Popularity Prediction on Twitter

EE219 Project 5 Report

Instructor: Professor Roychowdhury

Member 1: Jingzi Zhang

Member 2: Zhensong Wei

Member 3: Yi Zheng

Member 4: Ziwen Chen

Date of Submission: Mar. 22, 2017

Introduction

Social network apps have become an important part of the society. A good example is Twitter, which people use daily to share the moments, thoughts and get information. A useful practice in social network analysis is to predict future popularity of a subject or an event. Twitter thus becomes a good platform for such analysis due to its public discussion models. Specifically, we want to predict the future popularity of a hashtag based on current and previous tweet activities.

In this project, we formulated and solved such a problem with a dataset 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. Predictions for other hashtags can be made from a regression model trained by data from some of the related hashtags. We designed and chose good features within a time window for different hashtags to develop accurate and robust popularity prediction systems. In addition, we also developed a fan base prediction system to predict the location of the author of tweet given only the textual content. Finally, we proposed our own project and managed to solve the problem.

Results and Discussion

i) Popularity Prediction

1. Dataset overview

First, we wanted to get a general sense of the data. Thus, we extracted some basic statistics for each hashtag: average number of of tweets per hour, average number of followers of users posting the tweets, and average number of retweets. Table 1.1 shows the corresponding results. Different hashtags have different total number of tweets. It can be seen that average number of tweets per hour increases with total number of tweets, and is correlated with average number of retweets. Also, there is no strong relationship between average number of followers of users posting the tweets and the average number of retweets. For example, #nfl has the largest number of average followers, but the average number of retweets per hour is not the largest.

We also tried to catch the general distribution of popularity around the event by plotting number of tweets per hour over time for #nfl and #Superbowl, and Figure 1.1 shows the corresponding results. It can be observed that there are two peaks of the number of tweets in each plot. The first peak happened two weeks before the Super bowl day 2015, which might be caused by a special event, where the major singer for the Super Bowl was announced. Thus, people tended to talk about it during that period. The second peak corresponds to the day of Super Bowl 2015. It is reasonable that most of the tweets occurred around this day, explaining the second peak.

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Total number of tweets	188136	26232	259024	489713	826951	1348767
Avg number of tweets per hour	274.96	38.35	378.65	499.42	1419.89	1401.25
Avg number of followers of users posting tweets	2375	1294	4377	1650	2235	3592
Avg number of retweets per hour	2.02	1.40	1.54	1.78	2.51	2.39

Table 1.1: statistics for different hashtag

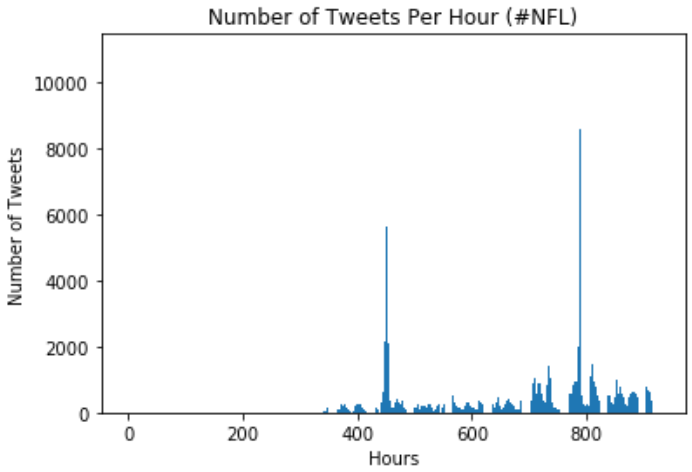

Figure 1.1(a): Number of tweets over hour (#NFL)

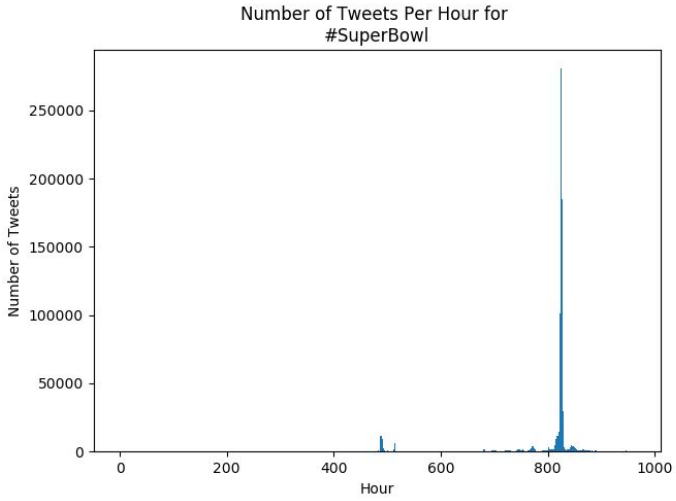

Figure 1.1(b): Number of tweets over hour (#superbowl)

2. Linear regression with 5 features

To predict the popularity from previous data, we fitted a linear model with 5 features extracted from the tweet data from the previous time window. The features are :

1. The number of tweets
2. Total number of retweets
3. Sum of the number of followers of the users posting the hashtag
4. Maximum number of followers of the users posting the hashtag
5. Time of the day (which takes 24 values that represent hours of the day with respect to a given time reference).

Preliminary study showed that the constant term is not significant for most of the hashtag, and the performance degrades when the constant term is included. Thus, for this project, constant term is not included in the linear fitting.

We also tried two different kinds of time windows. First, only the features from the previous hour were used for prediction. Second, the features from the previous 3 hours were used for fitting. More features from the current and previous tweet activities can possibly increase the training accuracy of the fitted models.

We evaluated the accuracy by looking at R-squared, which is a measure of how close the actual data are to the fitted regression. R-squared = 0 indicates that none of the variations are explained by the model, and R-squared = 1 means that all of the variations are explained by the model.
The significance of a feature was evaluated by t-value and p-value. Larger t-value and smaller p-value indicate that the feature is more significant. For a 95% confidence interval, we determined a feature to be significant if its p-value is less than 5%.

Summary for #gohawks

OLS Regression Results

Dep. Variable:	y	R-squared:	0.523
Model:	OLS	Adj. R-squared:	0.521
Method:	Least Squares	F-statistic:	212.4
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	7.27e-153
Time:	16:21:40	Log-Likelihood:	-7793.1
No. Observations:	972	AIC:	1.560e+04
Df Residuals:	967	BIC:	1.562e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.3560	0.110	3.245	0.001	0.141	0.571
x2	-0.1865	0.043	-4.310	0.000	-0.271	-0.102
x3	0.0006	7.05e-05	7.938	0.000	0.000	0.001
x4	-0.0008	0.000	-6.325	0.000	-0.001	-0.001
x5	5.7079	1.744	3.274	0.001	2.286	9.130

Omnibus:	1926.720	Durbin-Watson:	2.364
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5102746.525
Skew:	14.605	Prob(JB):	0.00
Kurtosis:	356.752	Cond. No.	1.25e+05

Figure 2.1(a): Summary for #gohawks

Summary for #gopatriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.703
Model:	OLS	Adj. R-squared:	0.701
Method:	Least Squares	F-statistic:	321.0
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	5.06e-176
Time:	16:20:50	Log-Likelihood:	-4411.9
No. Observations:	683	AIC:	8834.
Df Residuals:	678	BIC:	8856.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.1520	0.223	-0.681	0.496	-0.590	0.286
x2	-0.0304	0.201	-0.152	0.879	-0.424	0.363
x3	0.0011	0.000	11.221	0.000	0.001	0.001
x4	-0.0011	0.000	-9.929	0.000	-0.001	-0.001
x5	0.6556	0.425	1.543	0.123	-0.179	1.490

Omnibus:	592.974	Durbin-Watson:	2.262
Prob(Omnibus):	0.000	Jarque-Bera (JB):	546269.961
Skew:	2.610	Prob(JB):	0.00
Kurtosis:	141.449	Cond. No.	2.75e+04

Figure 2.1(b): Summary for #gopatriots

Summary for #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.647
Model:	OLS	Adj. R-squared:	0.645
Method:	Least Squares	F-statistic:	337.2
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	3.25e-205
Time:	16:26:58	Log-Likelihood:	-7003.3
No. Observations:	926	AIC:	1.402e+04
Df Residuals:	921	BIC:	1.404e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.2760	0.110	11.634	0.000	1.061	1.491
x2	-0.2232	0.067	-3.325	0.001	-0.355	-0.091
x3	-7.563e-05	2.4e-05	-3.149	0.002	-0.000	-2.85e-05
x4	0.0001	3.32e-05	4.217	0.000	7.48e-05	0.000
x5	2.2786	1.190	1.915	0.056	-0.057	4.614

Omnibus:	1021.554	Durbin-Watson:	2.170
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1350478.612
Skew:	4.234	Prob(JB):	0.00
Kurtosis:	189.895	Cond. No.	2.20e+05

Figure 2.1(c): Summary for #nfl

Summary for #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.701
Model:	OLS	Adj. R-squared:	0.700
Method:	Least Squares	F-statistic:	457.4
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	9.71e-253
Time:	16:35:10	Log-Likelihood:	-8788.4
No. Observations:	980	AIC:	1.759e+04
Df Residuals:	975	BIC:	1.761e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.5190	0.101	14.983	0.000	1.320	1.718
x2	-0.4913	0.099	-4.977	0.000	-0.685	-0.298
x3	-2.453e-05	3.81e-05	-0.644	0.520	-9.93e-05	5.02e-05
x4	0.0002	8.01e-05	2.882	0.004	7.37e-05	0.000
x5	7.4667	4.543	1.644	0.101	-1.448	16.382

Omnibus:	1898.438	Durbin-Watson:	1.879
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3721619.623
Skew:	14.012	Prob(JB):	0.00
Kurtosis:	303.593	Cond. No.	4.74e+05

Figure 2.1(d): Summary for #patriots

Summary for #sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.826
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	549.6
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	1.02e-216
Time:	17:07:29	Log-Likelihood:	-5689.3
No. Observations:	582	AIC:	1.139e+04
Df Residuals:	577	BIC:	1.141e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.1837	0.053	22.402	0.000	1.080	1.287
x2	-0.4203	0.047	-8.993	0.000	-0.512	-0.329
x3	0.0002	3.08e-05	7.976	0.000	0.000	0.000
x4	-0.0003	7.14e-05	-4.562	0.000	-0.000	-0.000
x5	-6.4374	13.726	-0.469	0.639	-33.397	20.522

Omnibus:	1170.219	Durbin-Watson:	1.705
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2352093.169
Skew:	14.276	Prob(JB):	0.00
Kurtosis:	313.126	Cond. No.	2.46e+06

Figure 2.1(e): Summary for #sb49

Summary for #superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.707
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	461.6
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	3.38e-252
Time:	18:26:58	Log-Likelihood:	-9976.7
No. Observations:	963	AIC:	1.996e+04
Df Residuals:	958	BIC:	1.999e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	3.0713	0.265	11.572	0.000	2.550	3.592
x2	-1.1734	0.126	-9.336	0.000	-1.420	-0.927
x3	0.0001	2.97e-05	3.951	0.000	5.91e-05	0.000
x4	0.0001	0.000	1.115	0.265	-0.000	0.000
x5	-18.1847	18.811	-0.967	0.334	-55.099	18.730

Omnibus:	1830.273	Durbin-Watson:	2.085
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5225560.602
Skew:	13.175	Prob(JB):	0.00
Kurtosis:	362.913	Cond. No.	3.93e+06

Figure 2.1(f): Summary for #superbowl

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
R-squared (Accuracy)	0.523	0.703	0.647	0.701	0.826	0.707
total number of tweets	Sig	Non-Sig	Sig	Sig	Sig	Sig
total number of retweets	Sig	Non-Sig	Sig	Sig	Sig	Sig
total number of followers posting tweets	Sig	Sig	Sig	Non-Sig	Sig	Sig
maximum number of followers posting tweets	Sig	Sig	Non-Sig	Sig	Sig	Non-Sig
time of the day	Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig

Table 2.1: Summary of accuracy and significance of feature for different hashtags

Figure 2.1(a)-(f) show the summary of the fitted models from 1 hour windows, and a summary of accuracy and significance of features was included in Table 2.1, where Sig means the feature is significant and Non-Sig means the feature is not significant. Generally, the hashtag with larger number of tweet yields a more accurate model. This is a reasonable observation because a larger dataset includes more information about the model so that the linear fitting can capture the true nature of the distribution better. The exception is #gopatriots, whose dataset is the smallest while the accuracy is the second highest. The reason is that this dataset is the most linear one and the outliers are the least, leading to less errors in results.

For the features, time of the day is not significant for most of the hashtags, so we excluded it in the analysis later in this project. Other features exhibits different significance in the linear models of different hashtag. Therefore, it is important to select different features for each hashtag in order to capture the most important information and produce the most accurate results. This is done for the analysis in part 2-part 5 in this project.

The same procedures were repeated using 3-hour windows. Figure 2.2(a)-(f) show the summary of the corresponding linear models, and Table 2.2 summarizes the training accuracy and significance of features.

Summary for #gohawks

OLS Regression Results

Dep. Variable:	y	R-squared:	0.506
Model:	OLS	Adj. R-squared:	0.498
Method:	Least Squares	F-statistic:	65.04
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	2.08e-134
Time:	19:17:18	Log-Likelihood:	-7788.3
No. Observations:	969	AIC:	1.561e+04
Df Residuals:	954	BIC:	1.568e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.6881	0.120	-5.713	0.000	-0.924	-0.452
x2	0.0033	0.047	0.070	0.944	-0.090	0.096
x3	0.0009	7.79e-05	11.364	0.000	0.001	0.001
x4	-0.0012	0.000	-8.981	0.000	-0.001	-0.001
x5	6.4857	5.153	1.259	0.208	-3.626	16.598
x6	-0.0738	0.125	-0.592	0.554	-0.318	0.171
x7	0.0076	0.046	0.167	0.867	-0.082	0.097
x8	0.0004	8.2e-05	4.866	0.000	0.000	0.001
x9	-0.0007	0.000	-5.026	0.000	-0.001	-0.000
x10	0.6861	6.991	0.098	0.922	-13.033	14.405
x11	-0.1304	0.122	-1.071	0.285	-0.369	0.109
x12	0.1808	0.045	4.010	0.000	0.092	0.269
x13	-0.0001	8.51e-05	-1.656	0.098	-0.000	2.61e-05
x14	-1.609e-06	0.000	-0.011	0.991	-0.000	0.000
x15	-0.8592	5.156	-0.167	0.868	-10.977	9.258

Omnibus:	1888.770	Durbin-Watson:	1.290
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3936313.958
Skew:	14.166	Prob(JB):	0.00
Kurtosis:	313.952	Cond. No.	9.17e+05

Figure 2.2(a): Summary for #gohawks using 3-hour windows

Summary for #gopatriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.815
Model:	OLS	Adj. R-squared:	0.811
Method:	Least Squares	F-statistic:	195.4
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	6.98e-232
Time:	19:20:46	Log-Likelihood:	-4232.9
No. Observations:	680	AIC:	8496.
Df Residuals:	665	BIC:	8564.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-2.1468	0.232	-9.242	0.000	-2.603	-1.691
x2	0.5727	0.190	3.017	0.003	0.200	0.945
x3	0.0016	8.96e-05	18.368	0.000	0.001	0.002
x4	-0.0014	0.000	-13.656	0.000	-0.002	-0.001
x5	-0.1196	1.016	-0.118	0.906	-2.114	1.875
x6	0.2826	0.229	1.235	0.217	-0.167	0.732
x7	0.2133	0.185	1.152	0.250	-0.150	0.577
x8	-0.0003	0.000	-3.006	0.003	-0.001	-0.000
x9	0.0005	0.000	3.890	0.000	0.000	0.001
x10	-0.0387	1.377	-0.028	0.978	-2.742	2.664
x11	-2.9708	0.192	-15.501	0.000	-3.347	-2.594
x12	1.5489	0.169	9.149	0.000	1.216	1.881
x13	0.0015	0.000	13.239	0.000	0.001	0.002
x14	-0.0015	0.000	-12.727	0.000	-0.002	-0.001
x15	0.1272	1.009	0.126	0.900	-1.853	2.108

Omnibus:	1264.263	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1983428.159
Skew:	12.309	Prob(JB):	0.00
Kurtosis:	266.434	Cond. No.	2.01e+05

Figure 2.2(b): Summary for #gopatriots using 3-hour windows

Summary for #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.590
Model:	OLS	Adj. R-squared:	0.584
Method:	Least Squares	F-statistic:	87.21
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	3.99e-164
Time:	19:22:17	Log-Likelihood:	-7050.4
No. Observations:	923	AIC:	1.413e+04
Df Residuals:	908	BIC:	1.420e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.5246	0.137	3.831	0.000	0.256	0.793
x2	0.0503	0.074	0.680	0.497	-0.095	0.195
x3	-1.468e-05	2.9e-05	-0.507	0.613	-7.15e-05	4.22e-05
x4	6.334e-05	3.88e-05	1.631	0.103	-1.29e-05	0.000
x5	2.7586	3.551	0.777	0.437	-4.210	9.727
x6	-0.2345	0.153	-1.531	0.126	-0.535	0.066
x7	0.1802	0.074	2.420	0.016	0.034	0.326
x8	5.461e-05	3.11e-05	1.758	0.079	-6.35e-06	0.000
x9	-4.438e-05	4.07e-05	-1.091	0.276	-0.000	3.54e-05
x10	0.1344	4.782	0.028	0.978	-9.251	9.519
x11	-0.8927	0.137	-6.514	0.000	-1.162	-0.624
x12	0.1255	0.075	1.682	0.093	-0.021	0.272
x13	0.0003	2.85e-05	8.828	0.000	0.000	0.000
x14	-0.0003	3.82e-05	-7.157	0.000	-0.000	-0.000
x15	1.9098	3.517	0.543	0.587	-4.994	8.813

Omnibus:	1173.085	Durbin-Watson:	1.165
Prob(Omnibus):	0.000	Jarque-Bera (JB):	467945.085
Skew:	6.103	Prob(JB):	0.00
Kurtosis:	112.629	Cond. No.	1.51e+06

Figure 2.2(c): Summary for #nfl using 3-hour windows

Summary for #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.526
Model:	OLS	Adj. R-squared:	0.519
Method:	Least Squares	F-statistic:	71.15
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	2.89e-144
Time:	19:30:10	Log-Likelihood:	-8988.3
No. Observations:	977	AIC:	1.801e+04
Df Residuals:	962	BIC:	1.808e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.3252	0.172	7.714	0.000	0.988	1.662
x2	-0.4795	0.154	-3.122	0.002	-0.781	-0.178
x3	-8.681e-05	6.15e-05	-1.412	0.158	-0.000	3.38e-05
x4	0.0002	0.000	2.100	0.036	1.51e-05	0.000
x5	5.2363	16.520	0.317	0.751	-27.183	37.656
x6	-0.5116	0.198	-2.588	0.010	-0.900	-0.124
x7	0.1978	0.158	1.249	0.212	-0.113	0.508
x8	0.0002	6.24e-05	2.644	0.008	4.25e-05	0.000
x9	-6.923e-05	0.000	-0.635	0.526	-0.000	0.000
x10	-3.9690	22.220	-0.179	0.858	-47.574	39.636
x11	-0.1145	0.175	-0.654	0.513	-0.458	0.229
x12	0.5752	0.150	3.846	0.000	0.282	0.869
x13	-0.0004	5.88e-05	-6.252	0.000	-0.000	-0.000
x14	0.0005	0.000	4.400	0.000	0.000	0.001
x15	-2.4855	16.279	-0.153	0.879	-34.432	29.461

Omnibus:	1894.109	Durbin-Watson:	1.103
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3097656.411
Skew:	14.139	Prob(JB):	0.00
Kurtosis:	277.398	Cond. No.	3.49e+06

Figure 2.2(d): Summary for #patriots using 3-hour windows

Summary for #sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.729
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	101.2
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	7.50e-149
Time:	20:05:25	Log-Likelihood:	-5790.4
No. Observations:	579	AIC:	1.161e+04
Df Residuals:	564	BIC:	1.168e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.0752	0.105	10.196	0.000	0.868	1.282
x2	-0.6302	0.091	-6.905	0.000	-0.810	-0.451
x3	0.0005	5.99e-05	8.054	0.000	0.000	0.001
x4	-0.0003	0.000	-2.650	0.008	-0.000	-6.88e-05
x5	-4.2699	48.192	-0.089	0.929	-98.927	90.387
x6	0.1806	0.136	1.327	0.185	-0.087	0.448
x7	-0.3419	0.096	-3.561	0.000	-0.531	-0.153
x8	0.0002	6.2e-05	2.558	0.011	3.68e-05	0.000
x9	-0.0003	0.000	-2.653	0.008	-0.000	-7.43e-05
x10	-11.5038	65.104	-0.177	0.860	-139.380	116.372
x11	0.1981	0.130	1.520	0.129	-0.058	0.454
x12	0.0558	0.089	0.629	0.530	-0.119	0.230
x13	-0.0002	5.43e-05	-3.062	0.002	-0.000	-5.96e-05
x14	8.12e-06	0.000	0.078	0.938	-0.000	0.000
x15	15.1341	47.826	0.316	0.752	-78.806	109.074

Omnibus:	1078.676	Durbin-Watson:	0.986
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1002925.985
Skew:	12.369	Prob(JB):	0.00
Kurtosis:	205.386	Cond. No.	1.91e+07

Figure 2.2(e): Summary for #sb49 using 3-hour windows

Summary for #superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.792
Method:	Least Squares	F-statistic:	244.5
Date:	Sun, 19 Mar 2017	Prob (F-statistic):	1.47e-312
Time:	22:22:51	Log-Likelihood:	-9774.9
No. Observations:	960	AIC:	1.958e+04
Df Residuals:	945	BIC:	1.965e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-2.1103	0.275	-7.673	0.000	-2.650	-1.571
x2	1.5057	0.139	10.795	0.000	1.232	1.779
x3	3.261e-05	4.08e-05	0.798	0.425	-4.75e-05	0.000
x4	-0.0005	0.000	-3.751	0.000	-0.001	-0.000
x5	4.3127	44.467	0.097	0.923	-82.953	91.579
x6	-5.8054	0.298	-19.489	0.000	-6.390	-5.221
x7	2.5229	0.134	18.805	0.000	2.260	2.786
x8	-3.397e-05	4.75e-05	-0.715	0.475	-0.000	5.93e-05
x9	-0.0010	0.000	-7.577	0.000	-0.001	-0.001
x10	12.6181	60.107	0.210	0.834	-105.340	130.576
x11	-3.9056	0.294	-13.287	0.000	-4.482	-3.329
x12	1.1846	0.123	9.635	0.000	0.943	1.426
x13	0.0001	3.85e-05	3.816	0.000	7.13e-05	0.000
x14	0.0006	0.000	4.560	0.000	0.000	0.001
x15	22.1905	44.169	0.502	0.616	-64.490	108.871

Omnibus:	1198.974	Durbin-Watson:	1.066
Prob(Omnibus):	0.000	Jarque-Bera (JB):	504284.113
Skew:	5.895	Prob(JB):	0.00
Kurtosis:	114.661	Cond. No.	2.99e+07

Figure 2.2(f): Summary for #superbowl using 3-hour windows

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
R-squared (Accuracy)	0.506	**0.815**	0.590	0.526	0.729	**0.795**
total number of tweets (1st hour)	Sig	Sig	Sig	Sig	Sig	Sig
total number of retweets (1st hour)	Non-Sig	Sig	Non-Sig	Sig	Sig	Sig
total number of followers posting tweets (1st hour)	Sig	Sig	Non-Sig	Non-Sig	Sig	Non-Sig
maximum number of followers posting tweets (1st hour)	Sig	Sig	Non-Sig	Sig	Sig	Sig
time of the day (1st hour)	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig
total number of tweets (2nd hour)	Non-Sig	Non-Sig	Non-Sig	Sig	Non-Sig	Sig
total number of retweets (2nd hour)	Non-Sig	Non-Sig	Sig	Non-Sig	Sig	Sig
total number of followers posting tweets (2nd hour)	Sig	Sig	Non-Sig	Sig	Sig	Non-Sig
maximum number of followers posting tweets (2nd hour)	Sig	Sig	Non-Sig	Non-Sig	Sig	Sig
time of the day (2nd hour)	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig
total number of tweets (3rdhour)	Non-Sig	Sig	Sig	Non-Sig	Non-Sig	Sig
total number of retweets (3rd hour)	Sig	Sig	Non-Sig	Sig	Non-Sig	Sig
total number of followers posting tweets (3rd hour)	Non-Sig	Sig	Sig	Sig	Sig	Sig
maximum number of followers posting tweets (3rd hour)	Non-Sig	Sig	Sig	Sig	Non-Sig	Sig
time of the day (3rd hour)	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig	Non-Sig

Table 2.2: Summary of accuracy and significance of features using 3-hour windows

For 3-hour windowing analysis, the training accuracy generally becomes worse, except for #gopatriots and #superbowl. Thus, the number of tweet of these two hashtag is more related to previous data compared to other hashtags. This can be also observed from the significance of features, where most of the features in the 3rd previous hour are significance for #gopatriots and #superbowl, while the those are not for other hashtags.The significance for features in the 1st previous hour is generally the same. 3-hour windowing does not necessarily improve the performance and has no significant effects on results. Thus, we did not consider using it for other analysis later in this project. In other words, we will stick to 1-hour time window on the rest of this project.

In conclusion, the performance for the fitted models is good, and is better for larger dataset, except for #gopatriots. Besides, 3-hour windowing does not necessarily improve the performance of the fitted model with these 5 features. For the next part, we want to see if we can improve the performance by using more appropriate features.

3. Linear regression with new features

To produce better results, we adjusted the feature set for each hashtag individually according to the results from part 1 and the generated training accuracy. The significant features from 1-hour windowing analysis in part 2 are chosen for each hashtag, and some new features that we found useful in predictions from the related literatures are also taken into consideration. The new features are shown below:

1. the total number of impressions of the tweets
2. the total number of urls in the tweets
3. the total number of favorites of the tweets
4. the total number of friends of the users posting the tweets

For each hashtag, the model producing the highest accuracy is reported. Then, we want to see what features are the most important for each hashtag. We selected the top 3 significant features, according to their t values as well as the value of the coefficients. To do so, we first identified the significant features whose p-value is less than 5%. Then, we sorted the features according to their t values in descending order. At the mean time, if the coefficient of any features is close to 0, the feature would be removed from the list. We chose the top 3 features with largest t values from the updated list.

Figure 3.1-3.6 show the summary of the fitted model for 6 hashtags as well as the scatters of popularity (the predictant) over the top 3 features. More specifically, Section (a) of each figure is the summary of model, and Section (b)-(d) are the plots from the top 3 features. The features that were used for each hashtag, the significance of features, and the training accuracy are then summarized in Table 3.1.

Summary for #gohawks

OLS Regression Results

Dep. Variable:	y	R-squared:	0.606
Model:	OLS	Adj. R-squared:	0.603
Method:	Least Squares	F-statistic:	211.8
Date:	Mon, 20 Mar 2017	Prob (F-statistic):	3.85e-190
Time:	20:18:40	Log-Likelihood:	-7700.9
No. Observations:	972	AIC:	1.542e+04
Df Residuals:	965	BIC:	1.545e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	-0.3345	0.197	-1.701	0.089	-0.720	0.051
x2	-0.2833	0.057	-4.950	0.000	-0.396	-0.171
x3	1.405e-05	7.58e-05	0.185	0.853	-0.000	0.000
x4	-0.0003	0.000	-2.229	0.026	-0.000	-3.06e-05
x5	4.2039	0.475	8.844	0.000	3.271	5.137
x6	0.0844	0.022	3.808	0.000	0.041	0.128
x7	0.0016	0.000	7.129	0.000	0.001	0.002

Omnibus:	2241.940	Durbin-Watson:	2.085
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12196530.092
Skew:	20.643	Prob(JB):	0.00
Kurtosis:	550.215	Cond. No.	4.12e+04

Figure 3.1(a): Summary for #gohawks using 1-hour windows with new features

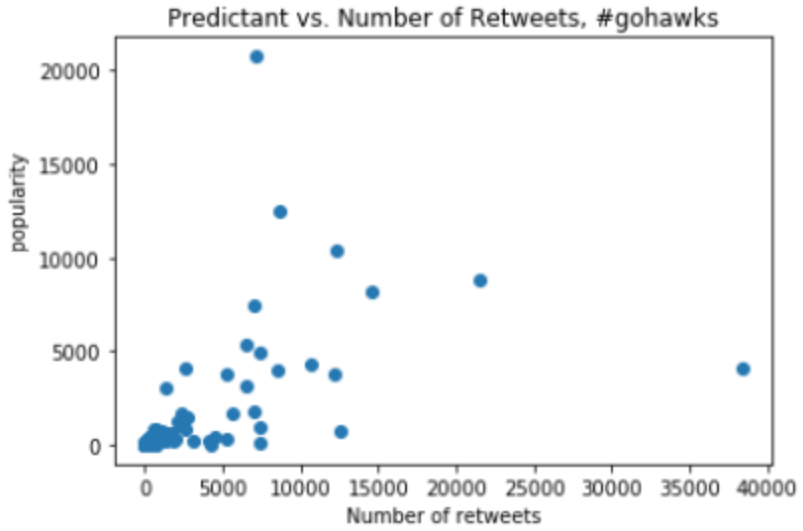

Figure 3.1(b): the scatter of popularity vs. number of retweets for #gohawks

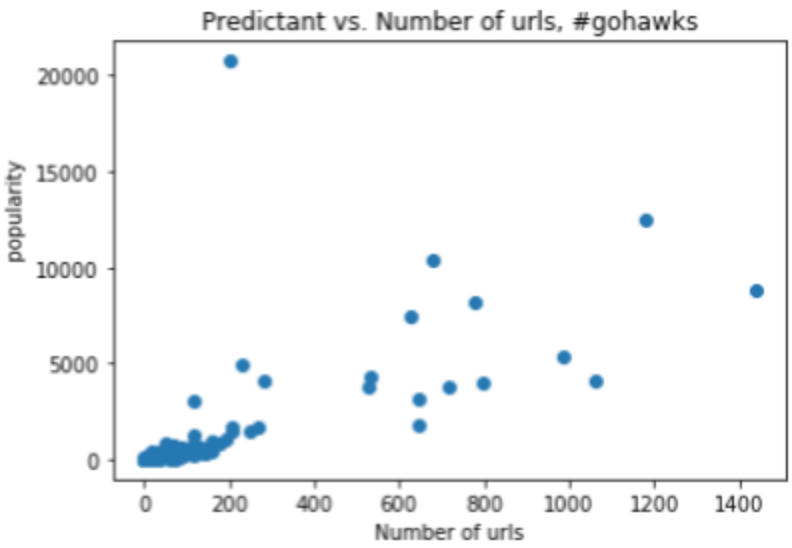

Figure 3.1(c): the scatter of popularity vs. number of urls for #gohawks

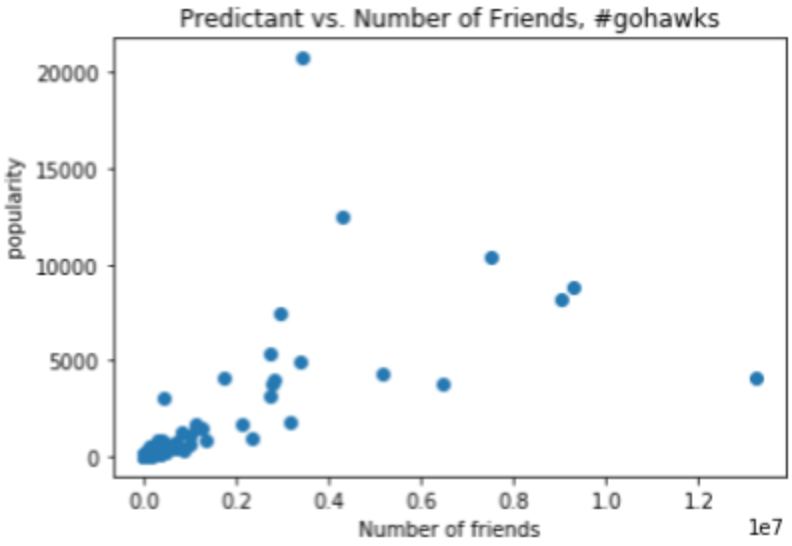

Figure 3.1(d): the scatter of popularity vs. number of friends for #gohawks

Summary for #gopatriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	503.6
Date:	Mon, 20 Mar 2017	Prob (F-statistic):	1.71e-225
Time:	20:19:11	Log-Likelihood:	-4297.0
No. Observations:	683	AIC:	8604.
Df Residuals:	678	BIC:	8627.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.0009	8.48e-05	10.119	0.000	0.001	0.001
x2	-0.0006	9.23e-05	-6.768	0.000	-0.001	-0.000
x3	-0.0002	4.53e-05	-4.194	0.000	-0.000	-0.000
x4	3.9948	0.630	6.343	0.000	2.758	5.231
x5	-13.0378	0.969	-13.450	0.000	-14.941	-11.135

Omnibus:	896.194	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	551631.474
Skew:	6.108	Prob(JB):	0.00
Kurtosis:	141.689	Cond. No.	1.03e+05

Figure 3.2(a): Summary for #gopatriots using 1-hour windows with new features

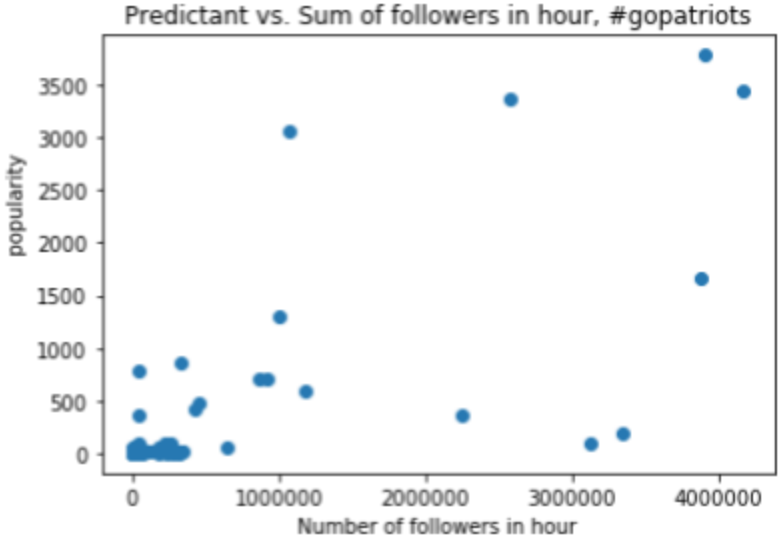

Figure 3.2(b): the scatter of popularity vs. number of followers for #gopatriots

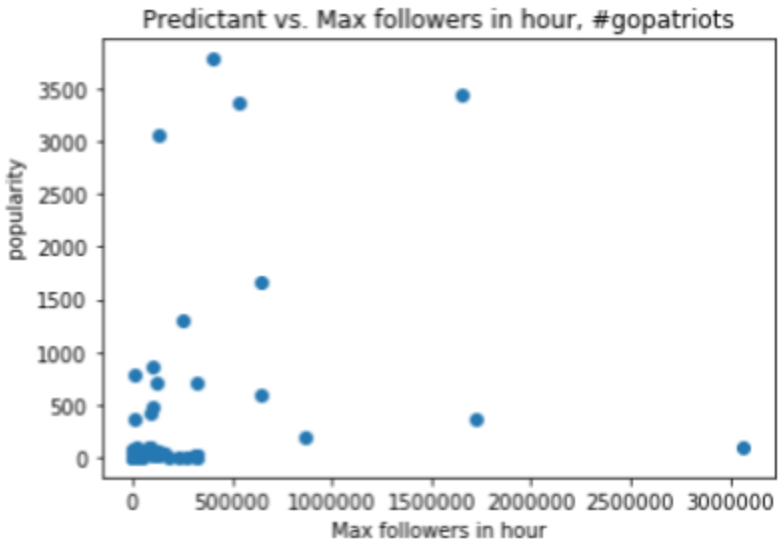

Figure 3.2(c): the scatter of popularity vs. maximum number of followers for #gopatriots

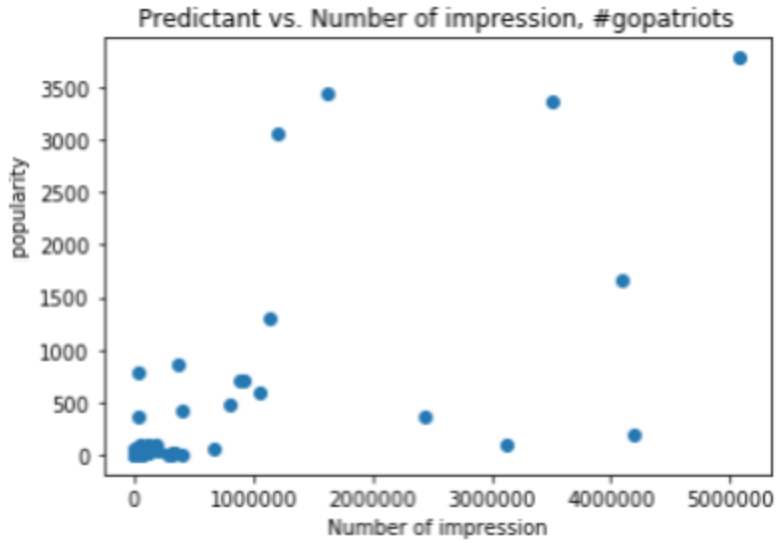

Figure 3.2(d): the scatter of popularity vs. number of impressions for #gopatriots

Summary for #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.665
Model:	OLS	Adj. R-squared:	0.663
Method:	Least Squares	F-statistic:	261.1
Date:	Mon, 20 Mar 2017	Prob (F-statistic):	1.63e-213
Time:	20:27:56	Log-Likelihood:	-6978.1
No. Observations:	926	AIC:	1.397e+04
Df Residuals:	919	BIC:	1.400e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.7715	0.160	4.831	0.000	0.458	1.085
x2	-0.0733	0.069	-1.068	0.286	-0.208	0.061
x3	-0.0001	2.77e-05	-4.550	0.000	-0.000	-7.17e-05
x4	0.0001	3.37e-05	4.059	0.000	7.07e-05	0.000
x5	3.304e-05	1.49e-05	2.223	0.026	3.87e-06	6.22e-05
x6	0.4884	0.107	4.574	0.000	0.279	0.698
x7	9.471e-05	0.000	0.677	0.499	-0.000	0.000

Omnibus:	1287.373	Durbin-Watson:	2.085
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1354906.603
Skew:	6.894	Prob(JB):	0.00
Kurtosis:	189.886	Cond. No.	5.31e+04

Figure 3.3(a): Summary for #nfl using 1-hour windows with new features

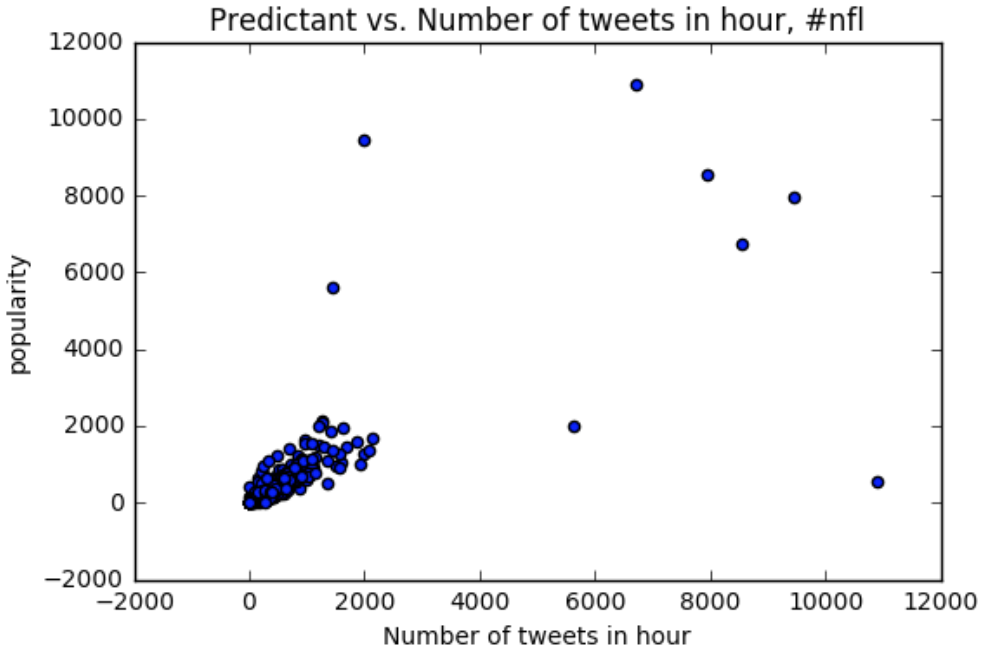

Figure 3.3(b): the scatter of popularity vs. number of tweets for #nfl

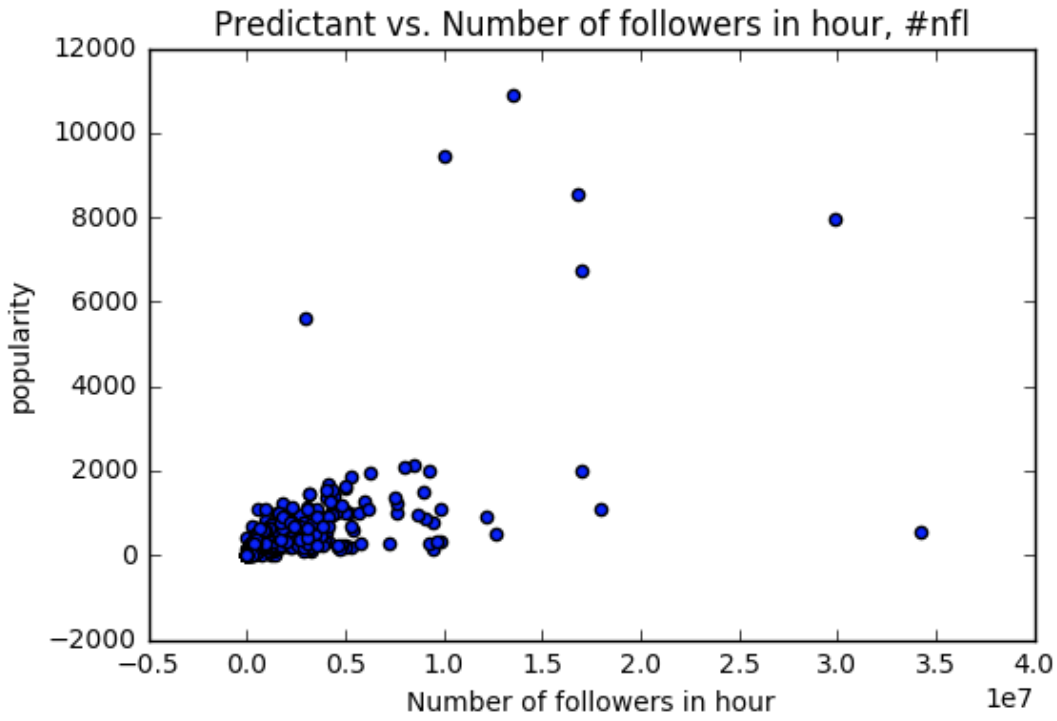

Figure 3.3(c): the scatter of popularity vs maximum number of followers for #nfl

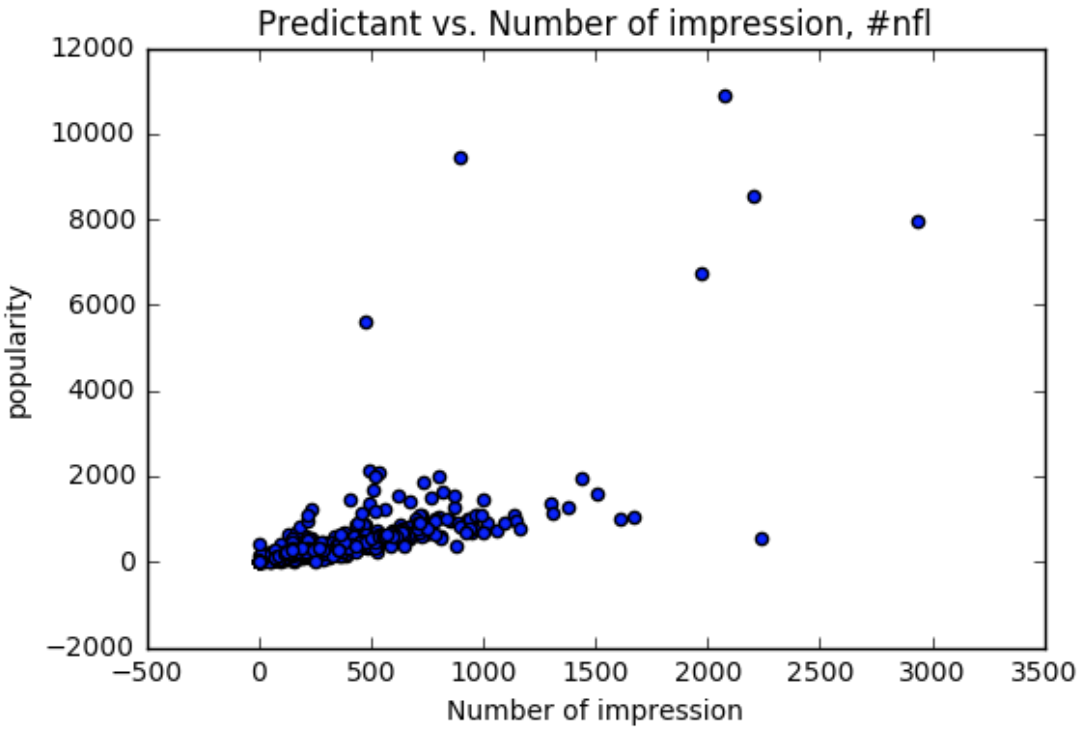

Figure 3.3(d): the scatter of popularity vs. number of impressions for #nfl

Summary for #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.758
Model:	OLS	Adj. R-squared:	0.756
Method:	Least Squares	F-statistic:	435.1
Date:	Tue, 21 Mar 2017	Prob (F-statistic):	1.65e-294
Time:	11:12:45	Log-Likelihood:	-8685.1
No. Observations:	980	AIC:	1.738e+04
Df Residuals:	973	BIC:	1.742e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-1.8767	0.337	-5.577	0.000	-2.537	-1.216
x2	0.1475	0.109	1.357	0.175	-0.066	0.361
x3	-0.0006	7.92e-05	-7.626	0.000	-0.001	-0.000
x4	0.0004	3.35e-05	11.703	0.000	0.000	0.000
x5	2.2233	0.207	10.716	0.000	1.816	2.630
x6	-0.5498	0.178	-3.096	0.002	-0.898	-0.201
x7	0.0007	0.000	2.887	0.004	0.000	0.001

Omnibus:	1935.977	Durbin-Watson:	1.766
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4645517.012
Skew:	14.556	Prob(JB):	0.00
Kurtosis:	339.036	Cond. No.	5.41e+04

Figure 3.4(a): Summary for #patriots using 1-hour windows with new features

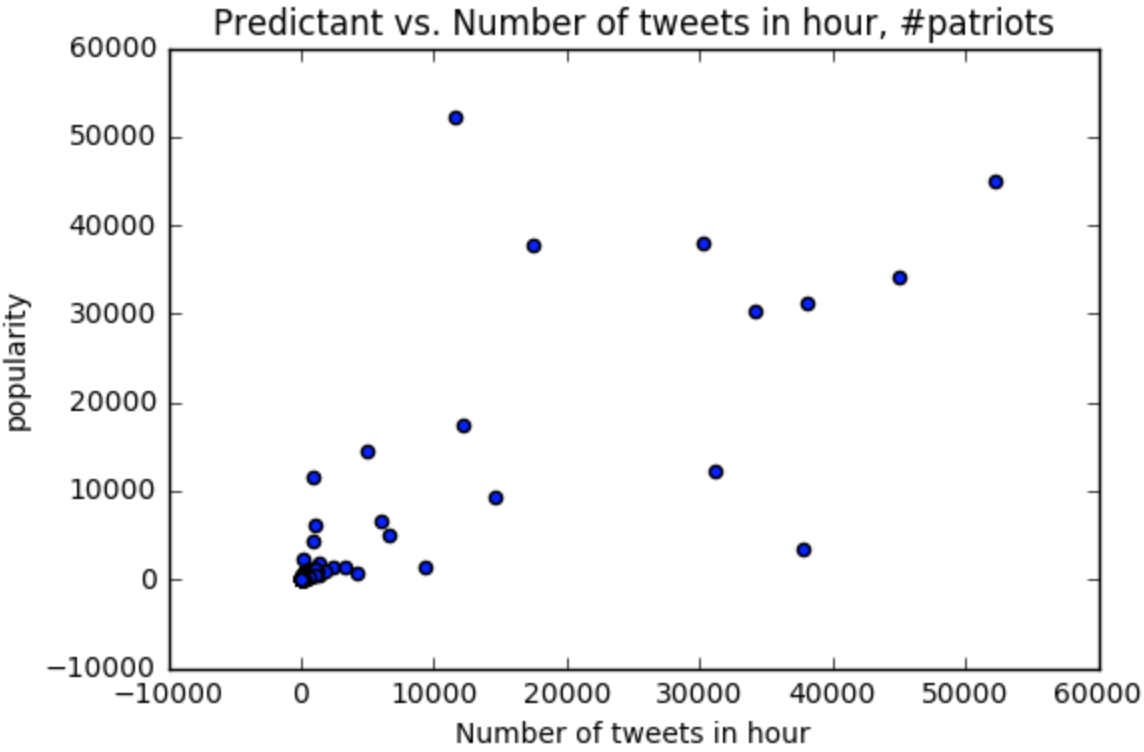

Figure 3.4(b): the scatter of popularity vs. number of tweets for #patriots

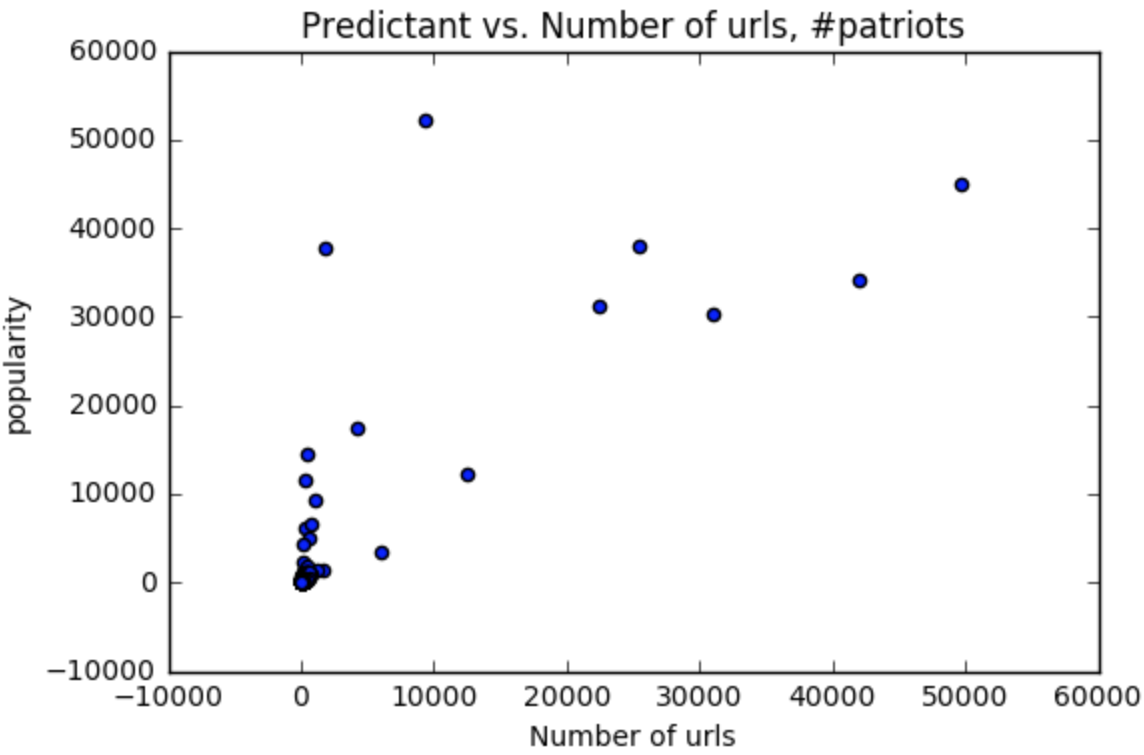

Figure 3.4(c): the scatter of popularity vs. number of urls for #patriots

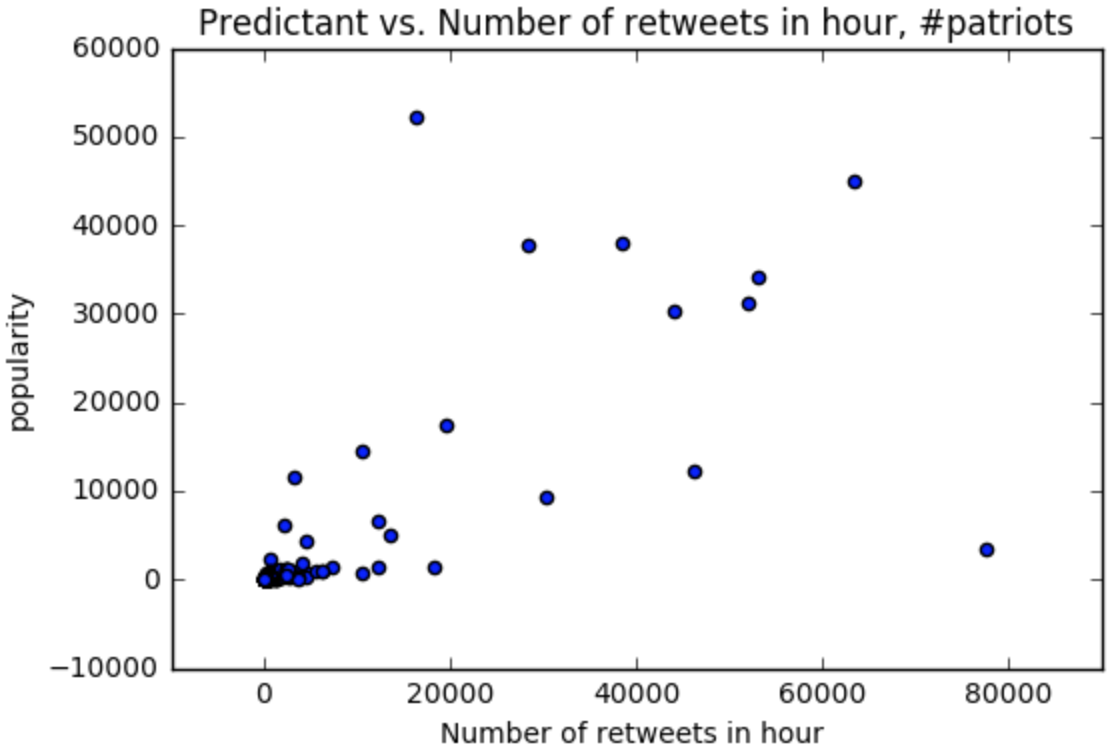

Figure 3.4(d): the scatter of popularity vs. number of retweets in hour for #patriots

Summary for #sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.850
Method:	Least Squares	F-statistic:	472.7
Date:	Tue, 21 Mar 2017	Prob (F-statistic):	8.78e-234
Time:	10:53:03	Log-Likelihood:	-5643.0
No. Observations:	582	AIC:	1.130e+04
Df Residuals:	575	BIC:	1.133e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-2.6085	0.487	-5.352	0.000	-3.566	-1.651
x2	0.1976	0.113	1.750	0.081	-0.024	0.419
x3	0.0003	4.3e-05	6.621	0.000	0.000	0.000
x4	-0.0005	7.22e-05	-6.476	0.000	-0.001	-0.000
x5	-6.358e-05	1.91e-05	-3.320	0.001	-0.000	-2.6e-05
x6	2.2877	0.346	6.621	0.000	1.609	2.966
x7	0.0022	0.000	4.923	0.000	0.001	0.003

Omnibus:	1040.160	Durbin-Watson:	1.782
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1458351.617
Skew:	11.045	Prob(JB):	0.00
Kurtosis:	247.234	Cond. No.	3.33e+05

Figure 3.5(a): Summary for #sb49 using 1-hour windows with new features

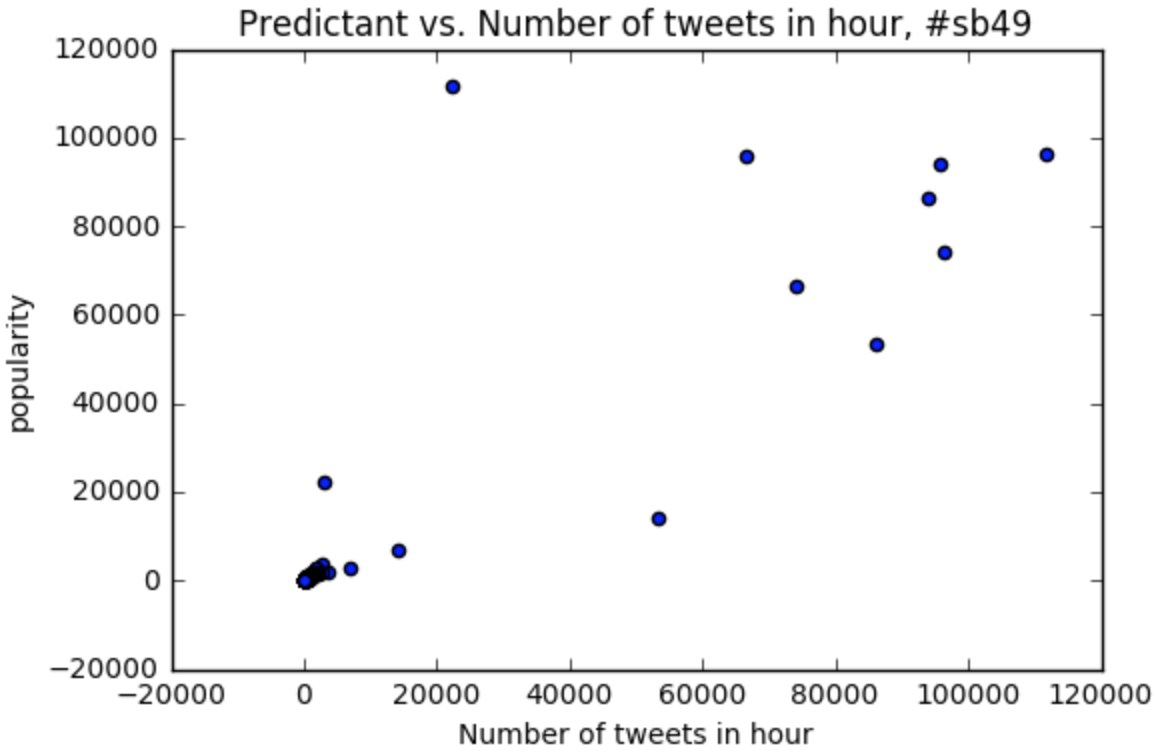

Figure 3.4(b): the scatter of popularity vs. number of tweets in hour for #sb49

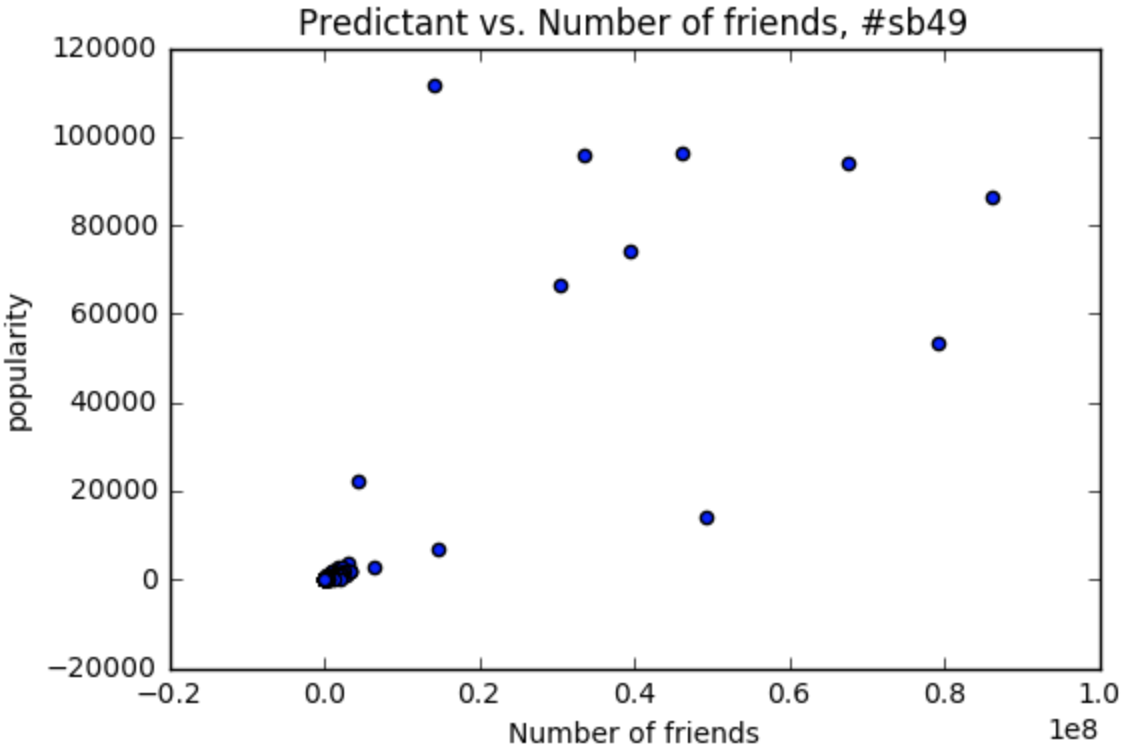

Figure 3.4(c): the scatter of popularity vs. number of friends for #sb49

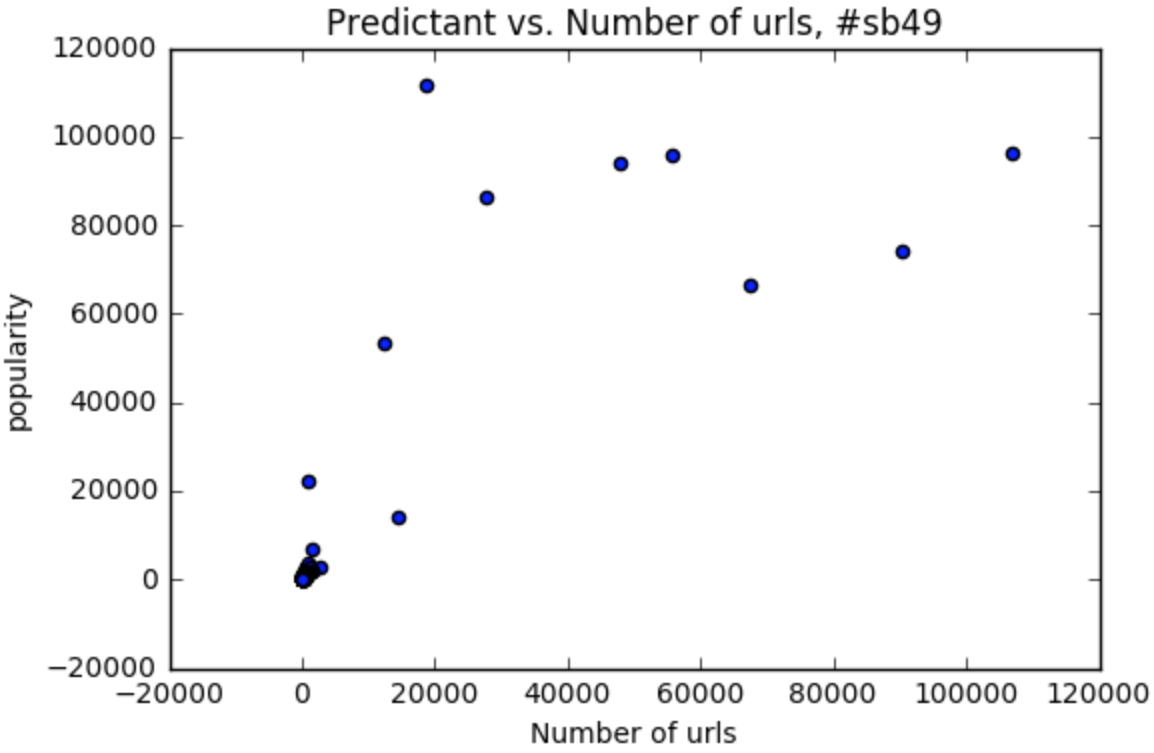

Figure 3.4(d): the scatter of popularity vs. number of urls for #sb49

Summary for #superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	809.4
Date:	Tue, 21 Mar 2017	Prob (F-statistic):	0.00
Time:	00:33:26	Log-Likelihood:	-9698.6
No. Observations:	963	AIC:	1.941e+04
Df Residuals:	957	BIC:	1.944e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-1.8043	0.305	-5.920	0.000	-2.402	-1.206
x2	0.3194	0.124	2.580	0.010	0.076	0.562
x3	-4.926e-05	3.44e-05	-1.431	0.153	-0.000	1.83e-05
x4	-0.0002	9.86e-06	-20.725	0.000	-0.000	-0.000
x5	7.1687	0.927	7.730	0.000	5.349	8.989
x6	0.0025	0.000	10.196	0.000	0.002	0.003

Omnibus:	1933.182	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7659536.883
Skew:	14.813	Prob(JB):	0.00
Kurtosis:	438.906	Cond. No.	5.40e+05

Figure 3.6(a): Summary for #superbowl using 1-hour windows with new features

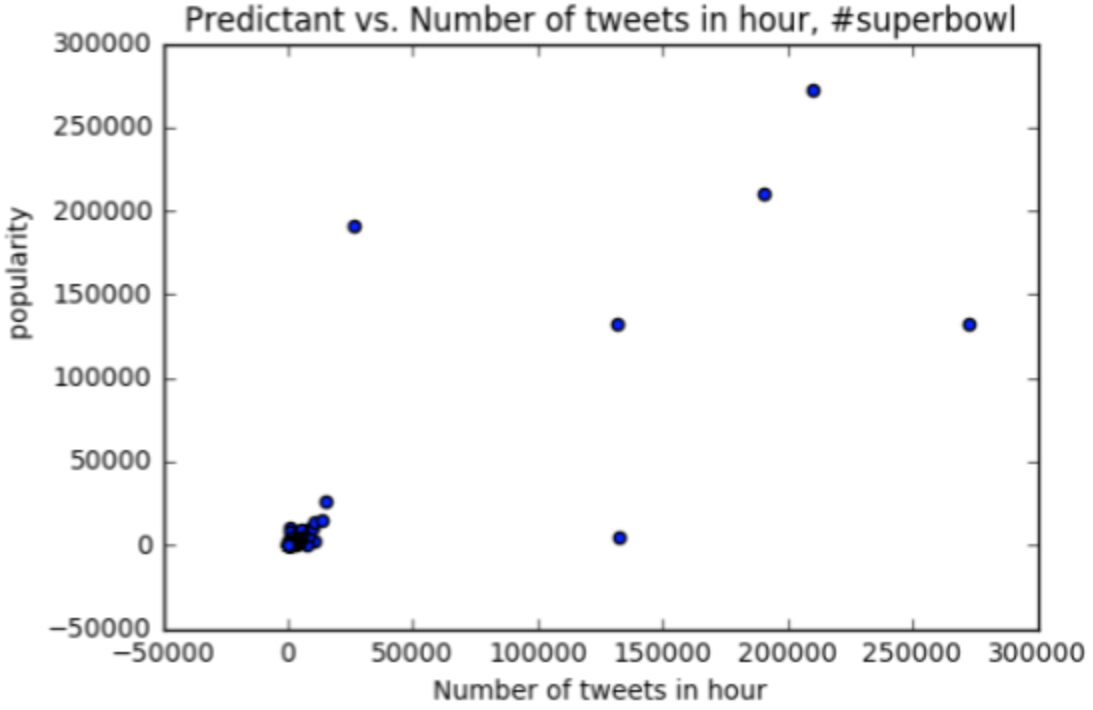

Figure 3.4(b): the scatter of popularity vs. number of tweets for #superbowl

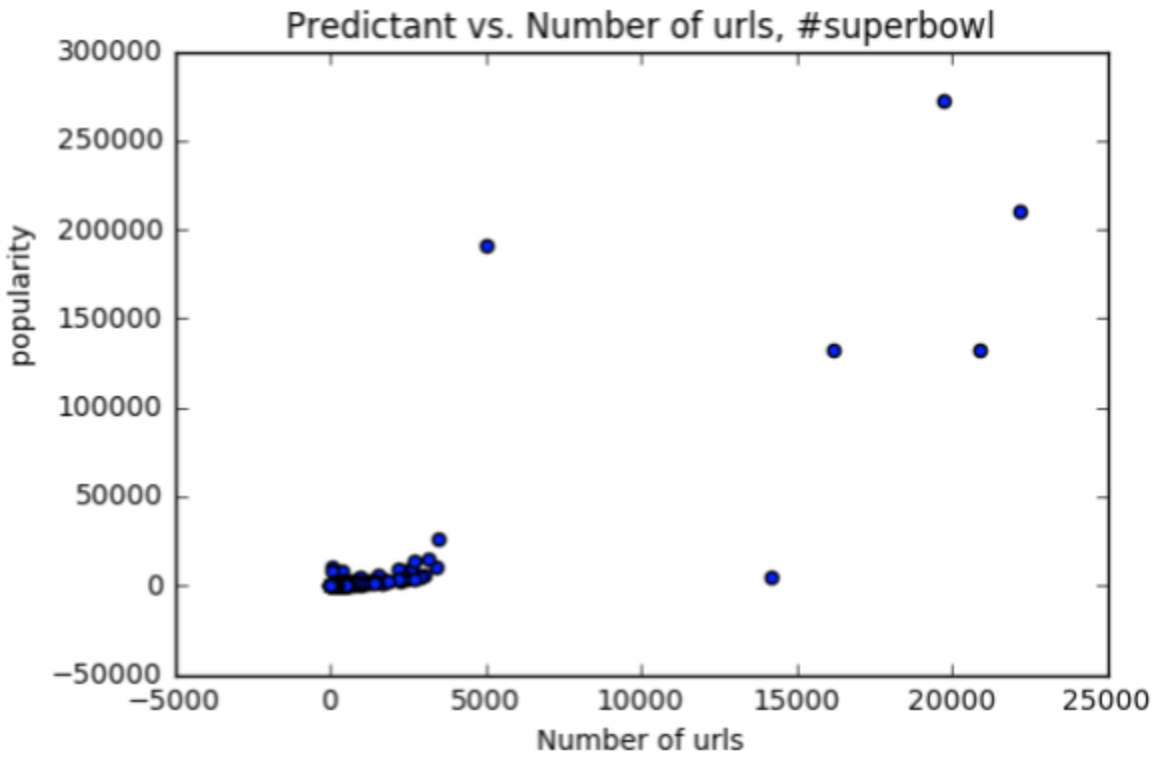

Figure 3.4(c): the scatter of popularity vs. number of urls for #superbowl

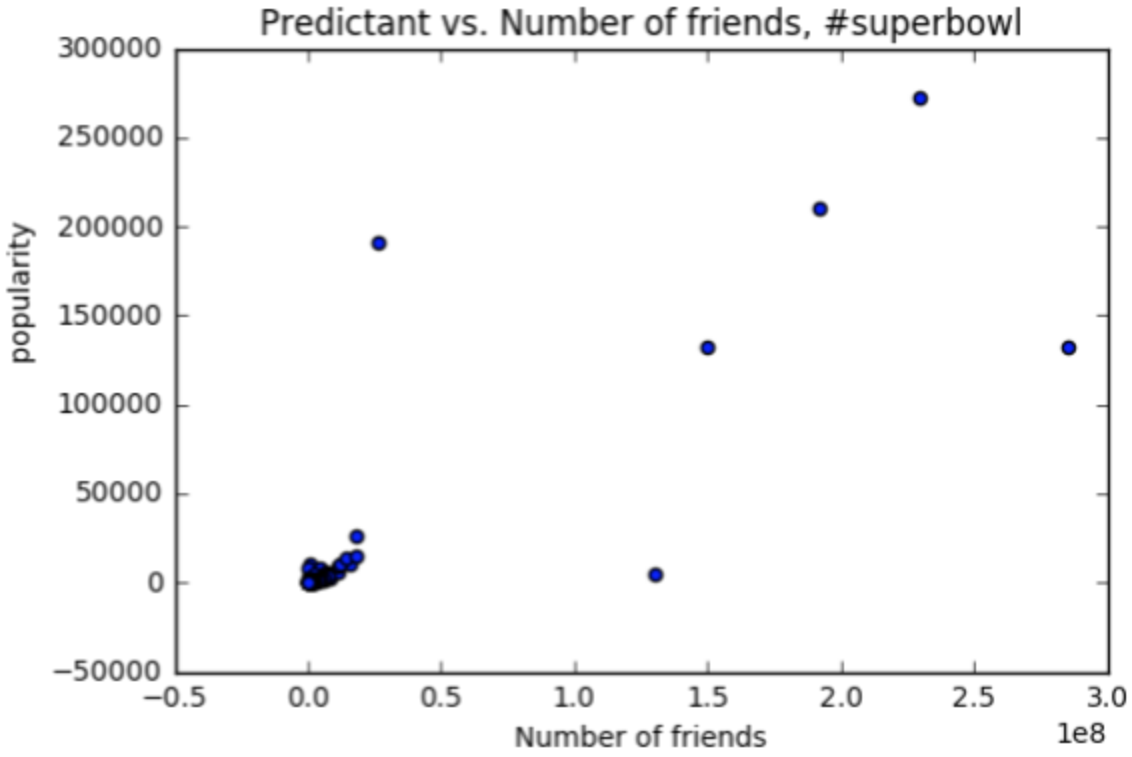

Figure 3.4(d): the scatter of popularity vs. number of friends for #sb49

	num_tweets_in_hour	num_retweets_in_hour	num_followers_in_hour	max_followers_in_hour	num_impression	sum_url	num_favorites	num_friends
#gohawks	□	**T**	□	□		**T**	□	T
#gopatriots			**T**	**T**	**T**	□	□	□
#nfl	**T**	□	**T**	□	**T**	□		□
#patriots	**T**	**T**		□	□	**T**	□	□
#sb49	**T**	□	□	□	□	**T**		**T**
#superbowl	**T**	□	□		□	**T**		**T**

Table 3.1 Top 3 and selected features in the model

In Table 3.1:

T : represents that the feature is one of the top 3 features

□ : represents that the feature is used in the regression model

From the scatter plots, it can be observed that the significant features tend to have a linear relationship with the number of tweets in the next hour, except for several outliers, which can affect the performance of linear regression. The appearance of outliers is reasonable because sometimes the number of tweets goes down even if the number of tweets was high in the previous day due to the fact that the event just passed on that day. More variations around the super bowl day and changing behaviors of people around that time can also account for the appearance of outliers.

The number of urls and the number of tweets are the most significant features for different hashtags while the number of favorites is significant for none of the hashtag. Other features have fair significance. They are used for different hashtags and are significant for at least one hashtag.

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
R-squared (Accuracy)	**0.606** (0.506)	**0.788** (0.815)	**0.665** (0.590)	**0.758** (0.526)	**0.862** (0.729)	**0.835** (0.795)

Table 3.2: Summary of accuracy with new features

In Table 3.2, the gray numbers in the parentheses are the R-squared values from part 2 without adding new features to the model. It can be observed that the accuracies were improved in the new model for the most of the hashtags but #gopatriots. Whereas the accuracy of #gopatriots is still decent. The accuracy of #patriots has a significant improvement. Therefore, the new feature sets are more related to the actual number of tweets in the next hour and thus give us better results.

4. Cross-validation and linear regression for 3 time periods

Further, to evaluate the robustness of the fitted models, we performed 10-fold cross-validation on the models obtained in part 3. More specifically, the feature data were splitted into 10 parts, and 10 tests were run, during each of which 9 parts were used for training and the remaining 1 part was used for testing. The average prediction error $|N_{predicted}-N_{real}|$ was reported for each model.

Table 4.1 summarizes the average prediction errors for different hashtags from 1-hour windowing analysis. The tweets with the hashtag #gopatriots produces the smallest cross-validation error, while the largest dataset of #superbowl is has the largest error of 1213.5. In summary, the cross-validation error increases with the size of the dataset, and exhibits decent performance and robustness.

Average prediction errors	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
1-hour window	187.78	33.65	122.85	729.11	973.32	1213.50

Table 4.1: summary of average prediction errors for different hashtags

The regression models can be more accurate in prediction if we take the time into consideration. Since we know the super bowl's date and time, we can create different regression models for different periods of time. We splitted the tweets into 3 parts based on their first posting date, and the 3 time periods for the 3 parts are:

1. Before Feb. 1, 8:00 a.m. (inactive period)
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m. (active period)
3. After Feb. 1, 8:00 p.m. (after the active period)

Then, we trained 3 regression models for these 3 time periods for each hashtag and performed cross-validation to obtain the average prediction errors for evaluation. The results average prediction errors for different hashtags in 3 time periods are summarized in Table 4.2

Average prediction errors	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
period 1	178.95	10.81	80.61	150.00	46.31	180.41
period 2	3304.49	2313.94	2009.71	11263.47	52759.26	67360.14
period 3	21.21	5.31	113.89	79.18	159.81	222.19

Table 4.2: summary of average prediction errors for different hashtags in 3 time periods

The general trend here is still the same when we did analysis on the entire dataset without considering time. In other words, larger dataset leads to larger average prediction error in this time-specific analysis. An exception is period 1, where #sb49 has a large dataset while generates small error. This may be due to the fact that this hashtag is more inactive than others during the inactive period. Interestingly, errors are different in different time period, where the models for period 2 yields the largest error, and the errors in period 1 and period 3 are similarly small. This is a reasonable observation considering the fact that period 2 is the most active period where the number of tweets boosted and the activities were harder to predict. Overall, our model has decent performance for period 1 and period 3, and is not good at predicting popularity for period 2 due to more variations during the active period.

5. Predictions for time-specific test data with time-specific models

To further test the accuracy of the fitted models in 3 time periods, we run our models with a set of provided test data to make predictions for the next hour. There are 10 set of data, each of them contains an unknown hashtag's tweets for 6-hour window and is specified by its period number. Then the corresponding model in that time period is used to predict the popularity in the next hour for each hashtag. Table 5.1-5.10 summarizes the real values and the predicted values for each set of data, where the real values are marked BLUE. The first hour is not included since there is no data that can be used for its prediction. Note that sample8_period1.txt only contains data in 5 hours, so Table 5.8 only have 4 columns of test results for number of tweets.

It can be observed that models from different hashtags made different predictions on the future popularity, and one of them would produce the most accurate results. This hashtag whose model made the most accurate predictions is most likely to be the hashtag whose tweets the test file contains. Thus, we can also find the most likely hashtag for each test file based on its models' accuracy. Table 5.1-5.10 also summarizes the average absolute error $|N_{predicted}-N_{real}|$ for every hashtag predicting the popularity for each test file, and the most likely hashtag is marked YELLOW for each test file.

Overall, the average prediction error is small for data in period 1 and period 3, while it becomes large for testing data in period 2 because the number of tweets boosted in this period and the pattern is not easy to catch. Whereas the models are good at predicting the general trend of popularity except for capturing a sudden burst in number. In short, the performance is decent for period 1 and period 3, and it needs to be improved for period 2.

Results for sample1_period1.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	82	68	94	171	178	N/A
#gohawks	984	523	-674	441	509	545
#gopatriots	1065	245	2028	293	562	720
#nfl	187	60	129	55	116	65
#patriots	166	47	913	87	42	229
#sb49	116	62	59	68	91	53
#superbowl	130	133	-28	118	176	58

Table 5.1: results for sample1_period1.txt

Results for sample2_period2.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	9361	10374	20066	81958	82923	N/A
#gohawks	9556	7108	5259	8134	23481	30307
#gopatriots	11057	22039	11463	16501	127254	26350
#nfl	-2627	-4502	-1760	28557	49226	27158
#patriots	19509	-5615	35296	-75517	-302398	116833
#sb49	25415	36347	38792	47670	20888	31415
#superbowl	8876	6752	14312	30000	-50998	39148

Table 5.2: results for sample2_period2.txt

Results for sample3_period3.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	550	610	888	616	523	N/A
#gohawks	-2713	-2930	-11540	-7680	-6977	7005
#gopatriots	40	446	-1574	-435	21	938
#nfl	404	488	449	770	491	179
#patriots	137	605	539	859	443	218
#sb49	277	260	473	554	459	233
#superbowl	336	421	491	676	510	175

Table 5.3: results for sample3_period3.txt

Results for sample4_period1.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	257	236	266	267	201	N/A
#gohawks	483	646	527	227	372	222
#gopatriots	1770	908	1073	799	981	861
#nfl	360	186	235	217	238	54
#patriots	605	233	51	-1	46	198
#sb49	422	245	174	236	228	65
#superbowl	277	204	167	141	153	65

Table 5.4: results for sample4_period1.txt

Results for sample5_period1.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	508	353	362	281	213	N/A
#gohawks	93	1007	487	-225	326	362.6
#gopatriots	1347	2029	1322	1279	903	1032.6
#nfl	325	451	317	386	229	89.4
#patriots	134	65	-26	-221	-10	355.0
#sb49	272	289	250	262	201	88.6
#superbowl	202	472	208	283	190	120.8

Table 5.5: results for sample5_period1.txt

Results for sample6_period2.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	12931	60619	52699	41019	37307	N/A
#gohawks	4138	87409	498725	430014	323143	231288
#gopatriots	18213	56756	331857	293122	207543	142128
#nfl	-1533	85013	520126	457576	348945	246896
#patriots	-65636	31823	43922	-80006	21542	50586
#sb49	8376	14797	37900	22715	23428	19472
#superbowl	7570	283917	1670667	1416735	1063619	849731

Table 5.6: results for sample6_period2.txt

Results for sample7_period3.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	102	66	60	55	120	N/A
#gohawks	46	-11	-111	12	33	87
#gopatriots	308	258	161	146	120	118
#nfl	122	96	66	58	50	26
#patriots	101	90	73	61	50	23
#sb49	123	101	69	58	43	29
#superbowl	131	112	74	63	54	33

Table 5.7: results for sample7_period3.txt

Results for sample8_period1.txt					
Number of tweets	hour2	hour3	hour4	hour5	Average Prediction Error
Real value	72	56	41	11	N/A
#gohawks	-30	-124	-61	-100	124
#gopatriots	181	265	204	152	156
#nfl	46	72	54	43	22
#patriots	-39	-81	-56	-57	103
#sb49	22	33	24	17	24
#superbowl	19	13	17	2	32

Table 5.8: results for sample8_period1.txt

Results for sample9_period2.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	1734	1619	1582	1857	2790	N/A
#gohawks	12191	12831	11004	10836	15119	10480
#gopatriots	9595	9995	8820	7909	47650	14877
#nfl	7925	6571	4345	6161	4999	4084
#patriots	12752	17106	19012	13472	-16976	15063
#sb49	1764	2709	3569	2674	5906	1408
#superbowl	35984	34618	30046	29941	40446	32291

Table 5.9: results for sample9_period2.txt

Results for sample10_period3.txt						
Number of tweets	hour2	hour3	hour4	hour5	hour6	Average Prediction Error
Real value	54	68	62	58	61	N/A
#gohawks	28	10	4	40	40	36
#gopatriots	169	146	209	168	163	110
#nfl	68	54	66	56	54	8
#patriots	75	52	66	56	55	10
#sb49	72	63	82	60	59	9
#superbowl	65	62	81	69	65	10

Table 5.10: results for sample10_period3.txt

ii) Fan Base Prediction

6) Prediction of Location

This tweet dataset contains a lot of information that can be used to explore user's' personal habits, preference and location. In this question, we want to use the text of tweet from #Superbowl to predict the user is coming either from Washington State or Massachusetts State. The first step here is to extract all textual content of the tweets with the hashtag #Superbowl whose authors are either from Washington or Massachusetts State. The number of tweets from this two states is plotted as a bin chart in Figure 6.1.

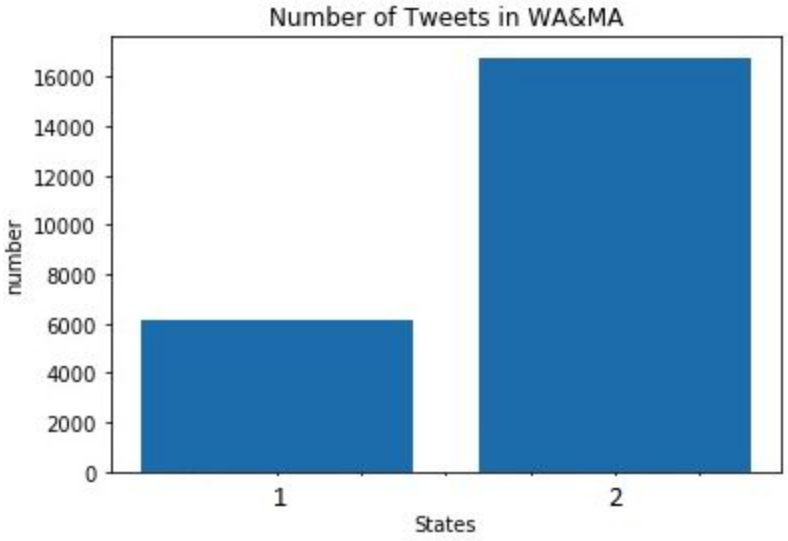

Figure 6.1: Number of tweets from WA and MA

In the figure 6.1, the left bin represents the number of tweets contains #superbowl from Washington State and right one is from Massachusetts States. It is obvious that tweets from Massachusetts from is much more than that from Washington. In this case, we managed to balance the number of data by cutting the tweets from Massachusetts down to 6111 tweets for better classifier performance. We randomly picked 5000 tweets from each of States and they were combined to generate the training set. Meanwhile, the remaining 1111 tweets from both location are combined to produce the testing set.

To convert the textual content into manipulatable data, we did the same thing as done in Project 2 to first vectorize the data into a Term Frequency-Inverse Document Frequency (TF-IDF) matrix. Punctuations and stop words were excluded from the analysis, and different stems of a word are considered as only one word. After obtaining the TF-IDF matrix, we also perform

Latent Semantic Indexing to reduce the dimension of the matrix for better performance. We therefore projected the original TF-IDF matrix onto a 50 dimensional space using the first 50 document eigenvectors from singular value decomposition. The 50-dimensional data was then fed into a classifier to yield the classification results/

We first tried a Support Vector Machine (SVM) classifier. The Received Operating Chacteristic Curve (ROC), confusion matrix and precision-recall metrics are shown in Figure 6.2. The accuracy of SVM:

Accuracy = 0.6827182718271827

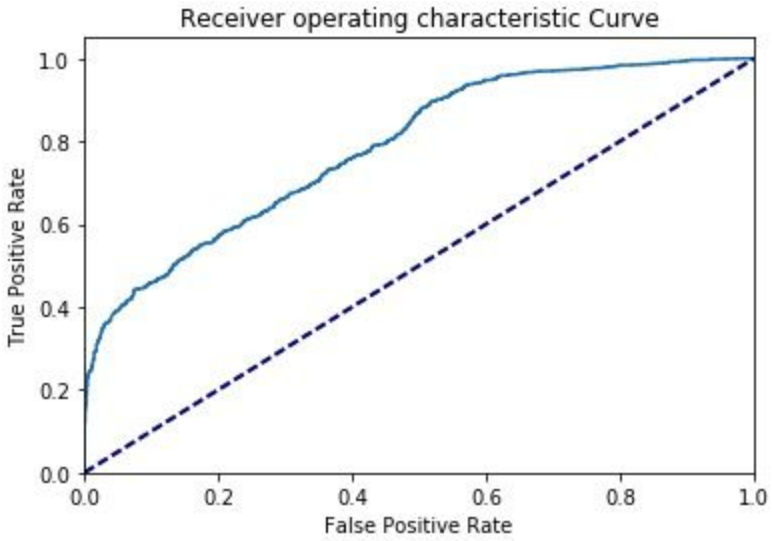

Figure 6.2(a): ROC of SVM

	precision	recall	f1-score	support
Washington	0.64	0.85	0.73	1110
Massachusetts	0.78	0.52	0.62	1112
avg / total	0.71	0.68	0.68	2222

Figure 6.2(b): Metrics of SVM report

```
[945 165]
[536 576]
```

Figure 6.2(c): Confusion Matrix of SVM
First row/column: Washington, second row/column: Massachusetts

From the figure 6.2, it can be easily concluded that SVM does not have a good performance. From the confusion matrix, it is obvious that it could successfully predict most part of tweets from Washington State but failed to predict tweets from Massachusetts.

The second algorithm we adopted is a logistic regression classifier. Its ROC, confusion matrix and metrics are shown in Figure 6.3. The accuracy of Logistic Regression:

Accuracy = 0.6971197119711972

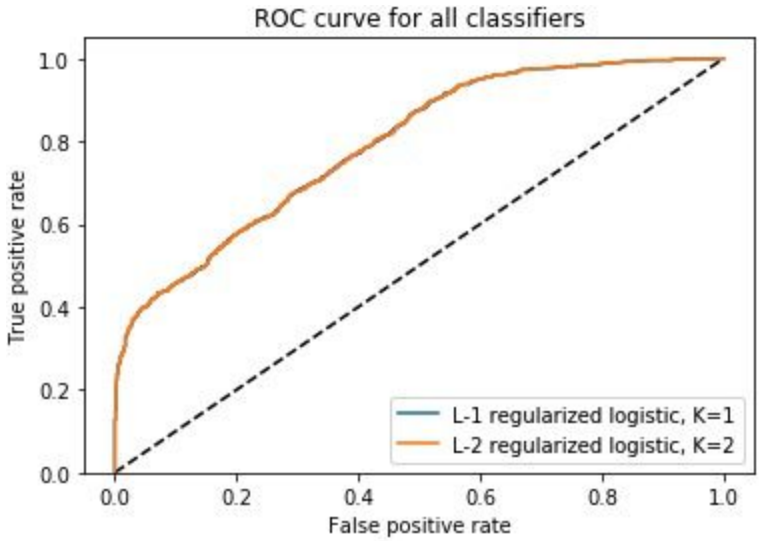

Figure 6.3(a): ROC of Logistic Regression

	precision	recall	f1-score	support
Washington	0.69	0.66	0.68	1110
Massachusetts	0.68	0.71	0.69	1112
avg / total	0.69	0.68	0.68	2222

Figure 6.3(b): Metrics of Logistic Regression

```
[733 377]
[323 789]
```

Figure 6.3(c): Confusion Matrix Logistic Regression
First row/column: Washington, second row/column: Massachusetts

From the figure 6.3, we can say that Logistic Regression is a better classifier than SVM since it has a much higher correctness on predicting tweets from Massachusetts.

The last algorithms is Multinomial Naive Bayes Algorithm. Its ROC, confusion matrix and metrics are shown in Figure 6.4.The accuracy of Naive Bayes:

Accuracy = 0.72997299729973

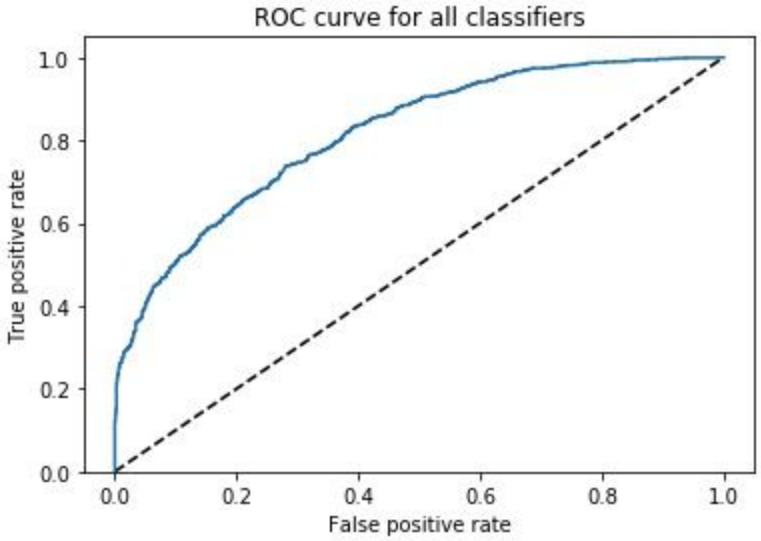

Figure 6.4(a): ROC of Multinomial Naive Bayes

	precision	recall	f1-score	support
Washington	0.73	0.71	0.72	1110
Massachusetts	0.72	0.74	0.73	1112
avg / total	0.73	0.73	0.73	2222

Figure 6.4(b): Metrics of Multinomial Naive Bayes

```
[789 321]
[289 823]
```

Figure 6.4(c): Confusion Matrix of Multinomial Naive Bayes.
First row/column: Washington, second row/column: Massachusetts

From the confusion matrix of Naive Bayes method, we can say the that it has the best predict performance compared to the other two algorithms.

Finally, we plotted of the ROC curve together, and the ROC curve of all classifier is shown in Figure 6.5.

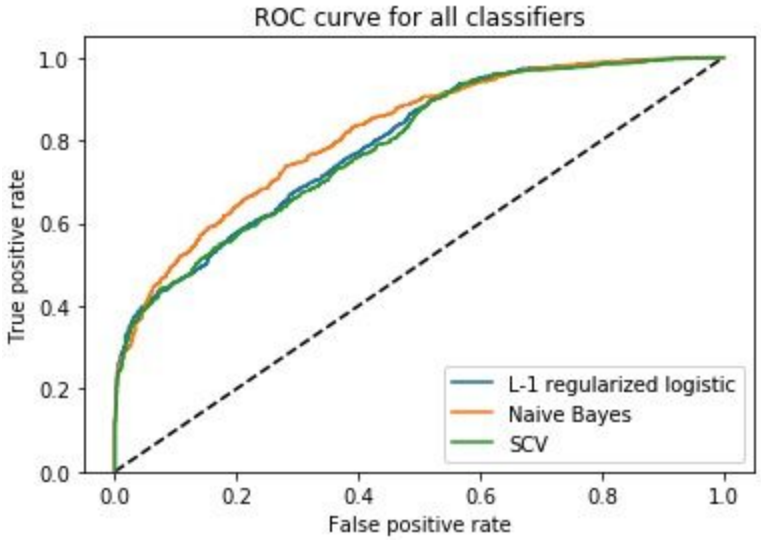

Figure 6.5: ROC of all algorithms

In the figure 6.5, all the ROC curves are plotted together. The ROC of Naive Bayes has a better True positive Rate at the low False positive Rate, and is closest to the upper left corner, indicating better performance.

In addition, the accuracy of different classifiers is compared in Table 6.1. The accuracy for Naive Bayes classifier is the highest among the three choices.

Classifier	Accuracy
SVM	0.68
Logistic Regression	0.70
Naive Bayes	0.73

Table 6.1: accuracy for different classifiers

Therefore, we can conclude that Naive Bayes classifier outperformed other two classifiers for this specific dataset.

In conclusion, the best classifier is **Naive Bayes** classifier, whose accuracy is:

Accuracy = 0.73

and the corresponding ROC curve. confusion matrix as well as precision-recall metrics have been shown in **Figure 6.4**.

iii) Define Your Own Project

7. 1 Clustering Analysis

Clustering is a powerful unsupervised learning method to group data into several clusters within which data are similar. In this part, we want to verify our classification results in part 6 using k-means clustering. The data we use is the same as that in part 6, which is the tweets with the hashtag #superbowl whose authors come from either Washington state or Massachusetts state.

We first used the same vectorizer in part 6 to convert the textual content into a TF-IDF matrix. Then, we performed Non-negative Matrix Factorization (NMF) to reduce the dimension of the features for better clustering results. Note that we did not use LSI in this part because Project 4 shows that NMF yields the best results in such analysis. We evaluated the performance by looking at 4 measures as those in Project 4: homogeneity score, completeness score, adjusted rand score, and adjusted mutual information score. Homogeneity score is a measure of how purely clusters contain only data points from a single class. On the other hand, completeness score measures how purely a single class is clustered together. Adjusted rand score computes similarity between the clusterings by considering all pairs of samples that are assigned in the same or different clusters in the predicted and true clusterings. Finally, adjusted mutual information measures mutual information between the cluster label distribution and the group truth label distribution. We also looked at the confusion matrix for evaluation.

The number of ambient components was swept from 1 to 50, and the resulted statistics are shown in Figure 7.1.1.

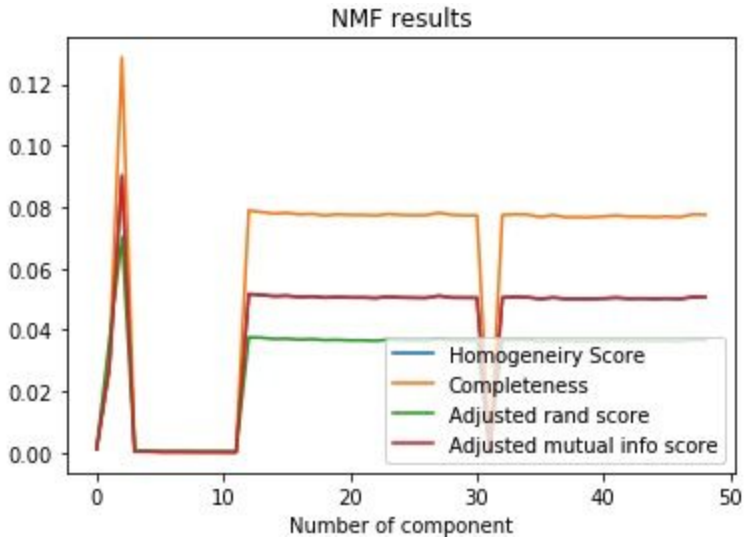

Figure 7.1.1: NMF Result

From figure 7.1.1, we observed that NMF has the best scores when n=2, and the confusion matrix is shown in figure 7.1.2.

```
[2502 2488]
[3464 1546]
```

Figure 7.1.2: NMF confusion matrix
First row/column: Washington, second row/column: Massachusetts

The best values of the measures are less than 0.12, indicating poor performance. Also, The confusion matrix shows that NMF clustering method did not offer a good result. Therefore, we want to visualize the data in a 2-D space to get some ideas on how to improve the clustering results.

Figure 7.1.3 shows the visualization of data in a 2-D space, and the colors stand for the true labels of locations. It can be observed that these two clusters almost overlap with each other, and there is more likely to be 3 natural clusters instead of 2 for this dataset. Thus, we also swept the number of clusters from 3 to 8 using K-means clustering and NMF, and the results are shown in Figure 7.1.4, where different colors represent different cluster. It can be observed that the data are not grouped well, indicating poor performance.

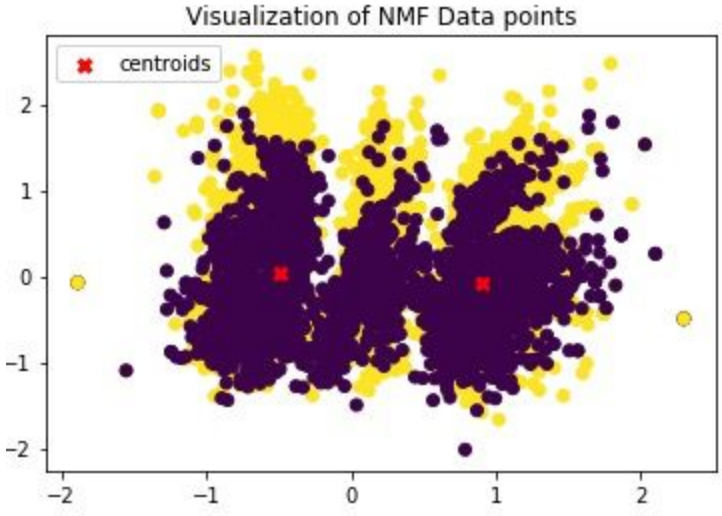

Figure 7.1.3: Visualization of NMF

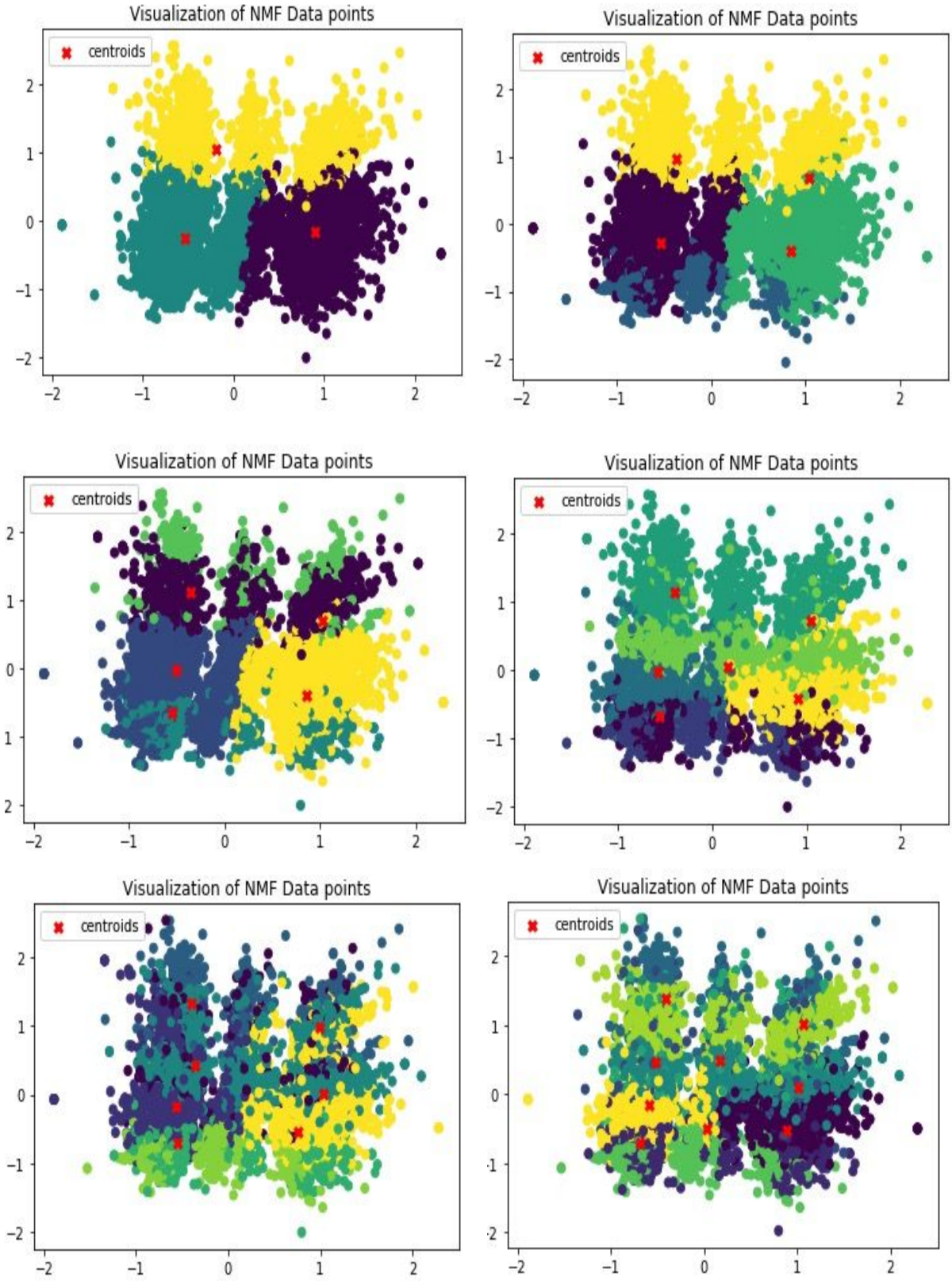

Figure 7.1.4: Visualization of NMF with different number of clusters

In conclusion, the performance of K-means clustering on the dataset is not promising for these number of clusters, and the dimension reduction method we tried cannot separate the data points. It can be due to the fact that the tweets from same location did not share common terms. Another reason can be that super bowl is an event targeting nationwide or worldwide audience such that people from different locations can have similar points of view about the game, leading to similar clustering results. Therefore, more data processing techniques or other kinds of vectorizers need to be performed in order to cluster the data into more meaningful groups.

7. 2 Time Zone Analysis

For this part, we look into the tweet dataset to find some useful information related to the users' location with respect to the time of the tweet. The user without a location is deleted and the distribution of locations of a particular hashtag is analyzed. #gopatriots is chosen as the target hashtag, and the bar plot of the most common locations is shown below in Figure 7.2.1:

Figure 7.2.1: the number of users posting tweets in different locations

From the above graph, it can be seen that Boston, Brazil and Mexico are the three most popular locations for posting tweet related to #gopatriots.

Before we calculate the average post time of users in each location to find the relations, we want to take a guess on it according to the different time zones these users are in. A time zone is a region of the globe that observes a uniform standard time for legal, commercial, and social purposes. Time zones tend to follow the boundaries of countries and their subdivisions because it is convenient for areas in close commercial or other communication to keep the same time.

From the Time Zone map in Figure 7.2.2, it's clear to see that Brazil and Boston are in similar time zones, while Mexico is about 2 time zones away from them. Therefore, the average time of post from Mexico should be about 2 hours earlier than the post time from Boston and Brazil, while Boston and Mexico's post times should be similar.

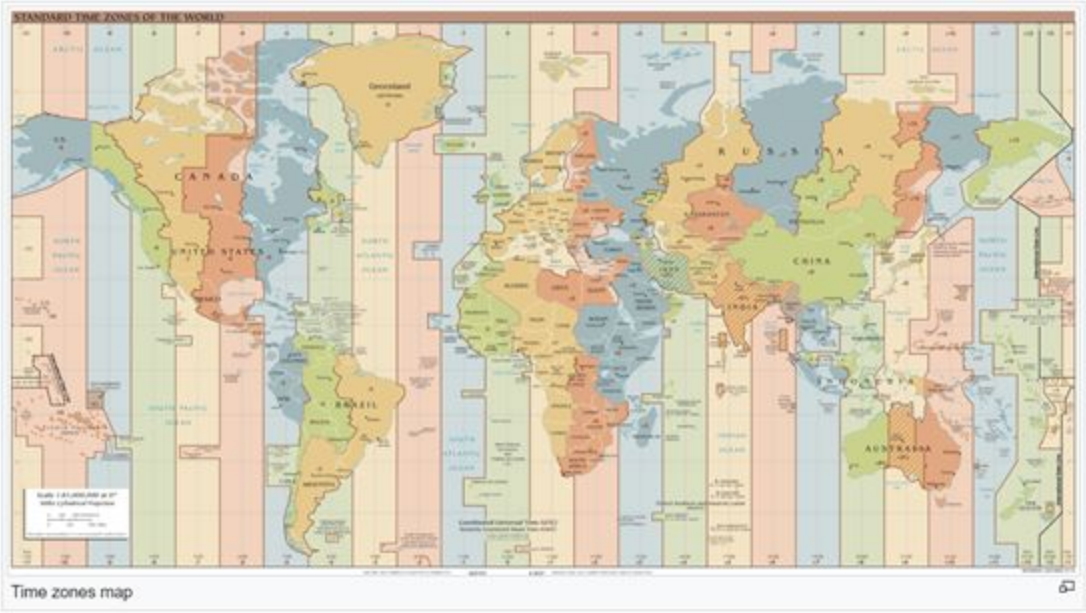

Figure 7.2.2: the Time Zone map

Then, we calculate the starting post time of each user from the three locations to see if our guess is correct, the graph is shown below in Figure 7.2.3:

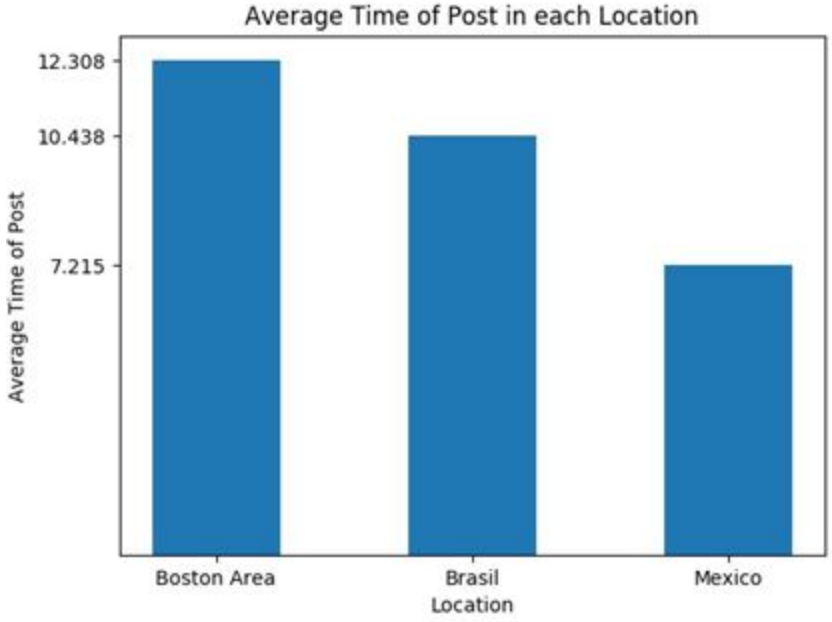

Figure 7.2.3: average time of posting tweets in different locations

It can be seen that the average post time of people from Mexico is 6am UTC, while the Boston and Brazil's is around 11-12am UTC. Therefore, the guess is basically correct and the source of error might be caused from the following aspects:

1. Many people don't have the location information in their tweet, therefore the size of samples isn't large enough to give an accurate analysis.
2. There is more than one time zone in both Brazil and Mexico, therefore the calculated average posting time is hard to be evaluated accurately.
3. The super bowl event might cause a tweet time changing for people in these 3 locations, therefore the posting time difference might not be able to follow the time zone difference directly.

Conclusion

We managed to predict the popularity of different hashtag in the next hour using the current and previous tweet activities with linear regression models. The regression models are designed with some useful features learnt from literature and our own analysis. The models were also proved to be robust with 10-fold cross-validation, and the average prediction error generally increases with the size of dataset. We also tried to split the data into three time periods and analyzed them separately. The accuracy is good for period 1 and period 3, and is bad for period 2 due to boosted number of tweets and unexpected variations. This pattern can be also observed when the models were used to predict the number of tweets for a testing dataset.

For the users posting tweets with #superbowl from Washington state and Massachusetts state, we also performed classification analysis with what we learnt in Project 2. Naive Bayes classifier produces the best performance due to the probabilistic nature of textual data.

We also defined our own project in clustering and time zone analysis. The first proposed project in clustering is not successful due to the nature of data, while the second project succeeds in analyzing the relationship between locations and their time zones from the first posting date.

Overall, our models are not perfectly and we should look for other methods to improve them in the future. Further, the twitter dataset contains tons of information, from which we can try to do various kinds of analysis. Twitter provides us with endless opportunities for data analysis.