

EE 219: Large-Scale Data Mining: Models and Algorithms

Project 5: Popularity Prediction on Twitter

Akshay Sharma (504946035)

Anoosha Sagar (605028604)

Nikhil Thakur(804946345)

Rahul Dhavalikar (205024839)

Part 1: Popularity Prediction

Problem 1.1

Download the training tweet data and calculate the following statistics for each hashtag:

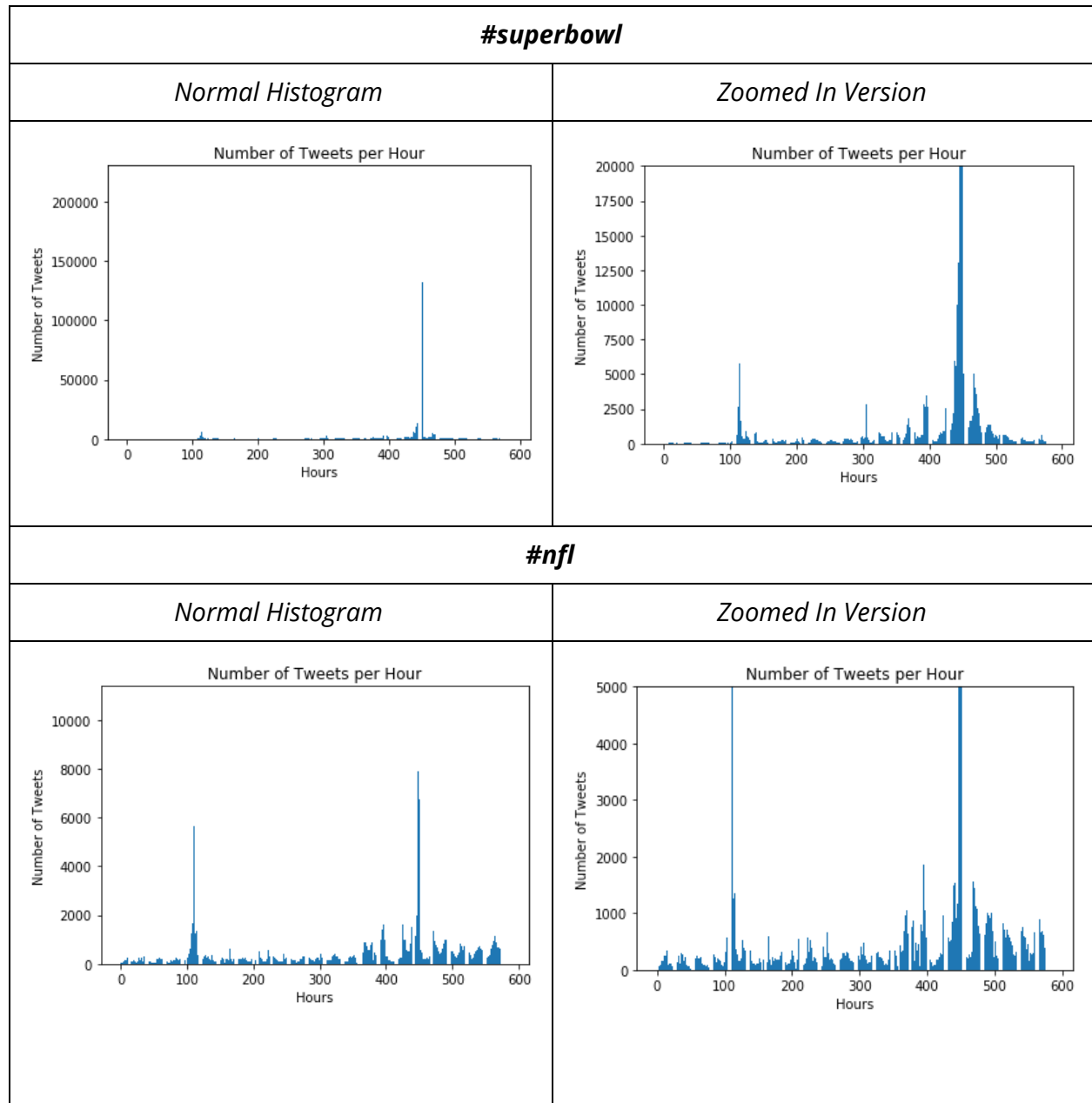
- Average Number of Tweets per Hour
- Average Number of Followers of Users Posting the Tweets
- Average Number of Retweets

Hashtag	Average Number of Tweets per Hour	Average Number of Followers of Users Posting the Tweets	Average Number of Retweets
#gohawks	324.9326424870466	2203.931767444827	2.014617085512608
#goPatriots	45.62086956521739	1401.8955093016164	1.4000838670326319
#nfl	441.26746166950596	4653.252285502502	1.5385331089011056
#patriots	834.2640545144804	3309.978828415827	1.7828156491659402
#sb49	1418.4408233276158	10267.31684948685	2.5111487863247035
#superbowl	2297.7291311754684	8858.974662784603	2.3882723999030224

Observations:

- The average number of tweets are highest for hashtag **'superbowl'**.
- Average number of followers are highest for the users posting tweets with hashtag **'sb49'**.
- On an average, tweets with hashtag **'sb49'** have the highest retweets.

Plot “number of tweets in hour” over time for #SuperBowl and #NFL (a histogram with 1-hour bins). The tweets are stored in separate files for different hashtags and files are named as tweet [#hashtag].txt.



Observations

- The number of tweets for hashtag **'superbowl'** experience a dramatic increase in the 450th hour. This time span involves the match day and hence the increase in the tweets.
- The number of tweets for hashtag **'nfl'** see high peaks at around 105th hour and 450th hour.

Problem 1.2

For each hashtag, fit a Linear Regression model using the following 5 features to predict number of tweets in the next hour, with features extracted from tweet data in the previous hour.

The features you should use are:

- Number of tweets (hashtag of interest)
- Total number of retweets (hashtag of interest)
- Sum of the number of followers of the users posting the hashtag
- Maximum number of followers of the users posting the hashtag
- Time of the day (which could take 24 values that represent hours of the day with respect to a given time zone)

For each hashtag, you should train a separate model.

For each of your models, report your model's training accuracy (training RMSE) and R-squared measure. Also, analyse the significance of each feature using the t-test and P-value. You may use the library statsmodels.api in Python.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. It is the percentage of the response variable variation that is explained by a linear model. Below we have reported the Training RMSE and R-squared measure for all the models that were fitted for each hashtag.

Hashtag	Training Accuracy (Training RMSE)	R-Squared Measure
#gohawks	972.4960929758814	0.473
#gopatriots	185.01584609874894	0.632
#nfl	581.4221228225172	0.564
#patriots	2526.2761166644177	0.670
#sb49	4470.451263196001	0.805
#superbowl	8003.558922638213	0.802

Next, we calculate the t-Value and p-Value for each hashtags for all the features. The lesser the p-Value, greater is the significance of the feature. So the best features for each hashtags are as below,

- **#gohawks:** The order of significance of the features in decreasing order are
 - Number of Tweets
 - Total number of retweets
 - Sum of number of followers of
 - Time of Day
 - Maximum number of followers of users.
- **#gopatriots:** The order of significance of the features in decreasing order are
 - Total number of retweets
 - Maximum number of followers of users
 - Sum of number of followers of

-
- Number of Tweets
 - Time of Day
 - **#nfl:** The order of significance of the features in decreasing order are
 - Number of Tweets
 - Sum of number of followers of
 - Total number of retweets
 - Maximum number of followers of users
 - Time of Day
 - **#patriots:** The order of significance of the features in decreasing order are
 - Number of Tweets
 - Maximum number of followers of users
 - Total number of retweets
 - Time of Day
 - Sum of number of followers of
 - **#sb49:** The order of significance of the features in decreasing order are
 - Number of Tweets
 - Total number of retweets
 - Maximum number of followers of users
 - Sum of number of followers of
 - Time of Day
 - **#superbowl:** For this hashtag, there are 4 hashtags which have almost a 0 p-Value. This shows that they are equally important.
 - Number of Tweets, Maximum number of followers of users, Total number of retweets, Sum of number of followers of
 - Time of Day

Features	#gohawks			#gopatriots		
	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>
<i>Number of tweets</i>	7.250	0.000	1.35814990e-12	-0.314	0.754	0.75361812
<i>Total number of retweets</i>	-2.886	0.004	4.05276848e-03	2.282	0.023	0.02286726
<i>Sum of number of followers of the users</i>	-2.064	0.039	3.94236710e-02	1.237	0.217	0.21655161
<i>Maximum Number of followers of the users</i>	0.112	0.911	9.10611320e-01	-1.908	0.057	0.05692787
<i>Time of Day</i>	0.344	0.731	7.30987682e-01	-0.123	0.902	0.90216103

Features	#nfl			#patriots		
	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>
<i>Number of tweets</i>	5.103	0.000	4.54237347e-07	12.867	0.000	1.64541115e-33
<i>Total number of retweets</i>	-2.606	0.009	9.38544525e-03	-1.484	0.138	1.38293237e-01
<i>Sum of number of followers of the users</i>	3.289	0.001	1.06793815e-03	-0.014	0.989	9.88609009e-01
<i>Maximum Number of followers of the users</i>	-2.494	0.013	1.29127257e-02	1.615	0.107	1.06947800e-01
<i>Time of Day</i>	-0.005	0.996	9.95750562e-01	-0.454	0.650	6.49821571e-01

Features	#sb49			#superbowl		
	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>	t-Value	p-Value <i>From summary</i>	p-Value <i>From results.pvalues</i>
Number of tweets	12.478	0.000	8.48163623e-32	28.925	0.000	1.44495499e-114
Total number of retweets	-2.437	0.015	1.50943940e-02	-8.039	0.000	5.09291742e-015
Sum of number of followers of the users	1.323	0.186	1.86357283e-01	-7.019	0.000	6.25161177e-012
Maximum Number of followers of the users	1.997	0.046	4.63149084e-02	5.343	0.000	1.31635953e-007
Time of Day	-0.654	0.513	5.13056746e-01	-0.506	0.613	6.13329520e-001

Problem 1.3

Design a regression model using any features from the papers you find or other new features you may find useful for this problem. Fit your model on the data of each hashtag and report fitting accuracy and significance of variables.

For each of the top 3 features in your measurements, draw a scatter plot of predicted (number of tweets for next hour) versus value of that feature, using all the samples you have extracted, and analyze it.

For this problem, we added all the below features which reduced the training RMSE by an appreciable amount. These features were extracted from various fields from the tweet, user, and metrics objects in the Twitter API data given to us.

Features Used:

Feature	Description
<i>Number of Tweets</i>	Total number of tweets in the current hour
<i>Number of Retweets</i>	Total number of retweets of the tweets posted in the current hour
<i>Sum of number of followers of the users</i>	Total number of followers of the authors of the tweets posted in the current hour
<i>Maximum Number of followers of the users</i>	The maximum number of followers of an author from all the tweets in the current hour
<i>Number of URLs</i>	The total number of URLs appearing in the tweets in the current hour
<i>Number of Hashtags</i>	The total number of hashtags appearing in the tweets in the current hour
<i>Status Count</i>	The total number of status counts (the number of Tweets (including retweets) issued by the user) of the authors of the tweets in the current hour
<i>Friend Count</i>	The total number of friends (the number of users this account is following) of the authors of the tweets in the current hour
<i>Favorite Count</i>	The total number of favorites (The number of Tweets this user has liked in the account's lifetime) of the authors of the tweets in the current hour
<i>Impressions</i>	The total number of impressions of the tweets in the current hour
<i>Number of Users</i>	The total number of unique users of the tweets in the current hour
<i>Unique Languages</i>	The total number of unique languages of the tweets in the current hour
<i>Unique Verified Users</i>	The number of unique verified users which tweeted in the current hour
<i>Emoticon Count</i>	The total number of emoticons in the tweets for the current hour
<i>Uppercase Characters</i>	The total number of uppercase characters in all the tweets in the current hour
<i>Hour of Day</i>	Time of the day

<i>Special Character Count</i>	The total number of special characters in all the tweets in the current hour
<i>Length of Tweet</i>	The total length of all the tweets in the current hour
<i>Ranking Score</i>	The sum of ranking scores of all the tweets in the current hour
<i>Momentum</i>	The sum of the momentum of all the tweets in the current hour

Below are the training RMSE after using the above features.

Hashtag	Training Accuracy	R-Squared Measure
#gohawks	663.2387380919895	0.755
#gopatriots	54.041288833466034	0.969
#nfl	422.69118462165807	0.770
#patriots	1799.8552160746183	0.832
#sb49	2680.933275730391	0.930
#superbowl	4349.6281589213295	0.942

Feature	#gohawks		#gopatriots	
	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>
<i>Number of Tweets</i>	0.000	1.82E-14	0.000	1.7405E-04
<i>Number of Retweets</i>	0.047	4.75E-02	0.011	1.1286E-02
<i>Sum of number of followers of the users</i>	0.000	3.28E-11	0.000	8.2852E-35
<i>Maximum Number of followers of the users</i>	0.000	9.34E-08	0.000	5.0441E-22
<i>Number of URLs</i>	0.000	3.39E-14	0.000	2.3336E-21

<i>Number of Hashtags</i>	0.366	3.66E-01	0.000	4.3180E-04
<i>Status Count</i>	0.003	2.88E-03	0.000	7.4489E-06
<i>Friend Count</i>	0.002	1.71E-03	0.403	4.0276E-01
<i>Favorite Count</i>	0.000	4.78E-17	0.000	1.3191E-12
<i>Impressions</i>	0.036	3.63E-02	0.000	6.3842E-22
<i>Number of Users</i>	0.000	7.36E-07	0.124	1.2354E-01
<i>Unique Languages</i>	0.080	7.98E-02	0.000	4.5599E-08
<i>Unique Verified Users</i>	0.805	8.05E-01	0.000	7.9239E-05
<i>Emoticon Count</i>	0.000	9.32E-08	0.000	1.4798E-09
<i>Uppercase Characters</i>	0.000	3.03E-09	0.000	5.8808E-64
<i>Hour of Day</i>	0.178	1.78E-01	0.781	7.8067E-01
<i>Special Character Count</i>	0.000	2.03E-05	0.000	2.5357E-15
<i>Length of Tweet</i>	0.001	8.70E-04	0.000	5.0496E-11
<i>Ranking Score</i>	0.000	7.53E-15	0.517	5.1740E-01
<i>Momentum</i>	0.000	1.01E-05	0.000	4.7430E-05

Feature	#nfl		#patriots	
	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>
<i>Number of Tweets</i>	0.005	5.0529E-03	0.000	1.6794E-07
<i>Number of Retweets</i>	0.729	7.2948E-01	0.000	6.1067E-07
<i>Sum of number of followers of the users</i>	0.180	1.7989E-01	0.158	1.5798E-01
<i>Maximum Number of followers of the users</i>	0.205	2.0486E-01	0.205	2.0513E-01
<i>Number of URLs</i>	0.001	7.7895E-04	0.000	2.3485E-28
<i>Number of Hashtags</i>	0.000	4.7946E-10	0.000	1.0804E-10
<i>Status Count</i>	0.711	7.1052E-01	0.299	2.9946E-01

<i>Friend Count</i>	0.002	2.2401E-03	0.051	5.0917E-02
<i>Favorite Count</i>	0.000	2.2800E-11	0.583	5.8319E-01
<i>Impressions</i>	0.592	5.9182E-01	0.155	1.5534E-01
<i>Number of Users</i>	0.000	5.0077E-11	0.006	5.5278E-03
<i>Unique Languages</i>	0.706	7.0595E-01	0.006	5.6109E-03
<i>Unique Verified Users</i>	0.000	1.5446E-04	0.288	2.8830E-01
<i>Emoticon Count</i>	0.341	3.4077E-01	0.000	9.5719E-05
<i>Uppercase Characters</i>	0.001	7.3287E-04	0.000	4.7711E-10
<i>Hour of Day</i>	0.155	1.5486E-01	0.523	5.2344E-01
<i>Special Character Count</i>	0.780	7.7982E-01	0.135	1.3467E-01
<i>Length of Tweet</i>	0.000	9.6059E-07	0.000	1.9128E-08
<i>Ranking Score</i>	0.190	1.8968E-01	0.000	5.1399E-15
<i>Momentum</i>	0.113	1.1313E-01	0.116	1.1567E-01

Feature	#sb49		#superbowl	
	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>	<i>p-Value</i> <i>From summary</i>	<i>p-Value</i> <i>From results.pvalues</i>
<i>Number of Tweets</i>	0.174	1.7378E-01	0.241	2.4130E-01
<i>Number of Retweets</i>	0.000	1.1956E-07	0.031	3.0521E-02
<i>Sum of number of followers of the users</i>	0.034	3.4259E-02	0.641	6.4053E-01
<i>Maximum Number of followers of the users</i>	0.000	1.6152E-07	0.094	9.4074E-02
<i>Number of URLs</i>	0.000	2.2420E-14	0.000	1.1152E-30
<i>Number of Hashtags</i>	0.001	9.5533E-04	0.000	2.2620E-12
<i>Status Count</i>	0.000	5.6035E-23	0.000	2.7220E-04
<i>Friend Count</i>	0.000	2.1697E-27	0.066	6.5586E-02

<i>Favorite Count</i>	0.000	5.4019E-24	0.000	2.3546E-12
<i>Impressions</i>	0.000	2.2121E-04	0.867	8.6726E-01
<i>Number of Users</i>	0.000	2.4649E-04	0.000	1.3094E-05
<i>Unique Languages</i>	0.000	3.2098E-05	0.000	3.6492E-07
<i>Unique Verified Users</i>	0.019	1.9166E-02	0.066	6.6345E-02
<i>Emoticon Count</i>	0.000	1.9323E-08	0.806	8.0571E-01
<i>Uppercase Characters</i>	0.196	1.9648E-01	0.000	7.4128E-17
<i>Hour of Day</i>	0.850	8.5015E-01	0.017	1.6993E-02
<i>Special Character Count</i>	0.000	8.6211E-07	0.000	3.8470E-41
<i>Length of Tweet</i>	0.000	4.3081E-05	0.369	3.6943E-01
<i>Ranking Score</i>	0.271	2.7149E-01	0.501	5.0128E-01
<i>Momentum</i>	0.000	1.0591E-09	0.029	2.8582E-02

Here we select the best 3 features using **SelectKBest** function and plot a graph of the Predictant values vs the values for each of the features. We can observe that the relationship is almost linear.

The best features that we selected for the hashtag 'gohawks' are,

- Number of URLs
- Number of Unique Languages
- Unique Verified Users

The best features that we selected for the hashtag 'gopatriots' are,

- Number of URLs
- Status Count
- Friend Count

The best features that we selected for the hashtag 'nfl' are,

- Number of Hashtags

- Uppercase Characters
- Special Character count

The best features that we selected for the hashtag 'patriots' are,

- Favourite Count
- Uppercase Characters
- Special Character count

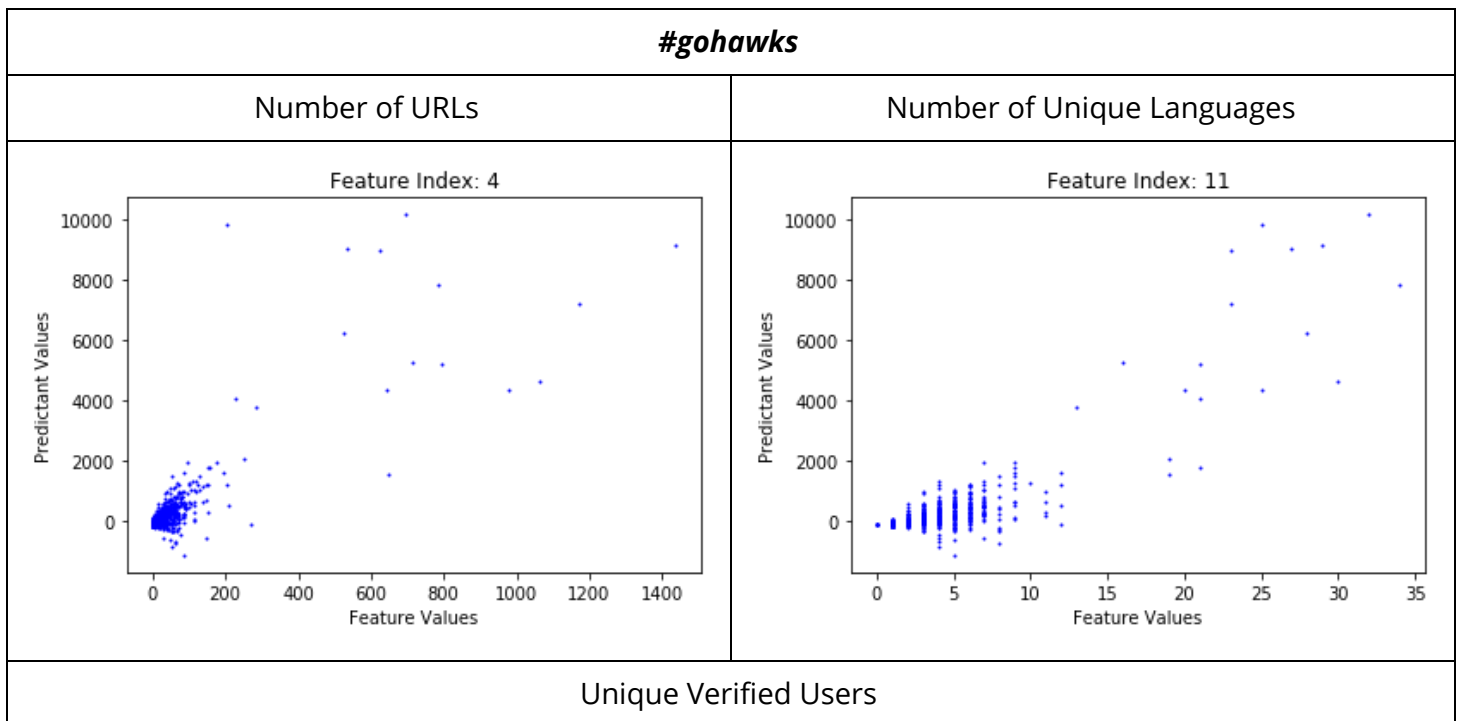
The best features that we selected for the hashtag 'sb49' are,

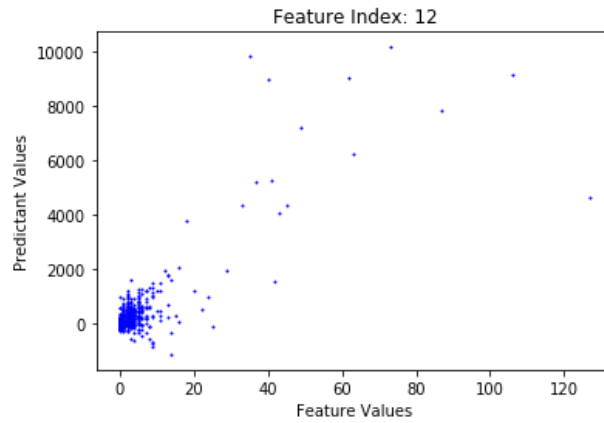
- Number of Users
- Uppercase Characters
- Special Character count

The best features that we selected for the hashtag 'superbowl' are,

- Number of Users
- Favourite Count
- Uppercase Characters

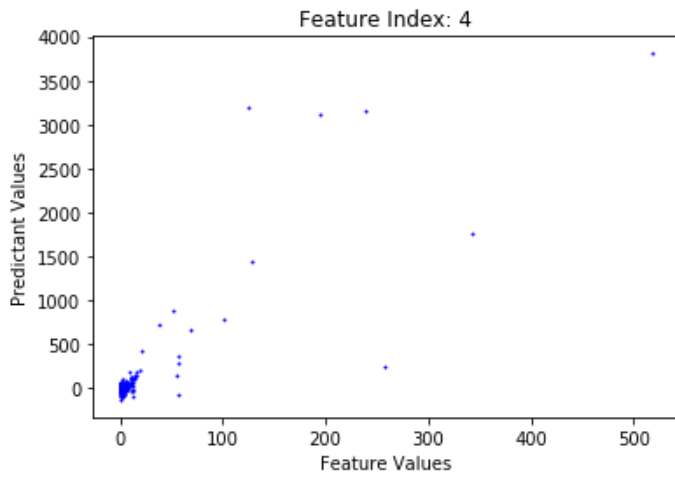
Best Feature Selection Using SelectKBest method



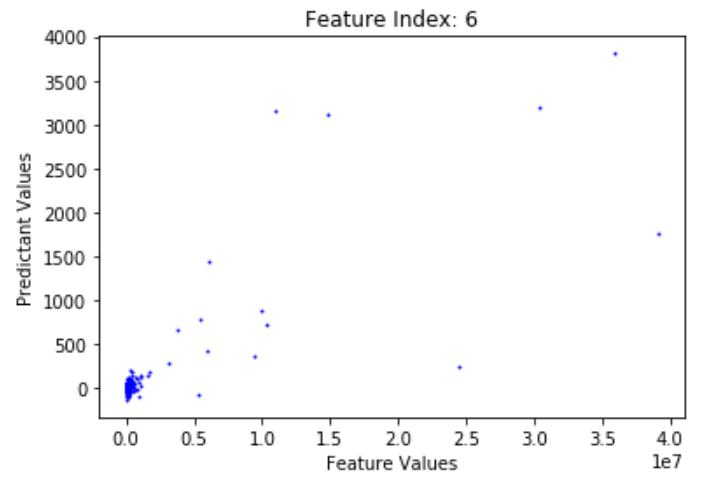


#gopatriots

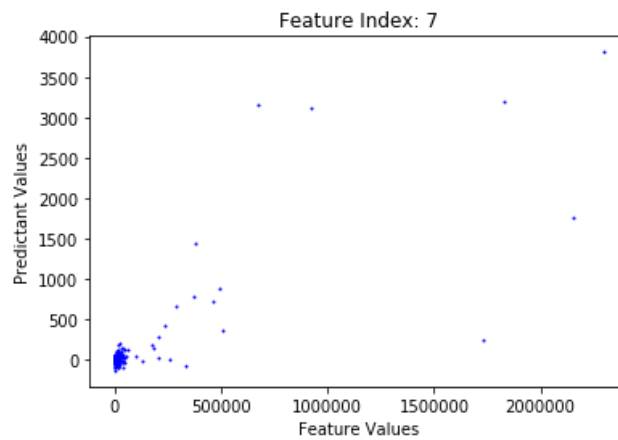
Number of URLs

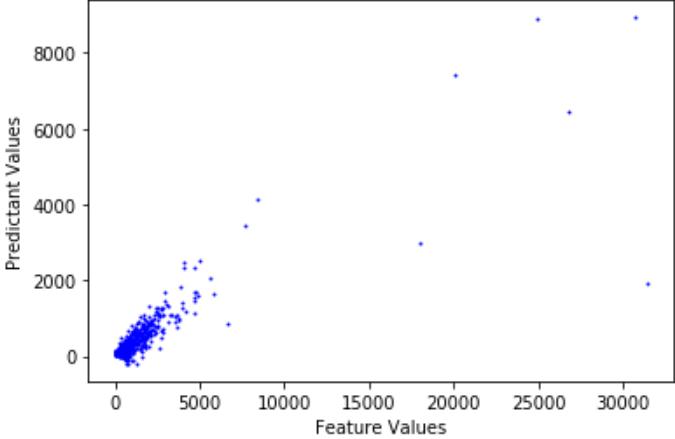
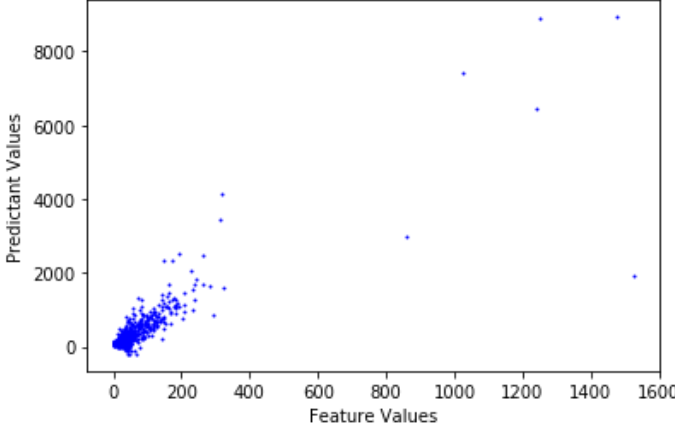
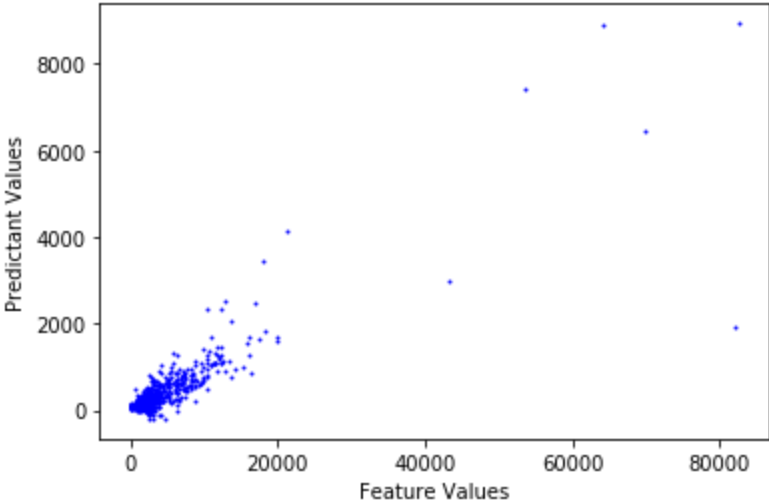


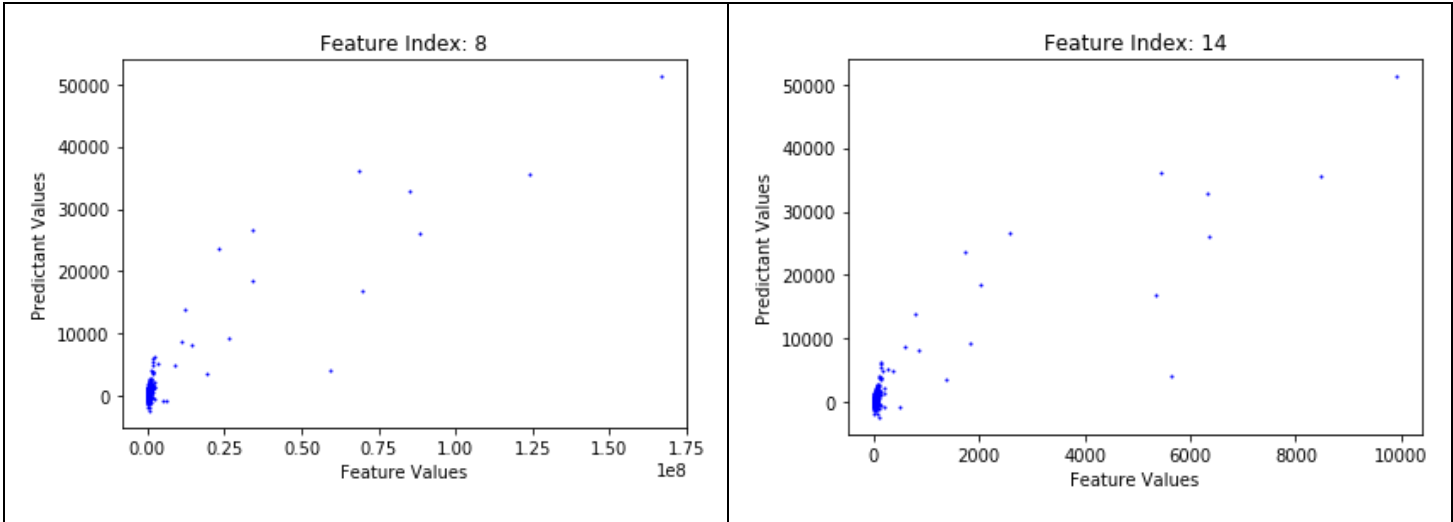
Status Count



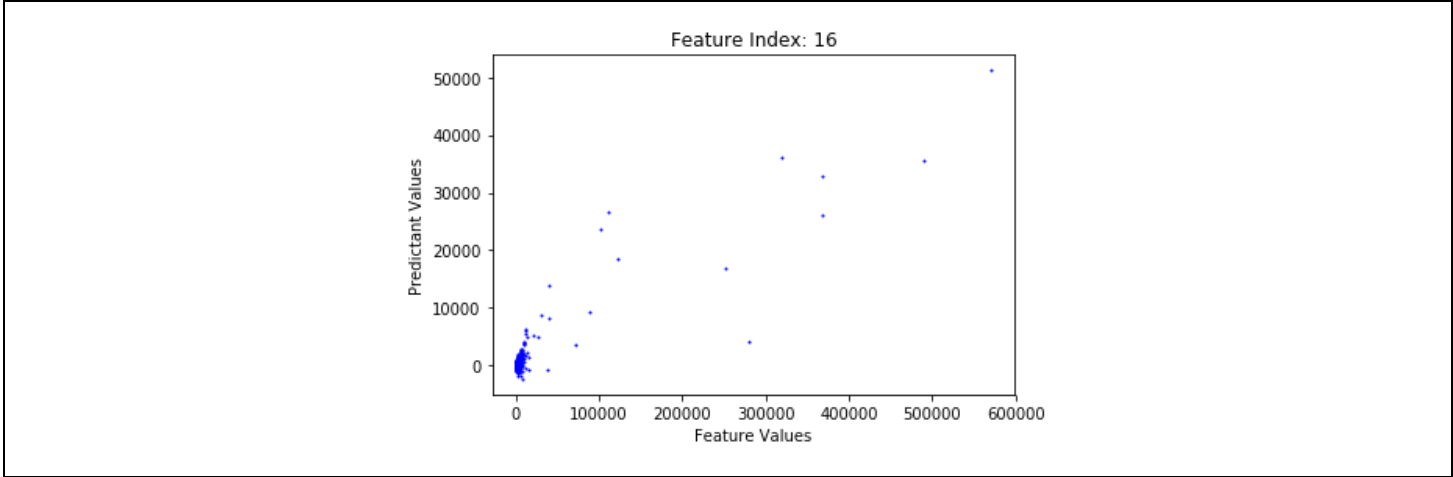
Friend Count



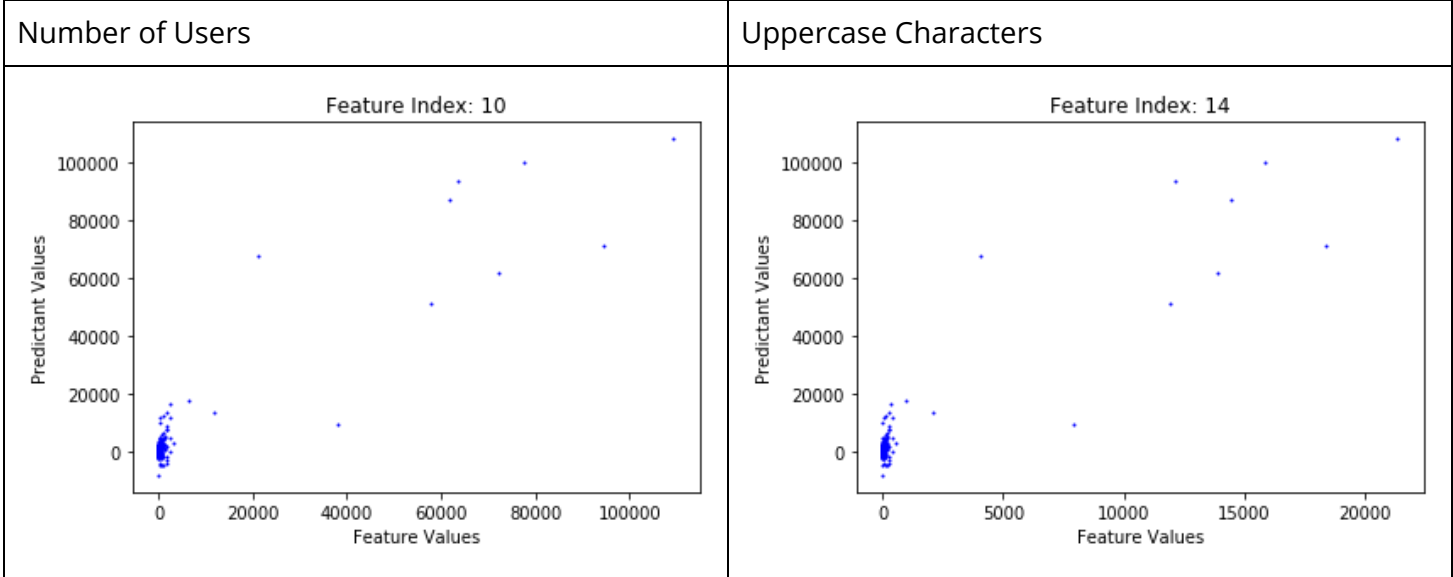
#nfl	
Number of Hashtags	Uppercase Characters
<div>Feature Index: 5</div> 	<div>Feature Index: 14</div> 
Special Character Count	
<div>Feature Index: 16</div> 	
#patriots	
Favorite Count	Uppercase Characters



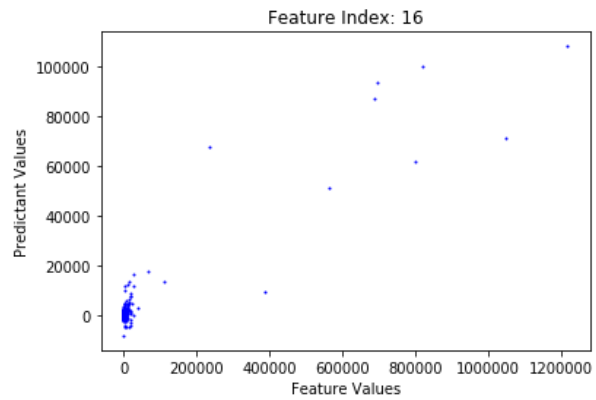
Special Character Count



#sb49

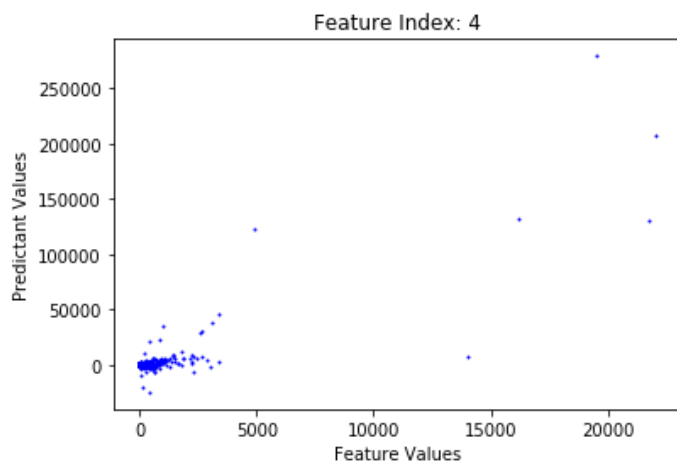


Special Character Count

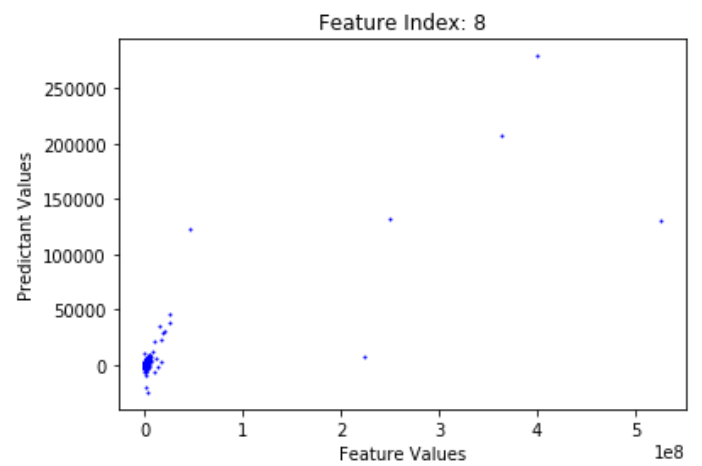


#superbowl

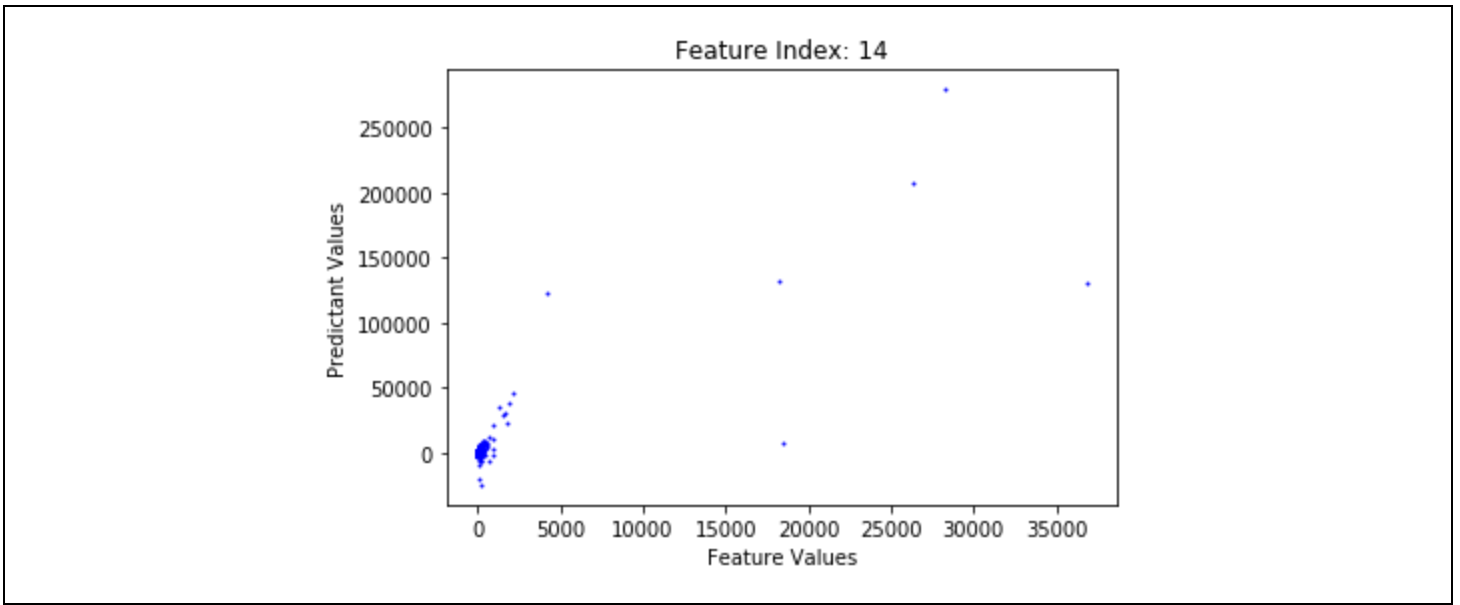
Number of URLs



Favorite Count



Uppercase Characters



Here we select the best 3 features using the **p-Values** of the features and plot a graph of the Predictant values vs the values for each of the features. We can observe that the relationship is almost linear.

The best features that we selected for the hashtag 'gohawks' are,

- Favourite Count
- Ranking Score
- Number of Tweets

The best features that we selected for the hashtag 'gopatriots' are,

- Uppercase Characters
- Sum of Number of Followers of the Users
- Maximum Number of Followers of the Users

The best features that we selected for the hashtag 'nfl' are,

- Favourite Count
- Number of Users
- Number of Hashtags

The best features that we selected for the hashtag 'patriots' are,

- Number of URLs
- Ranking Score
- Number of Hashtags

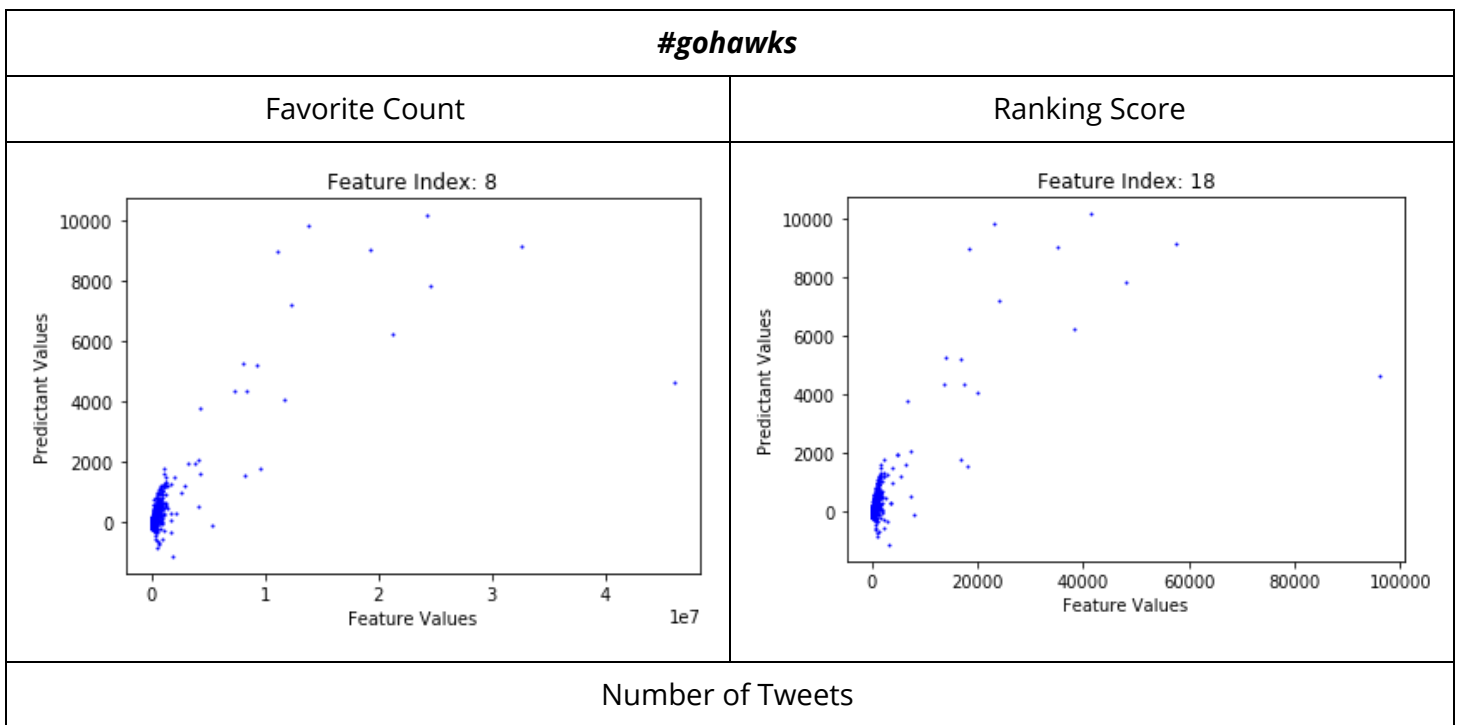
The best features that we selected for the hashtag 'sb49' are,

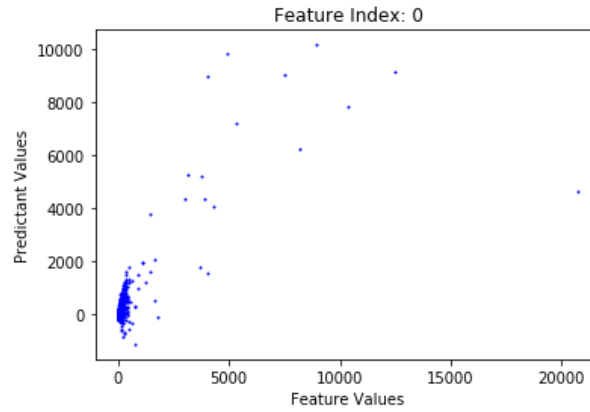
- Friend Count
- Favourite Count
- Status Count

The best features that we selected for the hashtag 'superbowl' are,

- Special Character Count
- Number of URLs
- Uppercase Characters

Selecting Best Features on the Basis of p-values

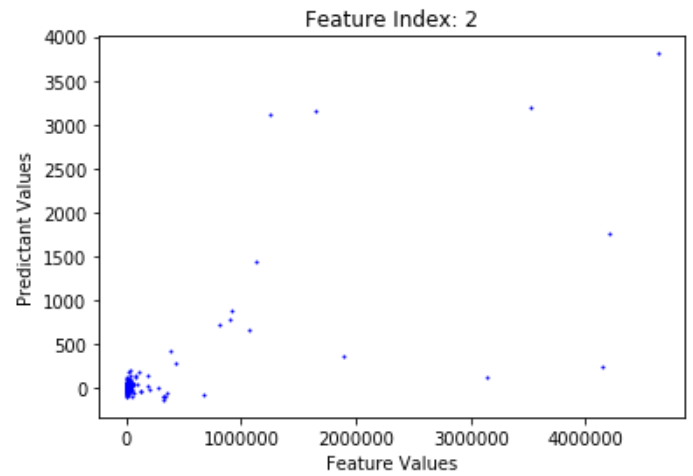
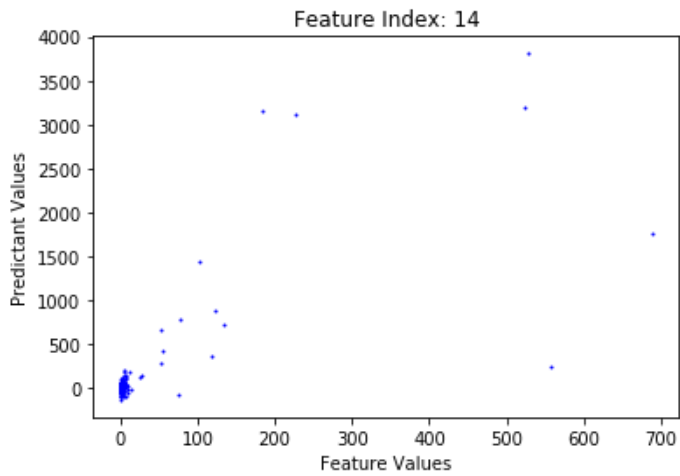




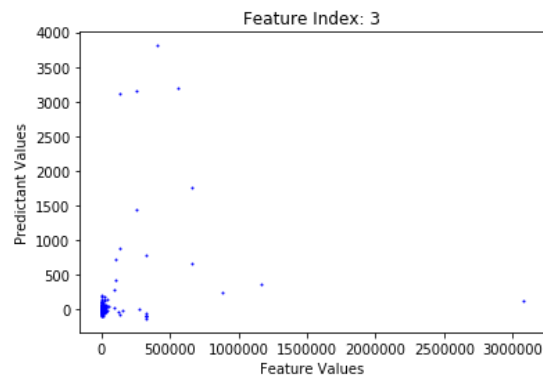
#gopatриots

Uppercase Characters

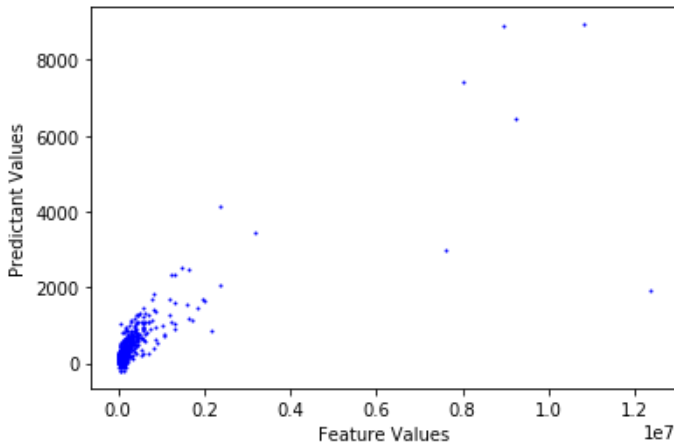
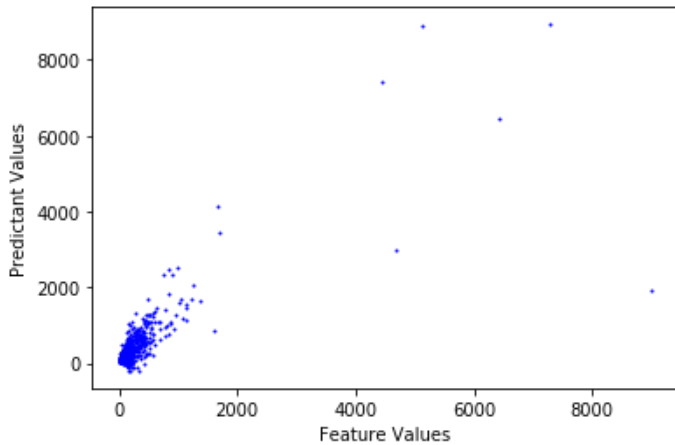
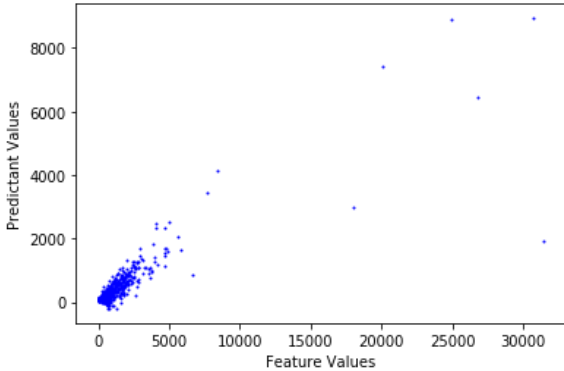
Sum of Number of Followers of the Users

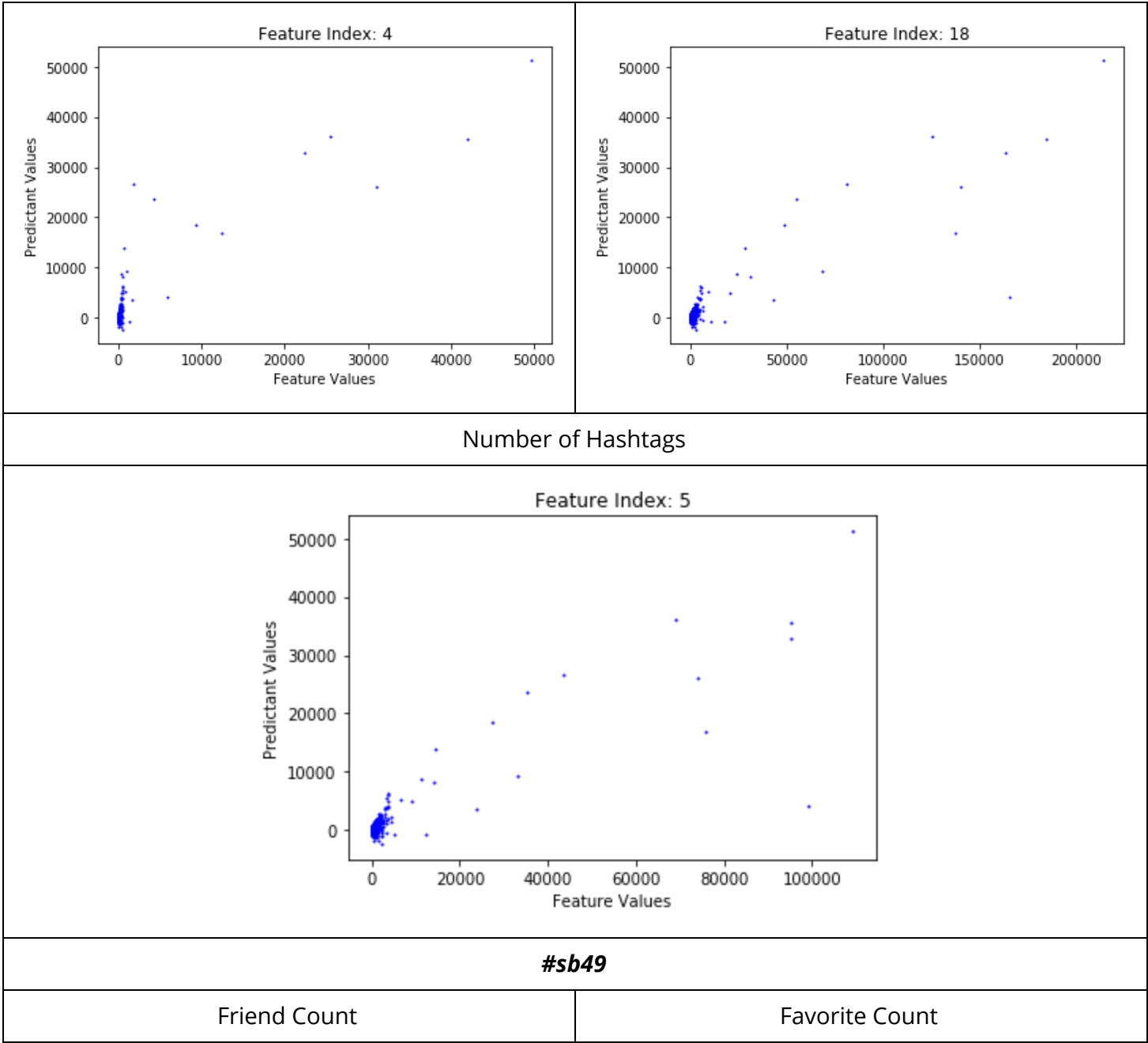


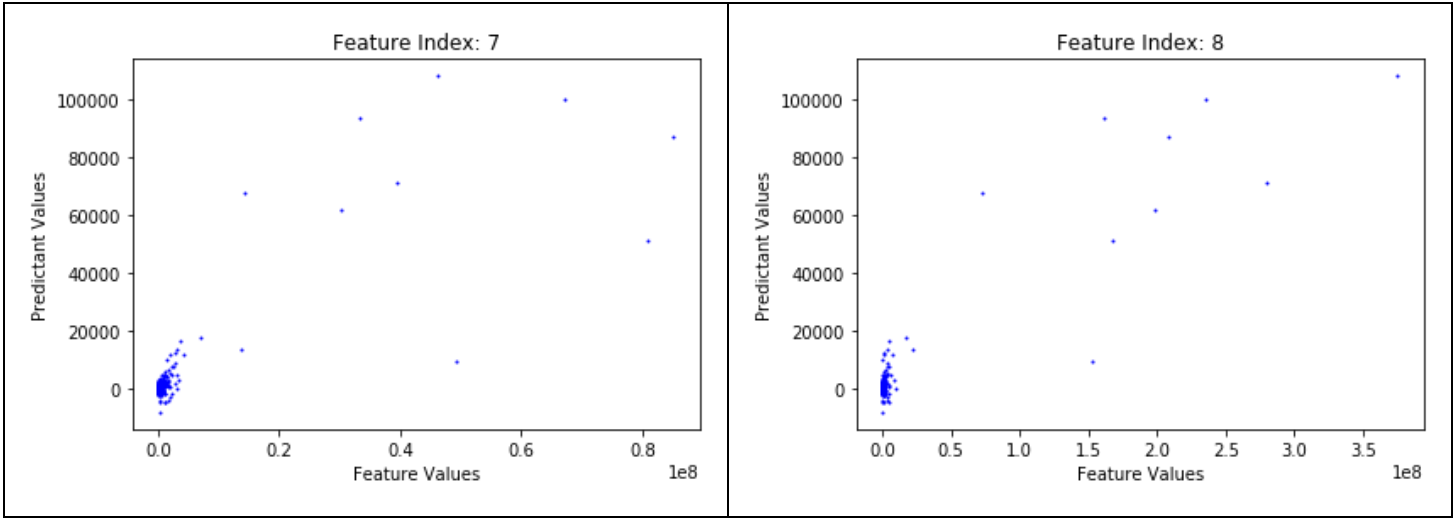
Maximum Number of Followers of the Users



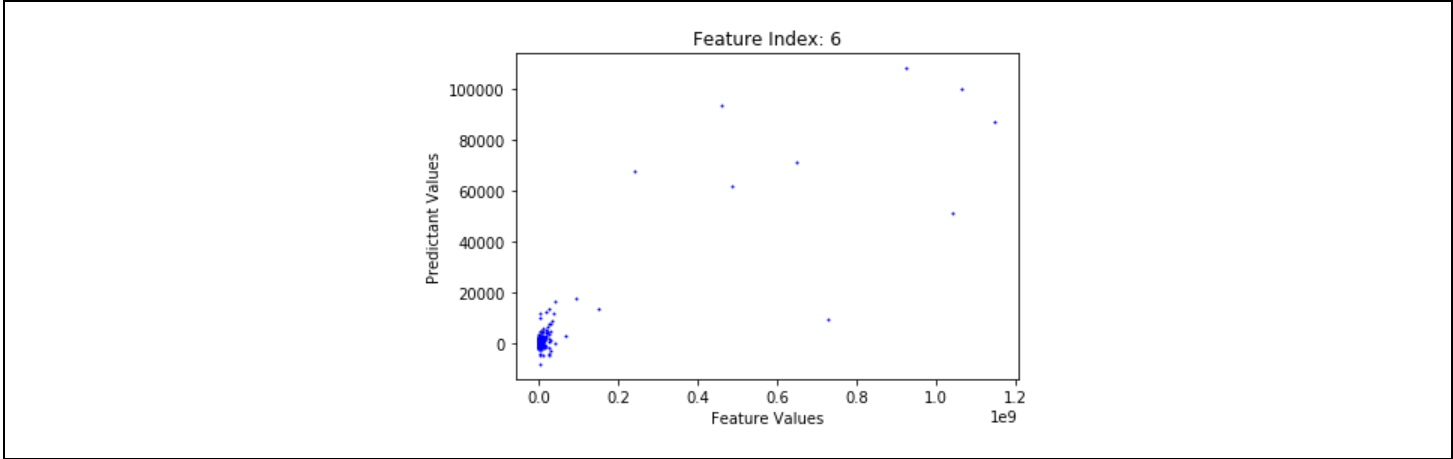
#nfl

Favorite Count		Number of Users	
<div>Feature Index: 8</div> 		<div>Feature Index: 10</div> 	
Number of Hashtags			
<div>Feature Index: 5</div> 			
#patriots			
Number of URLs		Ranking Score	

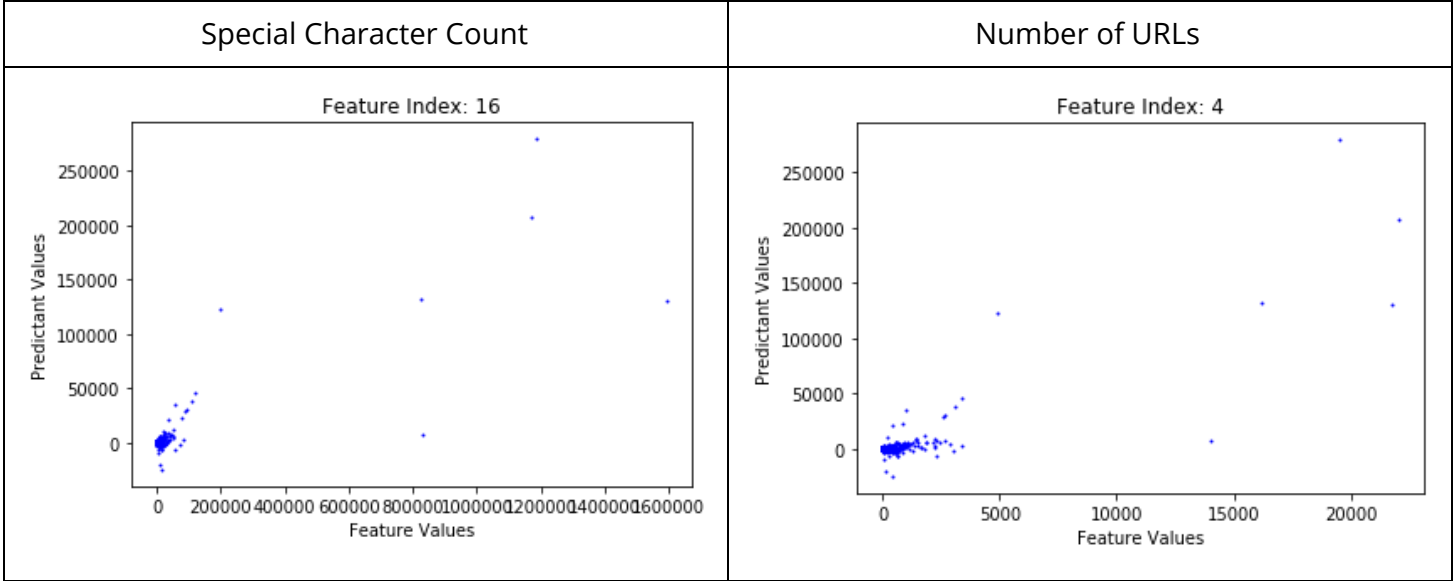




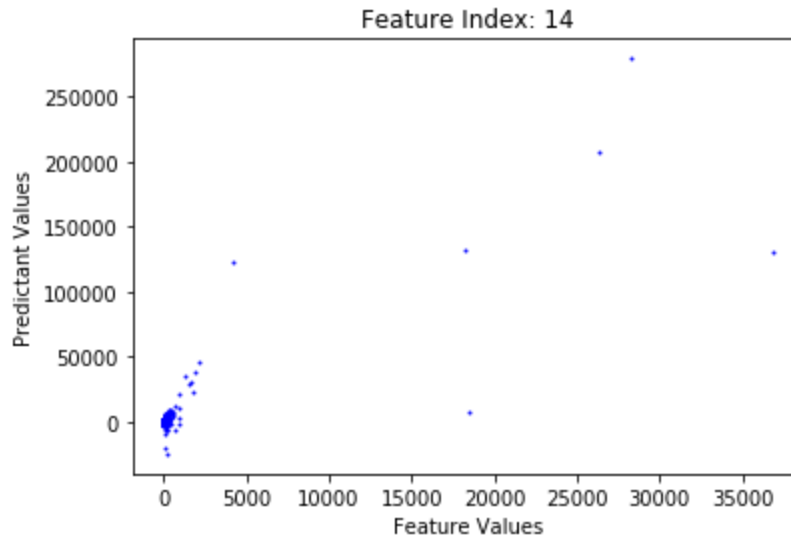
Status Count



#superbowl



Uppercase Characters



Problem 1.4

Since we know the Super Bowl's date and time, we can create different regression models for different periods of time. First, when the hashtags haven't become very active, second, their active period, and third, after they pass their high-activity time.

Train 3 regression models for these time periods (The times are all in PST):

1. Before Feb. 1, 8:00 a.m.
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.
3. After Feb. 1, 8:00 p.m.

For each hashtag, report the average cross-validation errors for the 3 different models. Note that you should do the 90-10% splitting for each model within its specific time window. I.e. Only use data within one of the 3 periods

above for training and testing each time, so for each period you will run 10 tests.

We have chosen Random Forest Regressor and K Neighbors Regressor as our two non linear models.

The parameters chosen for the Random Forest Regressor are:

- Number of estimators: 20
- Max features: 5
- Max depth: 4

The parameters chosen for the K Neighbors Regressor are:

- Number of Neighbors: 5

We have reported Training RMSE and Testing RMSE as the Cross Validation Errors for this part in the table below.

Hashtag	Model	Cross Validation Error					
		Period 1		Period 2		Period 3	
		Training RMSE	Testing RMSE	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE
#gohawks	Linear Regression	453.549710 2424257	4146.56939 720713	6.50090407 4106575e-0 9	4019.71143 35603314	11.5876752 85528842	521.349568 9915188
	Random Forest Regressor	407.729733 2154909	969.282869 8349082	1344.27900 7942247	3625.01420 84893955	26.8290503 93019447	77.1350603 071428
	K Neighbors Regressor	701.462072 1921294	870.361806 6333615	2764.86282 46031894	3561.99267 43327255	71.4026228 3896686	93.8957139 733444
#gopatriots	Linear Regression	31.1279033 39373198	63.8474519 27197106	9.30623087 5189484e-1 0	8886.69357 3324912	0.92253158 26606586	192.560358 62823623
	Random Forest Regressor	25.5158089 11693675	59.0068436 74629124	303.306423 32724454	619.714350 2251663	3.30695251 9004594	14.8335510 3105701
	K Neighbors	42.5709457	59.0677123	793.347199	1101.27351	12.1851250	16.8090856

	Regressor	41128355	17964386	127618	8250575	91182837	0791928
#nfl	Linear Regression	251.984017 58575207	444.468358 32242914	2.12313037 69324057e-08	10530.9659 01520514	130.446792 9470035	645.329656 7355821
	Random Forest Regressor	144.597372 70281903	283.031360 66275727	1053.77022 40304006	2649.50670 4732715	101.225564 94105886	175.625073 09959858
	K Neighbors Regressor	261.851316 84876103	329.852593 94387225	2614.09797 3367417	3170.05996 00007567	156.871986 01645997	190.077110 95734308
#patriots	Linear Regression	445.265168 7081368	857.975684 4991432	6.34190831 5568163e-08	78184.7269 9069651	70.7954188 7369925	459.643458 89365904
	Random Forest Regressor	328.958058 97366597	734.854039 8364116	6310.55479 45569815	15862.2428 15079295	75.8815525 843182	252.557755 05783325
	K Neighbors Regressor	620.558109 2526407	790.077900 07846	13652.0922 1730168	18267.6667 81119035	229.788550 8314379	287.419191 2425221
#sb49	Linear Regression	71.6980639 1596698	95.2909235 2057408	2.12837715 42645252e-07	206418.676 9925326	120.686806 63361361	429.057243 63239455
	Random Forest Regressor	56.8030111 373354	117.511050 40742449	16354.4092 39339624	34193.0382 9428051	126.638135 61553737	459.696926 387105
	K Neighbors Regressor	121.955871 92273727	154.346746 8329933	33682.1161 45522035	40723.0307 8770046	390.366535 926459	495.194070 11486226
#superbowl	Linear Regression	673.533388 75699	1544.43693 54086145	6.37056798 8470792e-07	581316.760 809673	292.071544 840946	2290.39261 4144926
	Random Forest Regressor	405.003192 63429094	872.634763 7200545	25804.7767 39116325	72116.9997 5920838	213.627652 73249013	407.206151 25612386
	K Neighbors Regressor	706.826198 5273781	897.635370 2582632	54908.5649 1204165	63435.1204 9034037	381.762429 07088385	473.789076 5466687

Hashtag	Best Model for Period 1	Best Model for Period 2	Best Model for Period 3
gohawks	K Neighbors Regressor	K Neighbors Regressor	Random Forest Regressor
gopatriots	Random Forest Regressor	Random Forest Regressor	Random Forest Regressor
nfl	Random Forest Regressor	Random Forest Regressor	Random Forest Regressor
patriots	Random Forest Regressor	Random Forest Regressor	Random Forest Regressor
sb49	Linear Regression	Random Forest Regressor	Linear Regression
superbowl	Random Forest Regressor	K Neighbors Regressor	Random Forest Regressor

Best Model for Period 1	Random Forest Regressor
Best Model for Period 2	Random Forest Regressor
Best Model for Period 3	Random Forest Regressor
Overall Best Model: Random Forest Regressor	

Also, aggregate the data of all hashtags, and train 3 models (for the intervals mentioned above) to predict the number of tweets in the next hour on the aggregated data.

Perform the same evaluations on your combined model and compare with models you trained for individual hashtags.

Here, we have trained 3 models on the combined data from the 6 hashtags for each period.

Model	Cross Validation Error					
	Period 1		Period 2		Period 3	
	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE	Training RMSE	Testing RMSE
Random Forest Regressor	1075.3056954 531914	2573.5928223 65316	38174.187204 96111	98382.388738 91965	374.10668866 256066	1041.7963419 621422

Observation:

- On preparing a model on the combined data for all the tags, we observe a much higher RMSE as compared to the individual models for each hashtag.
- This is expected as the number of tweets for each hashtag span a different range and may follow some distribution. The model used on all the combined tweets may not be able to fit this combined data.
- Hence, the individual model for each hashtag performs better achieving a lesser RMSE.

Problem 1.5

Download the test data3. Each file in the test data contains a hashtag's tweets for a 6-hour window (note that these hashtags are different from those in training data). Fit a model on the aggregate of the training data for all hashtags, and predict the number of tweets in the next hour for each test file. The file names show sample number followed by the period number the data is from. E.g. a file named sample5 period2.txt contains tweets for a 6-hour window that lies in the 2nd time period described in part 4. One can be creative here, and use the data from all previous 6 hours for making more

accurate predictions (as opposed to using features from the previous hour only).

Q: Report the model you use. For each test file, provide your predictions on the number of tweets in the next hour.

For this problem, we have combined all our training data and grouped by 'First Post Date' and created three models corresponding to each time period. For prediction, the model used here is the Random Forest Regressor obtained from the previous part. The result will vary depending on the type of features being used for the model.

The test data was also grouped according to 'First Post Date'

The predictions for the final hour of each file are as follows:

File Name	Window Period	Prediction of the Last Hour
sample10_period3.txt	3	103.88382112
sample1_period1.txt	1	128.65803446
sample2_period2.txt	2	147071.3
sample3_period3.txt	3	706.90925796
sample4_period1.txt	1	270.53596959
sample5_period1.txt	1	179.66642824
sample6_period2.txt	2	148723.75
sample7_period3.txt	3	103.88382112
sample8_period1.txt	1	46.96192049
sample9_period2.txt	2	133140

Part 2: Fan Base Prediction

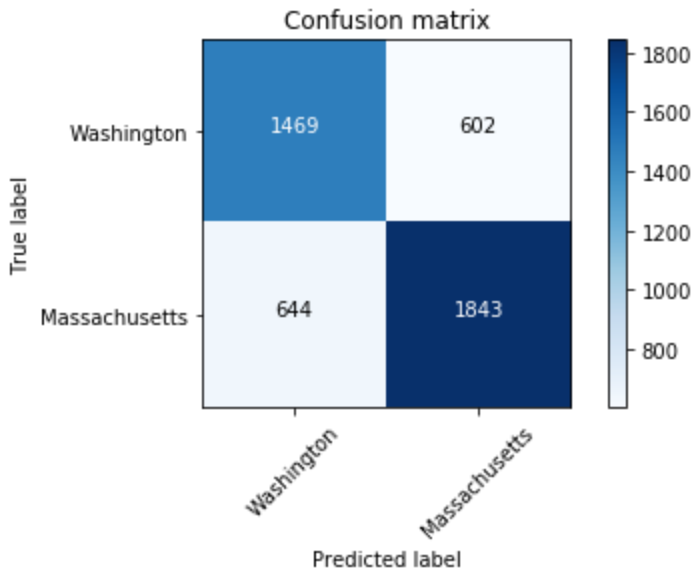
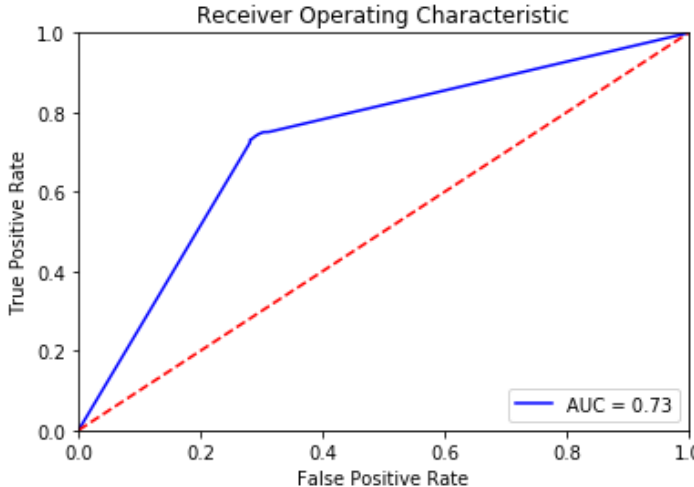
In this task we are trying to predict the location for a user using just the text of the tweet.

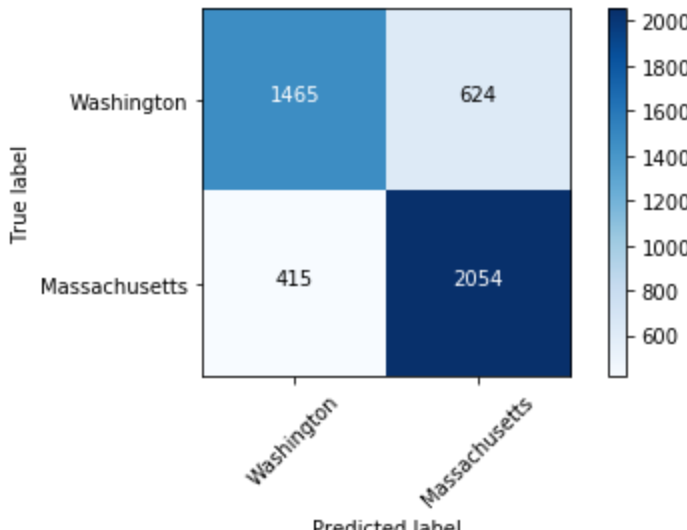
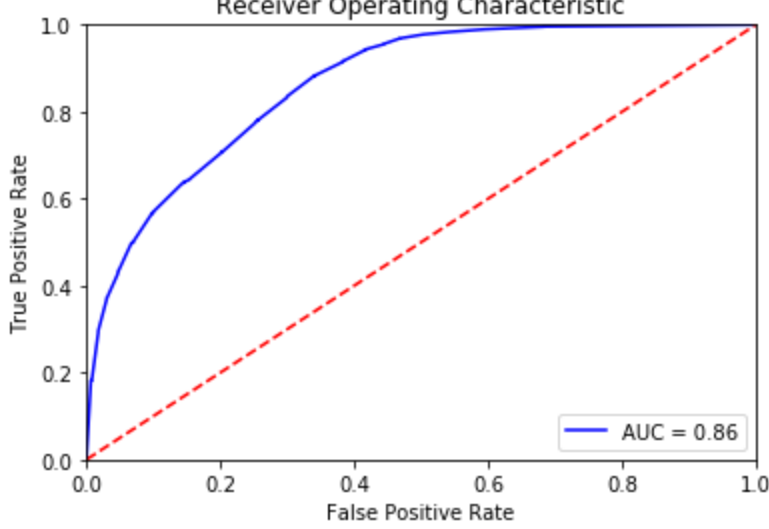
- We use `tweet['title']` for the tweet text and `tweet['tweet']['user']['location']` for the location true label
- The location true label first needs to be converted to either of the Washington class(Class 0) or the Massachusetts class(Class 1)
- The location data also contains the names of cities in the the 2 states so we have first used regular expressions to process this location to convert it one the 2 classes.
- At the end of processing, we have got a well balanced dataset with
 - **Total Washington(Class 0) labels: 20984**
 - **Total Massachusetts(Class 1) labels: 24590**
- For processing the tweet text data, we have used the techniques that we have mastered in Project 1
- We first preprocess the tweet text data to remove some unnecessary stuff.
- For the countvectorizer, we have used `min_df=2`.
- We have experimented with both LSI and NMF techniques of dimensionality reduction to reduce the features from TF-IDF transformer to 50 components
- For the actual classification task we have used 3 models:
 - **DecisionTreeClassifier:** Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features
 - **RandomForestClassifier:** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
 - **KNeighborsClassifier:** Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class

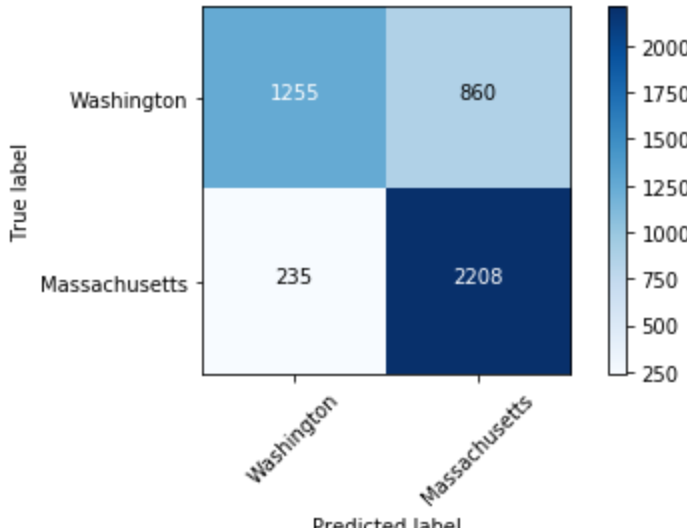
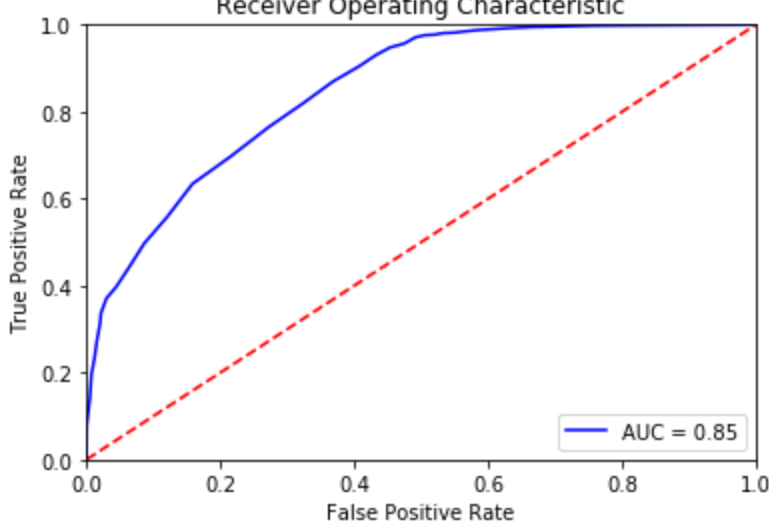
which has the most representatives within the nearest neighbors of the point.

- We first use a grid search technique along with 5 fold cross validation to get the best parameters for our models.
- We then split the dataset into 90% training and 10% testing and use the best model that we have found to find the performance of our modelling
- Below are the results for all the models for both our approaches.
- On analyzing the results of all the 3 models, we see that results for all the 3 models are comparable with the RandomForestClassifier performing the best.
- Both the LSI and NMF dimensionality reduction techniques perform equally well with NMF performing slightly better.

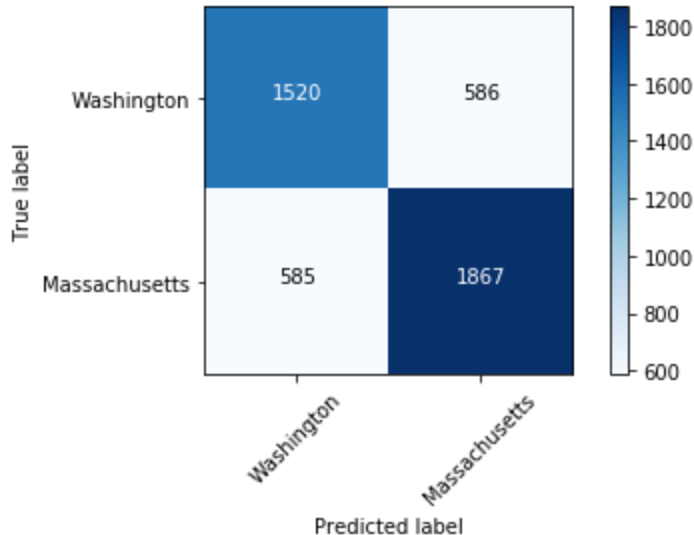
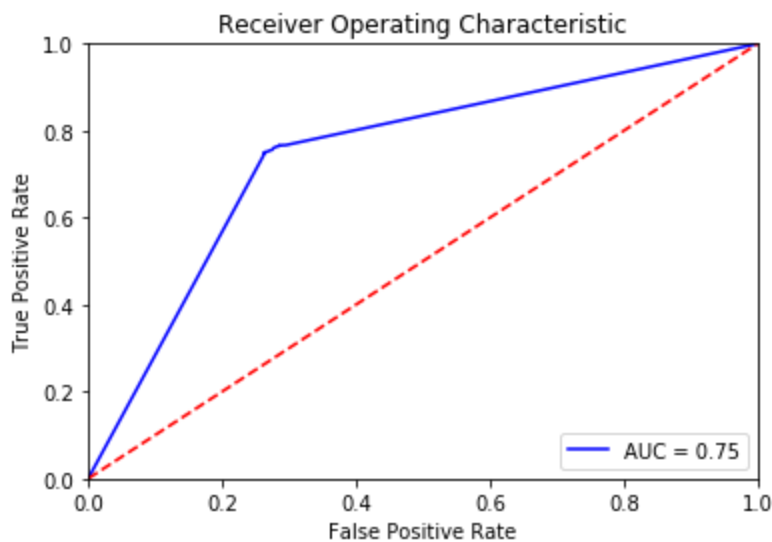
min_df=2 with SVD dimensionality reduction technique(50 components)

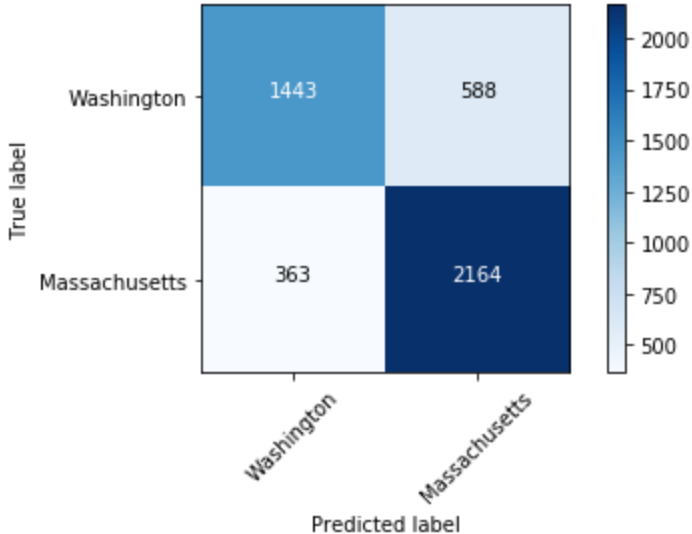
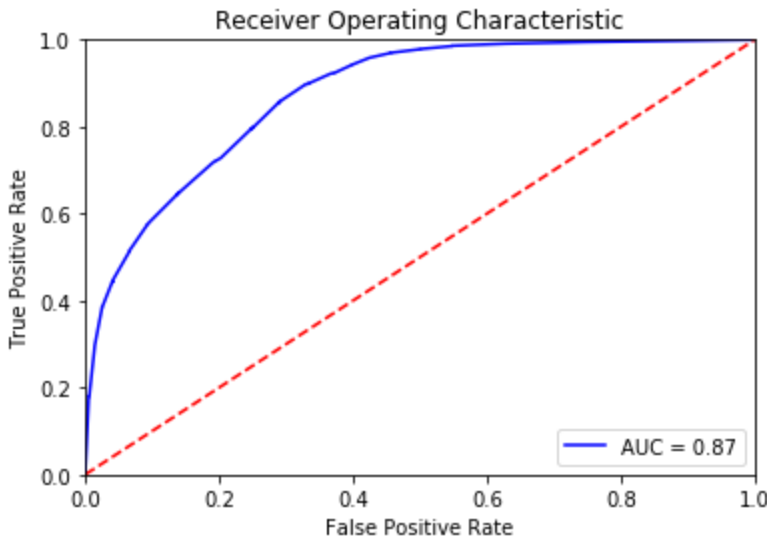
Model 1	DecisionTreeClassifier
Best Args	max_features = 19
Accuracy	72.66
Precision	75.37
Recall	74.10
F1	74.73
Confusion Matrix	 <p>Confusion matrix</p> <p>True label</p> <p>Washington</p> <p>Massachusetts</p> <p>Predicted label</p> <p>Washington</p> <p>Massachusetts</p> <p>1469</p> <p>602</p> <p>644</p> <p>1843</p> <p>1800</p> <p>1600</p> <p>1400</p> <p>1200</p> <p>1000</p> <p>800</p>
ROC Curve	 <p>Receiver Operating Characteristic</p> <p>True Positive Rate</p> <p>False Positive Rate</p> <p>AUC = 0.73</p>

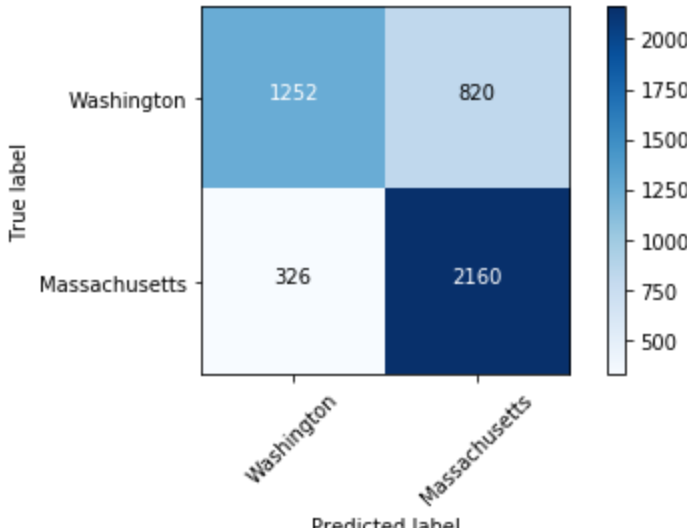
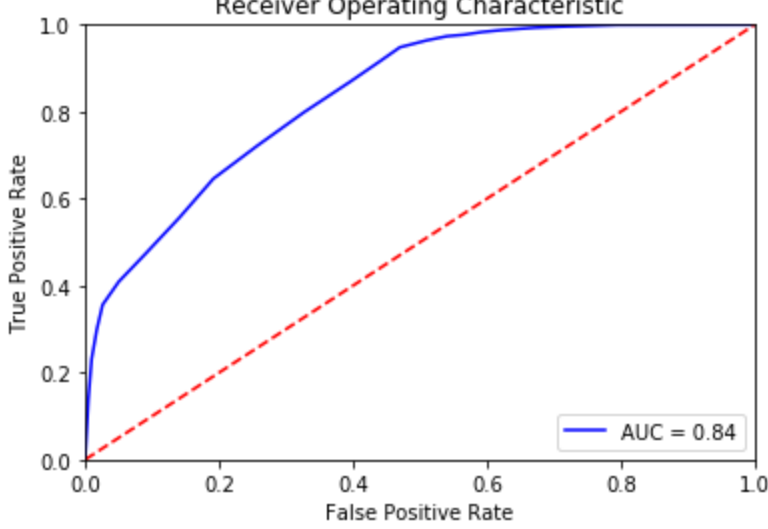
Model 2	RandomForestClassifier									
Best Args	n_estimators=19, max_features=41									
Accuracy	77.20									
Precision	76.69									
Recall	83.19									
F1	79.81									
Confusion Matrix	<div><p>Confusion matrix</p><p>The confusion matrix is a 2x2 heatmap. The y-axis is labeled 'True label' with categories 'Washington' and 'Massachusetts'. The x-axis is labeled 'Predicted label' with categories 'Washington' and 'Massachusetts'. The cells contain the following counts: True Washington & Predicted Washington: 1465; True Washington & Predicted Massachusetts: 624; True Massachusetts & Predicted Washington: 415; True Massachusetts & Predicted Massachusetts: 2054. A color bar on the right indicates the count scale from 600 to 2000.</p><table><tr><th></th><th>Washington</th><th>Massachusetts</th></tr><tr><th>Washington</th><td>1465</td><td>624</td></tr><tr><th>Massachusetts</th><td>415</td><td>2054</td></tr></table></div>		Washington	Massachusetts	Washington	1465	624	Massachusetts	415	2054
	Washington	Massachusetts								
Washington	1465	624								
Massachusetts	415	2054								
ROC Curve	<div><p>Receiver Operating Characteristic</p><p>The ROC curve plot shows the True Positive Rate (y-axis) versus the False Positive Rate (x-axis). A solid blue curve represents the model's performance, starting at (0,0) and ending at (1,1). A dashed red diagonal line represents the baseline performance of a random classifier. The area under the blue curve is labeled as AUC = 0.86.</p><p>True Positive Rate</p><p>False Positive Rate</p><p>AUC = 0.86</p></div>									

Model 3	KNeighborsClassifier									
Best Args	n_neighbours = 35									
Accuracy	75.97									
Precision	71.96									
Recall	90.38									
F1	80.13									
Confusion Matrix	<div><p>Confusion matrix</p><table><tr><th></th><th>Washington</th><th>Massachusetts</th></tr><tr><th>Washington</th><td>1255</td><td>860</td></tr><tr><th>Massachusetts</th><td>235</td><td>2208</td></tr></table></div>		Washington	Massachusetts	Washington	1255	860	Massachusetts	235	2208
	Washington	Massachusetts								
Washington	1255	860								
Massachusetts	235	2208								
ROC Curve	<div><p>Receiver Operating Characteristic</p><p>AUC = 0.85</p></div>									

min_df=2 with NMF dimensionality reduction technique(50 components)

Model 1	DecisionTreeClassifier
Best Args	max_features = 39
Accuracy	74.308
Precision	76.11
Recall	76.1419
F1	76.1264
Confusion Matrix	<div><p>Confusion matrix</p><p>True label</p><p>Washington</p><p>Massachusetts</p><p>Predicted label</p><p>Washington</p><p>Massachusetts</p><p>1520</p><p>586</p><p>585</p><p>1867</p><p>1800</p><p>1600</p><p>1400</p><p>1200</p><p>1000</p><p>800</p><p>600</p></div>
ROC Curve	<div><p>Receiver Operating Characteristic</p><p>True Positive Rate</p><p>False Positive Rate</p><p>AUC = 0.75</p></div>

Model 2	RandomForestClassifier									
Best Args	n_estimators=19, max_features=9									
Accuracy	79.1355									
Precision	78.6337									
Recall	85.6351									
F1	81.9852									
Confusion Matrix	<div><p>Confusion matrix</p><p>The confusion matrix is a 2x2 heatmap titled 'Confusion matrix'. The y-axis is labeled 'True label' with categories 'Washington' and 'Massachusetts'. The x-axis is labeled 'Predicted label' with categories 'Washington' and 'Massachusetts'. The cells contain the following counts: True Washington & Predicted Washington: 1443; True Washington & Predicted Massachusetts: 588; True Massachusetts & Predicted Washington: 363; True Massachusetts & Predicted Massachusetts: 2164. A color bar on the right indicates the count scale from 500 to 2000.</p><table><tr><th></th><th>Washington</th><th>Massachusetts</th></tr><tr><th>Washington</th><td>1443</td><td>588</td></tr><tr><th>Massachusetts</th><td>363</td><td>2164</td></tr></table></div>		Washington	Massachusetts	Washington	1443	588	Massachusetts	363	2164
	Washington	Massachusetts								
Washington	1443	588								
Massachusetts	363	2164								
ROC Curve	<div><p>Receiver Operating Characteristic</p><p>The ROC curve plot is titled 'Receiver Operating Characteristic'. The y-axis is 'True Positive Rate' and the x-axis is 'False Positive Rate', both ranging from 0.0 to 1.0. A solid blue curve represents the model's performance, starting at (0,0) and ending at (1,1). A dashed red diagonal line represents random performance. A legend in the bottom right corner indicates 'AUC = 0.87'.</p></div>									

Model 3	KNeighborsClassifier									
Best Args	n_neighbors=21									
Accuracy	74.8573									
Precision	72.4832									
Recall	86.8865									
F1	79.0340									
Confusion Matrix	<div><p>Confusion matrix</p><table><tr><th></th><th>Washington</th><th>Massachusetts</th></tr><tr><th>Washington</th><td>1252</td><td>820</td></tr><tr><th>Massachusetts</th><td>326</td><td>2160</td></tr></table></div>		Washington	Massachusetts	Washington	1252	820	Massachusetts	326	2160
	Washington	Massachusetts								
Washington	1252	820								
Massachusetts	326	2160								
ROC Curve	<div><p>Receiver Operating Characteristic</p><p>AUC = 0.84</p></div>									

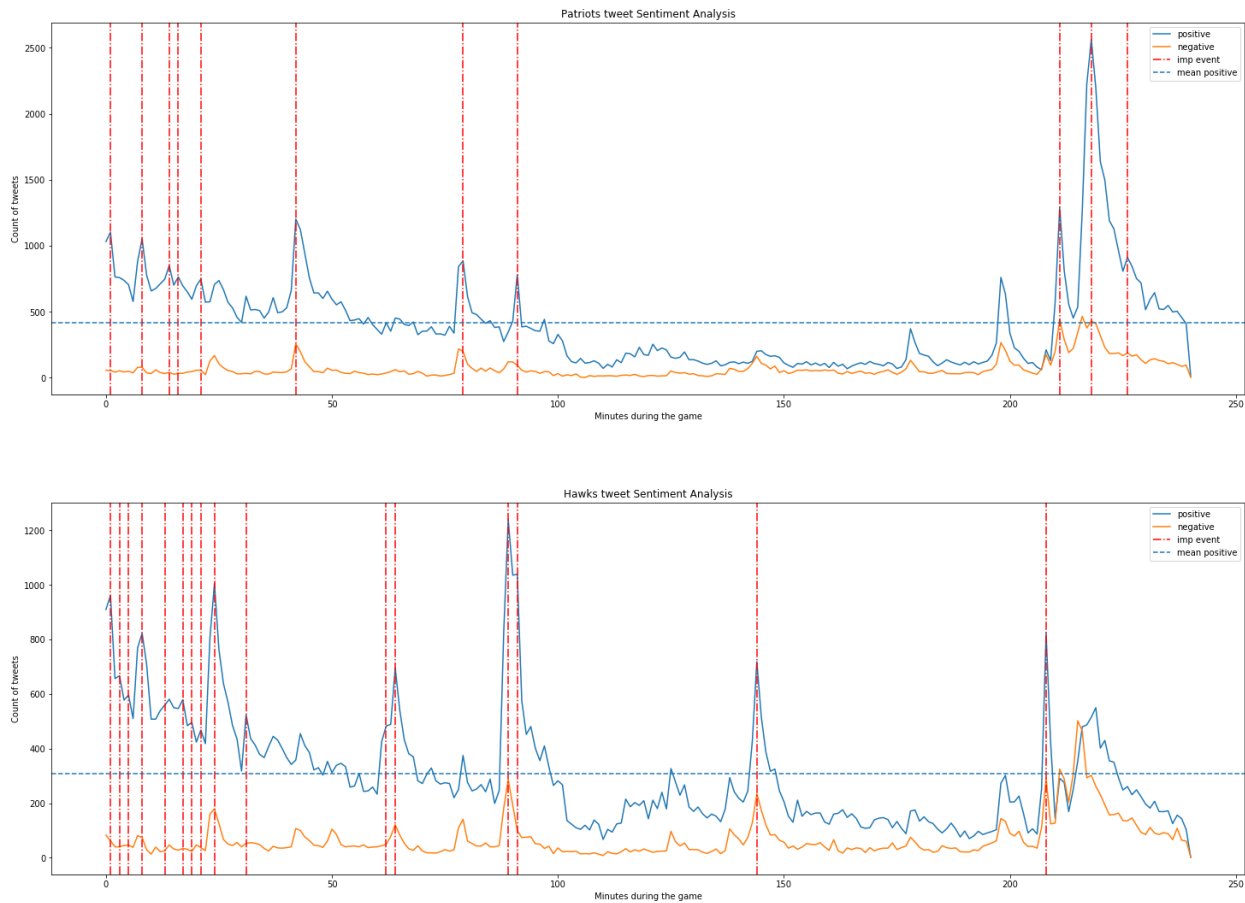
Problem 3

This was a very interesting part where we could effectively do anything using the rich dataset of super bowl tweets. We came up with many ideas of what could be done and have designed and implemented 3 tasks in this problem as below.

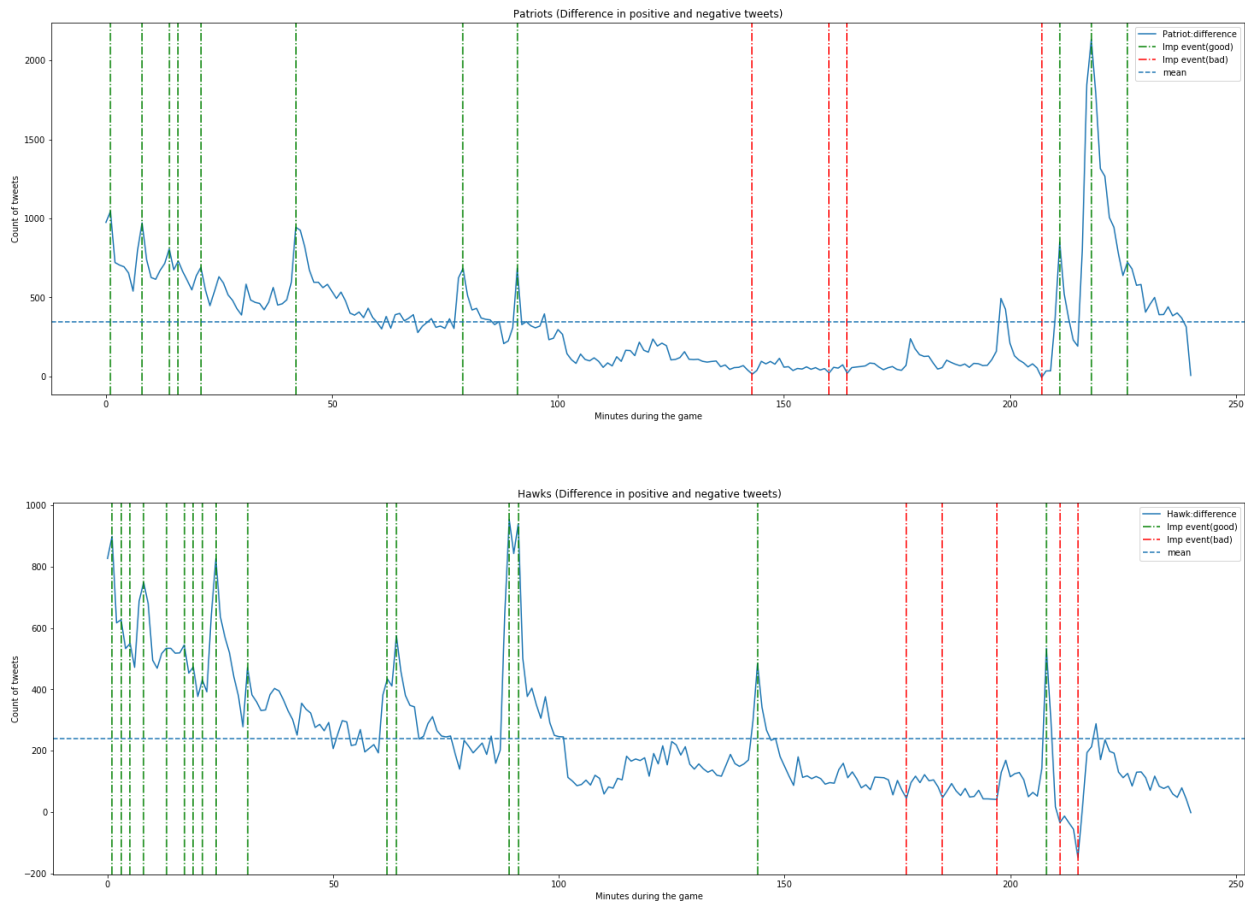
First Task: Important Events analysis

In this task, we have analyzed the tweet data and sentiments from the tweet data to detect when some important events that have taken place during the duration of the game.

- We first extracted the tweets that were posted during the game. The game happened roughly between 3:30 PM and 7:30 PM on 1st Feb 2015. After extracting the tweets for this duration we segregated the tweets for both the teams, Hawks and Patriots.
- After extracting the tweets for both the teams, we ran sentiment analysis on the tweets and got the positive, negative and neutral count of tweets for both the teams.
- For sentiment analysis, we have used the VADER Sentiment Analysis library. Of all the sentiment analysis libraries we have used VADER because it has been trained on social media platforms and twitter datasets as well. The library also takes emojis into consideration while trying to find the sentiments of the tweets.
- With the use of these counts we tried to analyse the key events that happened during the game for both the teams.
- The graph below shows the positive and negative tweets for each team.



- Looking at the above graphs we realised that people have a natural tendency to tweet positive stuff more than the negative stuff.
- We have then used the difference between the positive and negative tweets for further analysis. We plot the graph of this difference to get some interesting results.
- The key idea here is to find the highs and lows of this difference for both the teams and then compare and correlate the results.
- So using the per minute difference graph, we have found the min and max peaks in this graph for both the teams. The time when one team has a max peak and other team has a min peak is the event during the game we are looking for.
- Following are the graphs showing this difference graph along with the min and max peaks.
- The graphs also show some important event that are described below.



- After seeing the graph of the good and bad important events for both the teams we will now analyse their meaning.
- When there is a good important event for both the teams at the same time, that means there is a common important event for both the teams. This is seen at the start of the game, where fans of both the teams are super excited at the start of the game. This is also seen at the mid of the graph which corresponds to the halftime show where fans of both teams are equally delighted and posting a lot of positive tweets about their favorite singers and performers who performed during the halftime show.
- When one of team has an important good event and the other team at the same time has an important bad event, that would imply that there is some other key event that has taken place, most likely one of the team has scored a goal or touchdown, that has caused such changes in the sentiments. This can be seen in the rightmost part of the graph(the end of 4th quarter) which actually correspond to the

2-3 touchdowns which happened at the end of the game, one of which with a very high peak for Patriots and very low peak for the Hawks was the game turning touchdown which helped the Patriots win the Superbowl.

- We checked the results we have got from our graph with the actual highlights of the game to see that our results have successfully detected the important events.

News Articles from the web

Super Bowl XLIX: New England Patriots 28-24 Seattle Seahawks - as it happened!

- **New England Patriots win fourth Super Bowl**
- **Controversial play call dooms Seahawks bid at back-to-back titles**
- **Tom Brady named MVP**

Key action

- **Patriots win their first Super Bowl in 10 years**
- **Malcolm Butler intercepts a one-yard Seahawks pass on the goal-line with 20 seconds left**
- **Patriots quarterback Tom Brady throws four touchdown passes**
- **Seahawks came back from 7-0 and 14-7 down to lead 24-14**
- **Patriots were 10 points down midway through the fourth quarter**
- **Brady throws a winning three-yard pass to Julian Edelman with two minutes, two seconds left**

SUPER BOWL 2015 New England Patriots 28-24 Seattle Seahawks: Malcolm Butler's last-gasp interception gives Tom Brady an elusive fourth NFL championship victory

- New England Patriots beat Seattle Seahawks 28-24 to win Super Bowl XLIX in Glendale, Arizona
- Malcolm Butler makes vital interception in the dying seconds to deny Seahawks
- Patriots quarterback Tom Brady wins his fourth Super Bowl ring in dramatic match
- [Katy Perry performed half-time show at the University of Phoenix Stadium](#)

Second Task: People that were most tweeted

In this task we have tried to get information regarding the people who were most tweeted. Given a stream of tweets we try to find who the tweets are talking about and whether the masses are tweeting about these people, players, celebrities in a positive or negative way.

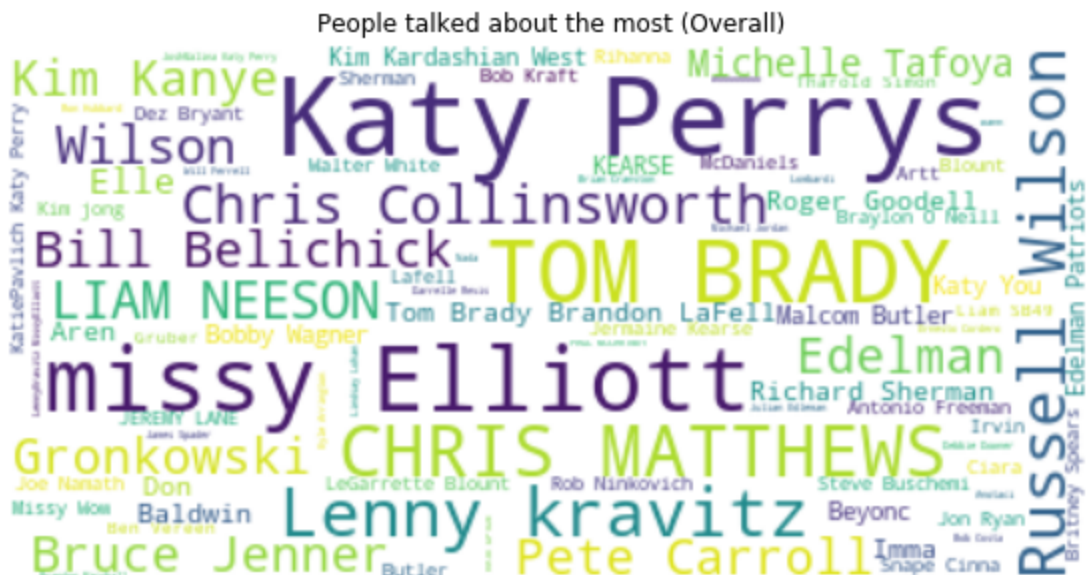
The motivation behind this task, is that during nearly every game in almost all the sports, there is some key player that performs good or even bad which decides the outcome of the event. We also can find any important people that were present during the event.

- First, we extracted the Person names from the data using the Named Entity Recognition(NER) tagger from the Stanford CoreNLP library. The reason we used Stanford CoreNLP is that it is known to have a high precision among most other libraries.
- Some preprocessing had to be done before extracting the entities as tweet data contains a lot of slang, emojis etc.
- After extracting the entities there were some duplicates in which the complete name of the person wasn't mentioned in the tweet or there were typos in the name of the same person. Example, some tweets mentioned '**Russell Wilson**' and some mention just '**Wilson**', some tweets mentioned **Katy Perry** and some others mentioned just **Katy** or even **Katie, Kati, KatyPary** etc.
- So to resolve this problem we devised an intelligent algorithm to merge the similar entities using Levenshtein distance between the entities to find similarities between the entities.

- Below are the top people we have got.

Katy Perrys	Russell Wilson
missy Elliott	Chris Collinsworth
TOM BRADY	LIAM NEESON
CHRIS MATTHEWS	Edelman
Lenny kravitz	Pete Carroll

- We observed that the top entities we got were mostly celebrities who either performed or attended the event. This makes sense because in this case supporters from both the teams tweeted about various famous personalities like Katy Perry and Missy Elliott and hence they were among the top people that we got from the tweets.
- Apart from the celebrities, we also have the quarterbacks of both the teams Tom Brady and Russell Wilson in the top people who are tweeted.
- We visualized the results in a better way in a word cloud which covers the top 100 people.



News Articles from the web



Super Bowl 2015: Katy Perry packs it all into fiery halftime show

Los Angeles Times - Feb 2, 2015

Folks from that camp no doubt enjoyed "Roar," with its lyrical nod to Survivor's "Eye of the Tiger," and "I Kissed a Girl," which **Perry** transformed into a blistering grunge jam complete with cameo by a leather-jacketed Lenny Kravitz. (You got the sense that Kravitz, who hasn't had a big hit in ages, was there for ...

Katy Perry and friends put on an underwhelming Super Bowl ...

Blog - Entertainment Weekly (blog) - Feb 2, 2015

[View all](#)

Missy Elliott Will Join Katy Perry for Super Bowl Halftime Performance

Rapper will be a "surprise guest" for Perry's February 1st set, which also features rock legend Lenny Kravitz

NBC's Collinsworth shined as much as Tom Brady on Super Bowl Sunday



Boy with prosthetic legs from Super Bowl ad is 'differently-abled ...

Today.com - Jul 10, 2015

Braylon O'Neill is just like any other sports-loving child his age — except this 6-year-old got to star in an inspiring **Super Bowl** commercial that focused on what his ... The **Super Bowl** ad featuring **Braylon** shows the boy actively playing sports with artificial legs crafted with the help of Microsoft technology.

Third Task: Organizations that were most tweeted

In this task we have tried to get information regarding the organizations and brands which were most tweeted. Given a stream of tweets we try to find which organizations the tweets are talking about and whether the masses are tweeting about these organizations, brands, sponsors in a positive or negative way.

The motivation behind this task, is that for such a big event like SuperBowl there are many organizations. Brands and sponsors who are involved. These sponsors usually come up with great ideas for the events. Getting a good insight on how this is perceived by the audience can be possible with the twitter data.

- We used a similar approach as used in the previous task to use the Named Entity Recognition(NER) tagger from the Stanford CoreNLP library to tag Organizations.
- We used a similar preprocessing of tweet data.
- Similar to the previous task, to club similar entities like BBC and BBC News, Seahawks and Seattle Seahawks, etc, we used the Levenshtein ratio to find similarities between the entities.
- Below are the top Organizations we have found.

Les Seahawks	NBC NFL
Edelman Patriots	BBC Breaking
Seattle NFL	NISSAN
McDonald	Walmart
PSI	Microsoft

- As the teams-Seahawks and Patriots are the finalists in this SuperBowl it is clear that they would be in the top entities. We also see some other sponsors like McDonald, PSI, Walmart, Honda etc.

News Articles from the web

Nissan

Super Bowl 2015 appears to be the the year of the dadvertising. Nissan's 90-second ad is the *Boyhood* of spots, following a family over the course of a boy's childhood as he watches his Dad compete on the NASCAR circuit. The cinematography is good, it's pretty heartwarming, but the story-telling is only so-so.

McDonald's

McDonald's has incredible news for America: It will now accept hugs as payment for Chicken McNuggets. The Pay with Lovin' campaign lets select customers use "Lovin'" as currency. This includes compliments, silly dances, and calling your mom just to say "I love you." It's sweet.

Microsoft

Microsoft's "empowering" commercial tells the story of a little boy who hasn't let the fact that he has two prosthetic legs hold him back. With the help of Microsoft-powered technology, Braylon runs relay races, plays tennis and...warms your heart.

Remarks:

- Using just the tweet textual data and volume of data, a lot of analysis can be done. Sentiment analysis is definitely a good tool to get more insights in the data.
- All the 3 tasks we did as part of problem 3 are very flexible and not fixed to just the superbowl dataset. The solutions we have developed can be used and even extended for any datasets to extract and gain useful information.
- This can have a lot of applications especially in the advertising industry.