EE219 Project 5

Popularity Prediction on Twitter

Yunchu Zhang (805030502)
Wenshan Li (105026914)
Wei Du (005024944)
Zeyu Zhang (505030513)

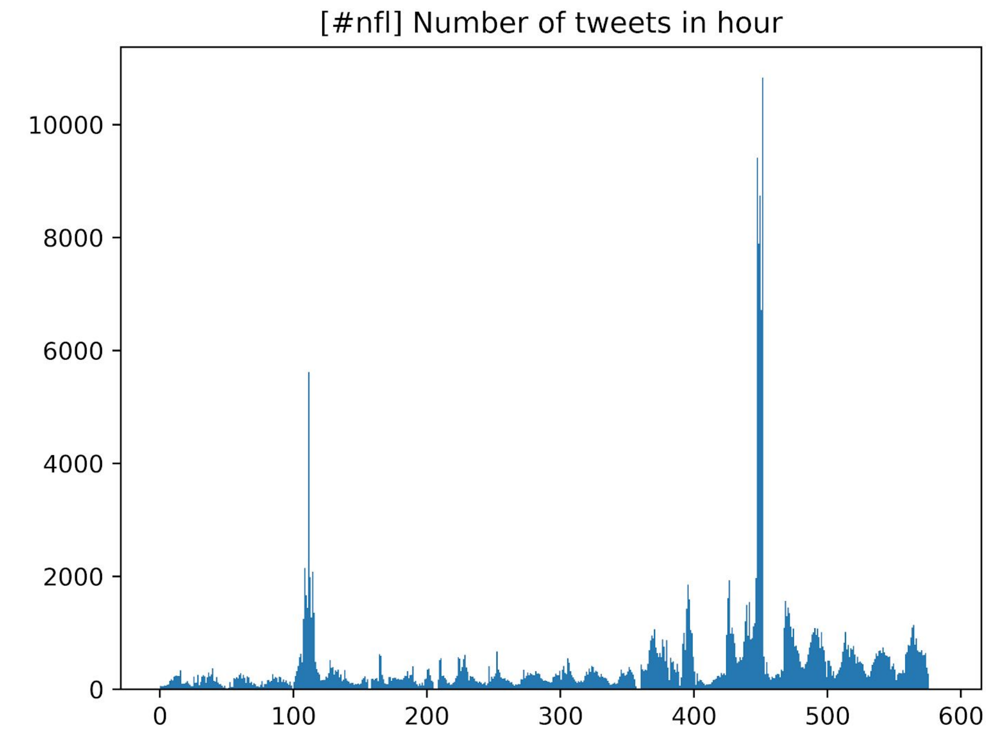
Part 1: Popularity Prediction

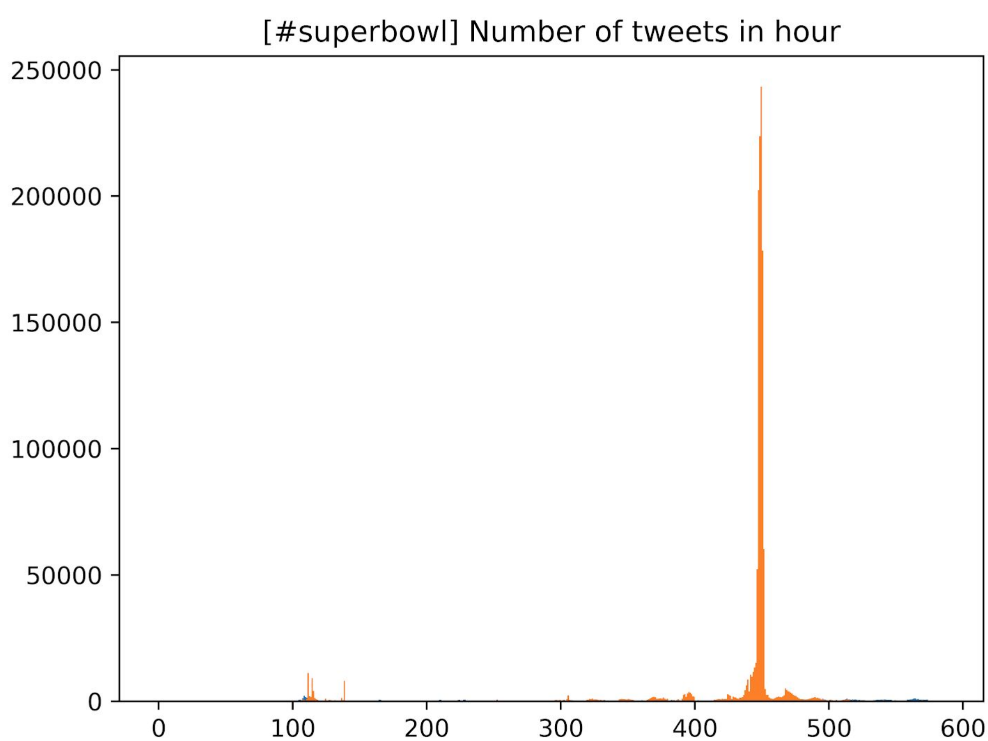
Problem 1.1

Statistics for each hashtag is shown in the Table below

hash tag	average tweets	average followers	average retweets
gohawks	324.933	2203.932	2.015
gopatriots	45.621	1401.896	1.4
nfl	441.267	4653.252	1.539
patriots	834.264	3309.979	1.783
sb49	1418.441	10267.317	2.511
superbowl	2301.65	8858.975	2.388

The histograms of “number of tweets in hour” over time for #NFL and #SuperBowl are shown below,


[#nfl] Number of tweets in hour
10000
8000
6000
4000
2000
0
0
100
200
300
400
500
600


[#superbowl] Number of tweets in hour
250000
200000
150000
100000
50000
0
0
100
200
300
400
500
600

Problem 1.2

The R-squared measurement for each hashtag is show in the Table below. Note, for the accuracy of the regression results, we decided to use the MAE (Mean Absolute Error) to measure the accuracy. MAE is an estimator of the relative quality of statistical models for a given set of data, which estimates the quality of each model.

	gohawks	**gopatriots**	**nfl**	**patriots**	**sb49**	**superbowl**
R-squared	0.519	0.611	0.646	0.716	0.844	0.869
MAE	212.427	33.268	174.138	471.231	836.2	1747.222

1) ***gohowks***

For hashtag ***#gohowks***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.3843	0.165	8.374	0.000	1.060	1.709
x2	-0.1456	0.039	-3.750	0.000	-0.222	-0.069
x3	-0.0002	8.37e-05	-2.963	0.003	-0.000	-8.35e-05
x4	0.0003	0.000	1.503	0.133	-7.84e-05	0.001
x5	6.9528	3.281	2.119	0.035	0.508	13.398

Based on the P-value, x1 (the number of tweets), x2 (the number of retweets), x3 (the sum of the number of followers) and x5 the time of the day are significant. Because their P-value is less than 0.05 by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

2) ***gopatriots***

For hashtag ***#gopatriots***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

Based on the P-value, x3 (the sum of the number of followers) and x4 (the number of maximum followers) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	−0.4179	0.264	−1.583	0.114	−0.937	0.101
x2	0.4516	0.231	1.955	0.051	−0.002	0.905
x3	0.0006	0.000	3.186	0.002	0.000	0.001
x4	−0.0007	0.000	−3.758	0.000	−0.001	−0.000
x5	0.9281	0.728	1.275	0.203	−0.502	2.358

3) ***nfl***

For hashtag ***#nfl***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.7608	0.135	5.618	0.000	0.495	1.027
x2	−0.1736	0.066	−2.633	0.009	−0.303	−0.044
x3	7.184e−05	2.62e−05	2.741	0.006	2.04e−05	0.000
x4	−6.813e−05	3.59e−05	−1.896	0.058	−0.000	2.45e−06
x5	7.4704	2.207	3.385	0.001	3.136	11.804

Based on the P-value, x1 (the number of tweets), x2 (the number of retweets), x3 (the sum of the number of followers) and x5 the time of the day are significant. Because their P-value is less than 0.05 by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

4) ***patriots***

For hashtag ***#patriots***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.2145	0.079	15.377	0.000	1.059	1.370
x2	-0.3371	0.068	-4.925	0.000	-0.472	-0.203
x3	3.479e-05	2.63e-05	1.325	0.186	-1.68e-05	8.64e-05
x4	0.0002	9.48e-05	1.682	0.093	-2.67e-05	0.000
x5	7.1287	8.269	0.862	0.389	-9.111	23.369

Based on the P-value, x1 (the number of tweets) and x2 (the number of retweets) are significant. Because their P-value is less than 0.05 by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

5) ***sb49***

For hashtag ***#sb49***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.2883	0.095	13.511	0.000	1.101	1.476
x2	-0.2949	0.087	-3.371	0.001	-0.467	-0.123
x3	2.865e-05	1.38e-05	2.069	0.039	1.46e-06	5.58e-05
x4	0.0002	4.26e-05	4.240	0.000	9.69e-05	0.000
x5	-17.1909	14.143	-1.215	0.225	-44.970	10.588

Based on the P-value, x1 (the number of tweets), x2 (the number of retweets), x3 (the sum of the number of followers) and x4 (the number of maximum followers) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

6) ***superbowl***

For hashtag ***#superbowl***, the measurement is shown below, where x1 represents the number of tweets, x2 represents the total number of retweets, x3 represents the sum of the number of followers of the users posting the hashtag, x4 represents maximum number of followers of the users posting the hashtag, x5 the time of the day. The t-test is shown in column t, and the P-valua is shown in column P>|t|.

	coef	std err	t	P>|t|	[0.025	0.975]
x1	2.5477	0.107	23.752	0.000	2.337	2.758
x2	-0.1549	0.035	-4.390	0.000	-0.224	-0.086
x3	-0.0002	1.08e-05	-20.224	0.000	-0.000	-0.000
x4	0.0011	0.000	10.441	0.000	0.001	0.001
x5	-56.3970	24.168	-2.334	0.020	-103.864	-8.930

Based on the P-value, x1 (the number of tweets), x2 (the number of retweets), x3 (the sum of the number of followers), x4 (the number of maximum followers) and x5 the time of the day are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected. Therefore, there should have strong relationships between these features and number of tweets in next hour.

Problem 1.3

In order to predict the number of tweets for a specific hashtag, I prefer to use some novel time series features. Because the number of tweets according to a specific hashtag is strongly associated time sequences, thereby time series features could be more effective.
This this problem, we chose 5 new features that is: 1) the total number of favourite count of tweets; 2) the total number of friends of the users posting the hashtag; 3) the total ranking score metric of tweets according to a specific hashtag; 4) the total influential metric of the author posting the hashtag; 5) the total impression metric of a tweets with a specific hashtag.

The intuition behind this is that the more influential of a author posting a hashtag, the more likely others will post a same hashtag.

The R-squared measurement for each hashtag is show in the Table below. Note, for the accuracy of the regression results, we decided to use the MAE (Mean Absolute Error) to measure the accuracy. MAE is an estimator of the relative quality of statistical models for a given set of data, which estimates the quality of each model.

	gohawks	**gopatriots**	**nfl**	**patriots**	**sb49**	**superbowl**
R-squared	0.587	0.694	0.751	0.729	0.840	0.845
MAE	188.068	35.774	168.896	552.841	608.22	1503.724

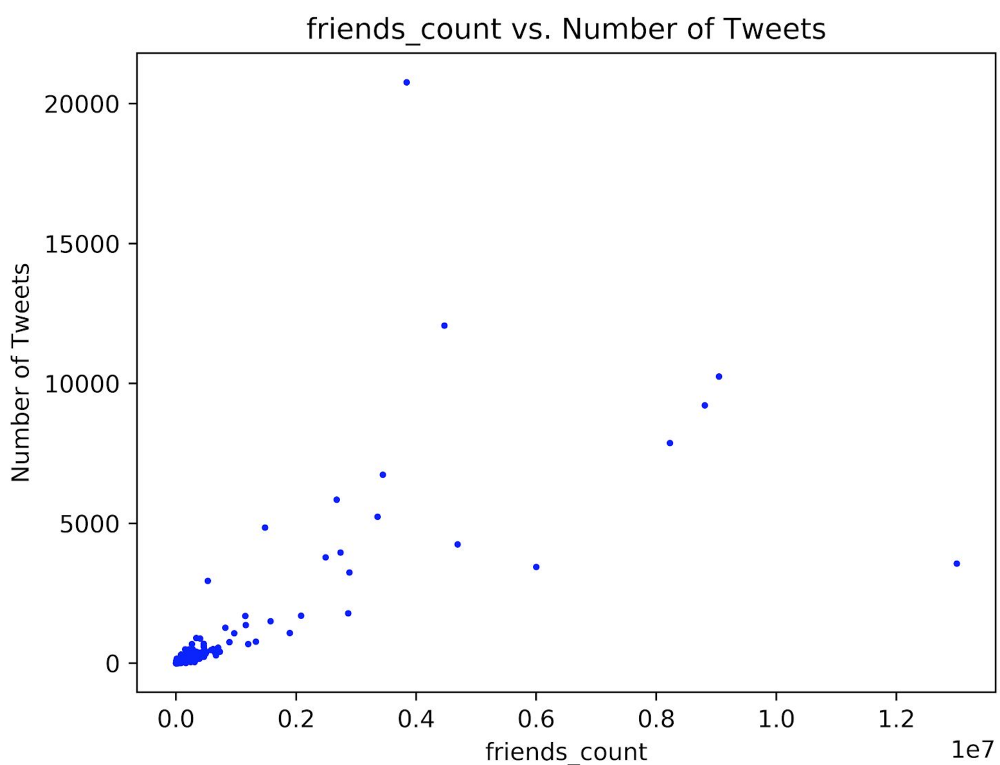
1) ***gohowks***

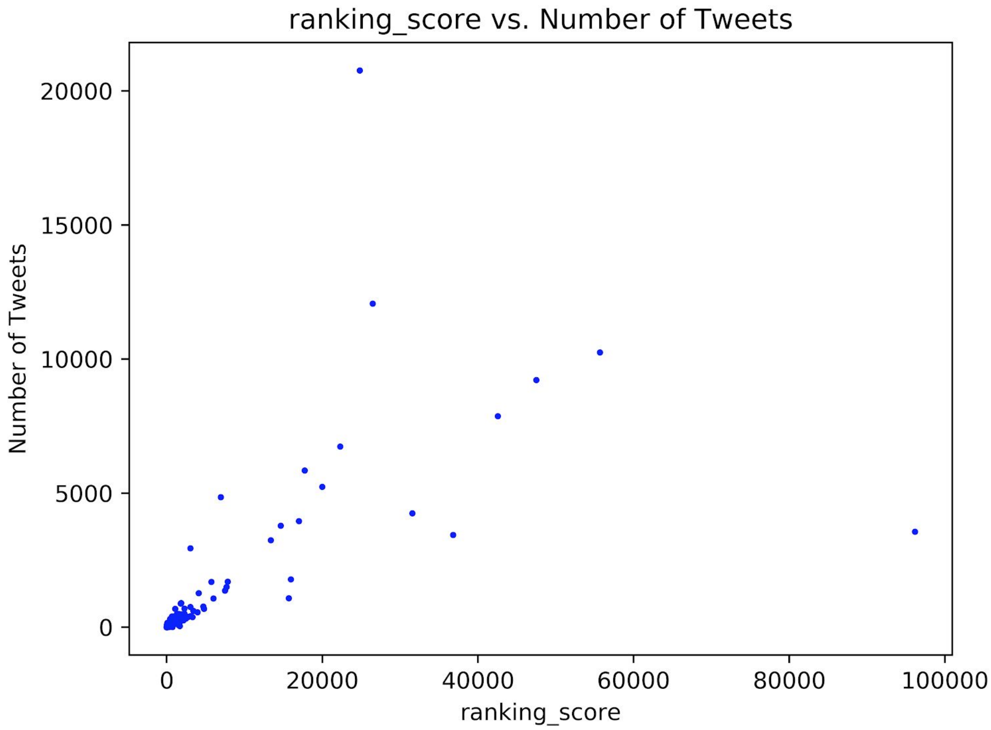
For hashtag ***#gohowks***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

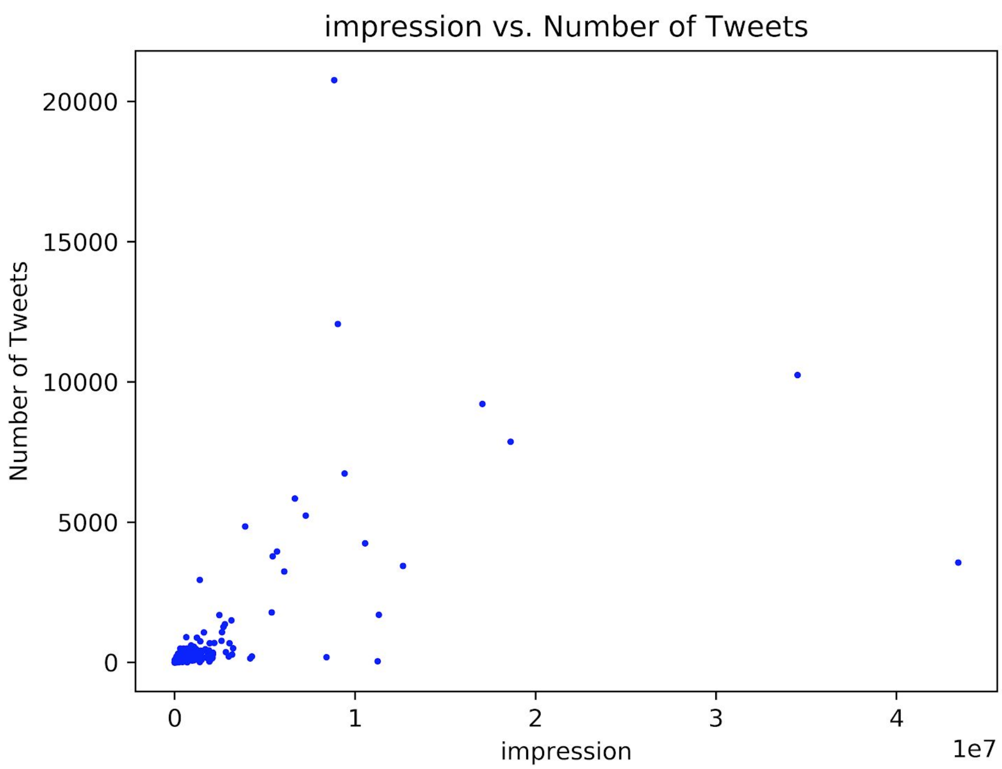
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.0262	0.026	-1.006	0.315	-0.077	0.025
x2	0.0027	0.000	11.106	0.000	0.002	0.003
x3	-0.2081	0.040	-5.167	0.000	-0.287	-0.129
x4	-0.5435	3.884	-0.140	0.889	-8.172	7.085
x5	-0.0001	4.07e-05	-3.260	0.001	-0.000	-5.27e-05

Based on the P-value, x2 (the number of friends), x3 (the sum of the ranking score) and x5 (the sum of the impression score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.




ranking_score vs. Number of Tweets
Number of Tweets
ranking_score
0
5000
10000
15000
20000
0
20000
40000
60000
80000
100000


impression vs. Number of Tweets
Number of Tweets
impression
0
5000
10000
15000
20000
0
1
2
3
4
1e7

In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

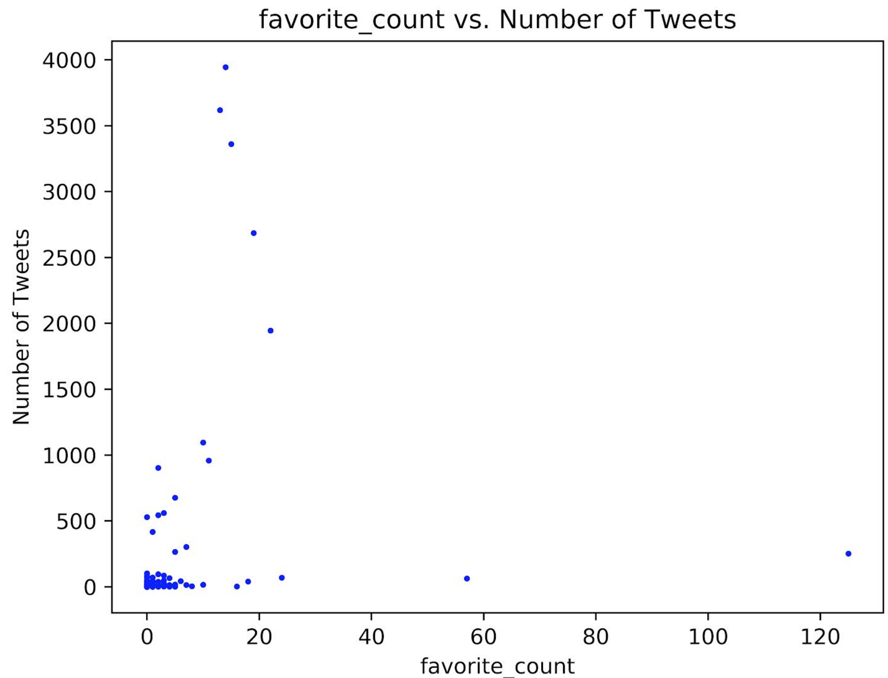
2) ***gopatriots***

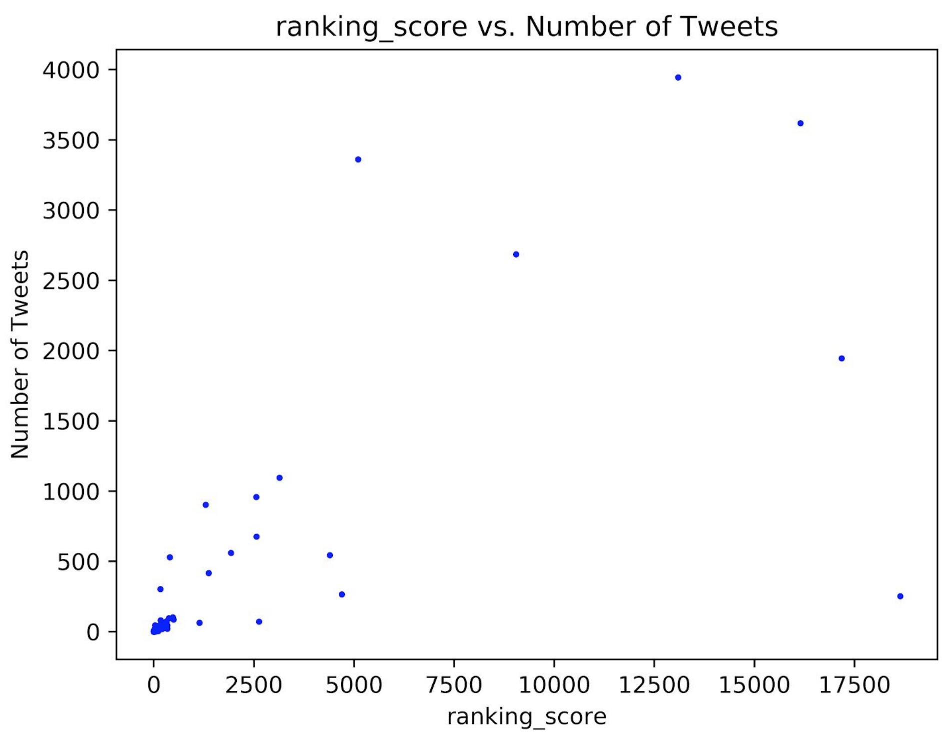
For hashtag ***#gopatriots***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

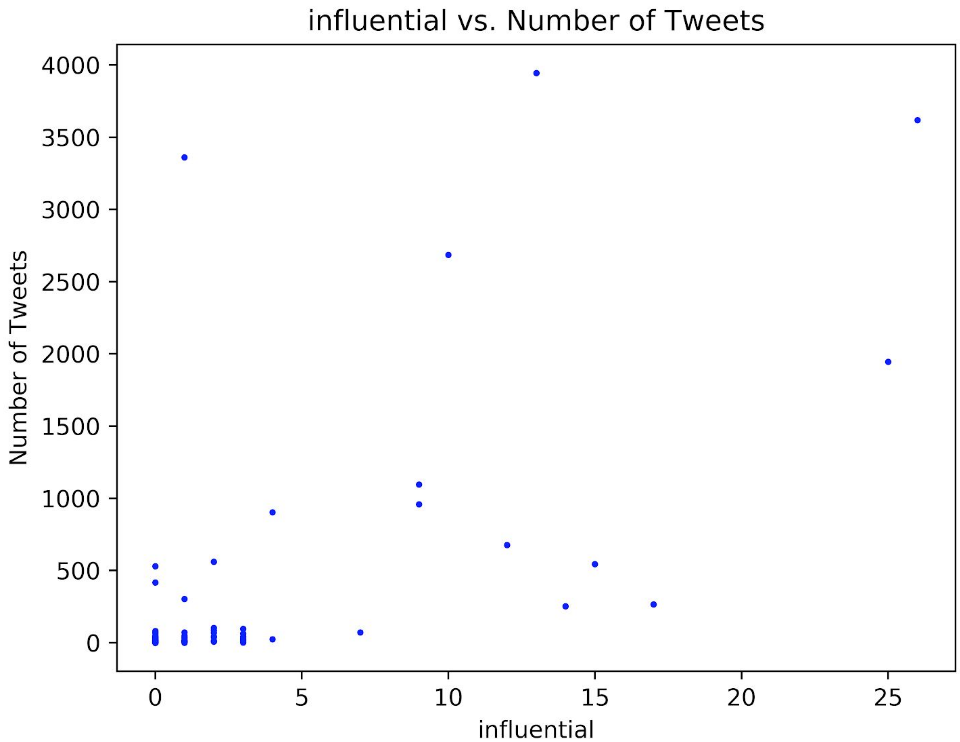
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-21.3333	1.659	-12.857	0.000	-24.592	-18.074
x2	0.0008	0.000	2.180	0.030	8.15e-05	0.002
x3	0.1546	0.044	3.543	0.000	0.069	0.240
x4	-22.2900	7.791	-2.861	0.004	-37.592	-6.988
x5	-1.883e-05	4.92e-05	-0.383	0.702	-0.000	7.78e-05

Based on the P-value, x1 (the total number of favorite count), x2 (the total number of friends count), x3 (the sum of the value of ranking score) and x4 (the sum of the value of the influential score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.


favorite_count vs. Number of Tweets
Number of Tweets
0
500
1000
1500
2000
2500
3000
3500
4000
0
20
40
60
80
100
120
favorite_count


ranking_score vs. Number of Tweets
Number of Tweets
0
500
1000
1500
2000
2500
3000
3500
4000
0
2500
5000
7500
10000
12500
15000
17500
ranking_score



In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

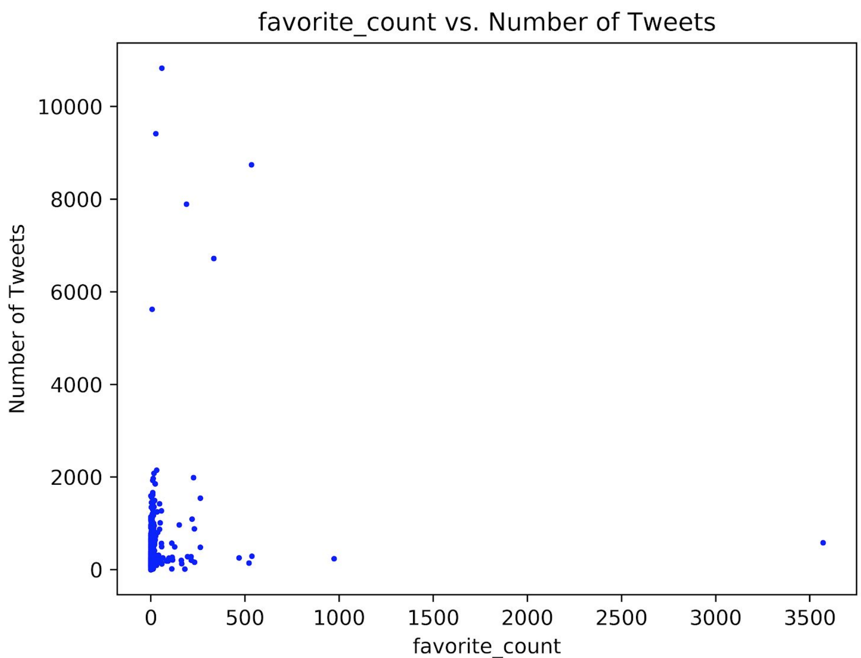
3) ***nfl***

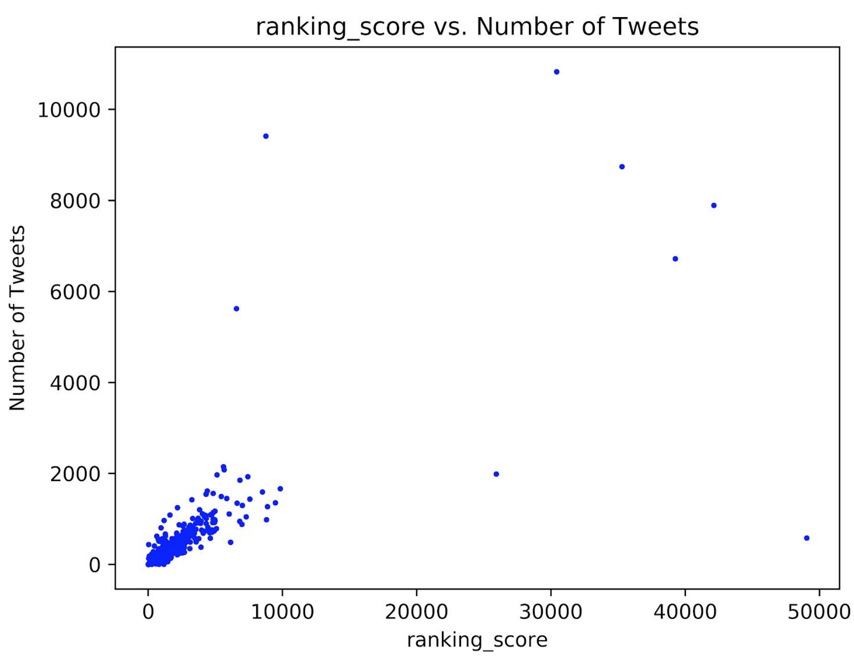
For hashtag ***#nfl***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

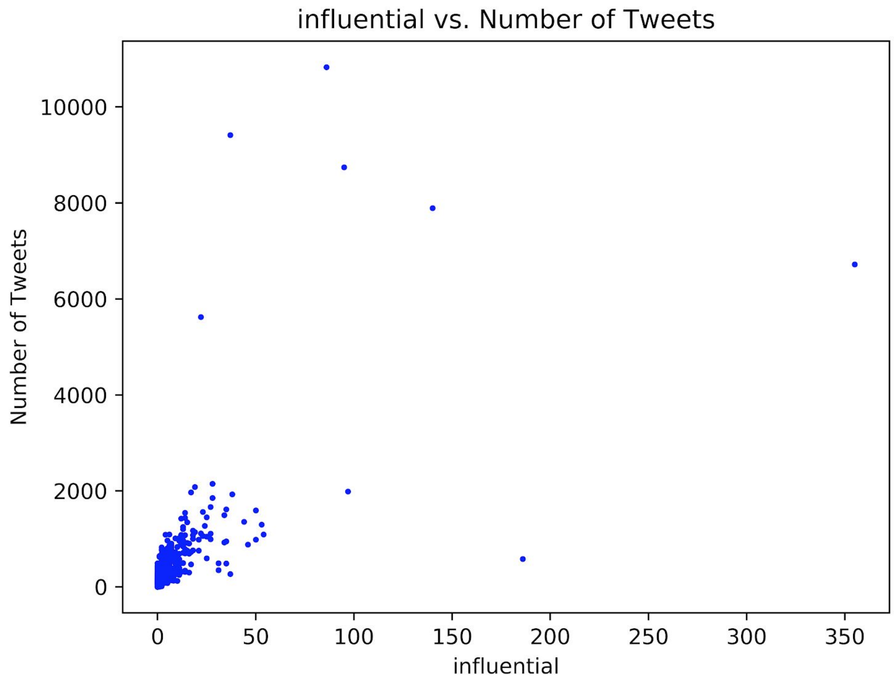
	coef	std err	t	P>|t|	[0.025	0.975]
x1	−2.5195	0.160	−15.722	0.000	−2.834	−2.205
x2	−0.0001	0.000	−1.022	0.307	−0.000	0.000
x3	0.2689	0.030	8.823	0.000	0.209	0.329
x4	−8.0972	2.189	−3.698	0.000	−12.397	−3.797
x5	2.266e−05	1.23e−05	1.843	0.066	−1.49e−06	4.68e−05

Based on the P-value, x1 (the total number of favorite count), x3 (the sum of the value of ranking score) and x4 (the sum of the value of the influential score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.







In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

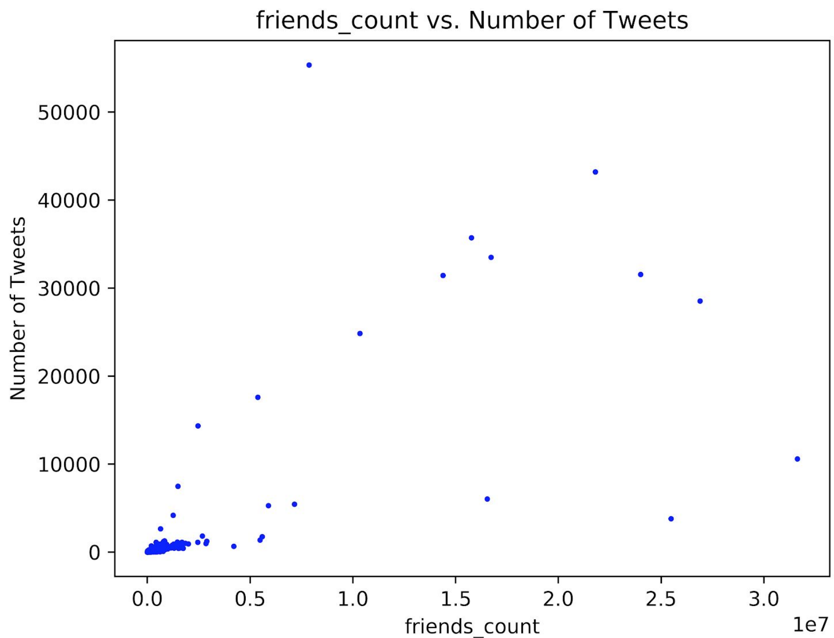
4) ***patriots***

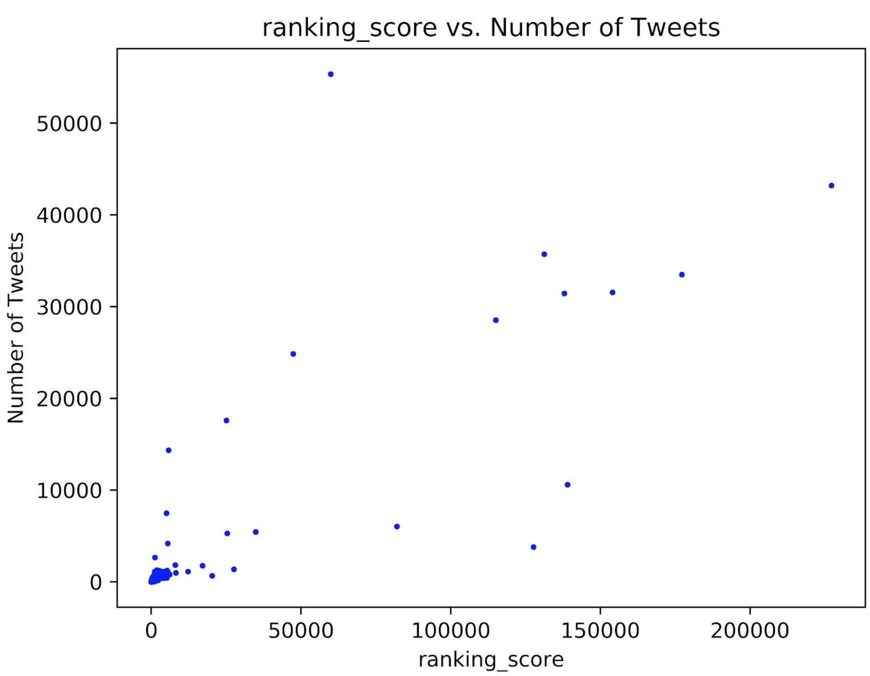
For hashtag ***#patriots***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

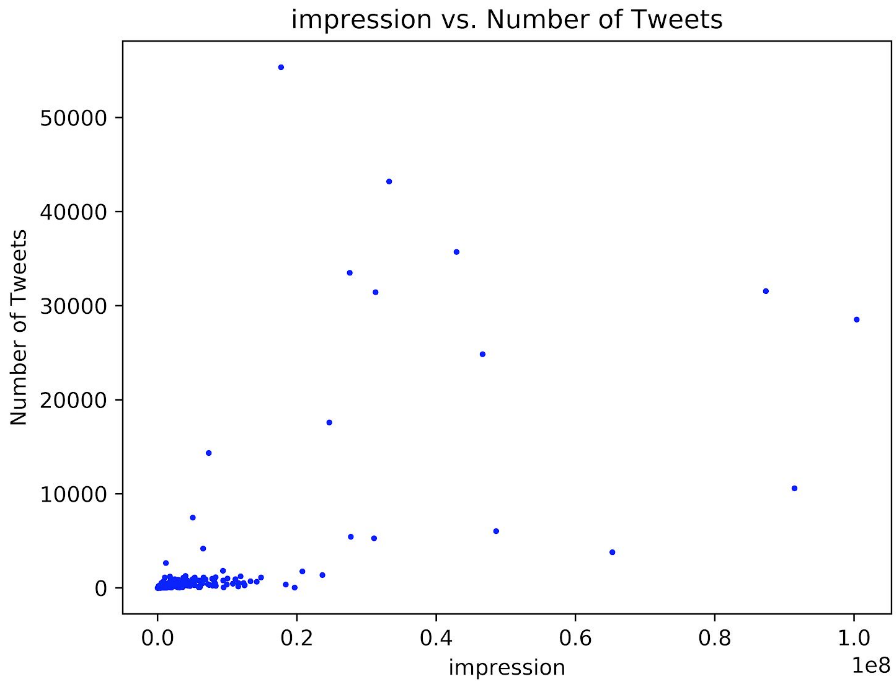
	coef	std err	t	P>|t|	[0.025	0.975]
x1	−0.1112	0.199	−0.559	0.576	−0.502	0.280
x2	−0.0015	0.000	−6.621	0.000	−0.002	−0.001
x3	0.3135	0.019	16.559	0.000	0.276	0.351
x4	−7.7581	4.648	−1.669	0.096	−16.888	1.372
x5	0.0003	3.98e−05	6.828	0.000	0.000	0.000

Based on the P-value, x2 (the total number of friends count), x3 (the sum of the value of ranking score) and x5 (the sum of the value of the impression score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.







In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

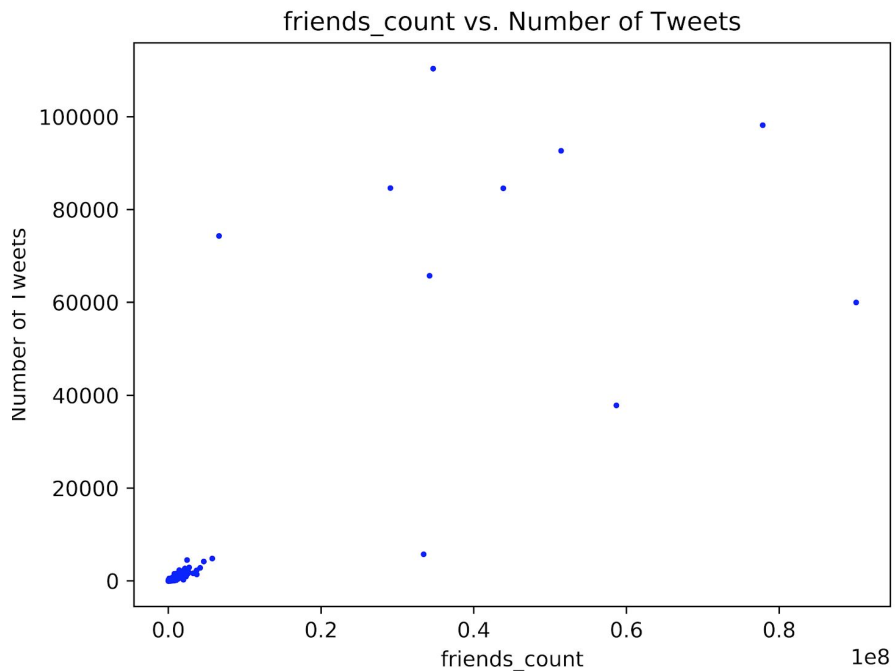
5) ***sb49***

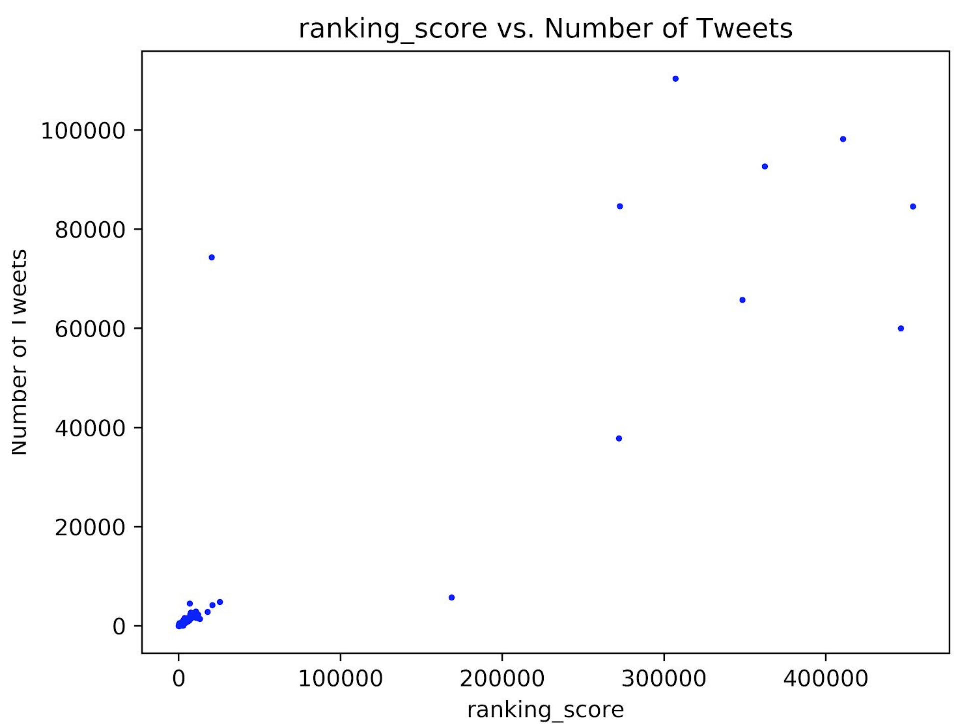
For hashtag ***#sb49***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

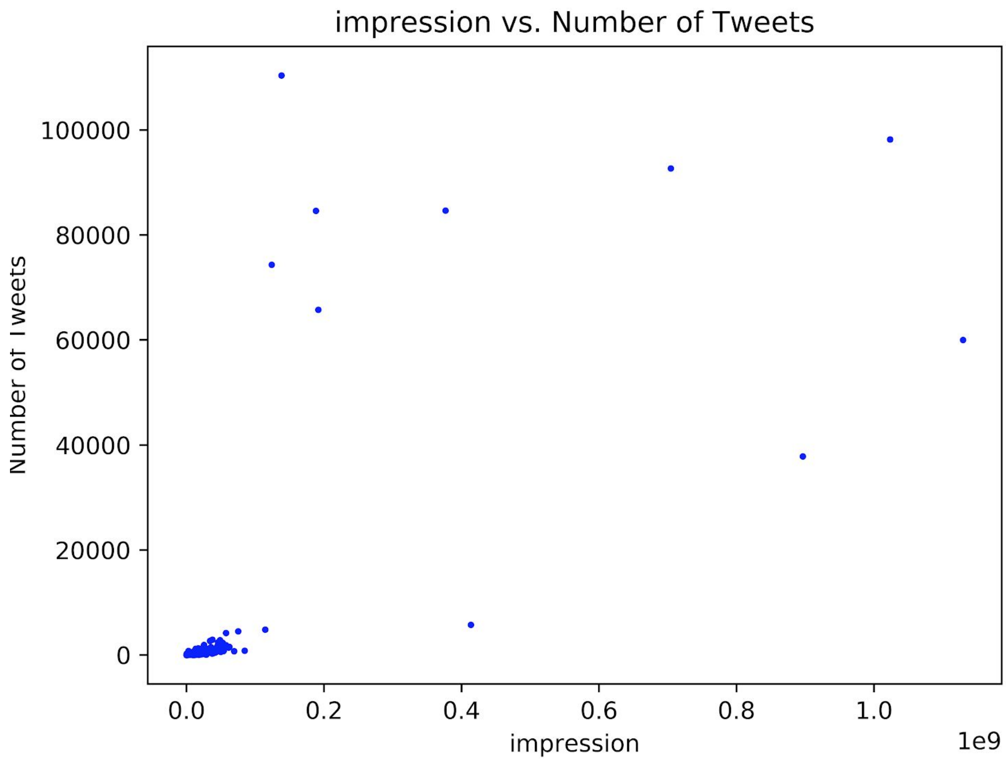
	coef	std err	t	P>|t|	[0.025	0.975]
x1	−0.1778	0.084	−2.113	0.035	−0.343	−0.013
x2	−0.0014	0.000	−3.563	0.000	−0.002	−0.001
x3	0.3499	0.033	10.535	0.000	0.285	0.415
x4	−3.3058	6.179	−0.535	0.593	−15.443	8.831
x5	5.482e−05	1.62e−05	3.389	0.001	2.3e−05	8.66e−05

Based on the P-value, x1 (the total number of favourite counts), x2 (the total number of friends count), x3 (the sum of the value of ranking score) and x5 (the sum of the value of the impression score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.







In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

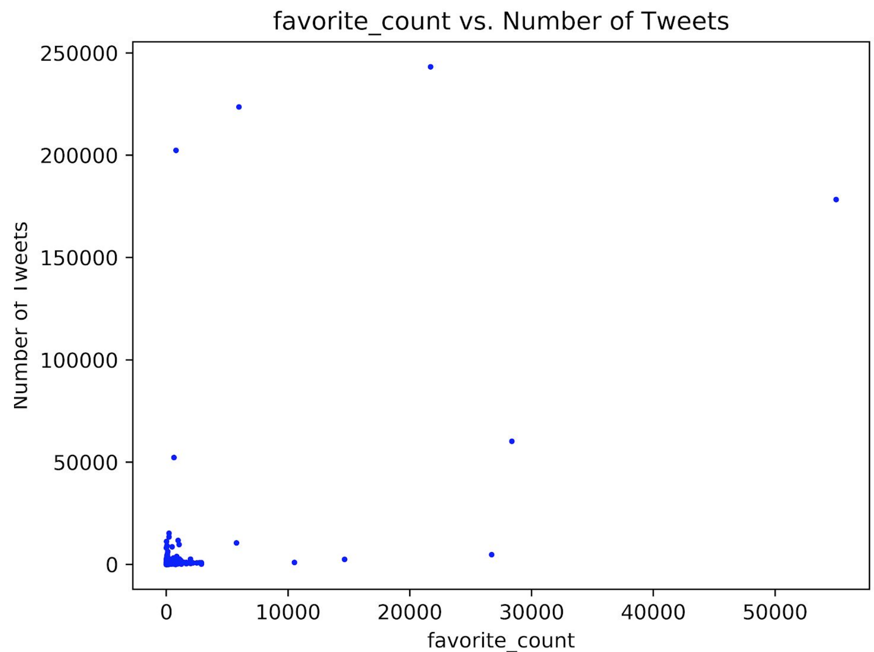
6) ***superbowl***

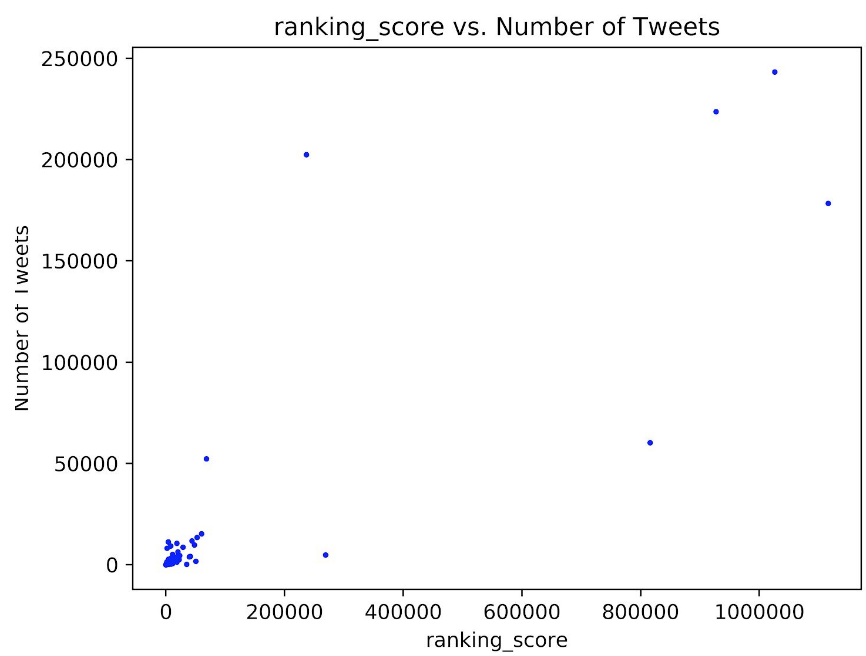
For hashtag #***superbowl***, the measurement is shown below, where x1 represents the total number of favourite count of tweets; x2 represents the total number of friends of the users posting the hashtag; x3 represents the total ranking score metric of tweets according to a specific hashtag; x4 represents the total influential metric of the author posting the hashtag; x5 represents the total impression metric of a tweets with a specific hashtag.

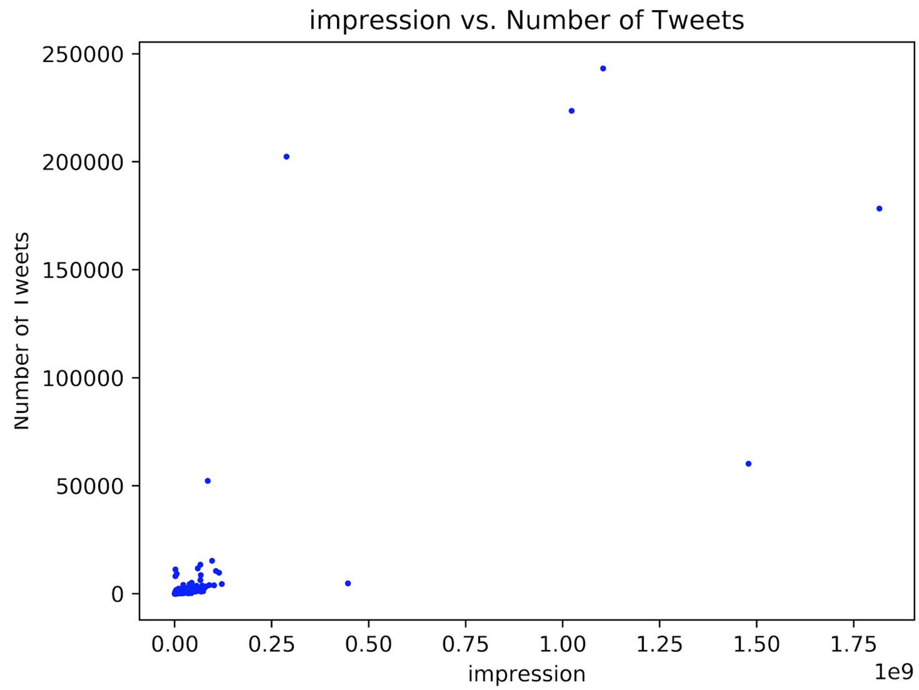
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-1.3029	0.260	-5.019	0.000	-1.813	-0.793
x2	-5.279e-05	0.000	-0.125	0.900	-0.001	0.001
x3	0.3889	0.076	5.141	0.000	0.240	0.537
x4	0.8270	3.472	0.238	0.812	-5.992	7.646
x5	-0.0001	1.55e-05	-7.010	0.000	-0.000	-7.82e-05

Based on the P-value, x1 (the total number of favourite counts), x3 (the sum of the value of ranking score) and x5 (the sum of the value of the impression score) are significant. Because their P-value is less than 0.05, by which we can say null hypothesis is rejected.

The scatter plot of predicted value (number of tweets for next hour) versus value of the top 3 feature measurement is shown below.







In the scatter plots above, the curve of predictant number versus the top 3 feature belong to a relatively linear relationship in most points. Also, there are few noise points in the plots which may not be considered in the whole plot.

Problem 1.4

Model: Ordinary Least Square (OLS)

Hashtag	Before Feb. 1, 8:00 a.m.	Feb. 1, 8:00 a.m. to 8:00 p.m.	After Feb. 1, 8:00 p.m.
gohawks	280.83	4652.999	1048.967
gopatriots	13.031	2826.796	1.904
nfl	127.61	8071.069	115.86
patriots	310.002	35004.707	175.293
sb49	877.285	24805.754	93.162
superbowl	281.993	159814.591	349.852

Here, we adopt mean absolute difference as our error measurement. The green part is the average cross-validation errors for the 3 different models with each individual hashtags. The yellow part is the average cross-validation errors for the 3 different models of the combined model.

Model: sklearn Linear Model Stochastic Gradient Descent (SGD)

Hashtag	Before Feb. 1, 8:00 a.m.	Feb. 1, 8:00 a.m. to 8:00 p.m.	After Feb. 1, 8:00 p.m.
gohawks	279.655	7152.49	325.741
gopatriots	13.802	2235.555	2.073
nfl	127.165	7746.597	118.199
patriots	317.033	25542.269	158.098
sb49	1013.384	26080.957	94.268
superbowl	304.68	177862.459	338.518

This model is a scikit-learn linear regression model, which tries to maximize the log likelihood function. Here, we adopt mean absolute difference as our error measurement. The green parts the average cross-validation errors for the 3 different models with each individual hashtags.

Model: Multi-layer Neural Network

Hashtag	Before Feb. 1, 8:00 a.m.	Feb. 1, 8:00 a.m. to 8:00 p.m.	After Feb. 1, 8:00 p.m.
gohawks	242.331	6136.48	36.375
gopatriots	13.602	1793.101	5.073
nfl	277.267	4990.022	524.034
patriots	316.726	30249.871	163.072
sb49	547.82	38310.21	321.255
superbowl	472.383	101087.678	732.144

This model is a multi-layer neural network model, in which a tanh activation function and lbfgs solver are adopted. Here, we adopt mean absolute difference as our error measurement. The green parts the average cross-validation errors for the 3 different models with each individual hashtags.

Part II: Combined Model

Model	Before Feb. 1, 8:00 a.m.	Feb. 1, 8:00 a.m. to 8:00 p.m.	After Feb. 1, 8:00 p.m.
COMBINED-OLS	319.631	20819.888	77.546
COMBINED-SGD	335.222	24031.615	83.62
COMBINED_MLNN	295.75	31103.111	298.459

Based on our experimental results, for the aggregated hashtags, OLS model performed the best in the last 2 intervals (i.e., Feb 1st, 8am to 8pm, After Feb 1st, 8pm). For the first interval (i.e., Before Feb 1st, 8am), multi-layer neural networks perform the best.

Problem 1.5

I have tried 3 different models (i.e., OLS, SGD, Multi-layer Neural Networks) in this section with a 10-Fold cross validation. The overall average performance is shown below,

Model	Before Feb. 1, 8:00 a.m.	Feb. 1, 8:00 a.m. to 8:00 p.m.	After Feb. 1, 8:00 p.m.
COMBINED-OLS	311.285	19892.546	137.835
COMBINED-SGD	15621.176	57641.152	865.355
COMBINED_MLNN	300.765	31101.961	310.664

Based on our previous results, in this section we chose OLS as our regression model to predict the number of tweets for next hour, since OLS has the best overall performance. Our prediction result is shown in the Table below. Since our TA said that only predict the result for the last hour, therefore only one result is shown in the table for each testing file.

File Name	Predicted Number of Tweets
sample1_period1.txt	52.27736

sample2_period2.txt	12.530241
sample3_period3.txt	596.576677
sample4_period1.txt	356.875386
sample5_period1.txt	15.381625
sample6_period2.txt	17.811189
sample7_period3.txt	49.116023
sample8_period1.txt	71.643865
sample9_period2.txt	9.132987
sample10_period3.txt	46.652107

Part 2: Fan Base Prediction

In this part, we have tried three different models, they are SVM, Multinomial Naive Bayes and Logistic Regression. For the feature extraction process, we adopted one-hot encoding and TF-IDF features with NMF dimension reduction. The result is shown below.

2.1 SVM Model

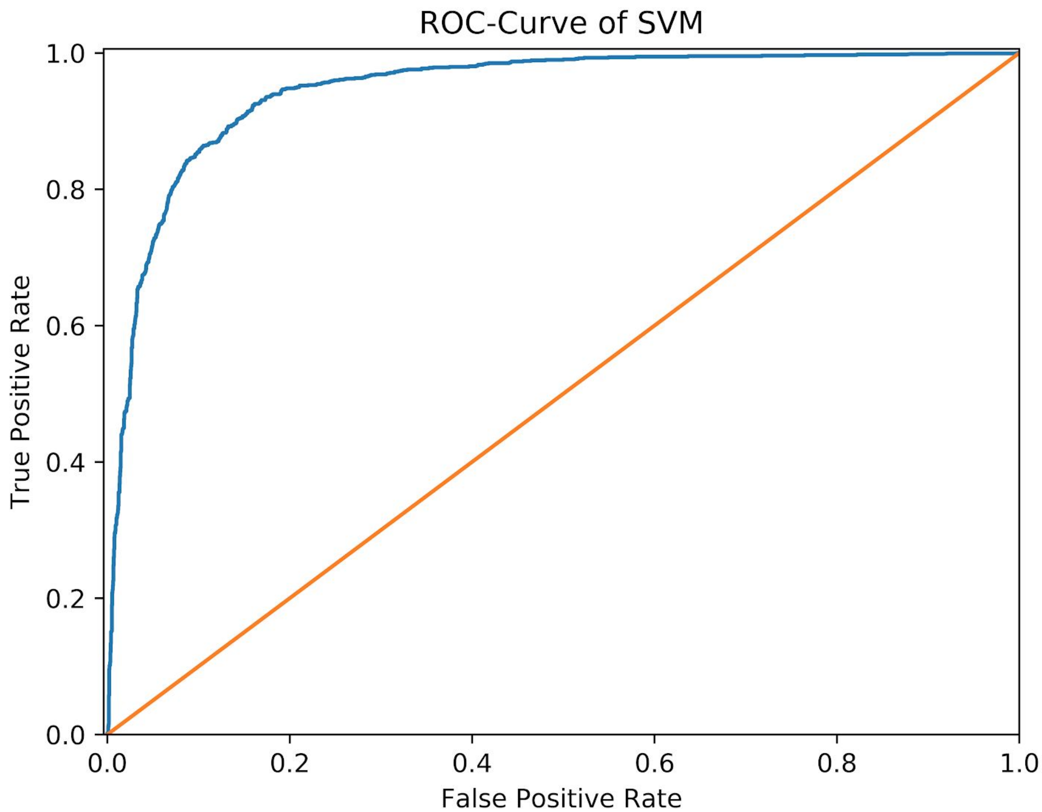
The evaluation metrics is shown in the table below. Here we treat Washington as Positive and Massachusetts and Negative.

Accuracy	0.881197
Recall	0.717767
Precision	0.859699

The confusion matrix is shown in the table below,

	Predicted MA	Predicted WA
Ground-truth MA	2855	149
Ground-truth WA	359	913

The ROC curve is shown below,



2.2 Multinomial Naive Bayes Model

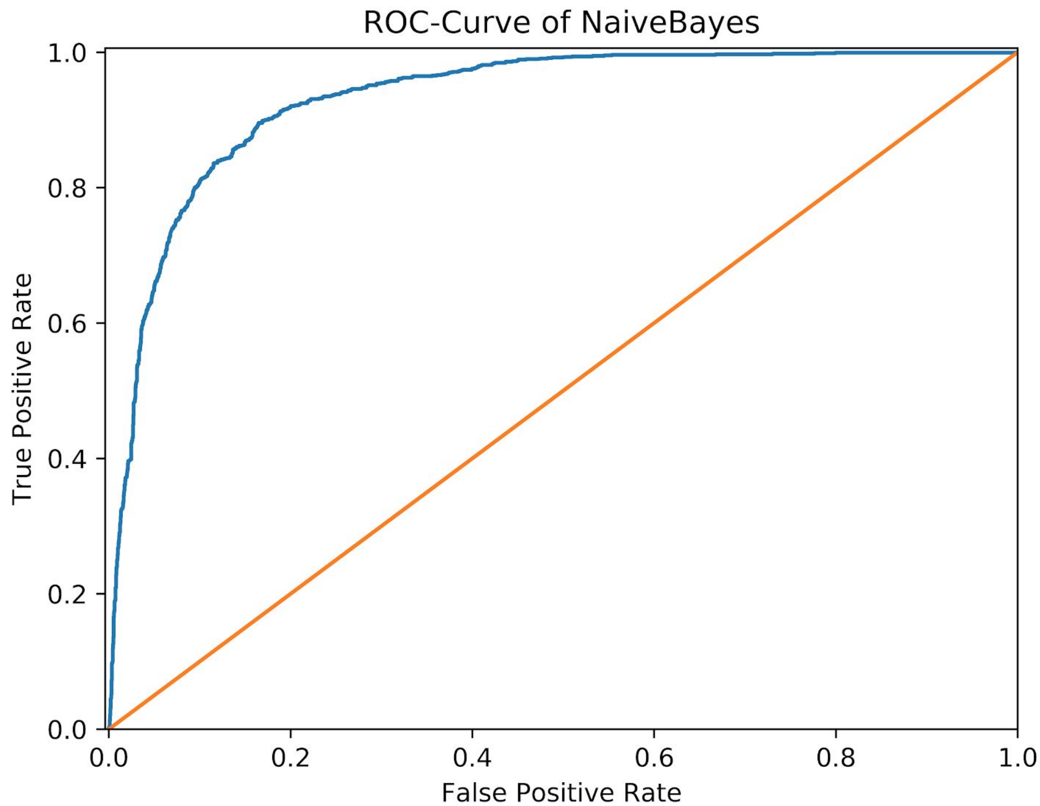
The evaluation metrics is shown in the table below. Here we treat Washington as Positive and Massachusetts and Negative.

Accuracy	0.741581
Recall	0.143868
Precision	0.919598

The confusion matrix is shown in the table below,

	Predicted MA	Predicted WA
Ground-truth MA	2988	16
Ground-truth WA	1089	183

The ROC curve is shown below,



2.3 Logistic Regression

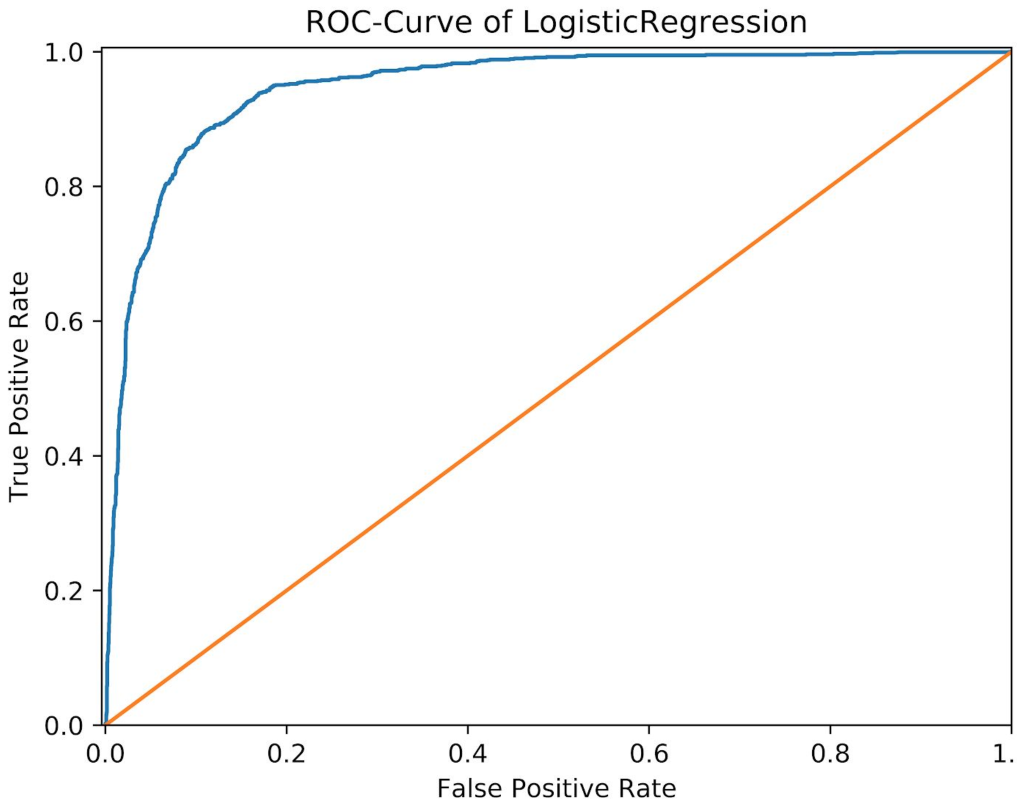
The evaluation metrics is shown in the table below. Here we treat Washington as Positive and Massachusetts and Negative.

Accuracy	0.881431
Recall	0.717767
Precision	0.860509

The confusion matrix is shown in the table below,

	Predicted MA	Predicted WA
Ground-truth MA	2856	148
Ground-truth WA	359	913

The ROC curve is shown below,



Part 3: Define Your Own Project

Task 1: Finding the most popular topics before, during and after the superbowl game

Idea: In this part, we want to find out the popular topics ***before***, ***during*** and ***after*** the SuperBowl Game.

To implement our idea, firstly, we need to split the #SuperBowl, #Gohawks, #Gopatriots dataset into 3 subsets (i.e., before the game, during the game and after the game) according the every tweet posting time. We, then, will utilize TF-ICF to find out the ranking score for every word.

Finally, we can find out the top-10 popular words by selecting the 10 most significant words. The result is shown below.

#Gohawks		
Before	During	After
gohawks	gohawks	gohawks
http	http	http
seahawks	seahawks	seahawks
game	sb49	season
12thman	superbowl	sb49
gbvssea	superbowlxlix	year
seattle	game	nfl
sb49	bowl	dangerusswilson
superbowl	seattle	superbowl
bowl	super	great

In #Gohawks topic, from the 3rd and 7th words we could assume that hawk's home is in Seattle. Also,we could see that the fans of hawks love there team very much, they want to be 12thman to fight with their team. In addition, the user named 'sb49' shows a high popularity. That maybe he does some brilliant comments through games.

#Gopatriots		
Before	During	After
gopatriots	gopatriots	gopatriots
http	http	http
patriot	superbowl	superbowl
superbowl	superbowlxlix	patriot
game	patriot	bowl
bowl	bowl	super

super	gopats	sb49
nfl	sb49	win
colt	super	brady
gopats	brady	gopats

In #Gopatriots topic, from the word ‘brady’ and ‘win’ we could assume that patriots won the game and Brady may become superhero in this game. Also, the user named ‘sb49’ shows a high popularity. That maybe he does some brilliant comments through games.

#SuperBowl		
Before	During	After
nfl	nfl	http
http	http	nfl
patriot	superbowl	seahawks
seahawks	seahawks	football
football	patriot	patriot
colt	superbowlxlix	bowl
superbowl	sb49	super
bowl	bowl	superbowl
new	super	wire
packer	game	sport

In #SuperBowl topic, from the word ‘http’ we could assume that maybe a lot of people foucs on this football game and after the generate of championship, more people went to see the news or videos about it.

Task 2: Predict which team should twitter users be a fan of based on their tweets context and sentimental analysis

Problem Statement

In this task, our goal is to predict which team should a twitter user be a fan of (e.g., we can predict twitter user A should be a fan of Team Hawks). We formalized such prediction problem into a classification problem. We can extract the tweets content and its corresponding hashtag from the tweets dataset. If the hashtag is #GoHawks then we assume this user is a fan of Hawks. If the hashtag is #GoPatriots, we then assume this user is a fan of Patriots. If the user is a fan of Hawks, then we say he/she belongs to class 1, if the user is a fan of Patriots, then we say he/she belongs to class 0. Then our goal is to predict which class that a user belongs to given the user's tweet content.

Approach

For the feature extraction part, as we mentioned before, we can extract the tweets content and its corresponding hashtag from the tweets dataset. Then we can use TFxIDF feature along with the one-hot encoding to represent the feature of a tweet content. For the efficiency, we also applied NMF dimensionality reduction method and any keep 50 features. Therefore, the dimension of the feature space is 50.

The the labeling part, since the dataset does not provide the label directly. However, we can generate the label by ourselves. For example, if the user is a fan of Hawks, then we say he/she belongs to class 1, if the user is a fan of Patriots, then we say he/she belongs to class 0.

Once we have the features extracted from the original tweet content, we can try different classification models, and evaluate their performances. In order to evaluate the overall performance cross validation is adopted, specifically, in this task we applied 10-Fold cross validation.

Experimental Results

SVM Model

The evaluation metrics is shown in the table below. Here we treat fan of Hawks as Positive and fan of Patriots and Negative.

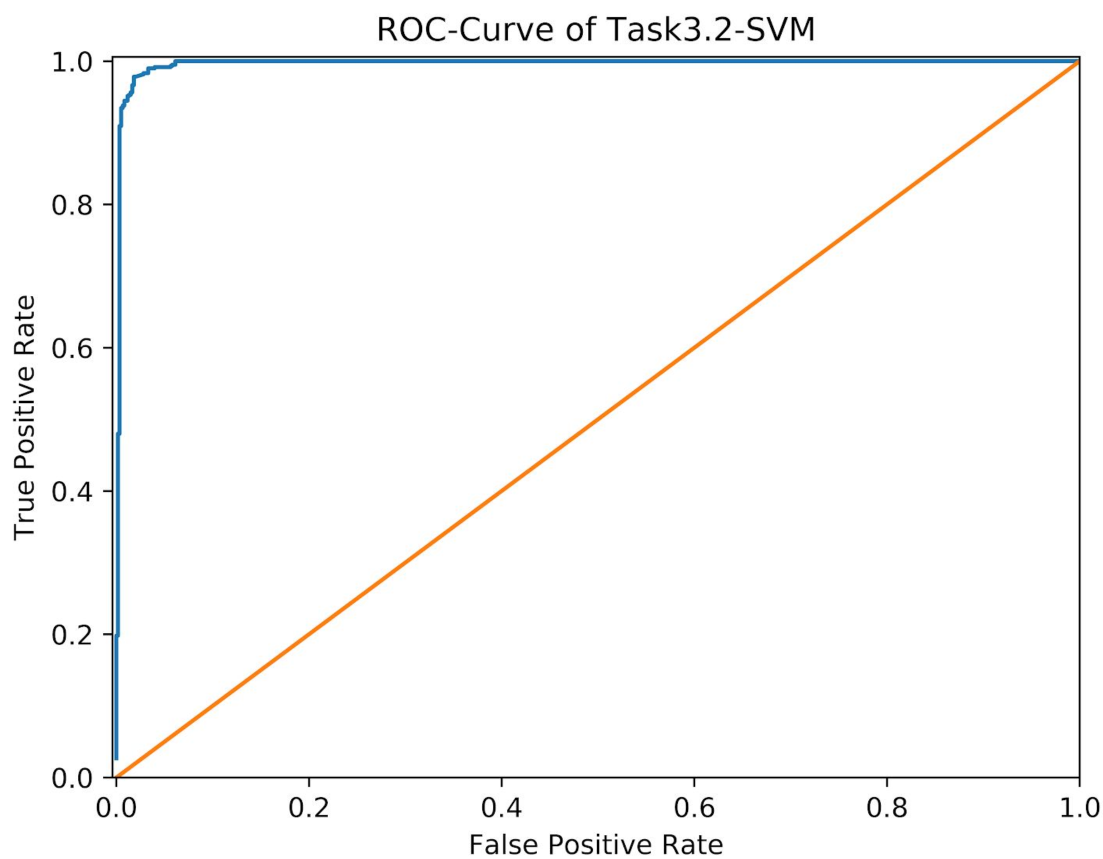
accuracy	0.970833

recall	0.958054
precision	0.982788

The confusion matrix is shown in the table below,

	Predicted Patriots	Predicted Hawks
Ground-truth Patriots	594	10
Ground-truth Hawks	25	571

The ROC curve is shown below,



Naive Bayes Model

The evaluation metrics is shown in the table below. Here we treat fan of Hawks as Positive and fan of Patriots and Negative.

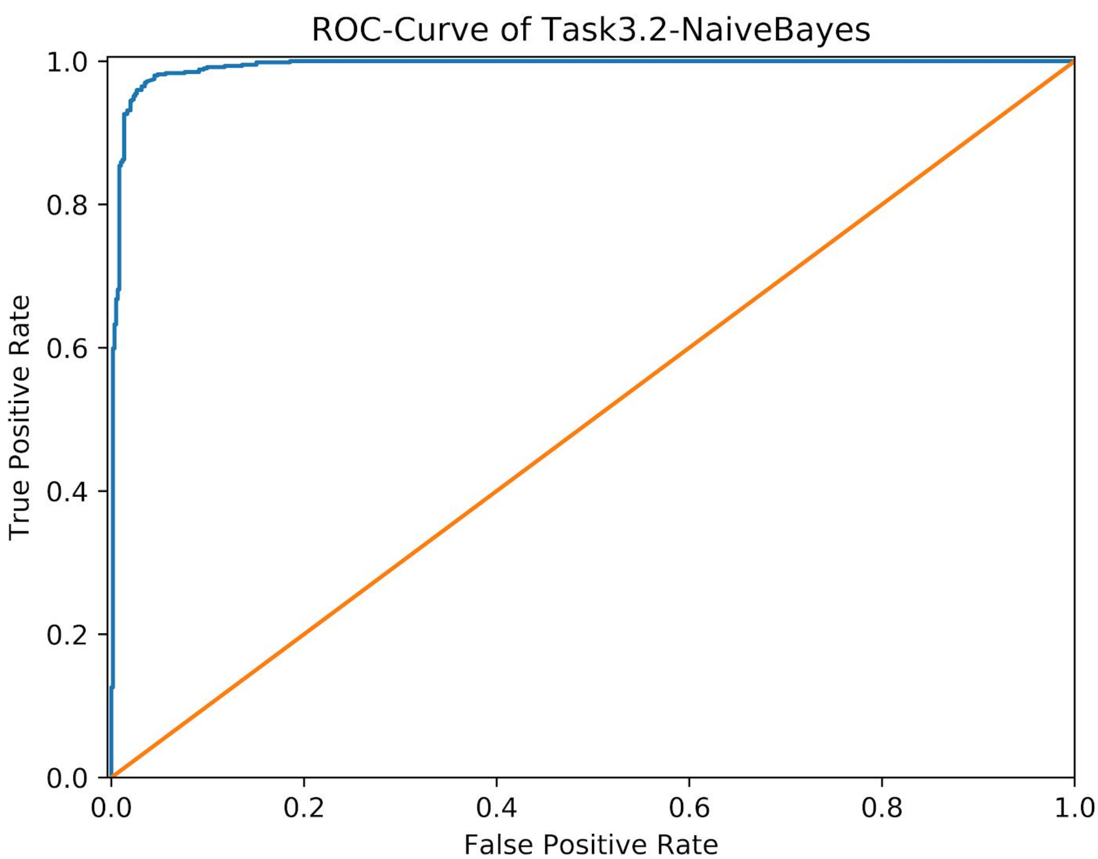
accuracy	0.965

recall	0.974832
precision	0.955592

The confusion matrix is shown in the table below,

	Predicted Patriots	Predicted Hawks
Ground-truth Patriots	577	27
Ground-truth Hawks	15	581

The ROC curve is shown below,



Logistic Regression Model

The evaluation metrics is shown in the table below. Here we treat fan of Hawks as Positive and fan of Patriots and Negative.

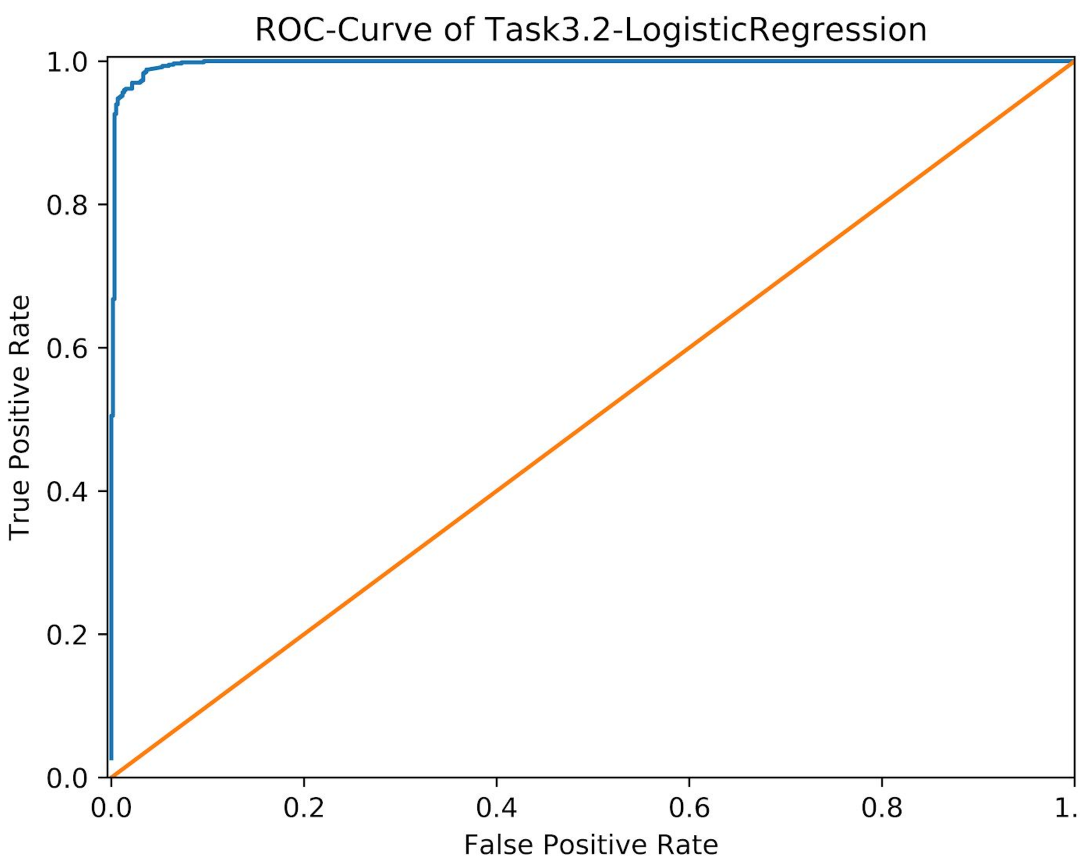
accuracy	0.974167

recall	0.981544
precision	0.966942

The confusion matrix is shown in the table below,

	Predicted Patriots	Predicted Hawks
Ground-truth Patriots	584	20
Ground-truth Hawks	11	585

The ROC curve is shown below,



Observation & Conclusion

Based on our observations, we find out all of the 3 models (i.e., SVM, Naive Bayes) has a very good performance. At the beginning of this experiment, we just keep the original tweet contents (did not perform any modification on the original dataset). Since we have a surprisingly good performance, we just assume, probably, twitter user just includes some super important keyword like “GoHawks” or “GoPatriots” in their tweets. In order to verify our assumption, we look into the dataset, and find out our assumption does not hold true, since there are only a very small proportion of user include such keywords in their tweets. Thereby, we can make a conclusion that our models have a good overall performance.