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

## Project 5
## Popularity Prediction on Twitter
## Winter 2017

Guanqun Mao, Jianing Liu
204777289 804759999

March 22nd 2017

**Section A: Introduction**

Twitter provides a platform to predict popularity of future topics and events. Knowing current and previous tweet activities with a hashtag(#), we can predict whether this topic will become more trendy in the future and by how much.

In this project, we analyze the data collected by querying popular hash-tags related to the 2015 Super Bowl from Twitter, from two weeks before the game to a week after the game. We will use this data set to train a regression model and then use this model to predict behaviors of other hashtags. The test data consists of tweets with a hashtag in a specified time frame. We will use our model to predict number of tweets with these hashtags posted within one hour immediately after the given time frame.

**Section B: Popularity Prediction**

**Part 1 Tweet Data Statistics**

In this part, we download the training tweet data to calculate for each hashtag the following statistics: average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets.

In order to keep track of the hour count, we use an hour-window approach, since the tweets are sorted in the order of their posting time (firstpost_date). The first tweet is considered and the 1st hour-windows is created using the formula:

$$end_{time} = start_{time} + 3600$$

We iterate through each tweet and compare the post-time of the tweets with the end time of the current window. If it lies within the window, we increase the hour-count; if it doesn't we create a new window by using the above equation and adding 3600 (1 hour in UNIX time) to the end-time. A counter is kept to track the number of followers of the users (author/followers) and the number of re-tweets (metrics/citations/total) for each tweet.

The training data set has six hashtags: #gohawks, #gopatriots, #nfl, #patriots, #sb49 and #superbowl. The three statistics calculated for all six hashtags are shown in Table 1.1.

| Hashtag | Total Tweets | Avg. # Tweets/hr | Avg. # of Followers | Avg. # of Retweets |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| #gohawks | 188135 | 193.54379644 | 1596.44360357 | 2.01462779387 |
| #gopatriots | 26231 | 38.3832405913 | 1292.20316579 | 1.40013724219 |
| #nfl | 259023 | 279.550300946 | 4394.25396783 | 1.53853904866 |
| #patriots | 489712 | 499.420031779 | 1607.44073544 | 1.7828192897 |
| #sb49 | 826950 | 1419.88619047 | 2229.69488328 | 2.51115182296 |
| #superbowl | 1348766 | 1401.24455475 | 3675.33948006 | 2.38827417061 |

Table 1.1 Statistics of Six Hashtags

*Analysis of the Statistics*

1. Most tweeted hashtags per hour: #sb49, #superbowl
2. Most followers of users for hashtag: #nfl, #superbowl
3. All of the tweet data collected have tweets that are not re-tweeted or are re-tweeted by very few users therefore making the average re-tweet count about 2

In order to visualize the number of tweets in an hour, we plot the histograms for #superbowl and #nfl. We can see a steep-rise in both graphs at the same time which coincides with the hour of the event.

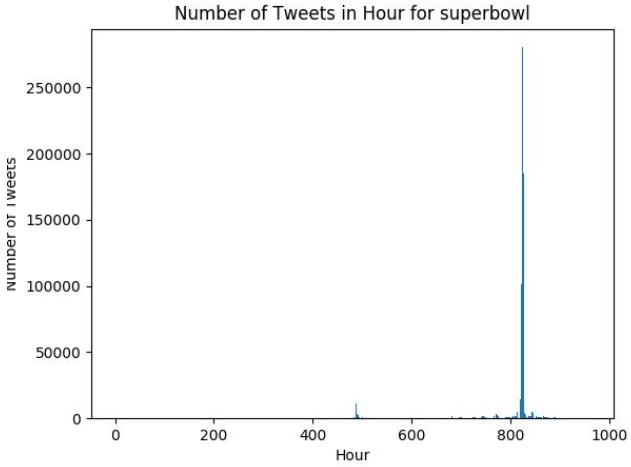


Figure 1.1 Histogram of #superbowl

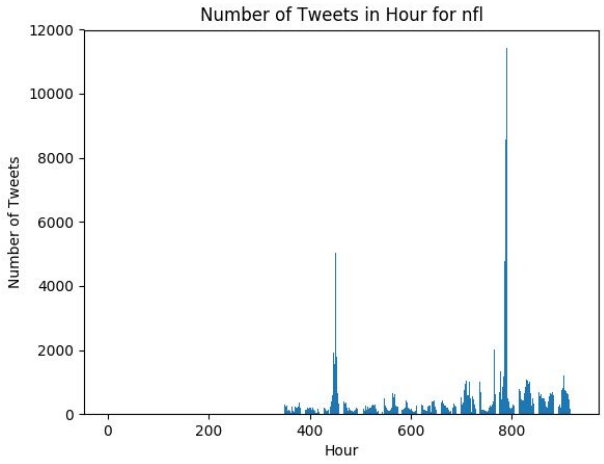


Figure 1.2 Histogram of #nfl

For #NFL, there is an additional peak - the last weekend before SuperBowl's final game. This implies that people tend to talk more about NFL the weekend before the final, which is intuitive to understand, as the final game is always the most exciting one and people tend to have more available time to tweet during the weekend.

**Part 2 Linear Regression**

We create a linear regression model using 5 features to predict number of tweets in the next hour, with features extracted from tweet data in the previous hour. The features we use are:

1. Number of tweets (Class Variable)
2. Total number of retweets (metrics/citations/total)
3. Sum of the number of followers of the users (authors/followers)
4. Maximum number of followers of the users
5. Time of the data - obtained using the post-time of the tweet

The same hour-window approach is used here to calculate all features. The output variable for each hour-window is the tweet for the next window. The model is trained using the OLS statsmodel library. T-test describes the significance of each feature and P-value describes the accuracy of each feature. The model accuracy are shown in Table 2.1.

| HashTag | Accuracy |
|---|---|
| #gohawks | 41.78 |

| | |
|---|---|
| #gopatriots | 43.16 |
| #nfl | 54.69 |
| #patriots | 43.14 |
| #sb49 | 58.46 |
| #superbowl | 66.01 |

Table 2.1 Model Accuracy of Hashtags

The p-value and t-value for each feature are shown in Table 2.2 and 2.3.

| hashtag/feature | # of Retweets | Sum of # of Followers of the Users | Max. # of Followers of Users | Time of the Data |
|---|---|---|---|---|
| #gohawks | 4.27343522 | 5.35547595 | -4.87196979 | 2.86729647 |
| #gopatriots | 11.24719656 | -8.38659191 | 7.27891607 | 0.17092002 |
| #nfl | 8.26674332 | 1.78715533 | -0.80934704 | 4.25396111 |
| #patriots | 18.05368198 | -7.60264257 | 4.26331804 | 0.71040951 |
| #sb49 | 17.64706329 | -12.37425173 | 8.22281528 | -1.85560124 |
| #superbowl | 31.35608131 | -26.50428285 | 16.14506699 | -1.80517618 |

Table 2.2 T-tests of Different Hashtags

| hashtag/feature | # of Re-tweets | Sum of # of Followers of the Users | Max. # of Followers of Users | Time of the Data |
|---|---|---|---|---|
| #gohawks | 2.11565802e-05 | 1.06694073e-07 | 1.29088771e-06 | 4.23004469e-03 |
| #gopatriots | 8.73217859e-27 | 3.51758538e-16 | 1.04826618e-12 | 8.64343706e-01 |
| #nfl | 4.77555207e-16 | 7.42411515e-02 | 4.18524590e-01 | 2.31480681e-05 |
| #patriots | 4.43245562e-63 | 6.80726659e-14 | 2.20983673e-05 | 4.77620115e-01 |

| | | | | |
|---|---|---|---|---|
| #sb49 | 4.68187282e-56 | 2.32906554e-31 | 1.32046476e-15 | 6.40201369e-02 |
| #superbowl | 4.89704326e-149 | 1.61268413e-116 | 4.85350481e-052 | 7.13613102e-002 |

Table 2.3 P-values of Different Hashtags

*Analysis of Staistics*

1. Based on the definition of p-value and t-value , we can see that the most contributing feature towards the regression model in all hashtag files is the Number of Retweets with a hashtag.
2. Most hashtags have fairly low accuracy. This may be due to the window size of one-hour, since in the initial hours, the average number of tweets are low and creating a good model for these sparse features is more tricky.

**Part 3 Regression Model with Extra Features**

In this part we need to design a new regression model with extra features and the original features we used in Part 2, based on our observation of the data and consultation with academic papers. The new features are:

1. Number of tweets (Class Variable)
2. Total number of re-tweets (metrics/citations/total)
3. Sum of the number of followers of the users (author/followers)
4. Maximum number of followers of the users posting the hashtag
5. Time of the data - obtained using the post-time of the tweet
6. Ranking score (metrics/ranking_score)
7. Impression count (metrics/impression) - measures the number of times a user is served a Promoted Tweet either in time-line or on search
8. Favorite count (tweet/favorite_count) - number of tweets favored by users
9. Number of users per hour (tweet/user/id) - number of users postng per hour
10. Number of long tweets per hour (title) - number of tweets with length > 100 characters

We use a total of 9 features to create the new regression model and follow the same approach in Part 2 - collecting features using one-hour window. The last hour window can't predict a tweet-count value, so it is removed when creating the model. The model accuracy is shown in Table 3.1

| HashTag | Accuracy |
|---|---|
| #gohawks | 78.39 |
| #gopatriots | 53.12 |
| #nfl | 64.84 |
| #patriots | 58.45 |
| #sb49 | 70.44 |
| #superbowl | 77.01 |

Table 3.1 Model Accuracy of Hashtags

We can see that the model accuracy has significantly increased for each hashtag. This is because the features are not sparse and now have a well-defined distribution throughout the SuperBowl. Metrics used in the tweet data are used to model the importance of the tweet for a given window and thus increases the accuracy. To visualize the contribution of features in the model, we create scatter plots for the top 3 features for each hashtag. The initial hours have fewer tweets, therefore all the plots show clustering of data points near low of tweets/hour.

#gohawks

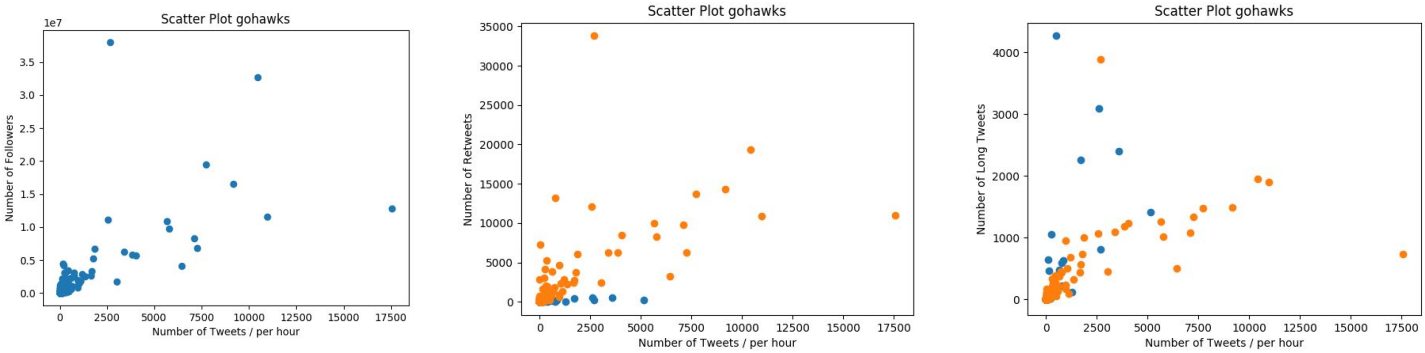


Figure 3.1 Top 3 Features for #gohawks (# of followers, # of re-tweets, # of long tweets)

#gopatriots

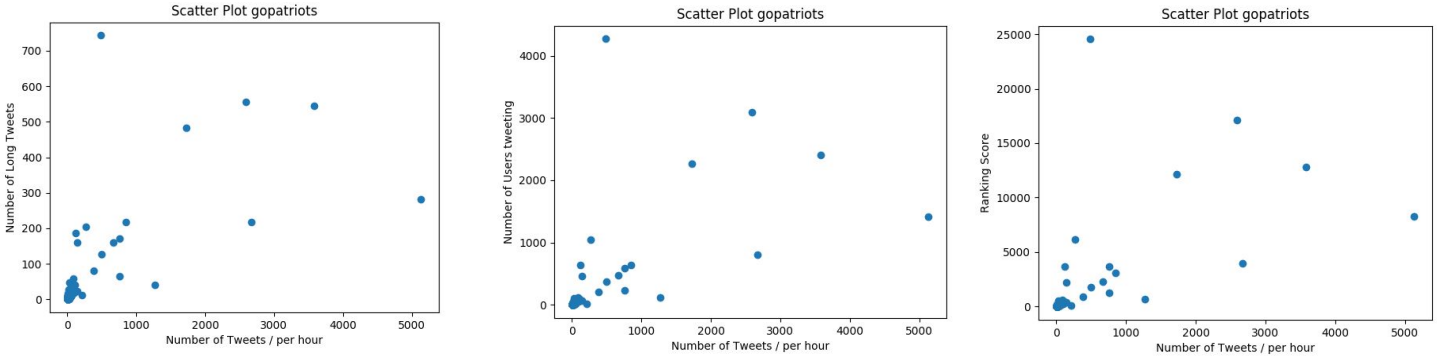


Figure 3.2 Top 3 Features for #gopatriots (# of long tweets, # users tweeting, ranking score)

#nfl

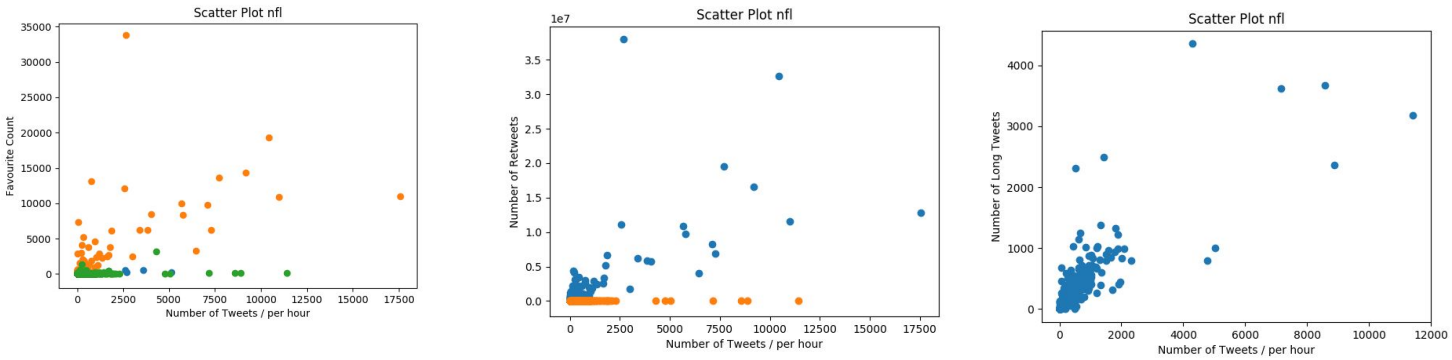


Figure 3.3 Top 3 Features for #nfl (favorite count, # of re-tweets, # of long tweets)

#patriots

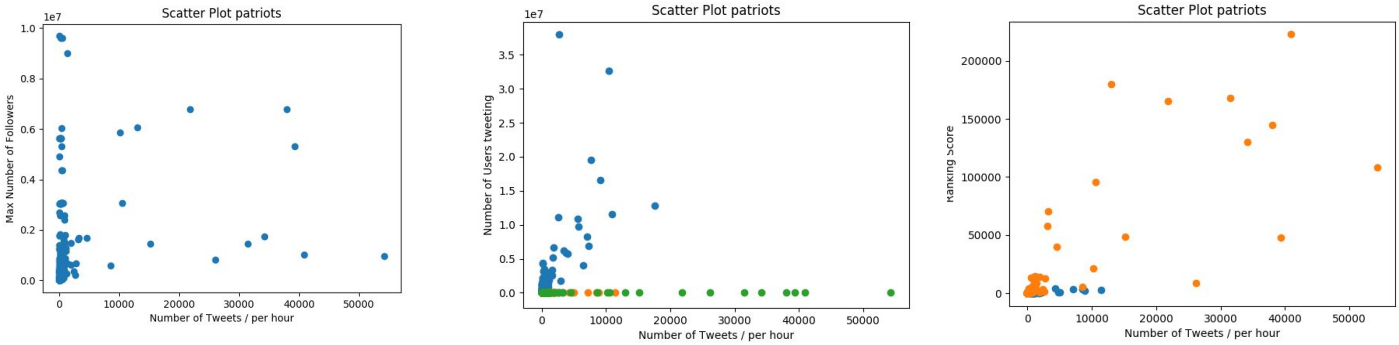


Figure 3.4 Top 3 Features for #patriots (max # of followers, # of users tweeting, ranking score)

#sb49

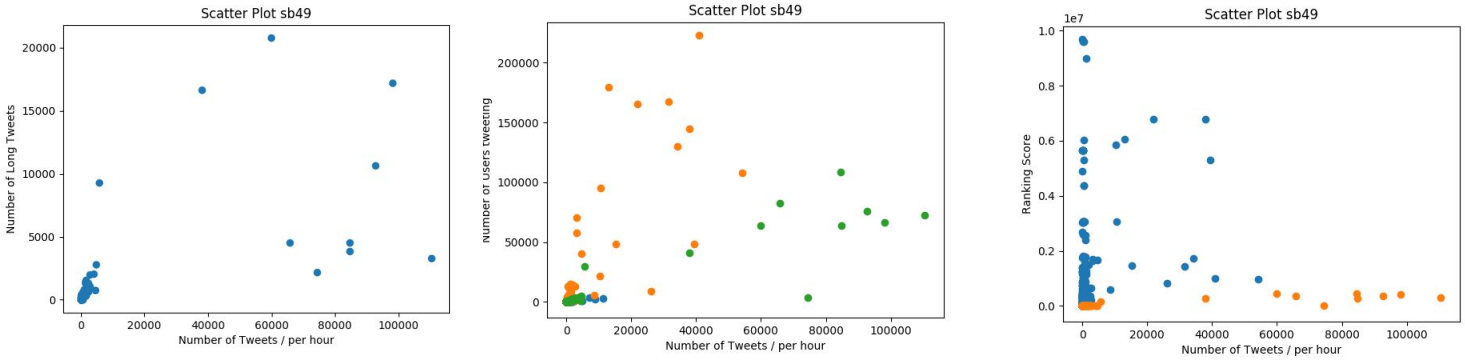


Figure 3.5 Top 3 Features for #sb49 (# of long tweets, # of users tweeting, ranking score)

#superbowl

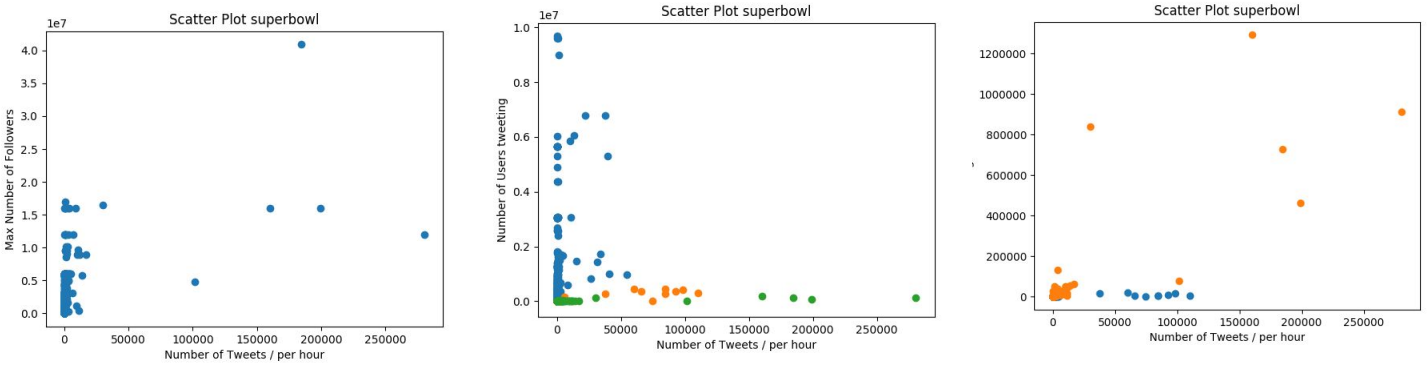


Figure 3.6 Top 3 Features for #superbowl (max # of followers, # of users tweeting, ranking score)

*Analysis of Scatter Plots*

| Hashtag | Analysis |
|---|---|
| #gohawks | There is a linear proportionality in the scatter plot, implying a good relationship between the top 3 features. |
| #gopatriots | All 3 features have almost identical scatter plots, with clustering around the origin. |
| #nfl | There is a constant relationship for features favorite count and # of long tweets, whereas there is a linear relationship for # of retweets. |
| #patriots | There is a constant relationship for max # of followers and # of users tweeting but a constant relationship for the last feature. |
| #sb49 | Similar to #patriots. |
| #superbowl | Max # of followers and ranking score show a clustered region with a tiny linear deviation, while # of users tweeting seems to have a constant relationship. Large number of instances fits a better regression model and thus leads to better accuracy. |

Table 3.2 Analysis of Top 3 Features

**Part 4 Cross Validation**

In the first half of Part 4, we use the same features used in Part 3 and perform a 10-fold cross validation on the data set. The accuracy result of different hashtags and over every fold are shown in Table 4.1.

| Fold Number | #gopatriots | #gohawks | #nfl | #patriots | #sb49 | #superbowl |
|---|---|---|---|---|---|---|
| 1 | 7.7824849 657829436 | 20.127615 680308633 | 23.921507 284201926 | 180.85575 184910365 | 101.45728 833000119 | 229.98050 534535864 |
| 2 | 8.4380414 376842001 | 46.514162 499367508 | 1.3767474 196434517 | 84.489863 890106108 | 150.84411 631333566 | 255.88116 080226337 |
| 3 | 10.145220 880328852 | 4.8140594 232694545 | 3.1815704 732446028 | 31.927206 639141783 | 159.24000 942898178 | 337.87083 509307303 |
| 4 | 204.98564 358880208 | 2.2450556 24512232 | 28.109584 165383914 | 52.189900 889932687 | 173.92132 419864723 | 397.13645 625529108 |
| 5 | 15.497373 086519449 | 117.97881 852948235 | 185.83362 314325186 | 265.85588 292731927 | 219.97513 460862953 | 361.33939 65810788 |
| 6 | 41.759471 335536269 | 629.26754 769887157 | 133.98044 846553503 | 997.12509 137624181 | 498.19533 197249837 | 2506.9280 498273729 |
| 7 | 19.302637 56919458 | 147.07959 033140153 | 93.183840 04337679 | 687.34151 160036799 | 1551.6943 239681038 | 1168.8498 417695128 |
| 8 | 18.391338 972596763 | 171.12006 998034039 | 194.82715 346973973 | 466.04622 179761867 | 9580.9973 555324887 | 2756.2481 524516911 |
| 9 | 30.380979 397477986 | 850.13197 21032404 | 524.83807 754497707 | 2046.5374 36429488 | 1273.7807 444704313 | 19664.687 30630253 |
| 10 | 247.47628 487133889 | 5.0990396 415662547 | 137.61212 970592561 | 176.49866 518473621 | 473.80186 642784986 | 1661.4691 721465799 |
| Average Error | 60.415947 6105 | 199.43779 3151 | 132.68646 8172 | 498.88675 3258 | 1418.3907 4953 | 2934.0390 8766 |

Table 4.1 Average Error of 10-Fold Cross Validation

*Analysis of Statistics*

1. We can see that there exists a relationship between the number of tweets with a hashtag and the average error of cross validation. The greater the number of tweets are, the higher the absolute average error the hashtag has.
2. For each hashtag, the error of one of the cross-validation fold is too high because the dataset is unevenly distributed. A fold might consider a split in which the test data has all high values for the class (tweets during the game time of Super Bowl) and training data has all low values (tweets before and after the game), therefore giving out a high error value for that fold (e.g. Fold 9 for #gopatriots).

**Part 4 Cross Validation with Time Periods**

In the second half of Part 4, we analyze the regression models created for different time frames during the Super Bowl. Three different time frames are chosen for our model:

1. Before 02/01, 8:00 am.
2. Between 02/01, 8:00 am and 8:00 pm.
3. After 02/01, 8:00 pm.

The tweets are segregated based on the time it was posted and then split into different windows of one hour. The models are tested using the 10-fold cross validation. The average errors are shown in Table 4.2.

| Hashtag | Before | Between | After |
|---|---|---|---|
| #gohawks | 16.2173277678 | 238.102571339 | 1760.68249394 |
| #gopatriots | 167.189881517 | 7022.1632622 | 2607.69218748 |
| #nfl | 75.919653634 | 753.944462613 | 533.593886399 |
| #patriots | 190.869282731 | 93528.0776912 | 9745.06592167 |
| #sb49 | 39.8330230088 | 51166.8783903 | 12012.4491342 |
| #superbowl | 203.754003253 | 12861.8776544 | 11834.395444 |

Table 4.2 Average Error of 10-Fold Cross Validation with Time Periods

*Analysis of Statistics*

1. The between time frame only has 12 one-hour time windows, the number of instances in this time frame for model creation is very low. Therefore the model created has high values of errors.
2. The before time frame has a greater number of instances, the model created has considerably lower average errors.

**Part 5 Making Predictions**

In this part, we download the test data and run our model to make predictions for the next hour in each case. Since the entire data are of 6-hour window, each testing data-set have less than 6 instances. Each period will be compared with the corresponding model created in Part 4 for each hashtag.

The test data has all the hashtags mixed, therefore we need to apply only the models that fit appropriately. Alternatively, we can apply all the models and check the error of predicted values of the first 6 hours to estimate the performance of the 7th hour. The predicted values for the th hour is shown in Table 5.1. The values with the least error with respect to the 6-hour data are highlighted to indicate the estimated predicted value.

| Hashtag | S1_P1 | S2_P2 | S3_P3 | S4_P1 | S5_P1 |
|---|---|---|---|---|---|
| #gohawks | 290.547238813 | 2383960.8263 | 34608.2486631 | 1602.20968681 | 393.140347339 |
| #gopatriots | 365.351800096 | -858140.983553 | -1814.66229845 | 93.0412206213 | 409.403244295 |
| #nfl | 174.141993643 | **2178909.08261** | -1567.13104741 | **284.860440598** | **280.906303176** |
| #patriots | 242.045484002 | 173637.896099 | 6571.39569339 | 219.402690772 | 231.576744421 |
| #sb49 | **111.779702732** | -1486543.40811 | **1488.88569158** | 143.832425435 | 172.105282545 |
| #superbowl | 15.3848727298 | -1283467.82453 | 1240.60660288 | 50.764836546 | 37.7006374376 |

Table 5.1 Predicted Value of the 7th Hour

| Hashtag | S6_P2 | S7_P3 | S8_P1 | S9_P2 | S10_P3 |
|---|---|---|---|---|---|
| #gohawks | -21993.4691141 | -87.6430194729 | -109.984042497 | 57966.7637008 | 58.7063487681 |
| #gopatriots | -3672.59481742 | -32.5831353284 | 295.12043583 | -20145.2250068 | -31.6957891454 |
| #nfl | **886782.04694** | 102.665249065 | 105.748577616 | **50099.6838592** | -17.3106902154 |
| #patriots | -91159.9062761 | -50.5818731978 | 151.477534788 | -26168.7791243 | 932.511267582 |
| #sb49 | -87872.1061525 | **197.338955324** | **40.0190262946** | -28214.7526228 | 1939.84535726 |
| #superbowl | -374372.274777 | 207.131864066 | 81.6244356427 | -25818.8441148 | **1482.71125146** |

Table 5.1 Predicted Value of the 7th Hour

*Analysis of Statistics*

1. The highlighted values in the table correspond to the predicted values for the 7th hour. The model with the least error is highlighted.
2. As previously stated in the between period, P2 has a training dataset of 12 instances, thus the values predicted for all the P2 test data have high variations.

**Section 3: Fan Base Prediction**

**Part 6 Location Prediction**

The textual content of a tweet often reveals some information about its offer. If the users tweeting on the same topic have different or opposing views, we can get more information from them. Therefore, fans of different teams during a game would express different emotions using different terms. We realize that the team a person supports has certain correlation with the person's location, thus we use the textual content of the tweet posted by a user to predict the

user's location. We consider all the tweets with #superbowl posted by people in Washington or Massachusetts.

We use the techniques employed in Project 2 to train our classifier to make location predictions. Specifically, we used 3 different classification algorithms in our approach: svm, gaussian naive bayes and logistic regression. The results are as follows.

Svm:

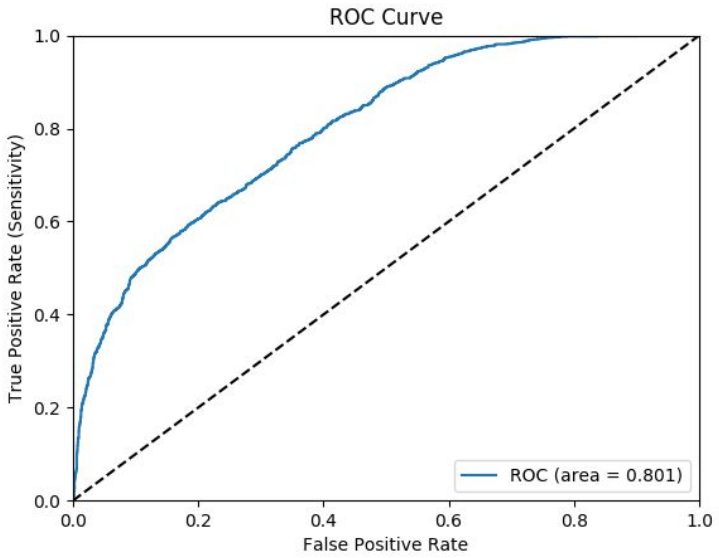


Figure 7.1 ROC Curve of SVM

| | Predicted Comp | Predicted Rect |
|---|---|---|
| Actual Comp | 2307 | 194 |
| Actual Rect | 953 | 703 |

Table 7.1.A Confusion Matrix of SVM

| | |
|---|---|
| Precision | 74% |
| Recall | 72% |
| Accuracy | 72% |

Table 7.1.B Statistics of SVM

Naive bayes:

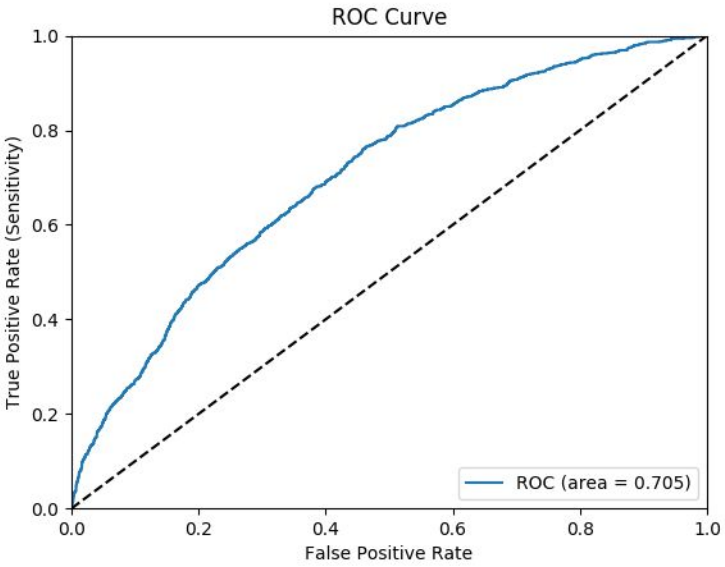


Figure 7.2 ROC Curve of Naive Bayes

| | Predicted Comp | Predicted Rect |
|---|---|---|
| Actual Comp | 1728 | 773 |
| Actual Rect | 670 | 986 |

Table 7.2.A Confusion Matrix of Naive Bayes

| | |
|---|---|
| Precision | 66% |
| Recall | 65% |
| Accuracy | 65% |

Table 7.2.B Statistics of Naive Bayes

Logistic regression:

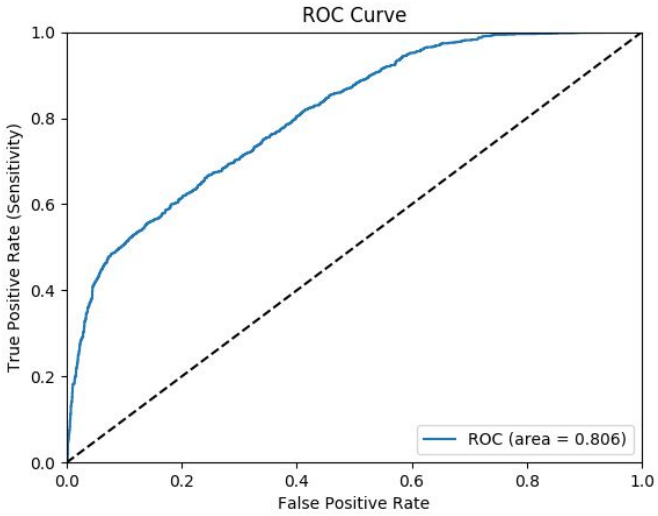


Figure 7.3 ROC Curve of Logistic Regression

| | Predicted Comp | Predicted Rect |
|---|---|---|
| Actual Comp | 2368 | 133 |
| Actual Rect | 948 | 708 |

Table 7.3.A Confusion Matrix of Logistic Regression

| | |
|---|---|
| Precision | 77% |
| Recall | 74% |
| Accuracy | 74% |

Table 7.3.B Statistics of Logistic Regression

*Analysis of Statistics*

1. The precision and accuracy of 3 classification algorithms are not very ideal, ranging from 65% to 77%. The low precision and accuracy of svm may due to the low balance of the datasets, when the difference between numbers of users in two locations is too large.
2. Naive bayes estimates the maximum likelihood probability of a class given a document with feature set based on the assumption that given the class, the features are statistically independent.

This assumption may be too simple for our huge dataset, therefore causing the low precision and accuracy.
3. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Intuitively, logistic regression gives the best result of all three algorithms and also has the largest AUC.
4. To improve the precision and accuracy values, in future we could use the combination of two algorithms to eliminate the disadvantages of applying only one.

**Part 7 Event Sequencing - Evaluating the Flow of the Events with Twitter**

*Problem Statement*

We are interested in what Twitter users say and what their emotions are at the time, which will be a sentiment analysis and can be used in a various occasions. For example, the candidates want to know what people's reactions are during a presidential campaign, which will give them insight of the next steps in their campaign. In this part, we try to detect the emotion changes from fans of both teams during and after the final game of the SuperBowl with the twitter data we have.

*Procedure*

We use the tweet matadata in the json files we are given, which represents the content of the tweet. The following preprocessing steps are needed before analysis:

1. Tokenize the tweet text - make all letters lowercase, remove the stop words and punctuations, and tokenize the text with regular expressions
2. Calculate the term frequency for each term in the tweet
3. Calculate the term co-occurrences for each term in the tweet

*Pointwise Mutual Information*

We define the Semantic Orientation (SO) of a word as the difference between its associations with positive and negative words. We want to calculate how close a word is to words like "good" and "bad". We measure closeness with Pointwise Mutual Information (PMI):

$$PMI(t_1, t_2) = log(\frac{P(t_1 \cap t_2)}{P(t_1) * P(t_2)})$$

The SO of a word is calculated against positive and negative words. Let $V^+$ be the set of positive words and $V^-$ be the set of negative words. The SO of a word is then:

$$SO(t) = \sum_{t' \in V^+} PMI(t, t') - \sum_{t' \in V^-} PMI(t, t')$$

We define the Document Frequency (DF) of a word to be the number of documents in which the word occurs, so the probabilities are:

$$P(t) = \frac{DF(t)}{|D|}$$

$$P(t_1 \cap t_2) = \frac{DF((t_1 \cap t_2)}{|D|}$$

*Results and Discussions*

We choose #gohakws and #patriots as the supporting hashtag for the two teams (the number of data in #gopatriots is too small). The data is split into one-hour time windows and we only consider the 24 hours around the game duration. We plot the semantic orientation of every hour for the tweets from fans of two teams and compare them.

Figure 7.1 shows the results. We can see that at first, hawks fans are more positive than patriots fans. However, 3 hours after the game started. we have a turning point at 6:30 pm is a turning point. We enlarge this time frame to have a closer look at it in Figure 7.2. We can see that after 6:30 pm, patriots fans became more positive than hawks fans. This makes sense because at that time, patriots fought back and eventually won the championship.

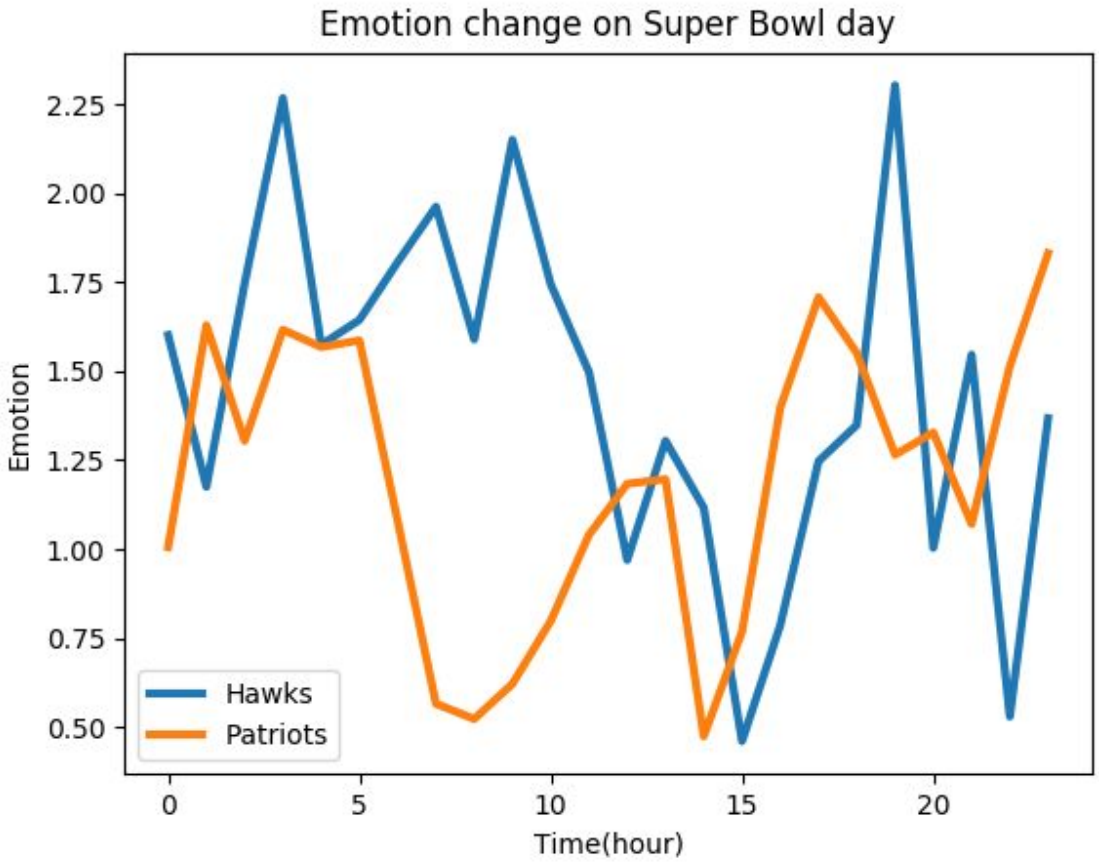


Figure 7.1 Emotion Changes in 24 Hours

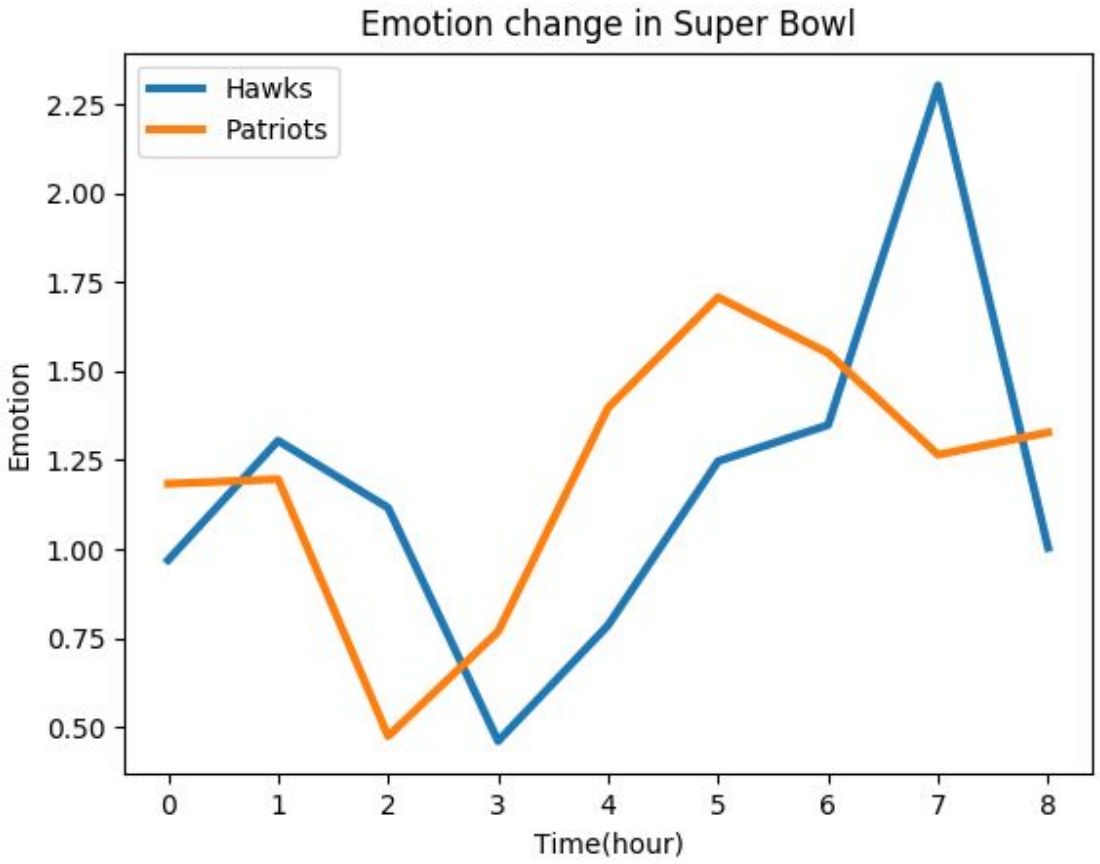


Figure 7.2 Emotion Changes from 0 hours to 8 hours after the game started

Table 7.1 shows some facts from the 2015 game. We can see that 190 minutes into the game, the hawks led the patriots by 24:14, after which point they no longer got points. The patriots fought back and won the game. This justifies why hawks fans started to go negative and patriots fans positive at around 6:30 pm, which is reflected in their tweets.

| Minutes After the Game Started | Patriots : Hawks |
|---|---|
| 88 | 7:0 |
| 105 | 7:7 |
| 120 | 14:7 |
| 130 | 14:14 |
| 175 | 14:17 |

| | |
|---|---|
| 190 | 14:24 |
| 227 | 21:24 |
| 243 | 28:24 |
| END | 28:24 |

Table 7.1 Game Facts of 2015 SuperBowl