Maximum number of followers	0.0011	0.000	10.278	0.000	0.001	0.001
Time of the day	-50.4217	24.332	-2.072	0.039	-98.211	-2.632

The training error(MAE) is 3623.623321665225

We can see with larger training data, the R-squared value also increases, which means that we have better accuracy. And for most hashtags, number of tweets in previous hour is important for next hour's prediction. The only exception is hashtag #gopatriots, it has maximum number of follower as the most important feature. This could be that this hashtag has much less tweets and someone with a great number of followers could have great influence in tweet number. And the feature time of the day seems to have large p values, which means this feature is not very important and we can exclude it from our model.

Problem 1.3

In this part, we designed a new model, still using Linear Regression, with new features we found from the paper *On the Real-time Prediction, Problems of Bursting Hashtags in Twitter*. After selecting features that are most relevant to the tweets number and to get a relatively low prediction error, we finally decided to use the features as below:

1. tweets_num: the total number of tweets in an hour;
2. retweers_num: the total number of retweets in an hour;
3. sum_followers: the sum of the number of followers of all users;
4. max_followers: the maximum number of followers of users;
5. time_of_day: one of the 24 hours in a day
6. URLs_num: the total number of URLs cited by the tweets in an hour;
7. authors_num: the total number of authors involved in an hour. In other words, active users in an hour
8. mentions_num: the total number of mentions in tweets in an hour;
9. ranking_score: the sum of ranking scores of tweets in an hour;
10. hashtags_num: the total number of tweets in the tweets.

With the new features, we extracted information from the hashtag files and created 1-hour time windows. Also, we did one-hot encoding on the feature "time_of_day" in order to get better results. Then we perform Linear Regression model on the dataset, predicted the number of tweets in the next 1-hour window with features extracted from tweet data in the previous 1-hour window, and calculated the RMSE of predicted value and true value. We also applied t-test to analyse the significance of each feature, using the library statsmodels.api in Python.

The RMSE and R-squared measures for each hashtag file are as below:

Tweet data	RMSE	R-squared
tweets_#gohawks	700.1174662302651	0.724
tweets_#gopatriots	100.15289056780102	0.894
tweets_#nfl	426.973937003711	0.765
tweets_#patriots	1840.1375511287124	0.823

tweets_#sb49	3143.147353956885	0.902
tweets_#superbowl	4260.876213460514	0.943

P-values of different features for each hashtag file are as below:

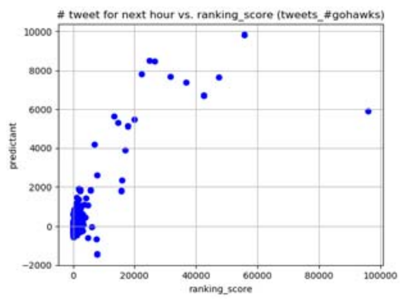
tweet data P-values features	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
tweets_num	9.329546e-48	3.377311e-03	4.495591e-01	2.423382e-32	7.879917e-16	2.017493e-81
retweets_num	1.577650e-04	2.681923e-50	8.783207e-05	7.979424e-19	2.450427e-02	5.337485e-38
sum_followers	1.363767e-05	9.143321e-19	3.272762e-01	1.112168e-11	1.952924e-21	2.258016e-24
max_followers	2.177873e-01	2.577612e-17	6.855078e-01	3.007636e-08	9.790209e-10	9.391348e-01
URLs_num	7.158163e-07	5.793266e-47	8.331178e-03	3.937402e-03	2.899248e-03	5.977705e-05
authors_num	2.269332e-09	1.587829e-26	9.487487e-34	2.655707e-02	1.475423e-02	9.846646e-68
mentions_num	1.687794e-09	4.884546e-14	5.904038e-06	7.729120e-13	4.434342e-03	1.528295e-50
ranking_score	3.314710e-48	1.729867e-06	3.861195e-01	2.331112e-28	1.463012e-14	2.509496e-80
hashtags_num	2.757178e-01	9.906028e-09	8.148776e-38	1.153019e-17	1.055765e-05	2.812801e-12

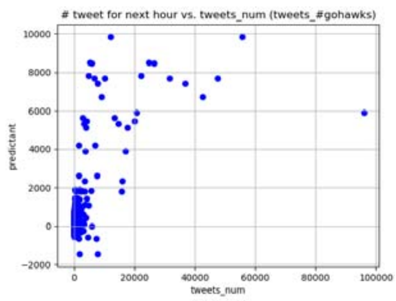
Here we listed the OLS Regression reports, top3 features of each hashtag file. And for each of the top 3 features, we plotted a scatter of predictant (number of tweets for next hour) versus value of that feature.

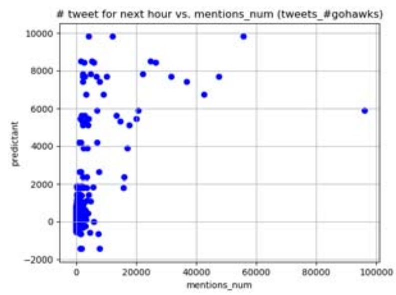
1. tweet_#gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.707
Method:                 Least Squares   F-statistic:                     44.66
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          1.19e-130
Time:                        00:06:28   Log-Likelihood:                -4614.7
No. Observations:                 579   AIC:                             9295.
Df Residuals:                     546   BIC:                             9439.
Df Model:                          32
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
tweets_num      -67.2442      4.190    -16.047      0.000     -75.476     -59.013
retweets_num     13.4966      3.547      3.805      0.000       6.529      20.464
sum_followers    -0.0003   6.86e-05     -4.390      0.000      -0.000      -0.000
max_followers     0.0002      0.000      1.234      0.218   -9.95e-05       0.000
URLs_num          7.5628      1.508      5.016      0.000       4.601      10.525
authors_num       4.5277      0.745      6.079      0.000       3.065       5.991
mentions_num      2.8669      0.468      6.130      0.000       1.948       3.786
ranking_score    13.5304      0.838     16.141      0.000      11.884      15.177
hashtags_num      0.3588      0.329      1.091      0.276      -0.287       1.005
0th_hour         54.3914    144.659      0.376      0.707    -229.764     338.547
1th_hour        -18.1857    144.375     -0.126      0.900    -301.784     265.413
2th_hour         16.6854    144.484      0.115      0.908    -267.128     300.499
3th_hour         14.8236    147.270      0.101      0.920    -274.461     304.108
4th_hour        -21.3342    147.266     -0.145      0.885    -310.611     267.943
5th_hour         -4.0843    147.593     -0.028      0.978    -294.005     285.836
6th_hour        -90.3962    148.223     -0.610      0.542    -381.552     200.760
7th_hour       -131.8647    149.518     -0.882      0.378    -425.565     161.835
8th_hour       -180.9324    152.608     -1.186      0.236    -480.702     118.837
9th_hour       -242.7575    152.420     -1.593      0.112    -542.158      56.643
10th_hour      -276.5702    154.535     -1.790      0.074    -580.126      26.986
11th_hour      -475.3761    155.993     -3.047      0.002    -781.795    -168.957
12th_hour      -452.4758    153.927     -2.940      0.003    -754.837    -150.115
13th_hour      -175.3495    152.959     -1.146      0.252    -475.809     125.110
14th_hour       575.9658    153.188      3.760      0.000     275.056     876.876
15th_hour      -289.4265    154.147     -1.878      0.061    -592.219      13.366
16th_hour       -67.9202    151.560     -0.448      0.654    -365.632     229.791
17th_hour       163.8076    152.011      1.078      0.282    -134.791     462.406
18th_hour      -156.5957    149.677     -1.046      0.296    -450.608     137.417
19th_hour        12.6608    155.421      0.081      0.935    -292.636     317.957
20th_hour        -2.0205    151.888     -0.013      0.989    -300.378     296.337
21th_hour       -52.0524    150.618     -0.346      0.730    -347.913     243.808
22th_hour         7.2817    148.972      0.049      0.961    -285.346     299.910
23th_hour        -2.1436    147.790     -0.015      0.988    -292.451     288.164
==============================================================================
Omnibus:                      933.001   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           675766.335
Skew:                           9.082   Prob(JB):                         0.00
Kurtosis:                     169.376   Cond. No.                     1.92e+07
==============================================================================
```

top 3 features are: ranking_score, tweets_num, mentions_num





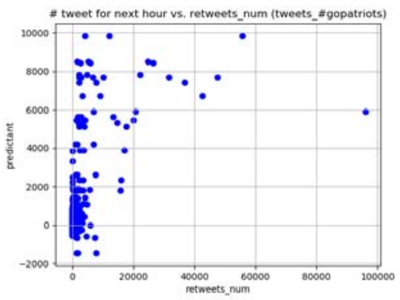


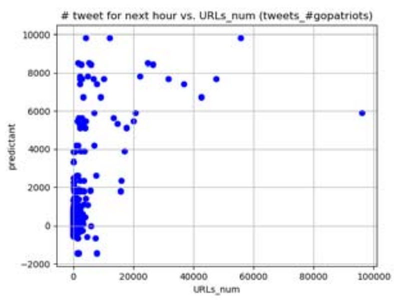
From the plots we can see that there's a linear distribution, which indicates a good relationships between the features.

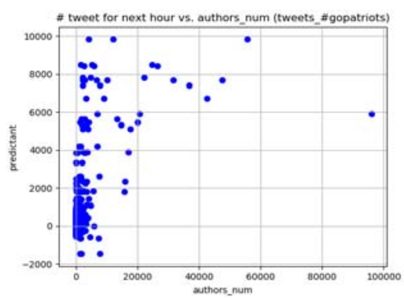
2. tweet_#gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.894
Model:                            OLS   Adj. R-squared:                  0.888
Method:                 Least Squares   F-statistic:                     143.3
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          2.45e-241
Time:                        00:06:32   Log-Likelihood:                -3464.7
No. Observations:                 575   AIC:                             6995.
Df Residuals:                     542   BIC:                             7139.
Df Model:                          32
Covariance Type:            nonrobust
==================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
tweets_num       -8.0301      2.727     -2.944      0.003     -13.388      -2.672
retweets_num    -43.2781      2.608    -16.592      0.000     -48.402     -38.154
sum_followers    -0.0020      0.000     -9.179      0.000      -0.002      -0.002
max_followers     0.0019      0.000      8.756      0.000       0.001       0.002
URLs_num         13.5502      0.853     15.892      0.000      11.875      15.225
authors_num      -6.6753      0.593    -11.248      0.000      -7.841      -5.510
mentions_num      3.1152      0.402      7.740      0.000       2.325       3.906
ranking_score     2.2835      0.472      4.836      0.000       1.356       3.211
hashtags_num      1.8346      0.315      5.823      0.000       1.216       2.453
0th_hour        -13.4494     21.067     -0.638      0.523     -54.832      27.934
1th_hour         -0.9546     21.060     -0.045      0.964     -42.325      40.415
2th_hour        -11.1825     21.073     -0.531      0.596     -52.578      30.213
3th_hour          1.2885     21.079      0.061      0.951     -40.119      42.696
4th_hour        -11.0857     21.099     -0.525      0.600     -52.532      30.360
5th_hour        -12.2316     21.084     -0.580      0.562     -53.649      29.186
6th_hour        -30.7719     21.122     -1.457      0.146     -72.263      10.720
7th_hour        -17.9052     21.133     -0.847      0.397     -59.417      23.607
8th_hour        -23.6883     21.136     -1.121      0.263     -65.206      17.829
9th_hour        -28.8476     21.277     -1.356      0.176     -70.642      12.947
10th_hour       -50.5863     21.388     -2.365      0.018     -92.600      -8.573
11th_hour       -15.3564     21.620     -0.710      0.478     -57.826      27.114
12th_hour        66.9282     21.341      3.136      0.002      25.006     108.850
13th_hour        11.5140     21.887      0.526      0.599     -31.480      54.508
14th_hour       -39.9123     21.823     -1.829      0.068     -82.781       2.956
15th_hour        37.0422     21.613      1.714      0.087      -5.414      79.498
16th_hour        -6.6506     21.945     -0.303      0.762     -49.759      36.458
17th_hour        -5.8325     21.765     -0.268      0.789     -48.586      36.921
18th_hour        -2.3024     21.274     -0.108      0.914     -44.093      39.488
19th_hour        -4.3360     21.109     -0.205      0.837     -45.801      37.129
20th_hour         6.1273     21.077      0.291      0.771     -35.276      47.531
21th_hour        15.0085     21.090      0.712      0.477     -26.419      56.436
22th_hour        -5.5452     21.062     -0.263      0.792     -46.918      35.828
23th_hour        -5.2516     21.517     -0.244      0.807     -47.519      37.016
==============================================================================
Omnibus:                      606.578   Durbin-Watson:                   2.125
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           119995.116
Skew:                           4.265   Prob(JB):                         0.00
Kurtosis:                      73.255   Cond. No.                     2.26e+06
==============================================================================
```

top 3 features are: retweets_num, URLs_num, authors_num





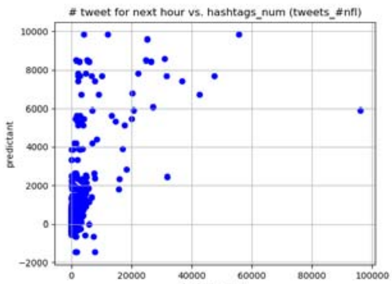


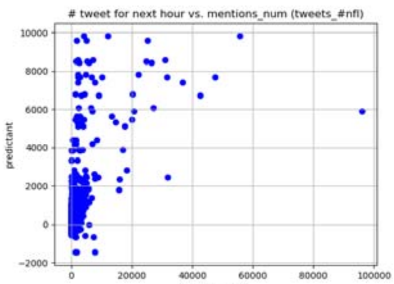
From the plots we can see that there's a linear distribution, which indicates a good relationships between the features.

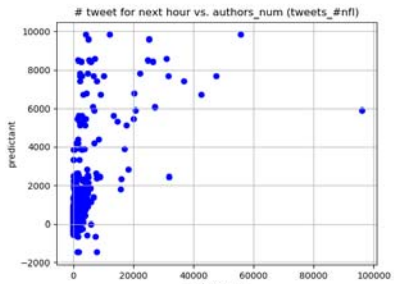
3. tweet_#nfl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.765
Model:                            OLS   Adj. R-squared:                  0.751
Method:                 Least Squares   F-statistic:                     56.25
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          8.37e-152
Time:                        00:06:53   Log-Likelihood:                -4388.2
No. Observations:                 587   AIC:                              8842.
Df Residuals:                     554   BIC:                              8987.
Df Model:                          32
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
tweets_num       -1.1317      1.496     -0.757      0.450      -4.070       1.806
retweets_num     -9.9461      2.517     -3.951      0.000     -14.891      -5.002
sum_followers -2.376e-05   2.42e-05     -0.980      0.327   -7.14e-05    2.38e-05
max_followers  1.302e-05   3.21e-05      0.405      0.686   -5.01e-05    7.62e-05
URLs_num         -0.4606      0.174     -2.648      0.008      -0.802      -0.119
authors_num      -4.4403      0.343    -12.962      0.000      -5.113      -3.767
mentions_num      2.9480      0.644      4.574      0.000       1.682       4.214
ranking_score     0.2665      0.307      0.867      0.386      -0.337       0.870
hashtags_num      1.1542      0.083     13.882      0.000       0.991       1.318
0th_hour        -85.0332     91.043     -0.934      0.351    -263.866      93.799
1th_hour        -76.5635     90.406     -0.847      0.397    -254.144     101.017
2th_hour        -68.3849     89.549     -0.764      0.445    -244.283     107.513
3th_hour         20.6908     89.562      0.231      0.817    -155.233     196.614
4th_hour         67.9376     90.042      0.755      0.451    -108.929     244.804
5th_hour        172.6458     89.976      1.919      0.056      -4.089     349.381
6th_hour        236.0467     91.980      2.566      0.011      55.375     416.719
7th_hour        180.6493     93.876      1.924      0.055      -3.748     365.047
8th_hour        163.9958     96.150      1.706      0.089     -24.867     352.859
9th_hour        120.3897     95.002      1.267      0.206     -66.219     306.998
10th_hour       187.8101     94.409      1.989      0.047       2.366     373.254
11th_hour       144.3960     99.966      1.444      0.149     -51.963     340.755
12th_hour        95.9948     99.206      0.968      0.334     -98.871     290.860
13th_hour        92.6347     97.977      0.945      0.345     -99.817     285.086
14th_hour       530.2282     98.621      5.376      0.000     336.512     723.945
15th_hour        50.4429     99.130      0.509      0.611    -144.275     245.160
16th_hour        18.9676     96.550      0.196      0.844    -170.681     208.616
17th_hour       189.4597     99.700      1.900      0.058      -6.376     385.295
18th_hour       119.1781     97.274      1.225      0.221     -71.892     310.248
19th_hour       -59.1401     96.239     -0.615      0.539    -248.179     129.898
20th_hour       -50.2045     95.079     -0.528      0.598    -236.963     136.554
21th_hour       -64.2906     95.030     -0.677      0.499    -250.955     122.373
22th_hour      -167.6198     93.279     -1.797      0.073    -350.844      15.605
23th_hour      -131.1583     93.127     -1.408      0.160    -314.083      51.766
==============================================================================
Omnibus:                      691.527   Durbin-Watson:                   2.155
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            89935.999
Skew:                           5.454   Prob(JB):                         0.00
Kurtosis:                      62.650   Cond. No.                     4.42e+07
==============================================================================
```

top 3 features are: hashtags_num, authors_num, mentions_num





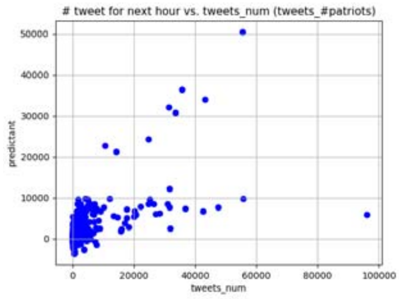


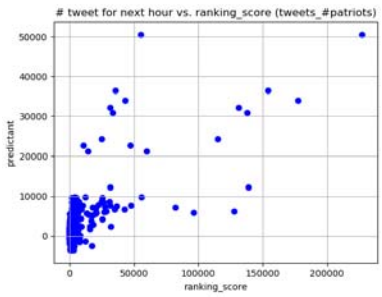
From the plots we can see that there’s a linear distribution, which indicates a good relationships between the features.

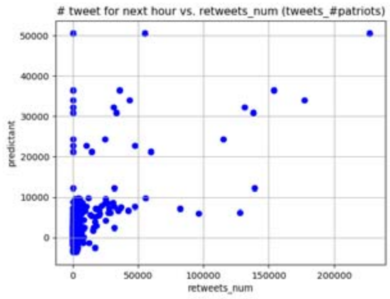
4. tweet_#patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.813
Method:                 Least Squares   F-statistic:                     80.43
Date:                Mon, 12 Mar 2018   Prob (F-statistic):          1.65e-185
Time:                        00:07:30   Log-Likelihood:                -5245.7
No. Observations:                 587   AIC:                         1.056e+04
Df Residuals:                     554   BIC:                         1.070e+04
Df Model:                          32
Covariance Type:            nonrobust
==================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
tweets_num      -63.0835      4.992    -12.636      0.000     -72.889     -53.278
retweets_num    -27.0591      2.945     -9.188      0.000     -32.844     -21.274
sum_followers     0.0004    5.7e-05      6.938      0.000       0.000       0.001
max_followers    -0.0006      0.000     -5.621      0.000      -0.001      -0.000
URLs_num         -5.0346      1.739     -2.895      0.004      -8.450      -1.619
authors_num       2.2108      0.994      2.224      0.027       0.258       4.163
mentions_num      7.0053      0.955      7.339      0.000       5.130       8.880
ranking_score    11.0880      0.949     11.687      0.000       9.224      12.952
hashtags_num      3.6577      0.413      8.852      0.000       2.846       4.469
0th_hour        197.8735    381.113      0.519      0.604    -550.729     946.476
1th_hour        -65.4957    382.060     -0.171      0.864    -815.958     684.967
2th_hour        -36.3362    380.504     -0.095      0.924    -783.744     711.072
3th_hour       -156.8004    380.059     -0.413      0.680    -903.333     589.732
4th_hour       -114.6024    381.599     -0.300      0.764    -864.160     634.955
5th_hour       -164.1821    383.748     -0.428      0.669    -917.962     589.597
6th_hour       -331.3956    388.221     -0.854      0.394   -1093.961     431.169
7th_hour       -669.8187    392.247     -1.708      0.088   -1440.291     100.654
8th_hour       -756.3629    394.680     -1.916      0.056   -1531.615      18.889
9th_hour       -313.3698    405.646     -0.773      0.440   -1110.163     483.423
10th_hour       972.2628    391.872      2.481      0.013     202.526    1741.999
11th_hour      -843.7053    401.738     -2.100      0.036   -1632.821     -54.590
12th_hour      -711.3069    403.419     -1.763      0.078   -1503.725      81.111
13th_hour      -581.7998    403.619     -1.441      0.150   -1374.610     211.010
14th_hour      -313.4766    407.543     -0.769      0.442   -1113.994     487.041
15th_hour      -562.1471    405.120     -1.388      0.166   -1357.906     233.612
16th_hour       -76.3642    404.922     -0.189      0.850    -871.734     719.006
17th_hour       140.2104    401.606      0.349      0.727    -648.646     929.067
18th_hour       345.9975    409.073      0.846      0.398    -457.527    1149.522
19th_hour      -362.4667    400.155     -0.906      0.365   -1148.473     423.540
20th_hour       330.5370    402.736      0.821      0.412    -460.540    1121.614
21th_hour      -309.1792    397.389     -0.778      0.437   -1089.753     471.395
22th_hour       151.0046    393.973      0.383      0.702    -622.860     924.869
23th_hour       225.2162    392.846      0.573      0.567    -546.433     996.866
==============================================================================
Omnibus:                     1036.548   Durbin-Watson:                   1.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1014783.857
Skew:                          10.956   Prob(JB):                         0.00
Kurtosis:                     205.510   Cond. No.                     6.69e+07
==============================================================================
```

top 3 features are: tweets_num, ranking_score, retweets_num





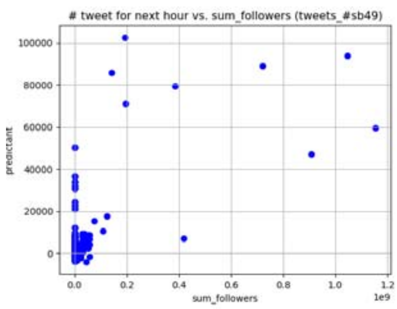


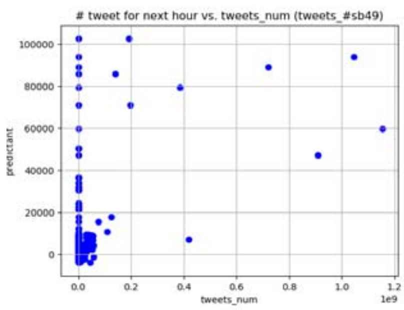
From the plots we can see that there’s a linear distribution, which indicates a good relationships between the features.

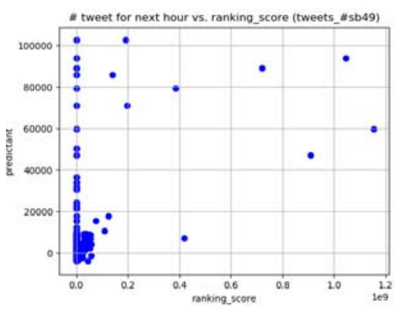
5. tweet_#sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     157.9
Date:                Mon, 12 Mar 2018   Prob (F-statistic):         5.72e-254
Time:                        00:08:30   Log-Likelihood:                -5522.1
No. Observations:                 583   AIC:                         1.111e+04
Df Residuals:                     550   BIC:                         1.125e+04
Df Model:                          32
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
tweets_num      -62.3620      7.510     -8.304      0.000     -77.114     -47.610
retweets_num     38.2823     16.974      2.255      0.025       4.940      71.624
sum_followers     0.0002   2.08e-05      9.917      0.000       0.000       0.000
max_followers    -0.0004   6.51e-05     -6.221      0.000      -0.001      -0.000
URLs_num          5.7079      1.908      2.992      0.003       1.960       9.456
authors_num      -2.3755      0.971     -2.446      0.015      -4.283      -0.468
mentions_num      3.0465      1.066      2.857      0.004       0.952       5.141
ranking_score    11.9501      1.512      7.906      0.000       8.981      14.919
hashtags_num      2.4295      0.546      4.447      0.000       1.356       3.503
0th_hour       -165.0774    648.509     -0.255      0.799   -1438.935    1108.780
1th_hour         48.1939    650.880      0.074      0.941   -1230.320    1326.708
2th_hour       -685.9166    652.544     -1.051      0.294   -1967.700     595.867
3th_hour       -831.8146    655.672     -1.269      0.205   -2119.742     456.113
4th_hour       -880.0948    661.754     -1.330      0.184   -2179.969     419.779
5th_hour       1536.0553    658.248      2.334      0.020     243.068    2829.042
6th_hour        234.3266    669.661      0.350      0.727   -1081.079    1549.732
7th_hour      -1771.5668    674.098     -2.628      0.009   -3095.688    -447.446
8th_hour      -1310.1479    668.042     -1.961      0.050   -2622.373       2.077
9th_hour       -885.5791    683.618     -1.295      0.196   -2228.400     457.242
10th_hour      -713.8312    678.722     -1.052      0.293   -2047.035     619.373
11th_hour      -334.9490    675.902     -0.496      0.620   -1662.615     992.717
12th_hour      -418.1016    682.315     -0.613      0.540   -1758.363     922.160
13th_hour      -797.8293    670.214     -1.190      0.234   -2114.322     518.663
14th_hour      -289.9327    676.372     -0.429      0.668   -1618.521    1038.655
15th_hour      -252.1275    667.650     -0.378      0.706   -1563.584    1059.329
16th_hour       296.4913    666.397      0.445      0.657   -1012.503    1605.486
17th_hour       396.5588    670.308      0.592      0.554    -920.118    1713.236
18th_hour       139.2750    662.873      0.210      0.834   -1162.797    1441.347
19th_hour       276.7323    663.295      0.417      0.677   -1026.170    1579.635
20th_hour       221.8280    661.866      0.335      0.738   -1078.266    1521.922
21th_hour        78.6283    660.838      0.119      0.905   -1219.446    1376.703
22th_hour        33.6887    661.351      0.051      0.959   -1265.395    1332.773
23th_hour       -60.6663    661.336     -0.092      0.927   -1359.720    1238.387
==============================================================================
Omnibus:                      991.551   Durbin-Watson:                   1.577
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           850934.405
Skew:                          10.136   Prob(JB):                         0.00
Kurtosis:                     189.062   Cond. No.                     4.91e+08
==============================================================================
```

top 3 features are: sum_followers, tweets_num, ranking_score





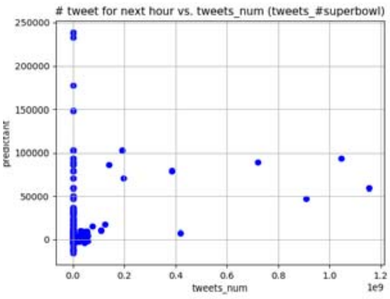


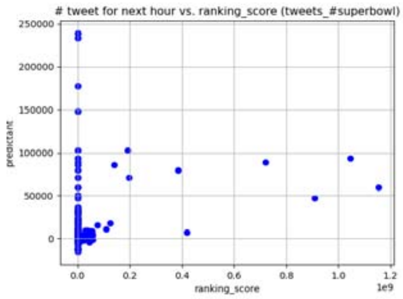
From the plots we can see that the data points cluster in the region near the origin.

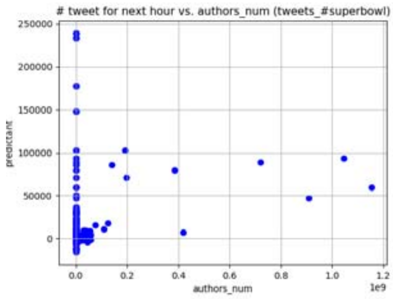
6. tweet_#superbowl

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           target_value   R-squared:                       0.943
Model:                            OLS   Adj. R-squared:                  0.940
Method:                 Least Squares   F-statistic:                     286.4
Date:                Mon, 12 Mar 2018   Prob (F-statistic):         1.36e-320
Time:                        00:10:03   Log-Likelihood:                -5728.8
No. Observations:                 586   AIC:                         1.152e+04
Df Residuals:                     553   BIC:                         1.167e+04
Df Model:                          32
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
tweets_num      -111.7023      4.908    -22.761      0.000    -121.342    -102.063
retweets_num      48.3817      3.474     13.925      0.000      41.557      55.207
sum_followers     -0.0001    1.1e-05    -10.690      0.000      -0.000   -9.61e-05
max_followers  -7.19e-06   9.41e-05     -0.076      0.939      -0.000       0.000
URLs_num           2.4611      0.608      4.045      0.000       1.266       3.656
authors_num       15.7667      0.785     20.073      0.000      14.224      17.310
mentions_num     -13.8255      0.833    -16.606      0.000     -15.461     -12.190
ranking_score     22.3118      0.990     22.547      0.000      20.368      24.256
hashtags_num       1.7899      0.250      7.147      0.000       1.298       2.282
0th_hour         650.9831    887.874      0.733      0.464   -1093.035    2395.001
1th_hour        1004.6848    881.765      1.139      0.255    -727.333    2736.703
2th_hour         172.0183    879.944      0.195      0.845   -1556.422    1900.459
3th_hour         112.2364    884.318      0.127      0.899   -1624.796    1849.269
4th_hour         663.0132    882.862      0.751      0.453   -1071.159    2397.186
5th_hour         423.9448    890.243      0.476      0.634   -1324.727    2172.616
6th_hour        -346.4221    887.821     -0.390      0.697   -2090.336    1397.492
7th_hour       -1179.6425    895.216     -1.318      0.188   -2938.082     578.797
8th_hour        -833.1994    897.610     -0.928      0.354   -2596.341     929.942
9th_hour        -708.9208    897.564     -0.790      0.430   -2471.973    1054.131
10th_hour       -865.3510    921.721     -0.939      0.348   -2675.854     945.151
11th_hour       -158.8165    922.615     -0.172      0.863   -1971.075    1653.442
12th_hour        -27.3412    917.905     -0.030      0.976   -1830.348    1775.666
13th_hour        803.5519    918.861      0.875      0.382   -1001.332    2608.436
14th_hour       3179.5406    929.410      3.421      0.001    1353.934    5005.147
15th_hour       -239.1008    938.458     -0.255      0.799   -2082.480    1604.279
16th_hour       1537.7769    934.802      1.645      0.101    -298.419    3373.973
17th_hour       1407.0925    932.097      1.510      0.132    -423.792    3237.977
18th_hour       -330.7317    926.100     -0.357      0.721   -2149.836    1488.373
19th_hour         48.6008    933.891      0.052      0.959   -1785.807    1883.009
20th_hour       -350.2424    908.766     -0.385      0.700   -2135.299    1434.814
21th_hour        566.5183    906.024      0.625      0.532   -1213.151    2346.188
22th_hour        574.2600    901.899      0.637      0.525   -1197.307    2345.827
23th_hour        564.9300    907.577      0.622      0.534   -1217.791    2347.651
==============================================================================
Omnibus:                      430.488   Durbin-Watson:                   1.368
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            75438.479
Skew:                           2.253   Prob(JB):                         0.00
Kurtosis:                      58.402   Cond. No.                     7.85e+08
==============================================================================
```

top 3 features are: tweets_num, ranking_score, authors_num







From the plots we can see that the data points cluster in the region near the origin.

	R-squared in 1.2	R-squared in 1.3
#gohawks	0.520	0.724

#gopatriots	0.611	0.894
#nfl	0.648	0.765
#patriots	0.716	0.823
#sb49	0.844	0.902
#superbowl	0.869	0.943

Compared to the R-squared values in 1.2, we can see that, for every hashtag after feature selection the R-squared value has increased a lot by at least 0.1, which means the feature selection works well.

Problem 1.4

(i)

We use the top 3 features for each hashtag we find in 1.3 and use linear regression model, knn model, and random forest model to predict in different time periods. Here are the results. We use mean-absolute-error as our cross validation error.

hashtag #gohawks	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	279.5100866847174	4836.515208262082	76.40498305413541
random forest	288.6823287526427	4159.425	78.97095998214749
knn	299.8025369978858	3556.805555555555	75.89957264957266

hashtag #gopatriots	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	18.73798918274914	1974.372195619929	11.87473711745062
random forest	23.35988803945437	1724.237	12.825644075118216
knn	20.52827108292224	1395.338888888888	11.613176638176636

hashtag #nfl	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	219.6876700733168	5958.541684944354	263.5734438768908

random forest	245.3372653276955	3160.248999999999	278.16917187931114
knn	231.7822821235611	4065.272222222222	264.6154456654457

hashtag #patriots	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	313.983817241189	19863.465011919576	252.7553923362671
random forest	377.9335644820295	17292.869	270.5190989010988
knn	340.2360582569885	16955.94444444444	263.8892551892552

hashtag #sb49	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	137.0538899105231	28496.66780689263	570.42997577721
random forest	149.3651573476322	38466.743	554.9786043956043
knn	138.6837796570354	38514.97777777778	542.8374236874237

hashtag #superbowl	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.
linear model	539.0432143236909	138201.4376614626	710.5318620215337
random forest	601.3395570824524	96899.72400000002	810.6253406593406
knn	539.2623737373737	92451.59999999999	727.5567155067155

We can see that for every hashtag, period 2 has the largest MAE, this is because at that time tweet number reaches a peak because of the event and we only have 12 hours of data at that time, thus it is no surprise that prediction of that period is not very reliable.

(ii)

In this part we created a txt file named ‘aggregated data’ and perform same evaluations on this data. Here are the results.

All hashtag combined	before Feb. 1, 8:00 a.m.	between Feb. 1, 8:00 a.m. and 8:00 p.m.	after Feb. 1, 8:00 p.m.

linear model	1478.153162653538	153670.7443885021	1675.055817364881
random forest	1553.30383615222	139443.413	1799.908989010988
knn	1444.762561663143	105825.7722222222	1751.500976800977

Compared with results of individual hashtags, we can see that MAE increased for every time period. Which means that in different hashtag the behavior may be different and one model will not generalize well.

Problem 1.5

In this part, the task is to predict the number of tweets of last hour in each test file. The test data we used here contains a hashtag's tweets for a 6-hour window (except sample8_period1, which has only a 5-hour window). The hashtags in test data are different from those in training data we previously used.

First, we aggregated all the files in tweet_data into a large train_merge.txt file and used it as train data. We created a 5-hour time window in this part, by grouping features from n~n+4 hours into a larger feature vector of 5 hours. The model we used here is the best model we found out from Part 1.4, which is **K-nn Regressor** with features **tweet_number, ranking_score, user_followers**. We fitted models in 3 periods as described in Part 1.4 separately on the aggregate of the training data for all hashtags, and predict the number of tweets in the 6th hour for each test file using features from the previous 5-hour window, and also in 3 periods separately. Then we calculated the RMSE to evaluate the accuracy of our prediction except test file sample8_period1, which has no true value for the 6th hour.

The predictions, true values and RMSE are listed as below:

test file	prediction	actual	rmse
sample1_period1	887.26	178.0	709.26
sample2_period2	151989.416667	82923.0	69066.41666666666
sample3_period3	995.7	523.0	472.70000000000005
sample4_period1	348.12	201.0	147.12
sample5_period1	743.0	210.0	533.0
sample6_period2	151989.416667	37293.0	114696.41666666666
sample7_period3	652.98	120.0	532.98

sample8_period1	357.66	N/A	N/A
sample9_period2	151989.416667	2790.0	149199.41666666666
sample10_period3	652.98	61.0	591.98

From the above table we can see that for most test file, the RMSE is large but still can be regarded as reasonable. However for some certain test file, the RMSE is extremely large, the reason of which may be overfitting or insufficient data samples.

Part 2: Fan Base Prediction

In this part, we need to realize a classifier to predict the location of the author of a tweet based on the textual information. In order to complete this assignment, we divided it into four parts. We will talk about the main function and obtained results of each part in the sections below.

2.1 Tweets Filtering

Because we just need to classify the tweets from WA State and MA State, we decide to pick these tweets from the entire JSON dataset and label them with 0 (Massachusetts) or 1 (Washington). In this way, we can save the computational time greatly when we use those data later. Meanwhile, we convert the texts from JSON to string and save them in the CSV file "filtered_tweets.csv" which will be used in the following part. With our selection method, we picked out 21,149 tweets in total, 6,817 tweets from Washington State and 14,332 from Massachusetts.

2.2 Tweet Vectors Generating

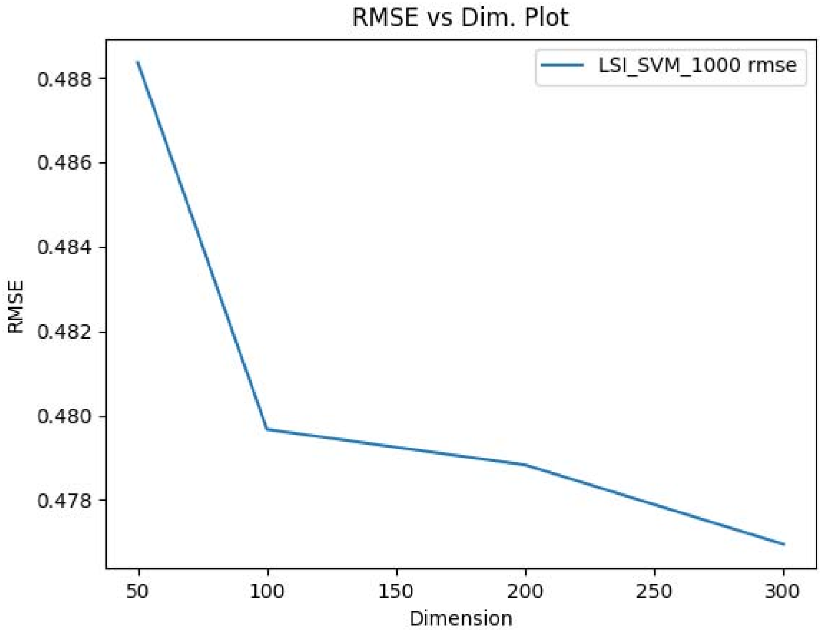
Like what we did in Project 1, we need to convert the textual information into numerical form which could be used to train a learner. First, we extracted the features and labels respectively. Then we represented those selected tweets with TFxIDF matrix. Besides, we need to drop the unnecessary vocabulary such as stop words, special characters and merge the words with the same stem into a single term. In this way, we vectorized each tweet and obtained a sparse matrix. The matrix dimension is (21149, 6811). The dimensionality reduction will be done in the next step. Similarly, we stored the obtained feature sparse matrix and labels in two CSV files and proceeded to next step.

2.3 Select Best Parameters

In the above two parts, we have preprocessed our data and saved them in files. In this part, we need to select the best dimensionality reduction method for different classification models. Considering that the function is quite time-consuming, the dimension set we selected is {50, 100, 200, 300}. We used SVM (hard and soft), Logistic Regression and Naive Bayes classifiers to predict the location in our project and computed their RMSE with cross-validation as the performance metric. The dimensionality reduction methods include LSI and NMF. We will list the results in all the cases in our project as follow. We find that the RMSE in NMF case is generally smaller than that in LSI with the same classifier.

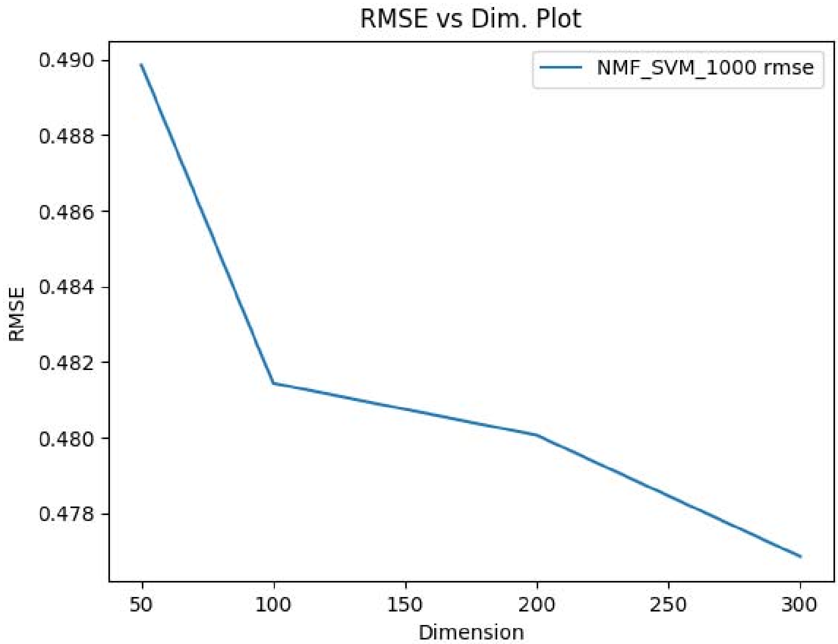
2.3.1 SVM (C = 1000), LSI

The RMSE against dimensionality plot is shown in the following figure. According to the plot, the best dimension in this case is 300, whose corresponding RMSE is 0.4769.



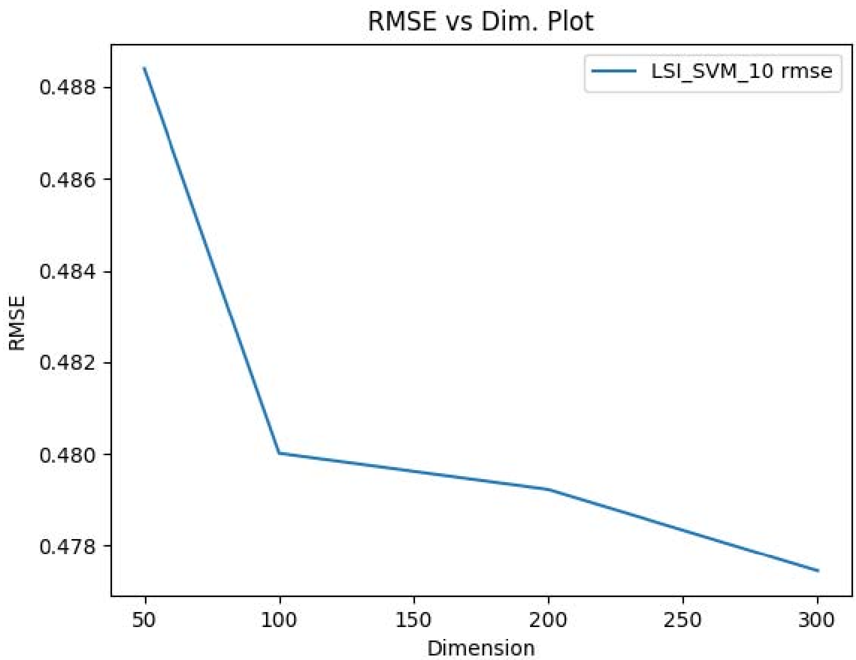
2.3.2 SVM (C = 1000), NMF

According to the plot, the best dimension in this case is 300, whose corresponding RMSE is 0.4769.



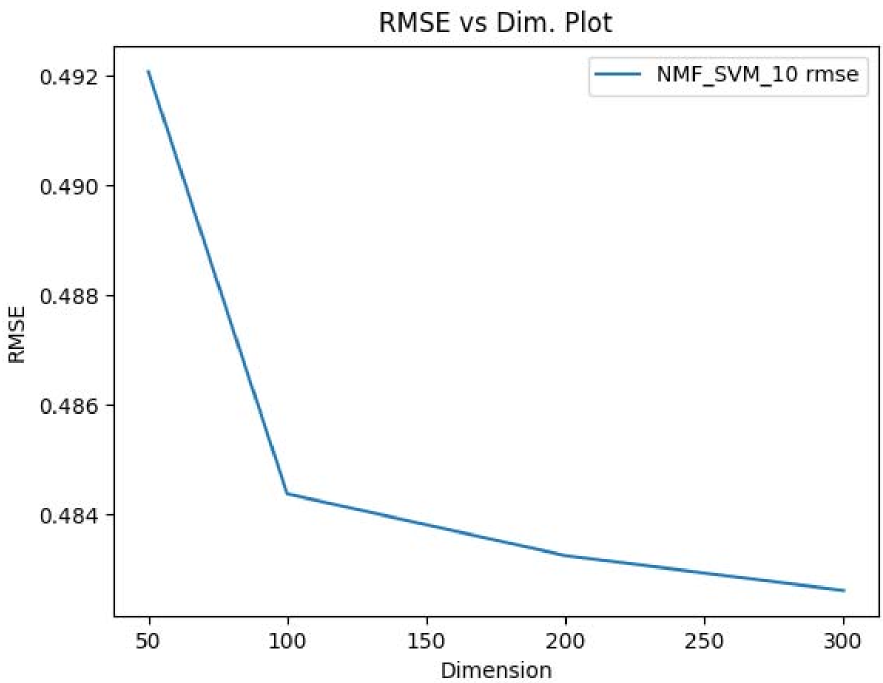
2.3.3 SVM (C = 10), LSI

According to the plot, the best dimension in this case is 300, whose corresponding RMSE is 0.4774.



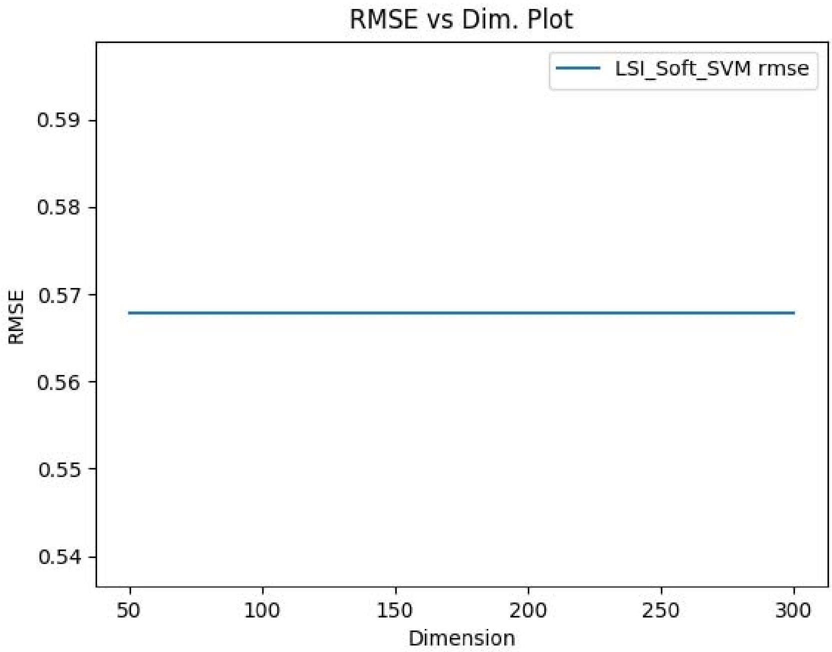
2.3.4 SVM (C = 10), NMF

According to the plot, the best dimension in this case is 300, whose corresponding RMSE is 0.4826.



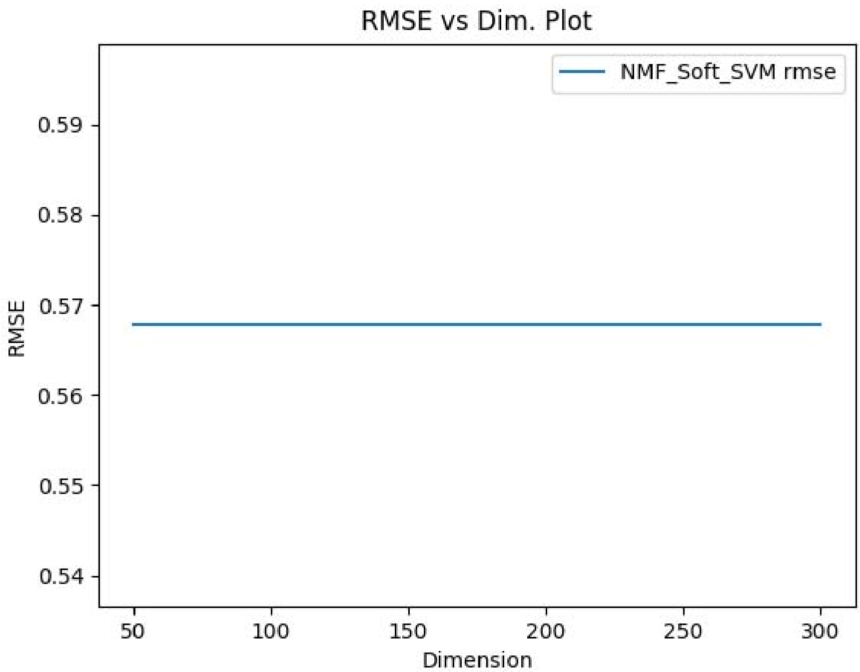
2.3.5 SVM (C = 0.001), LSI

According to the plot, the best dimension in this case is 50, whose corresponding RMSE is 0.5677.



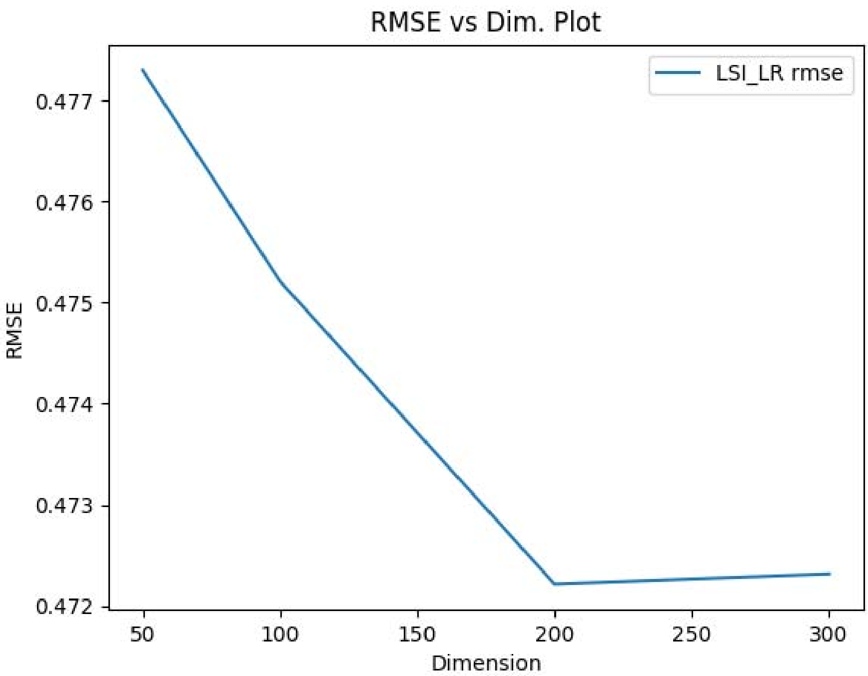
2.3.6 SVM (C = 0.001), NMF

According to the plot, the best dimension in this case is 50, whose corresponding RMSE is 0.5677.



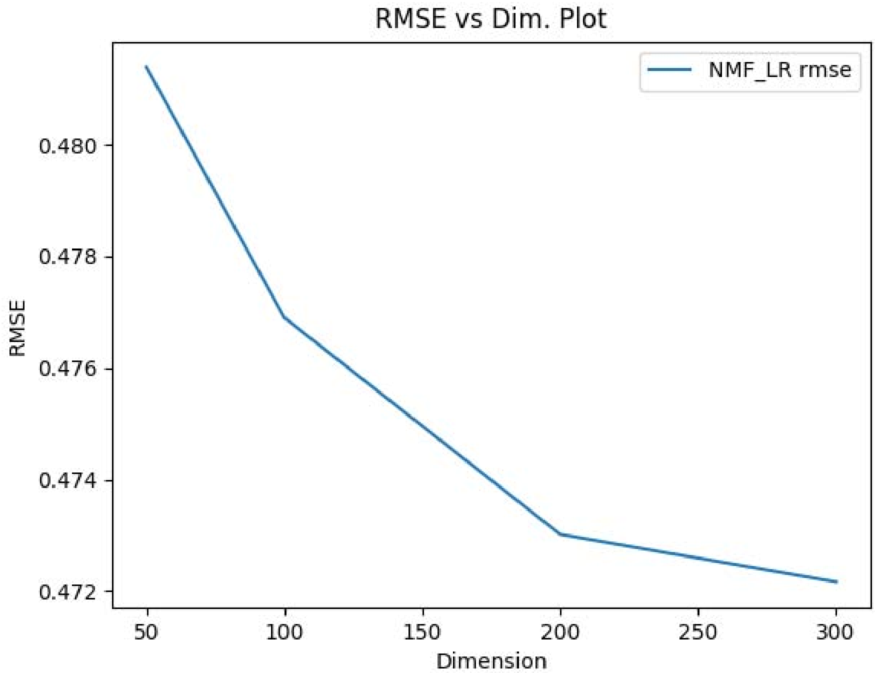
2.3.7 Logistic Regression, LSI

According to the plot, the best dimension in this case is 200, whose corresponding RMSE is 0.4722.



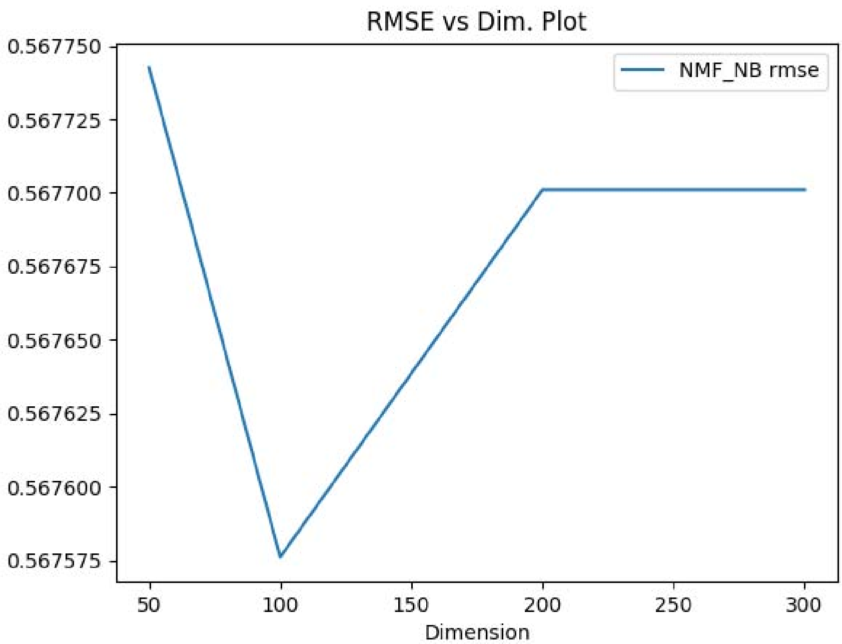
2.3.8 Logistic Regression, NMF

According to the plot, the best dimension in this case is 300, whose corresponding RMSE is 0.4722.



2.3.9 Naive Bayes, NMF

Because Naive Bayes classifier use non-negative values to make predictions, we only used NMF for Naive Bayes learner. According to the plot, the best dimension in this case is 100, whose corresponding RMSE is 0.5676.

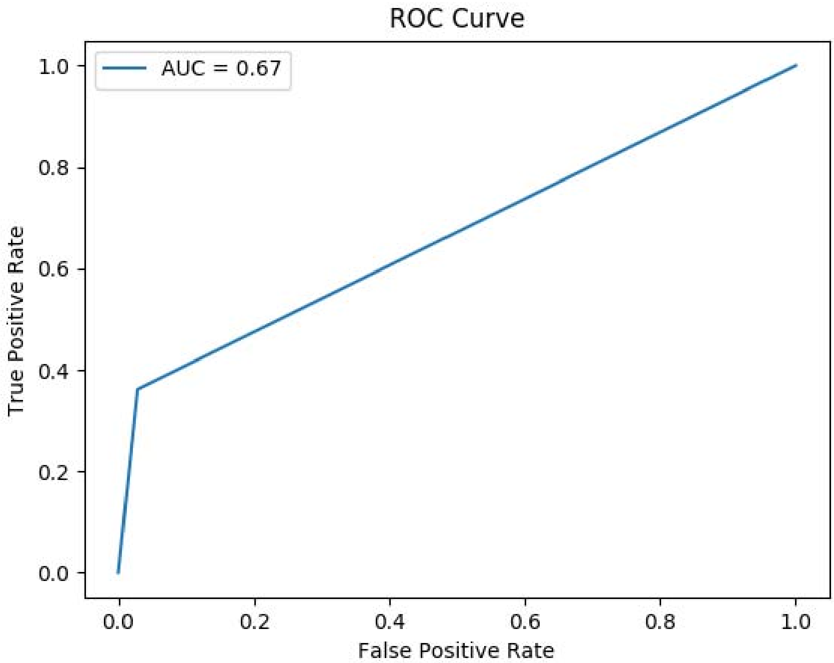


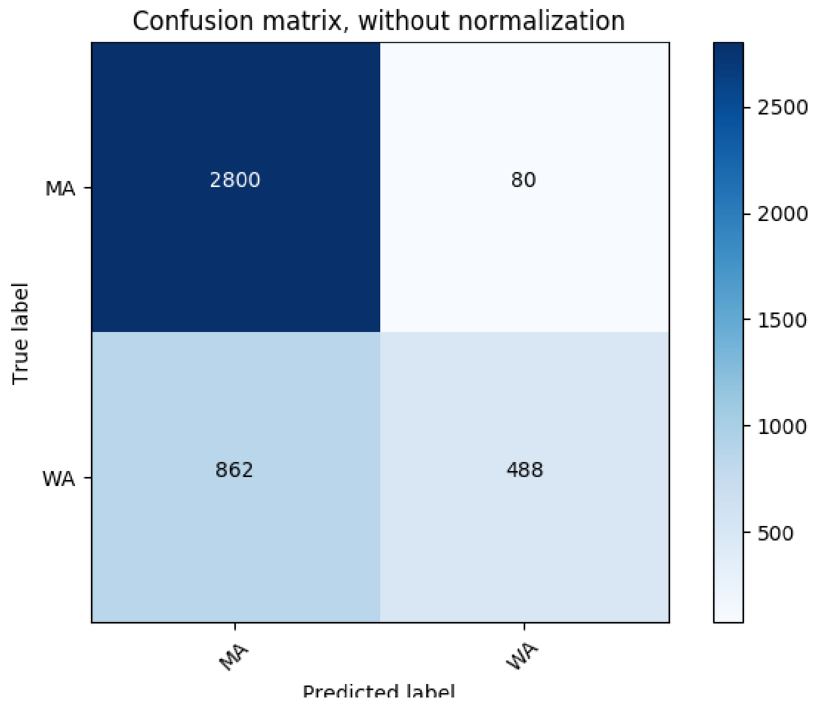
2.4 Classification Performance

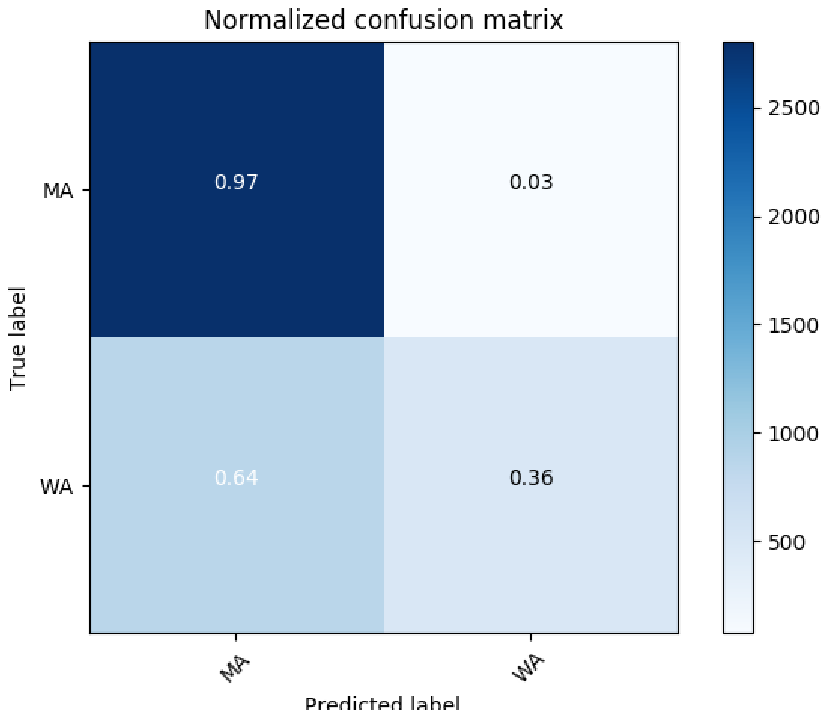
In previous sections, we have determined the best parameters for different models. In this part, we will set those classifiers with suitable parameters and study prediction performance. The performance metrics include accuracy, recall, precision, confusion matrix and ROC curve.

2.4.1 SVM (C = 1000), NMF

Accuracy: 0.777
Recall: 0.361
Precision: 0.859
Confusion Matrix (normalized and non-normalized) and ROC curve are shown in the following plots.

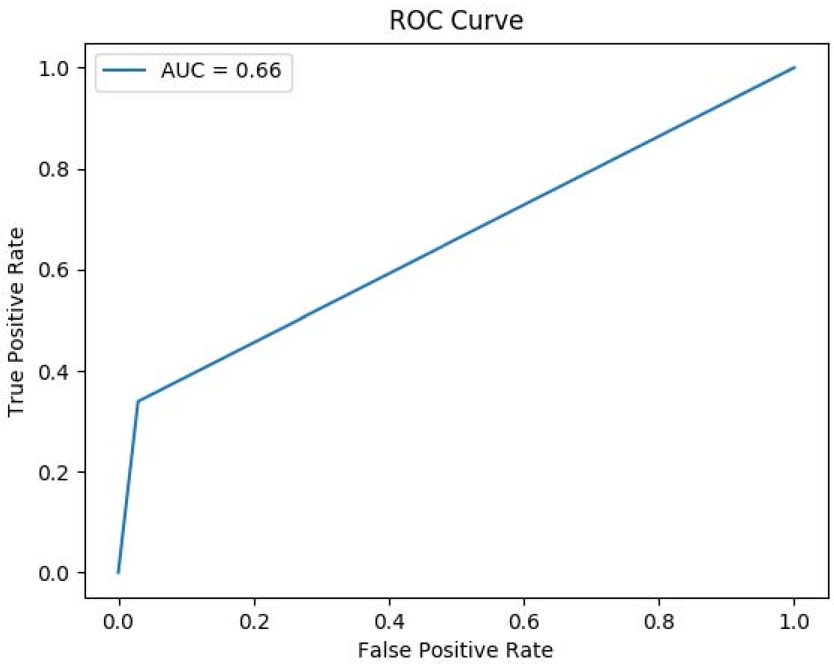


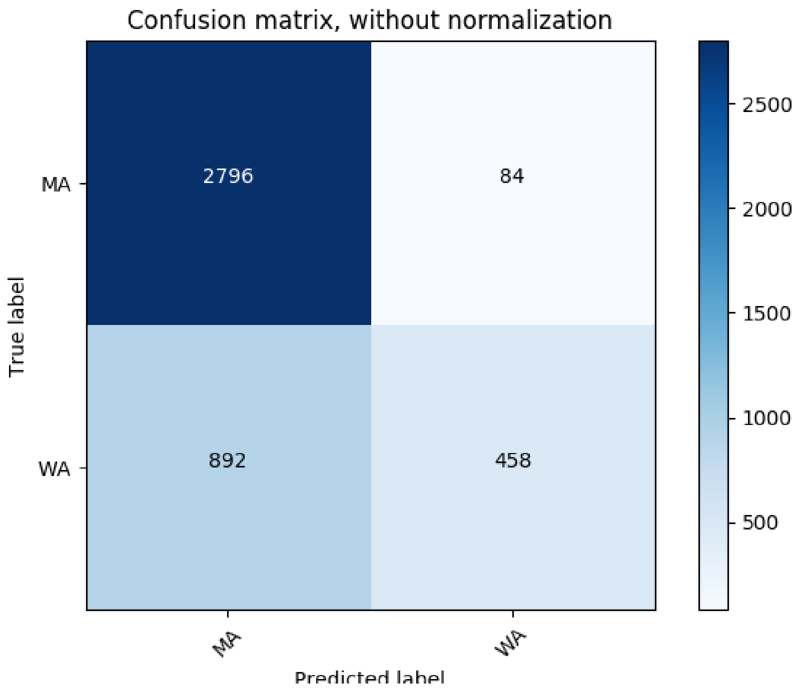

Confusion matrix, without normalization
2800
80
862
488
MA
WA
True label
MA
WA
Predicted label
2500
2000
1500
1000
500

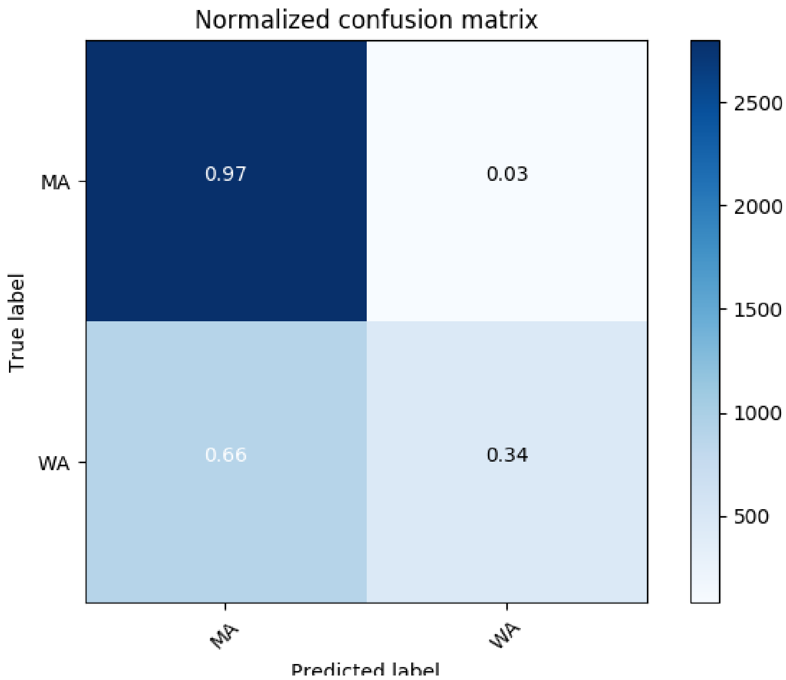

Normalized confusion matrix
0.97
0.03
0.64
0.36
MA
WA
True label
MA
WA
Predicted label
2500
2000
1500
1000
500

2.4.2 SVM (C = 10), NMF

Accuracy: 0.769
Recall: 0.339
Precision: 0.845
Confusion Matrix (normalized and non-normalized) and ROC curve are shown in the following plots.

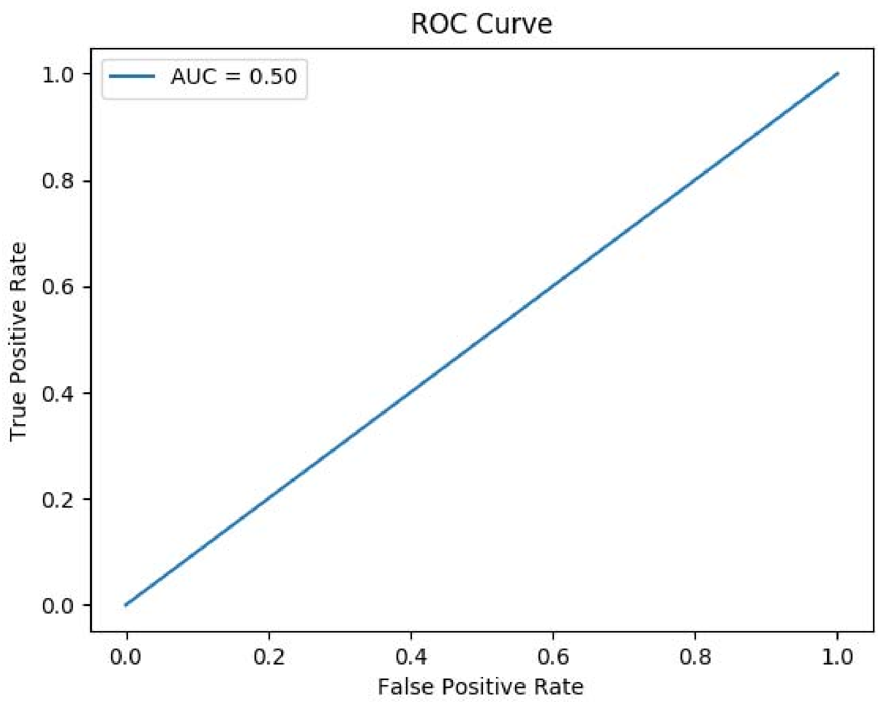


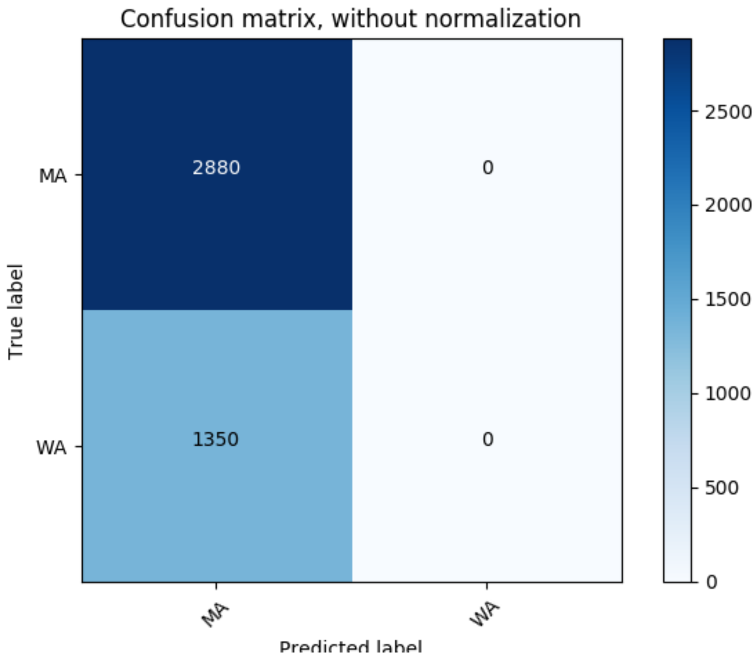

Confusion matrix, without normalization
True label
MA
WA
2796
84
892
458
MA
WA
Predicted label
2500
2000
1500
1000
500

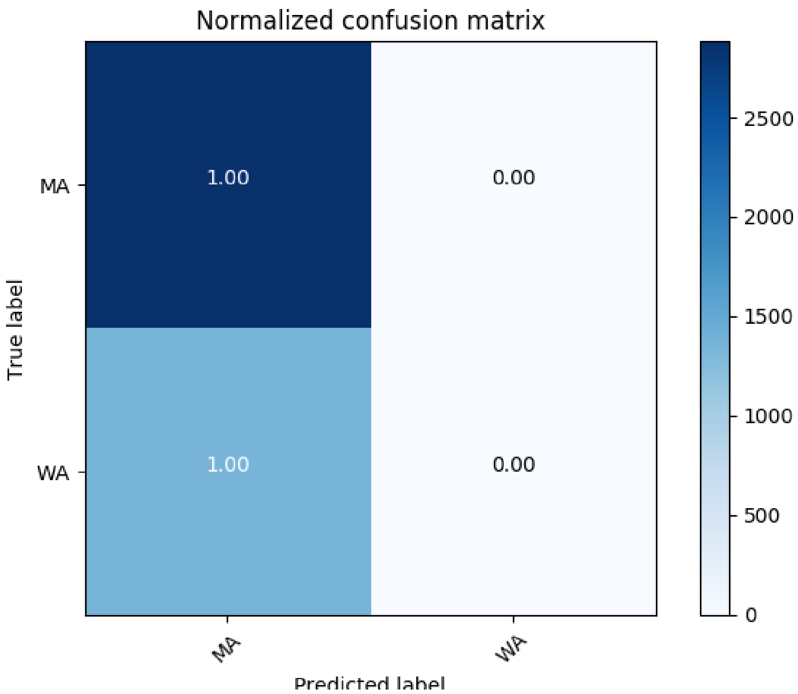

Normalized confusion matrix
True label
MA
WA
0.97
0.03
0.66
0.34
MA
WA
Predicted label
2500
2000
1500
1000
500

2.4.3 SVM (C = 0.001), NMF

Accuracy: 0.681
Recall: 0.0
Precision: 0.0
Confusion Matrix (normalized and non-normalized) and ROC curve are shown in the following plots.

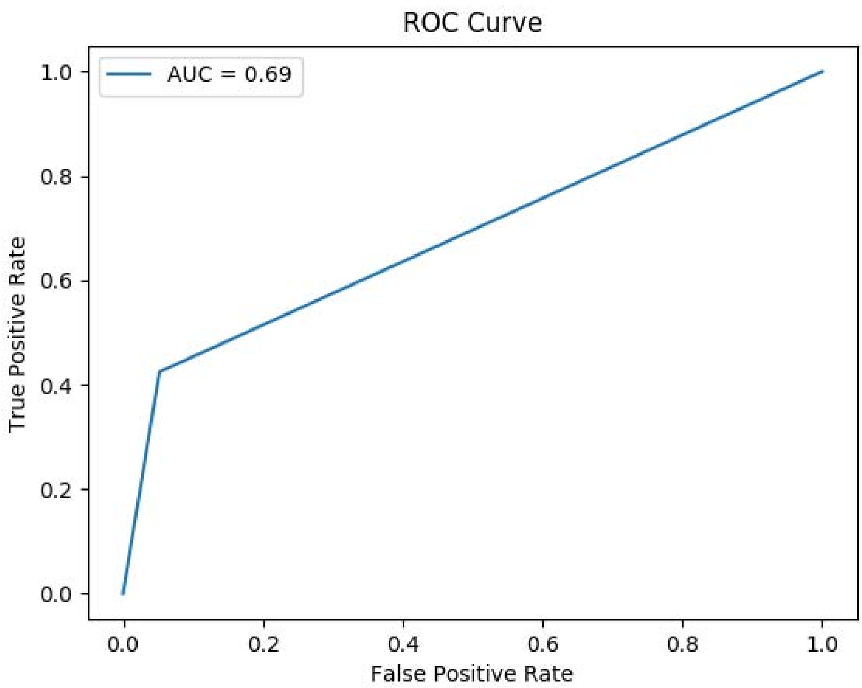


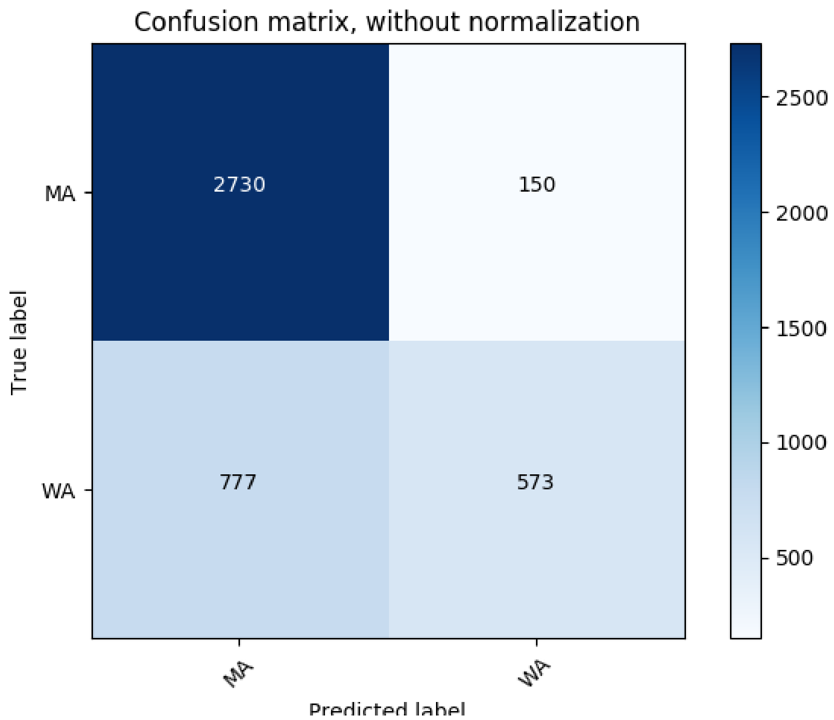

Confusion matrix, without normalization
True label
Predicted label
MA
WA
2880
0
1350
0
2500
2000
1500
1000
500
0

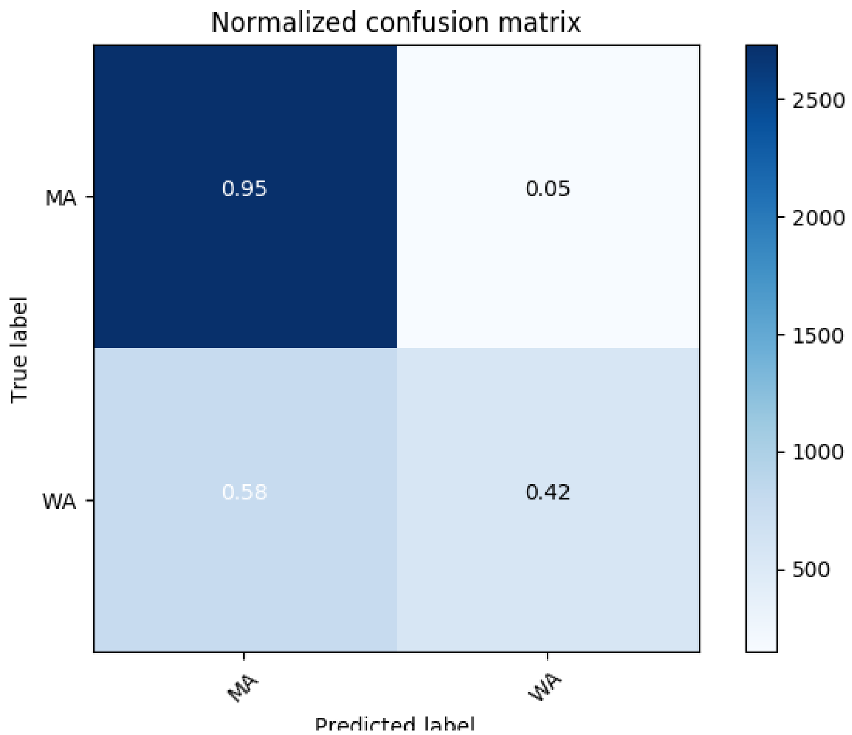

Normalized confusion matrix
True label
Predicted label
MA
WA
1.00
0.00
1.00
0.00
2500
2000
1500
1000
500
0

2.4.4 Logistic Regression, NMF

Accuracy: 0.781
Recall: 0.424
Precision: 0.793
Confusion Matrix (normalized and non-normalized) and ROC curve are shown in the following plots.

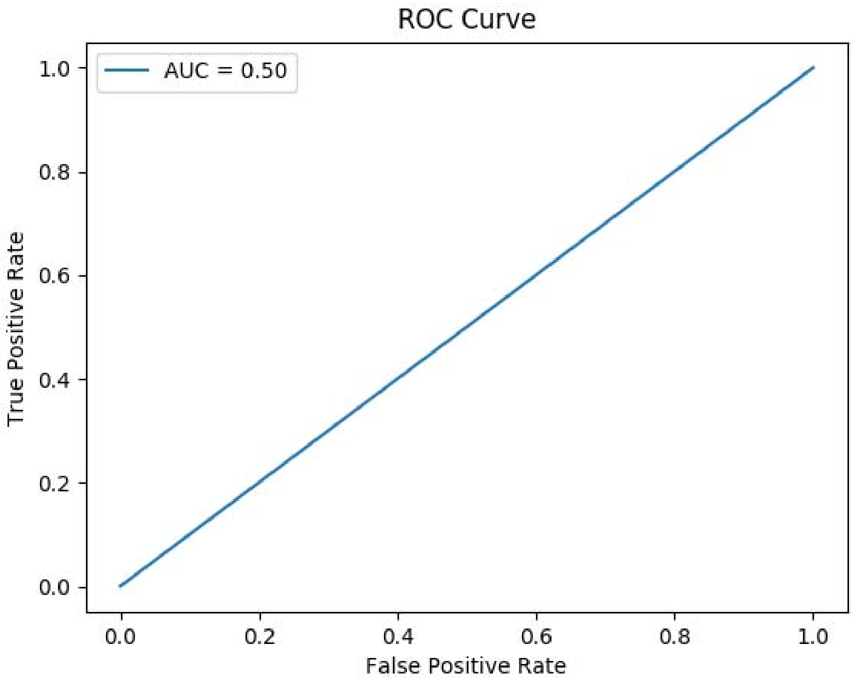


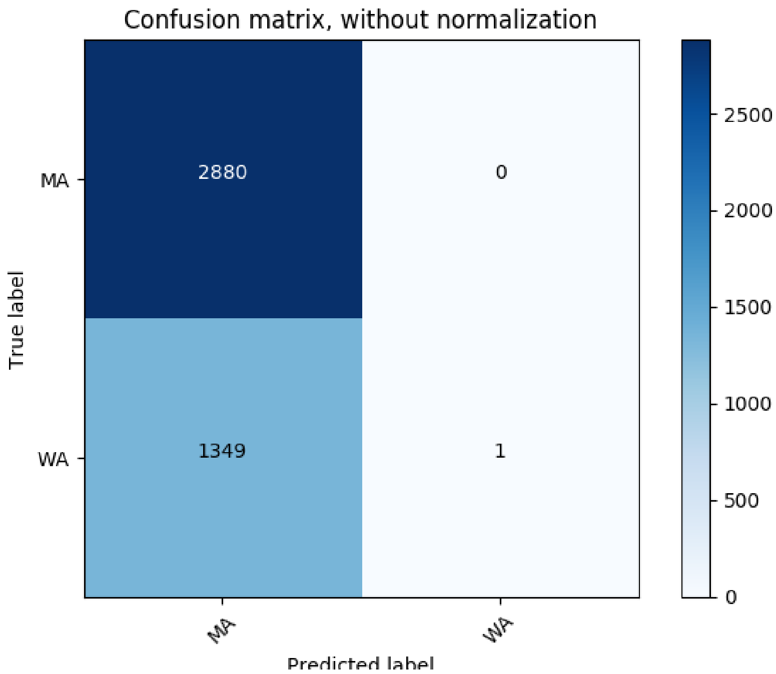

Confusion matrix, without normalization
True label
MA
WA
2730
150
777
573
Predicted label
2500
2000
1500
1000
500

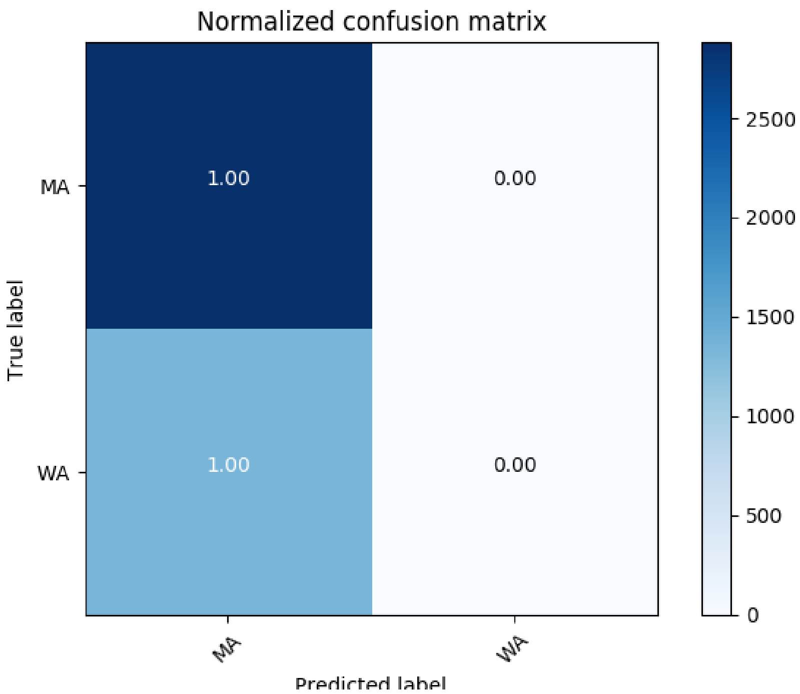

Normalized confusion matrix
True label
MA
WA
0.95
0.05
0.58
0.42
Predicted label
2500
2000
1500
1000
500

2.4.5 Naive Bayes, NMF

Accuracy: 0.681
Recall: 0.001
Precision: 1.0
Confusion Matrix (normalized and non-normalized) and ROC curve are shown in the following plots.







According to the above results, it is obvious that Logistic Regression classifier performs best among these models while Naive Bayes and Soft SVM perform worst. Although they can predict the tweets in MA correctly, their prediction of tweets in WA is totally wrong. In general, the prediction of the tweets of MA is better than that of WA for all the learners.

Part 3: Define Your Own Project

In previous part the prediction is about twitter number, and in this part we want to do prediction about users. Like the world-famous detective Sherlock Holmes who can deduce stranger's occupation based on details of clothing and striding, we can use user data to predict whether this user is active or relatively inactive. To be more specific, we are going to predict user's **tweet number**(which can be extracted by json_object['tweet']['user']['statuses_count']) from its account features. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object From this web page, we select several features and make prediction on each hashtag.
Here are the feature we use:

feature	meaning	why use it
followers_count	The number of followers this account currently has.	If an account has many users following it, it could be an active one.
friends_count	The number of users this account is following (AKA their "followings").	If an account has followed many users, it could be an active one.
favourites_count	The number of Tweets this user has liked in the account's lifetime.	If an account has liked many tweets, it could be an active one.
listed_count	The number of public lists that this user is a member of.	If an account has joined many lists, it could be an active one.
created_at	The UTC datetime that the user account was created on Twitter.	If an account was created a long period of time ago, it could have posted more tweets.
description	The user-defined UTF-8 string describing their account.	If an account has a description, it could be an active one.
default_profile	When true, indicates that the user has not altered the theme or background of their user profile.	If an account has altered the default profile, it could be an active one.
default_profile_image	When true, indicates that the user has not uploaded their own profile image and a default image is used instead.	If an account has altered the default profile image, it could be an active one.

And because there are 8 features, at that time knn may not be a very good choice, thus we decide to use random forest model to predict and cross validate to see the performance. We use mean-absolute-error as cross validation error. Here are the results for each hashtag.

For hashtag #gohawks:
The average cross-validation error(MAE) using random forest is 1855.4136309978542
The average tweet per user is 9191.520224731046

For hashtag #gopatriots:
The average cross-validation error(MAE) using random forest is 4212.566152401393
The average tweet per user is 9899.068084781946

For hashtag #nfl:
The average cross-validation error(MAE) using random forest is 2435.39503300255
The average tweet per user is 35674.22959262462

For hashtag #patriots:
The average cross-validation error(MAE) using random forest is 3636.3309333774328
The average tweet per user is 15175.413301260125

For hashtag #sb49:
The average cross-validation error(MAE) using random forest is 3452.5034799692417
The average tweet per user is 10324.564969387546

For hashtag #superbowl:
The average cross-validation error(MAE) using random forest is 2942.3201015621685
The average tweet per user is 13380.932903162666

We can see that for most hashtags the error is about 20% of the real average value. We can say the model works relatively well considering its parameters and complexity.