EE219 Project 5
Popularity Prediction on Twitter

Winter 2017

Boyang Cai 304330123
Manni Chen 304145309
Zhuoqi Li 004855607

Introduction

Twitter has been a super popular social media platform since its debut. It has become a one of most favorite way for the public to share opinion, attitudes and information toward trending topics. Due to the rich content carried on twitter, by performing an large-scale data analysis over what people have been sharing about, one would know the hot topic undergoing and people's attitude towards them. Additionally, it is also possible to perform a fairly-accurate prediction on the next trending topic via generalization and analysis of information from the past.

In this project, our objective is to predict the number of tweets posted in the next hour of a certain topic related to 2015 superbowel using tweets collected in a three-week time window. In order to perform the prediction, we will need to perform a linear regression on the collected data, and then use this same model to predict for the next hour.

Part 1. Data Statistics

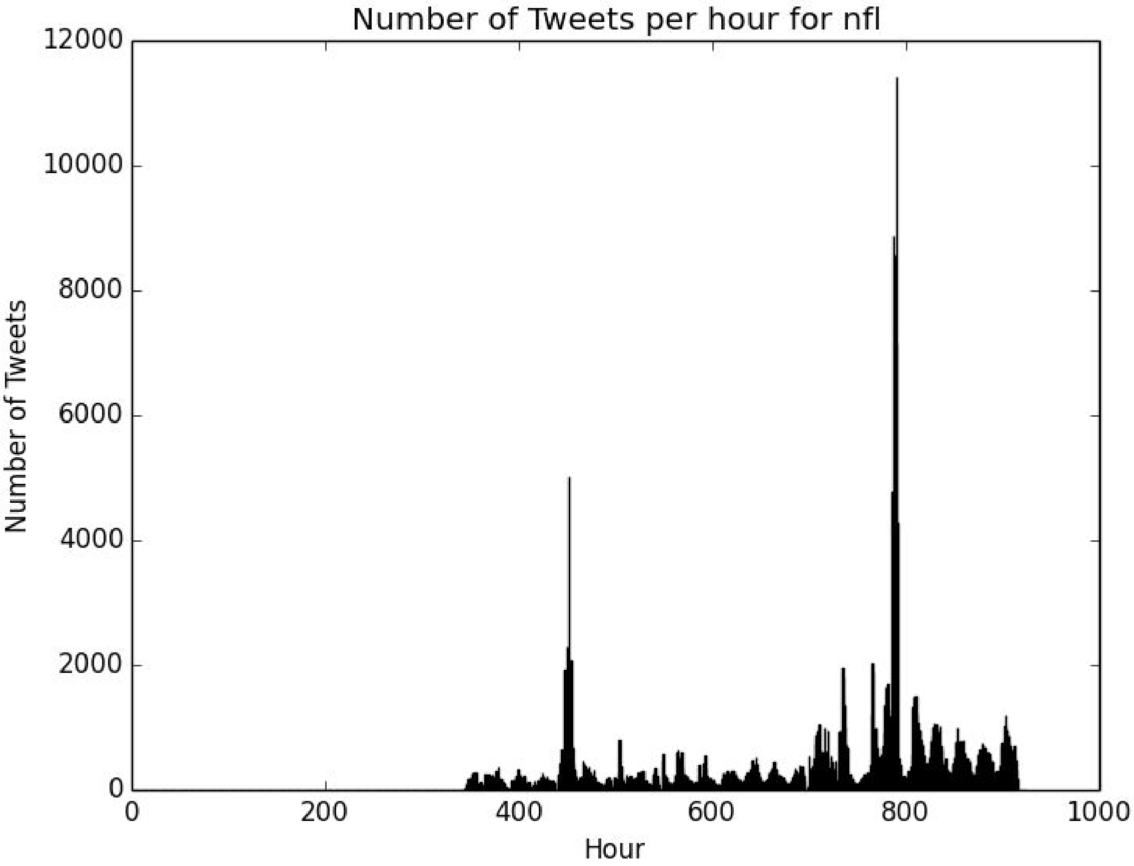
The following table showcases three statistics of tweet data files we are going to use later: average number of tweets per hour (Total number of tweets / total number of hours), average number of followers per unique user (Total number of followers / total unique user number) and average number of retweets per tweet (Total number of retweets / total number of tweets).

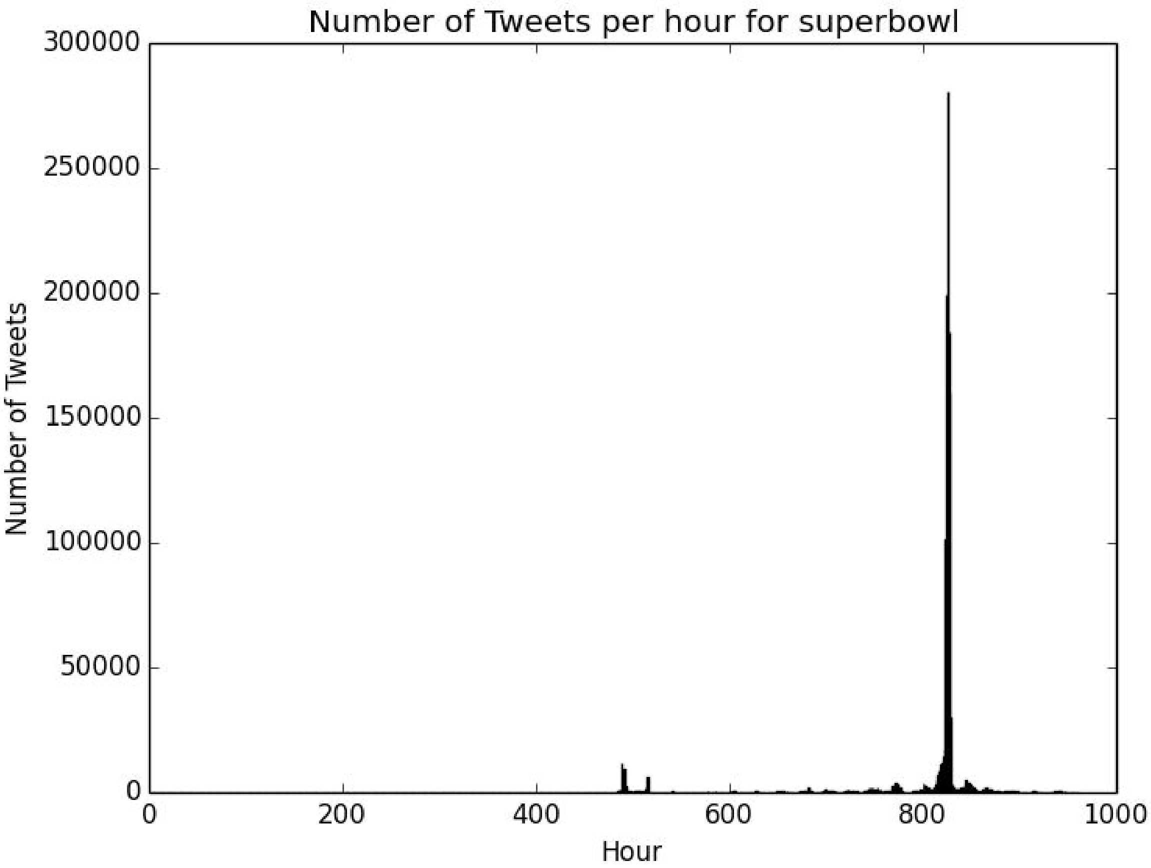
Hashtag	Avg # of tweets	Avg # of followers	Avg # of retweets
#gohawks	193.54	1596.44	2.01
#gopatriots	38.41	1292.2	1.4
#nfl	279.42	4394.25	1.54
#patriots	499.2	1607.44	1.78
#sb49	1420.12	2229.69	2.51
#superbowel	1400.56	3675.33	2.39

According to the table above, from these three features, hashtag #sb49 has the most average number of tweets posted each hour and the most average number of retweets per tweet. This stats shows that it is the most popular topics among our chosen hashtags. It actually makes sense because the topic “superbowl” should be the most discussed topic. Its variant hashtag #superbowl also shares similar popularity. Although its average number of tweets and retweets is slightly less but still largely exceeds those of other hashtags. The reason behind why #sb49 is more popular than #superbowl is probably because #sb49 is shorter and easier to type in tweet.

However, as for the average number of followers, #nfl has the most number. It indicates that for those who engaged in our chosen hashtags, there are more celebrities involved in the topic of "nfl". Or #nfl simply has a higher celebrity ratio among the population involved in the hashtag tweeting.

Two diagrams below display the number of tweets per hour for hastags #nfl and #superbowl over the timespan of our collected data. To better analyze information contained in the giant amount of tweets data collected, tweets data are broken into hourly frames to be analyzed. Since all tweets all collected in timely order, we set the hour of which the first tweet is collected to be the first hour and calculate data features over one-hour window. The following two plots show the number of tweets feature in hourly windows from the hour of the first tweet to the hour of the last tweet





From the #nfl diagram, there are two peaks which occurred after 450 hours and 790 hours since the first tweet hour. To better understand this phenomenon, we need to first convert this reference hour time to actual date time. The very first tweet was posted at 20:21 12/30/2014. After about 450 hours, it was on 1/18/2015, the day of Conference Championships. Furthermore, about 790 hours after the first post was in 2/1/2015, the day of the Superbowl. No doubt there were bursts of tweets about nfl during these two times.

For hashtag #superbowl, the graph reveal that there is a single spike of tweets at about 830 hours since the first tweet, which is also in 2/1/2015, the day of superbowl. Therefore, the burst of tweets in #superbowl is also well explained.

Part 2.

In this part, the objective is to fit a linear regression model over five different features of the collected tweet data in hourly window, and use this fitted model to predict the number of tweets in the next hour. These five features are:

1. **Number of tweets (previous hour)**
2. **Total number of Re-tweets**
3. **Total number of followers of unique users**
4. **Max number of follower**
5. **Time of the data in 24-hour clock**

To evaluate the regression performance, the following three parameters are used:

R-Squared Accuracy

The R-squared value is used as a accuracy measure. It is a measure of how well the model fit the data. A value of 1 indicates your model has perfectly fit the data, which implies data overfitting.

P-value

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that it can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

T-test values

T is simply the calculated difference represented in units of standard error. The greater the magnitude of T (it can be either positive or negative), the greater the evidence *against* the null hypothesis that there is no significant difference. The closer T is to 0, the more likely there isn't a significant difference.

The following table shows the Accuracy, P/T values for each hashtag file:

Hashtag	Accuracy	Time (hour)	Re-tweet count	Follower count	Tweets count	Max follower
#gohawks	0.483	0.365/0.906	1.55e-3/-3.178	2.24e-4/3.71	**1.78e-4/3.772**	**5.2e-5/-4.07**
#gopatriots	0.66	0.722/0.356	0.322/0.99	**4.05e-5/4.15**	8.94e-2/-1.70	**1.8e-6/-4.85**
#nfl	0.57	0.94/0.076	4.47e-2/-2.012	**1.52e-5/-4.36**	**4.4e-19/9.248**	3.05e-5/4.20
#patriots	0.713	0.816/0.233	**5.4e-24/-10.53**	6.4e-10/6.28	**3.5e-59/18.09**	1.67e-2/-2.4
#sb49	0.82	0.56/-0.58	**8.5e-15/-8.0**	7.87e-12/7	**1.65e-69/20.7**	1.37e-4/-3.84
#superbowl	0.741	0.91/0.118	0.86/0.174	**5.7e-25/-10.8**	5.72e-7/5.06	**2.3e-12/7.17**

				1		

Features with significant impact to the prediction model has been highlighted in the table above:

Hashtags	Significant features
#gohawks	Max Followers, Tweets Count
#gopatriots	Max Followers, Followers Count
#nfl	Tweets Count, Followers Count
#patriots	Tweets Count, Retweets Count
#sb49	Tweets Count, Retweets Count
#superbowl	Max Follower, Follower Count

As above two tables show, all chosen features except the day hour time feature contribute to our prediction model. The low p and t value of the time feature suggests that the model would not be affected much if this feature is dropped. Additionally, the ‘tweets count’ feature seems to have the most significance in the sense that it appears to be significant feature in four out of six hashtags.

Part 3.

Feature Selection

In this part, our objective is to introduce new features and use them along with our old features to perform regression analysis over hashtag files. Features we used in this part:

Old features:

1. **Number of tweets (previous hour)**
2. **Total number of Re-tweets**
3. **Total number of followers of unique users**
4. **Max number of follower**
5. **Time of the data in 24-hour clock**

New features:

1. **Sum of Ranking Score** - total amount of influence tweets on audience
2. **Impression count** - total number of time tweets served as promoted tweet
3. **Unique user count** - total number of unique users posting the tweet
4. **Verified user count** - total number of verified users posting the tweet

5. **User mention** - total number of mentioned user in tweets
6. **URL links count** - total number of URL included in tweets
7. **Long tweet count** - total number of tweets with 80 or more words
8. **Lists count** - total number of public lists user belong to
9. **Max list number** - the max number of public list a user belongs to
10. **User friends count** - total number of people the user is following

Therefore, we had included **15** features to be used in the regression analysis we were going to perform. Some of these features were suggested from this paper from the project spec and others we found from Twiter Developer Documentation.

However, these features listed above were not randomly picked but were rather chosen carefully after dozens of experiments on different features combination. Here are some of candidate features we did not include in our regression model because we found they actually increase the prediction error: **Statuses count, favorite count, max favorite number.**

Model Selection

From part two we learned that a simple OLS regression may not be enough. First of all, the accuracy score are relatively low especially for the first three hashtags. However, most importantly, according to the OLS summary report generated, the regression model generated in part two has a very large condition number of **4.15e6,** indicating that the model suffers a **strong multicollinearity problem**.

To solve this multicollinearity problem, we first tried with the **Lasso and Ridge Regression** (linear regression with regularization) with alpha equal to **0.01** as the new regression model to use. However, the prediction result is still quite large. We also tried **Polynomial and Neural network regression** and finally decide to use **Random Forest Regressor** as our model because it yields the least prediction error.

Model Accuracy

The following table shows the R squared accuracy of our random forest model and feature importance for each hashtag:

	#gohawks	**#gopatriots**	**#nfl**	**#patriots**	**#sb49**	**#superbowl**
Accuracy	0.941	0.942	0.924	0.954	0.964	0.921

The model accuracy of the random forest regression model displayed above is considerably better than that obtained from the linear model in the previous part. Now the accuracy for each hashtag is higher than 90%. This indicates that random forest regression is indeed a good prediction model for our datasets.

Feature importance

The following table displays importance score of each chosen feature to the hashtag file.

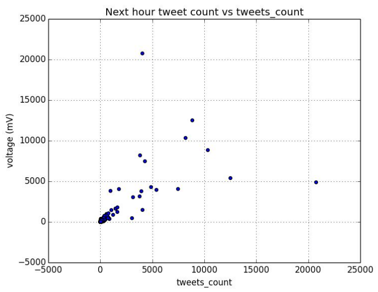
	#gohawks	**#gopatriots**	**#nfl**	**#patriots**	**#sb49**	**#superbowl**
Ranking	0.082	0.054	0.046	0.037	0.059	0.038
Friend	0.066	0.124	0.045	0.070	0.060	**0.112**
Impression	0.007	0.007	0.025	0.019	0.023	0.068
URL	**0.213**	0.069	**0.285**	**0.259**	**0.135**	**0.124**
List	0.118	0.023	0.042	0.032	**0.092**	**0.180**
Long tweet	0.042	**0.165**	**0.130**	0.049	0.005	0.097
Time	0.003	0.016	0.013	0.016	0.004	0.030
Tweets Count	**0.217**	0.061	**0.162**	0.036	0.079	0.037
Verified User	0	0	0	0	0	0
User Count	**0.141**	**0.136**	0.104	**0.156**	0.059	0.083
Max list	0.042	0.029	0.005	0.017	0.003	0.024
Retweets Count	0.004	0.056	0.034	0.004	0.020	0.081
Followers	0.038	0.017	0.022	0.057	0.057	0.046
Mention	0.022	**0.211**	0.063	**0.232**	**0.391**	0.075
Max Follower	0.004	0.031	0.023	0.016	0.011	0.003

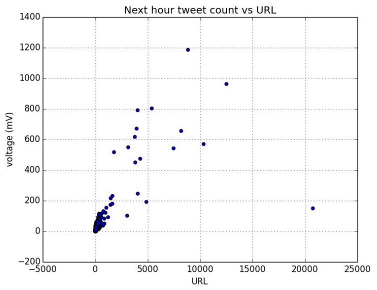
In the importance score table above, top three features in each hashtag have been highlighted. Top three most important features for all hashtags are URL, User Count and Mention Counts. The score table reveals that we could drop the ‘Verified User’ without affecting the model at all as it contributes nothing to any chosen hashtags. It also implies that feature ‘time’ and ‘Max Follower’ contributes little to the model as their importance scores are relatively very low. It may not affect the model much if we drop those features.

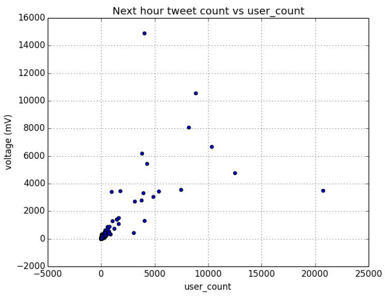
Tweets prediction vs top features

#gohawks:

Top three features: **Tweet Count, URL, User Count**

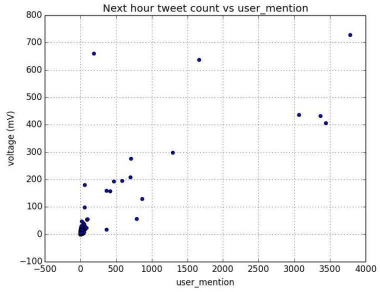


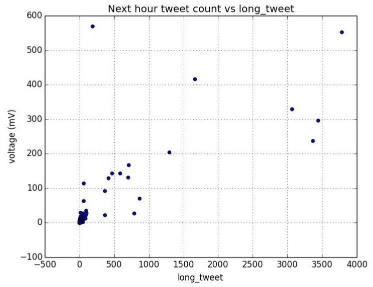


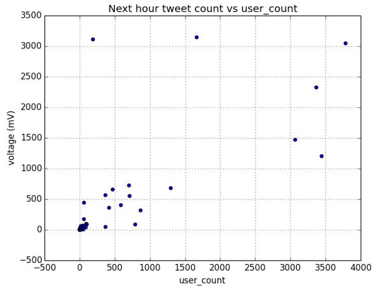


#gopatriots:

Top three features: **Mention, Long tweet, User Count**

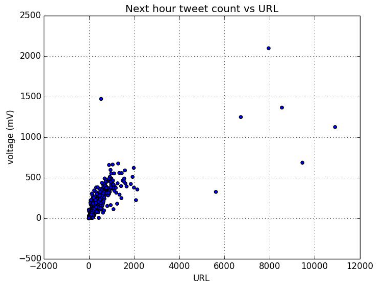


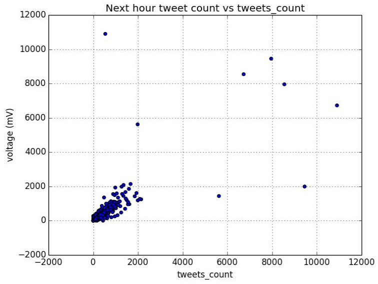


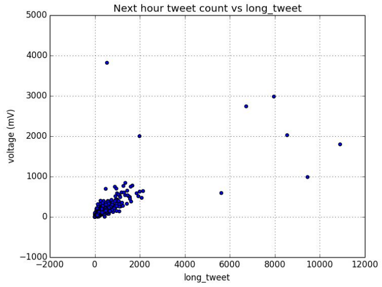


#nfl:

Top three features: **URL, Tweets Count, Long tweet**

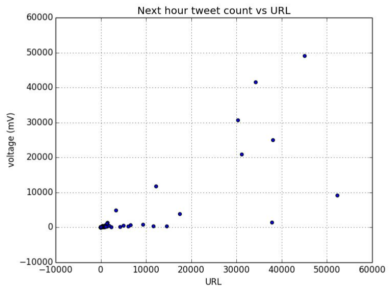


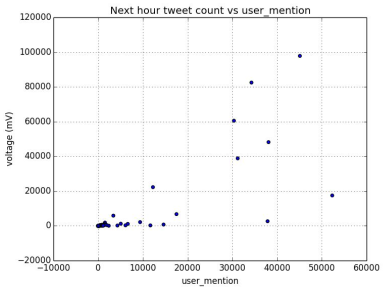


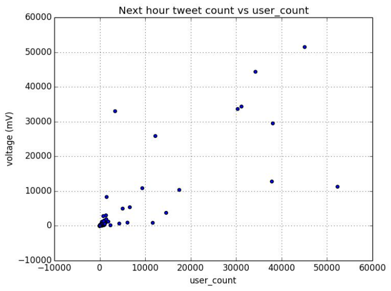


#patriots:

Top three features: **URL, Mention, Tweets Count**

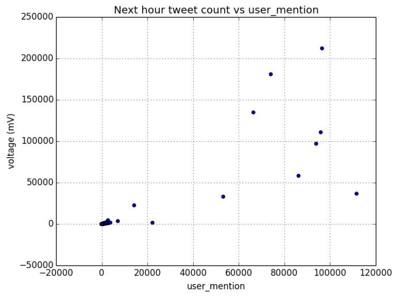


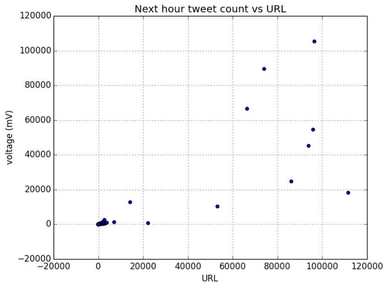


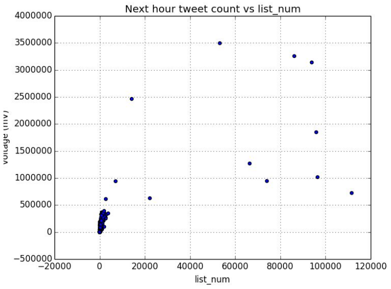


#sb49:

Top three features: **Mention, URL, List**

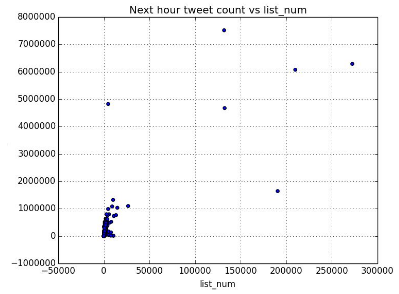


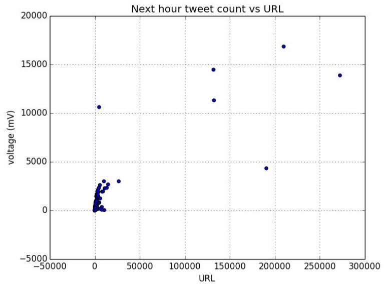


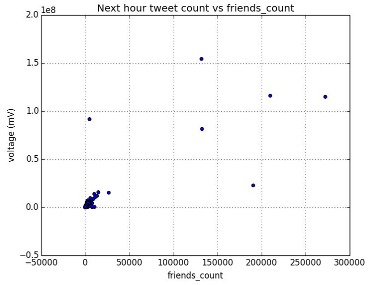


#superbowl:

Top three features: **List, URL, Friend Count**







Part 4.

Overall 10-fold-validation prediction error

The following table shows the prediction error for each fold during the 10-fold cross validation analysis and also the overall average error.

Fold	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
1	49.8	5.3	110.5	69.4	70.8	240.2
2	134.1	78.2	261.3	266.1	470.4	552.4

3	40.0	9.1	104.1	207.8	63.1	220.6
4	400.4	151.4	292.7	1489.3	709.8	4322.3
5	144.28	115.3	180.8	199.5	2346.4	1998.1
6	176.8	47.6	282.9	1057.5	2558.3	3533.7
7	64.8	13.5	109.6	365.8	59.2	327.0
8	319.5	24.3	216.9	145.5	342.8	333.5
9	42.1	4.4	74.6	51.8	33.8	103.2
10	85.3	6.9	88.3	309.4	254.8	249.0
Avg Error	**145.7**	**45.6**	**172.2**	**416.2**	**690.9**	**1188.0**

As the table reveals, prediction errors are actually very high in spite of the high accuracy rate obtained before. The overall prediction errors are especially high for hashtag #patriots, #sb49 and #superbowl.

Period-wise prediction error

	#gohawks	**#gopatriots**	**#nfl**	**#patriots**	**#sb49**	**#superbowl**
Before	121.5	12.9	112.0	177.6	48.5	245.27
Between	2329.4	1114.7	2567.9	15454.43	25831.8	58890.5
After	19.9	4.47	111.2	57.1	105.9	178.2

This table above breaks the timeline of the whole dataset into three periods and displays their corresponding prediction errors. The timespan was split into three periods - before superbowl, during the superbowl and after the superbowl day. As it exhibits, prediction error skyrockets for the ‘Between’ period while prediction errors for other two periods remain relatively low. This is probably because there are only 12 training sets available (12 hours) for the model to train for the ‘between’ period and the lack of training leads to inaccurate prediction.

Part 5.

The following table reveals the error percentage using data from hour 2 - 6 in the test file. We could then determine the prediction for the 7th hour using the model with the smallest prediction error percentage.

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
S1P1	50.21	256.8	116.56	57.12	**41.2**	52.93
S2P2	**50.63**	88.24	61.0	107.69	159.84	188.81
S3P3	34.53	92.47	**13.94**	19.47	45.61	25.36
S4P1	20.61	76.73	**7.84**	22.94	29.59	19.39
S5P1	33.63	**33.5**	59.73	38.64	37.68	33.79
S6P2	76.13	93.05	84.14	**29.51**	58.94	81.29
S7P3	86.28	41.23	132.84	**24.39**	49.08	90.91
S8P1	**49.40**	58.24	182.07	74.79	108.1	61.84
S9P2	167.65	**38.04**	78.27	1087.11	1671.93	1026.79
S10P3	39.19	41.61	41.53	**15.74**	33.59	78.45

The model to be used to predict for the 7th hour has been highlighted in the table above. We choose them because they yield the lowest prediction error for hour 2 - 6 for the test file.

	S1P1	S2P2	S3P3	S4P1	S5P1	S6P2	S7P3	S8P1	S9P2	S10P3
Model	#sb49	#gohawks	#nfl	#nfl	#gopatriots	#patriots	#patriots	#gohawks	#gopatriots	#patriots
7th hour	102.44	115.76	975.44	602.16	336.84	27361.14	60.26	55.5	2453.08	47.54

The table above shows the model to be used for each file and the popularity prediction for the 7th hour.

Part 6.

In this section, we only considered textual content in all tweets including #superbowl. However, our laptop cannot process this large quantity of tweets. Therefore, we selected the first 250000 tweets from the data. This number of samples are large enough make the classification similar to the original dataset and if the hardware requirement is allowed, the same algorithm and our code can be directly applied to process all the data.

For the classification, we applied three algorithms which we used in project2: Support Vector Machine, Logistic Regression and Multinomial Naive Bayes classifier. For each classification algorithm, we calculated the accuracy, precision, reported the confusion matrix and plotted the ROC curve. Results are summarized as below:

Accuracy:

	SVM	Logistic	Multi-NB
Accuracy	0.779070	0.763214	0.723044

Precision and recall values:

'0' is tweet from WA and '1' is tweet from MA

Support Vector Machine

	precision	recall	f1-score	support
0	0.89	0.76	0.82	1123
1	0.71	0.86	0.69	769
Avg / total	0.82	0.80	0.80	1892

Logistic Regression

	precision	recall	f1-score	support
0	0.89	0.76	0.82	1123
1	0.71	0.86	0.78	769
Avg / total	0.82	0.80	0.80	1892

Multinomial-NB Classifier

	precision	recall	f1-score	support
0	0.76	0.77	0.77	1123
1	0.66	0.65	0.65	769
Avg / total	0.72	0.72	0.72	1892

Confusion Matrix:

Support Vector Machine

	Predicted: No	Predicted: Yes
Actual: No	912	211
Actual: Yes	152	617

Logistic Regression

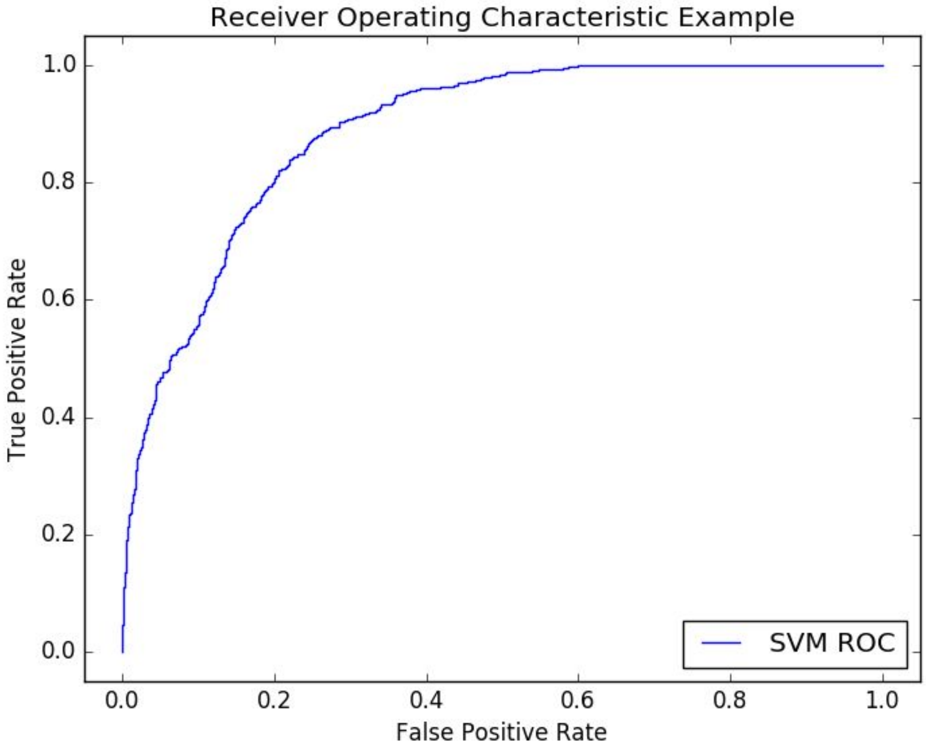
	Predicted: No	Predicted: Yes
Actual: No	852	271
Actual: Yes	107	662

Multinomial-NB Classifier

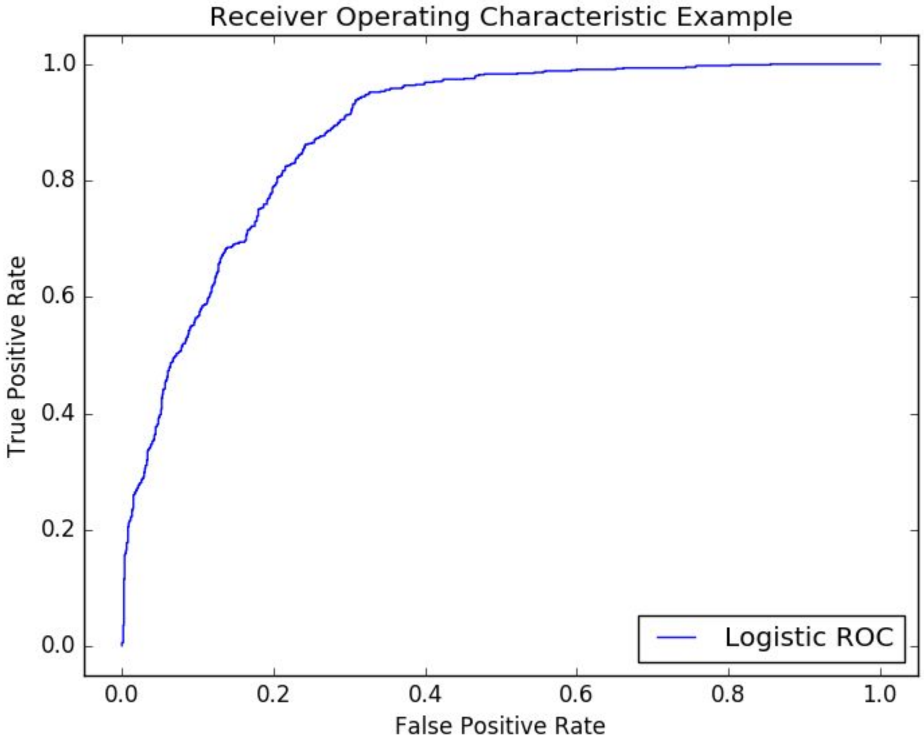
	Predicted: No	Predicted: Yes
Actual: No	865	258
Actual: Yes	269	500

ROC plot:

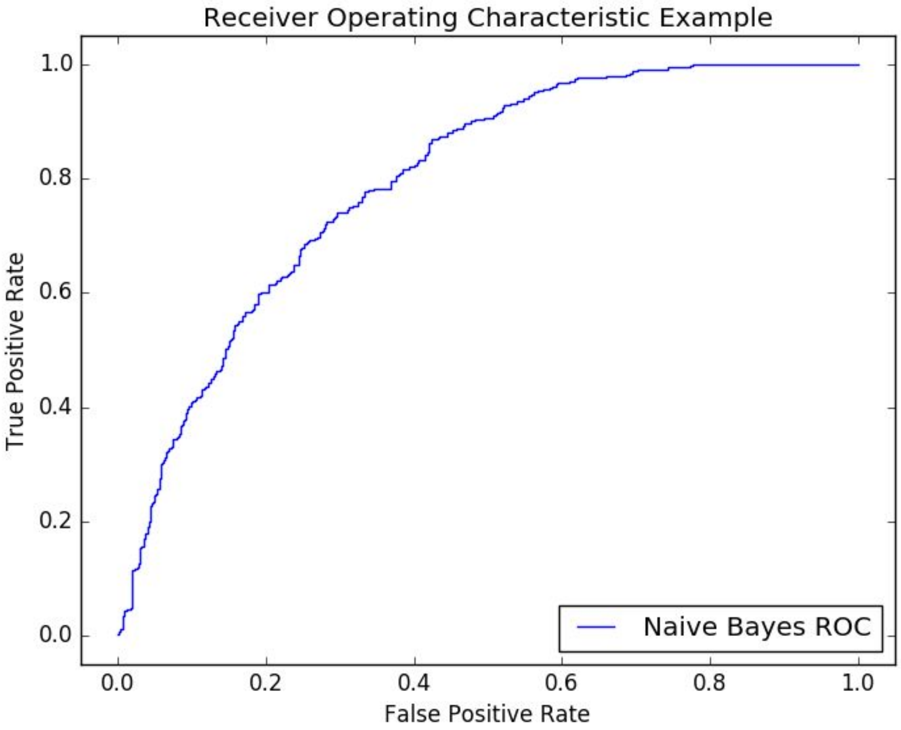
Support Vector Machine



Logistic Regression



Multinomial NM Classifier



From the results, we can see all three methods have similar performance. SVM and logistic regression have slightly better accuracy and precision than multinomial naive bayes classifier.

Part 7.

Tweet data is a very rich source of information which provides insights about the population posting the tweet and topics they are discussing about as well as tons of other metadata. In this part, we are focusing on another important aspect of the tweet data - the actual tweets content. By directly analyzing the content of the tweet, we are able to extract lots of useful information. One very desirable information usually extracted from tweet data analysis is the population sentiment. By applying NLP (natural language processing) techniques on the tweet, we could know people’s sentiment (positive, negative and neutral) behind it. This information is very useful as it directly reflects people’s attitude/opinion toward a certain topic.

Objective

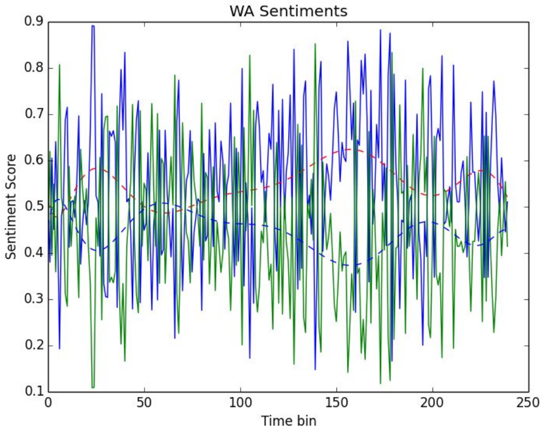
Our goal for this part of the project is to analyze tweets sentiments of the public at two geolocations of the two competing teams competing in the Superbowl game during the day of Superbowl 49 (Seattle/Washington and Boston/Massachusetts). Hopefully through the sentiment analysis, we could catch high and low moments during the day for people supporting these two teams and guess the match result from their sentiments.

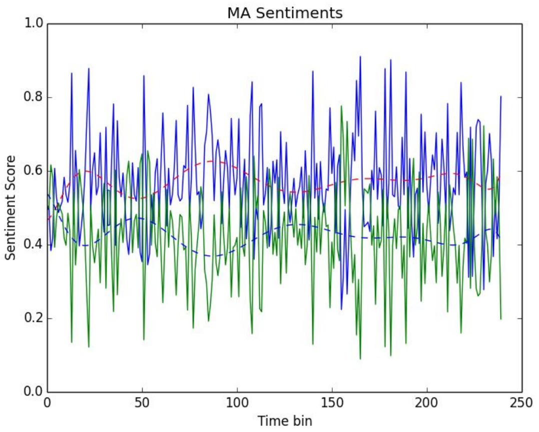
Approach

Since we do not have much time to create our own sentiment analyzer, we choose to use the Text-processing sentiment analysis API to perform the analysis.

1. **Extract all tweets during 2/1/2015 0:00 to 2/1/2015 23:59 from the #Superbowl tweet data file.**
2. **Filter tweets according to their geo-location, save Boston/MA and Seattle/WA tweets.**
3. **Break extracted tweet contents into 6-minute windows.**
4. **Perform text preprocessing on each time window, take out hashtags, URL, user mention, stopwords and undefined texts (numbers, undefined symbols and etc.).**
5. **Push the preprocessed text to sentiment analyzer API.**
6. **Save the sentiment analyzer API result and plot.**

Result Analysis





Two diagrams above exhibit sentiments score of people from Wa and Mass during the day of superbowl. Red dash line represent positive emotion score while the blue one represent negative sentiment. A score of 0.5 means the sentiment is neutral. There are some interesting observations we could extract from them. For example, people from Boston/Ma seems to be more excited earlier in the day while people from seattle/Wa are more excited later that day, especially after the game starts at 3:00 PM. This may have something to do with the time difference. Additionally, to our surprise, people in seattle/Wa did not express much of frustration or disappointment after the game. This is probably because rather than expressing their frustration, people tended to post cheering tweets after the loss.

However, by simply looking at these two sentiment score diagram we are unable to guess which team won the game eventually. Although there is a sudden decrease in the positive sentiment score for seattle/Wa after the game, it is not significant enough for us to reach to any conclusion. Same thing for the MA scores; there seems to be no burst of positive sentiments after the game, but there is a indeed increase in the positive score. We were expecting more

drastic results such as sudden decrease in positive score for Wa and burst of positive score for Boston in the sense that Patriots won the game.
There are a few factors that may help explain why the sentiment analysis result was not as good as we expected:

- Imperfect text preprocessing: Our preprocessing technique is actually very simple. Besides filtering out undefined characters and words, we did not perform filtering on the tweet content at all. There may exist tweets containing strong sentiment content but are unrelated to the match at all. For example, commercial advertisements usually use phases with strong sentiments and they do make a sizeable portion in the tweet data. By containing these data in the analysis, it may mask people's actual sentiments toward the match.
- Inappropriate sentiment analyzer: Because of the limit time we have, instead of building our own customized analyzer, we used an online API to perform the analysis. We may get better result if we could use an analyzer that is customized to our needs.