EE239AS - Project 4

Yuxin Jin (104195828), Jamie Lee (604589757),
David Hong (204588953) and Nick Cirillo (103834979)

March 18, 2016

1 Popularity Prediction on Twitter

With the increase in popularity of social media and communication, Twitter has emerged as a major platform for online networking which allows users to post and read 140-character messages called tweets. Tweets are publicly viewable via the website and users can subscribe to other user's tweets in the form "following" another user. A word, phrase or topic that is mentioned at a greater rate than others is said to be a "trending topic" and can often be recognized in the form of hashtags ie. #TrendingTopic. These topics become popular usually either through a specific event or topic that prompts people to discuss it or by purely users creating these discussions. This information easily allows users to obtain broad perspectives and recent updates. In addition to users, these trending and bursting topics have sparked recent interest in the scientific community to analyze and predict these topics.

In this project we seek to analyze such data by utilizing current and previous tweet activity for given hashtags in order to predict future tweet activity and behavior. Specifically we analyze tweet data for the 2015 Super Bowl (Seattle Seahawks vs. New England Patriots) from a period of two weeks before the game to a week after the game. We will use from related hashtags to train a regression model and use the same model to make predictions for other hashtags.

1.1 Part 1:

The training data was downloaded and statistics were calculated for each of the hashtags (#gohawks, #gopatriots, #nfl, #patriots, #sb49 and #superbowl). We report data on each hashtag for the total number of tweets, average number of tweets per hour, average number of retweets and average number of followers per users. Statistics for each hashtag can be found in the tables below.

The total tweet count per hour both both the #SuperBowl and the #NFL hashtags are also shown below. Both hashtags noticeably show initial trending prior to the SuperBowl, show significant bursting during the SuperBowl event and both die relatively quickly after the SuperBowl (Feburary 1, 2015).

Total # of Tweets	188135
Avg. # of Tweets per Hour	276.669118
Avg. # of Retweets	0.209164
Avg # of Followers of (77168) Users	1720.634084

Table 1: #gohawks statistics

Total # of Tweets	26231
Avg. # of Tweets per Hour	58.551339
Avg. # of Retweets	0.026838
Avg # of Followers of (18005) Users	1559.278200

Table 2: #gopatriots statistics

Total # of Tweets	259019
Avg. # of Tweets per Hour	420.485390
Avg. # of Retweets	0.050938
Avg # of Followers of (75167) Users	4399.303205

Table 3: #nfl statistics

Total # of Tweets	489710
Avg. # of Tweets per Hour	736.406015
Avg. # of Retweets	0.091462
Avg # of Followers of (326173) Users	1865.903974

Table 4: #patriots statistics

Total # of Tweets	826905
Avg. # of Tweets per Hour	1531.305556
Avg. # of Retweets	0.178023
Avg # of Followers of (590066) Users	2247.285607

Table 5: #sb49 statistics

Total # of Tweets	1348766
Avg. # of Tweets per Hour	2207.472995
Avg. # of Retweets	0.136686
Avg # of Followers of (689690) Users	4228.627878

Table 6: #superbowl statistics

2
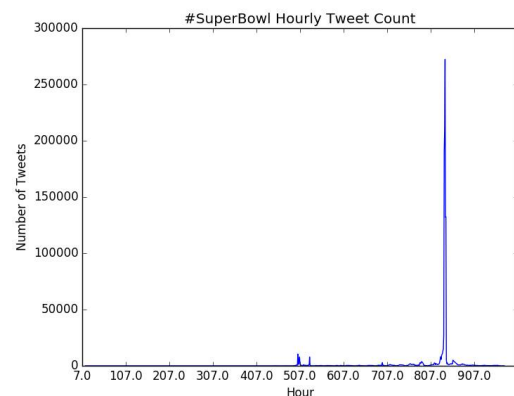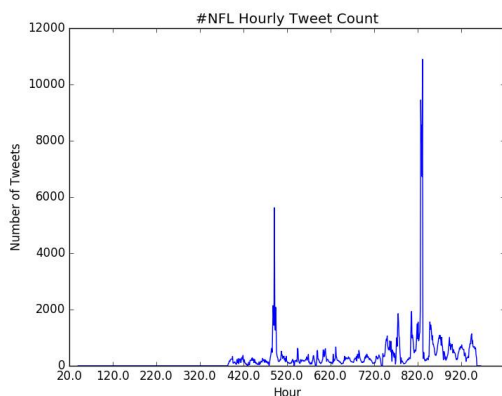
1.2 Part 2:

We now fit a linear regression model using 5 features to predict the numbers of tweets in a next hour using features extracted from a previous hour. The features used are: number of tweets, total number of retweets, sum of the number of followers of the users posting in the hashtag and the time of day.

The T-test feature values and P-value feature values are plotted below per hashtag. Additionally R^2 values are shown for 4 different linear methods per hashtag. The feature numbers correspond as follows. Feature 1: Total number of retweets, Feature 2: sum of the number of followers of the users posting in the hashtag, Feature 3: Maximum followers and Feature 4: Hour.

We used R^2 to explain our model's training accuracy. The best possible score is 1.0 however the value can be negative because it is possible the model can be arbitrarily worse. A constant model that always predicts the same value, disregarding the input features would achieve an R^2 of 0.0.

Different linear models were observed to have varying performance depending on the hashtag. For most of the hashtags the LARS Lasso method was observed to have the best possible score. Support Vector Regression was observed to have undesirable performance across all hashtags.

A t-test is a statistical examination of two population means. With relevance to our model the t-test represents how well that feature does with our model, the higher the t-test value the better.

The p-value represents a function of the observed results relative to our model and measures how extreme the observation is. Specifically, a small p-value indicates strong evidence against the null hypothesis and a large value indicates weak evidence against the null hypothesis.

T-test Feature 1	32.254497
T-test Feature 2	12339.208428
T-test Feature 3	2.607192
T-test Feature 4	nan
P-value Feature 1	0.185462
P-value Feature 2	0.000000
P-value Feature 3	0.182427
P-value Feature 4	nan

Table 7: #gohawks: T-test and P-value

Linear Regression R^2	-3.431273
Logistic Regression R^2	0.137352
Support Vector Regression R^2	-0.083056
LARS Lasso R^2	0.941716

Table 8: #gohawks: R^2 Scores for different linear regression models

T-test Feature 1	nan
T-test Feature 2	7148.513718
T-test Feature 3	3.428438
T-test Feature 4	nan
P-value Feature 1	nan
P-value Feature 2	0.000000
P-value Feature 3	0.181693
P-value Feature 4	nan

Table 9: #gopatriots: T-test and P-value

Linear Regression R^2	0.866516
Logistic Regression R^2	0.257283
Support Vector Regression R^2	-0.106411
LARS Lasso R^2	0.857947

Table 10: #gopatriots: R^2 Scores for different linear regression models

T-test Feature 1	75.215080
T-test Feature 2	1604.618371
T-test Feature 3	3.178866
T-test Feature 4	nan
P-value Feature 1	0.432376
P-value Feature 2	0.000003
P-value Feature 3	0.202118
P-value Feature 4	nan

Table 11: #nfl: T-test and P-value

Linear Regression R^2	-2.498618
Logistic Regression R^2	0.045448
Support Vector Regression R^2	-0.151815
LARS Lasso R^2	-2.148534

Table 12: #nfl: R^2 Scores for different linear regression models

T-test Feature 1	61.083197
T-test Feature 2	19299.952523
T-test Feature 3	1.501174
T-test Feature 4	nan
P-value Feature 1	0.226676
P-value Feature 2	0.000000
P-value Feature 3	0.388718
P-value Feature 4	nan

Table 13: #patriots: T-test and P-value

Linear Regression R^2	-2.291648
Logistic Regression R^2	0.057163
Support Vector Regression R^2	-0.089487
LARS Lasso R^2	0.912602

Table 14: #patriots: R^2 Scores for different linear regression models

T-test Feature 1	243.116177
T-test Feature 2	292994.277252
T-test Feature 3	0.744969
T-test Feature 4	nan
P-value Feature 1	0.289677
P-value Feature 2	0.000000
P-value Feature 3	0.472409
P-value Feature 4	nan

Table 15: #sb49: T-test and P-value

Linear Regression R^2	0.941199
Logistic Regression R^2	0.055304
Support Vector Regression R^2	-0.084825
LARS Lasso R^2	0.949693

Table 16: #sb49: R^2 Scores for different linear regression models

T-test Feature 1	659.440411
T-test Feature 2	30766.778560
T-test Feature 3	0.908156
T-test Feature 4	nan
P-value Feature 1	0.302070
P-value Feature 2	0.000000
P-value Feature 3	0.422320
P-value Feature 4	nan

Table 17: #superbowl: T-test and P-value

Linear Regression R^2	-0.102577
Logistic Regression R^2	0.042759
Support Vector Regression R^2	-0.112585
LARS Lasso R^2	-0.098248

Table 18: #superbowl: R^2 Scores for different linear regression models

1.3   Part 3:

In part 3 we design our own linear regression model using additional features at our discretion. In addition to the previous features used in Part 2. The features we added were the average number of tweets for each user, a threshold of the number of users who tweeted (threshold = 1), the number of users who tweeted greater than or equal to 3 times (threshold 3), the number of active users (users who tweet every hour in the past 6 hours) and the number of tweets with 100 characters or more (threshold = 100).

The T-test feature values and P-value feature values are plotted below per hashtag. Additionally R^2 values are shown for 4 different linear methods per hashtag. The feature numbers correspond as follows. Feature 1: Total number of retweets, Feature 2: sum of the number of followers of the users posting in the hashtag, Feature 3: Maximum followers, Feature 4: Hour, Feature 5: Average number of tweets for each user, Feature 6: Threshold of the number of users who tweeted (threshold 1), Feature 7: Threshold of the number of users who tweeted greater than or equal to 3 times (threshold 3) and Feature 8: the number of tweets with 100 characters or more (threshold = 100).

In our model we found the most significant features to be mainly: 1. The number of followers someone has, 2. The number of users tweeting that hashtag and to a lesser extent 3. the number of tweets that contain 100 characters or more.

Scatter plots show the significance our top three features (showing the number of tweets vs our feature) and all show linear relationships and therefore act as good predictors

T-test Feature 1	32.254497
T-test Feature 2	12339.208428
T-test Feature 3	2.607192
T-test Feature 4	nan
T-test Feature 5	2.305868
T-test Feature 6	26540.184413
T-test Feature 7	645.459722
T-test Feature 8	1392.503026
P-value Feature 1	0.185462
P-value Feature 2	0.000000
P-value Feature 3	0.182427
P-value Feature 4	nan
P-value Feature 5	0.451127
P-value Feature 6	0.000000
P-value Feature 7	0.000026
P-value Feature 8	0.000000

Table 19: #gohawks: T-test and P-value

Linear Regression R^2	0.830944
Logistic Regression R^2	0.137352
Support Vector Regression R^2	-0.083049
LARS Lasso R^2	0.978351

Table 20: #gohawks: R^2 Scores for different linear regression models

T-test Feature 1	nan
T-test Feature 2	7148.513718
T-test Feature 3	3.428438
T-test Feature 4	nan
T-test Feature 5	22.746685
T-test Feature 6	621064.374186
T-test Feature 7	nan
T-test Feature 8	7142.937408
P-value Feature 1	nan
P-value Feature 2	0.000000
P-value Feature 3	0.181693
P-value Feature 4	nan
P-value Feature 5	0.572824
P-value Feature 6	0.000000
P-value Feature 7	nan
P-value Feature 8	0.000000

Table 21: #gopatriots: T-test and P-value

Linear Regression R^2	0.989521
Logistic Regression R^2	0.257283
Support Vector Regression R^2	-0.106411
LARS Lasso R^2	0.916684

Table 22: #gopatriots: R^2 Scores for different linear regression models

T-test Feature 1	75.215080
T-test Feature 2	1604.618371
T-test Feature 3	3.178866
T-test Feature 4	nan
T-test Feature 5	6.831252
T-test Feature 6	2187.079982
T-test Feature 7	267.385951
T-test Feature 8	1835.564400
P-value Feature 1	0.432376
P-value Feature 2	0.000003
P-value Feature 3	0.202118
P-value Feature 4	nan
P-value Feature 5	0.183349
P-value Feature 6	0.000000
P-value Feature 7	0.000000
P-value Feature 8	0.000000

Table 23: #nfl: T-test and P-value

Linear Regression R^2	0.978169
Logistic Regression R^2	0.045448
Support Vector Regression R^2	-0.151815
LARS Lasso R^2	0.991687

Table 24: #nfl: R^2 Scores for different linear regression models

T-test Feature 1	61.083197
T-test Feature 2	19299.952523
T-test Feature 3	1.501174
T-test Feature 4	nan
T-test Feature 5	0.425384
T-test Feature 6	196192.985044
T-test Feature 7	4203.146057
T-test Feature 8	3147.501792
P-value Feature 1	0.226676
P-value Feature 2	0.000000
P-value Feature 3	0.388718
P-value Feature 4	nan
P-value Feature 5	0.633074
P-value Feature 6	0.000000
P-value Feature 7	0.003434
P-value Feature 8	0.000000

Table 25: #patriots: T-test and P-value

Linear Regression R^2	0.541453
Logistic Regression R^2	0.057163
Support Vector Regression R^2	-0.089484
LARS Lasso R^2	0.990319

Table 26: #patriots: R^2 Scores for different linear regression models

T-test Feature 1	243.116177
T-test Feature 2	292994.277252
T-test Feature 3	0.744969
T-test Feature 4	nan
T-test Feature 5	3.109246
T-test Feature 6	1401642.010906
T-test Feature 7	32016.027715
T-test Feature 8	8617.255121
P-value Feature 1	0.289677
P-value Feature 2	0.000000
P-value Feature 3	0.472409
P-value Feature 4	nan
P-value Feature 5	0.447412
P-value Feature 6	0.000000
P-value Feature 7	0.000000
P-value Feature 8	0.000000

Table 27: #sb49: T-test and P-value

Linear Regression R^2	0.996203
Logistic Regression R^2	0.055304
Support Vector Regression R^2	-0.084825
LARS Lasso R^2	0.997678

Table 28: #sb49: R^2 Scores for different linear regression models

T-test Feature 1	659.440411
T-test Feature 2	30766.778560
T-test Feature 3	0.908156
T-test Feature 4	nan
T-test Feature 5	4.231599
T-test Feature 6	169100.241200
T-test Feature 7	22053.868096
T-test Feature 8	9162.252714
P-value Feature 1	0.302070
P-value Feature 2	0.000000
P-value Feature 3	0.422320
P-value Feature 4	nan
P-value Feature 5	0.379419
P-value Feature 6	0.000000
P-value Feature 7	0.000000
P-value Feature 8	0.000000

Table 29: #superbowl: T-test and P-value

Linear Regression R^2	0.996959
Logistic Regression R^2	0.042759
Support Vector Regression R^2	-0.112585
LARS Lasso R^2	0.997060

Table 30: #superbowl: R^2 Scores for different linear regression models
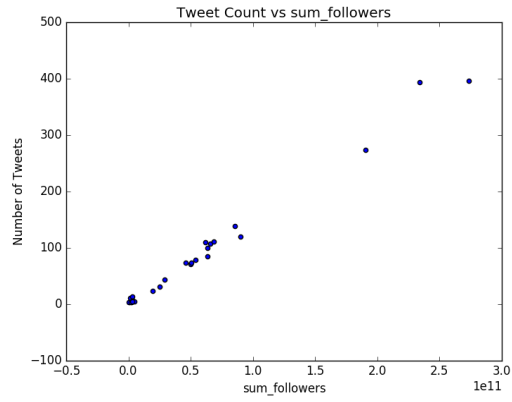
(a) Tweet Count vs Sum of Total Number of Followers
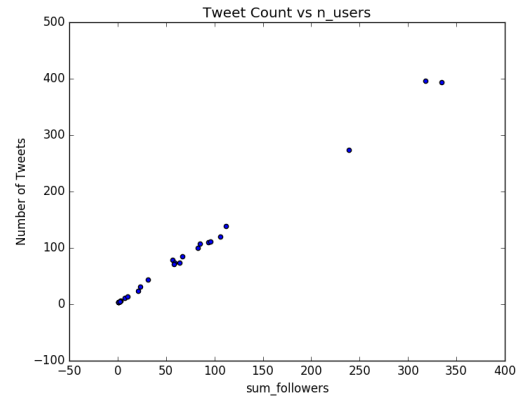
(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour
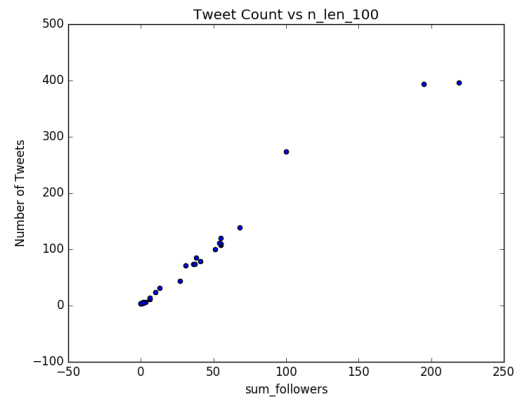
(c) Tweet Count vs Number of Tweets with 100 or More Characters

Figure 2: #gohawks: Scatter plots for top 3 features in measurements
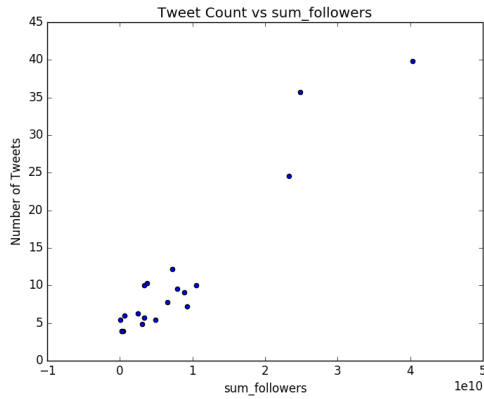
(a) Tweet Count vs Sum of Total Number of Followers
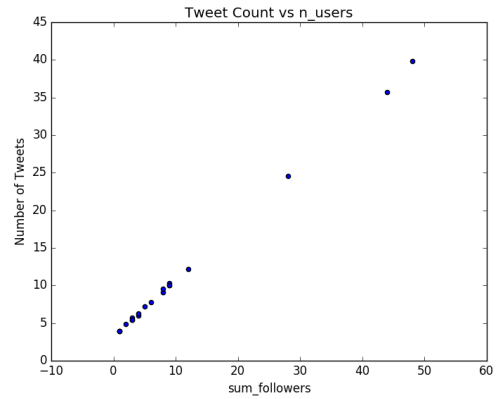
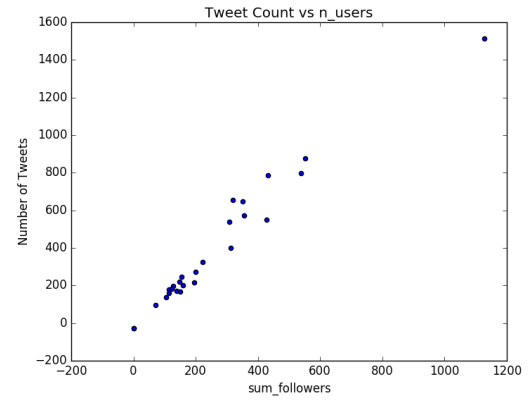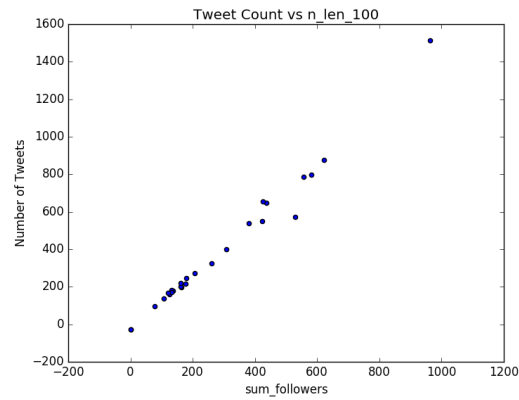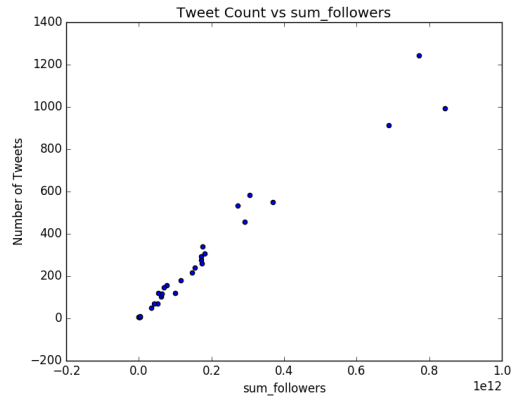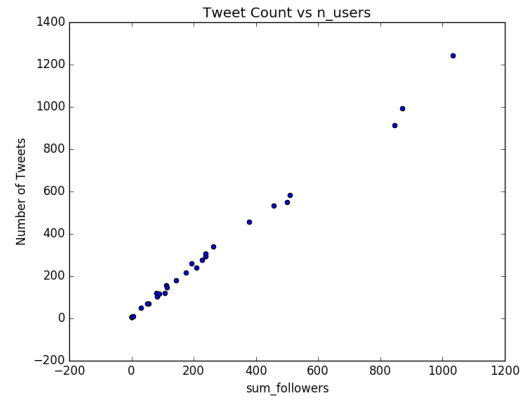(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour
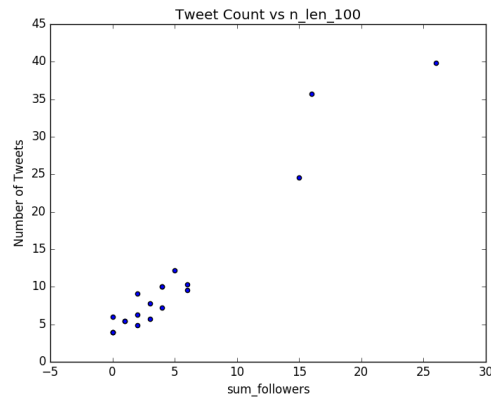
(c) Tweet Count vs Number of Tweets with 100 or More Characters

Figure 3: #gopatriots: Scatter plots for top 3 features in measurements
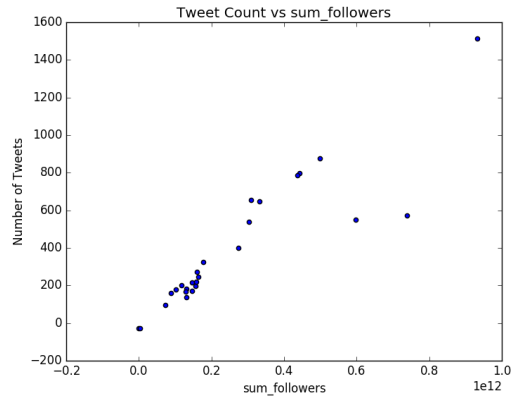
(a) Tweet Count vs Sum of Total Number of Followers

(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour

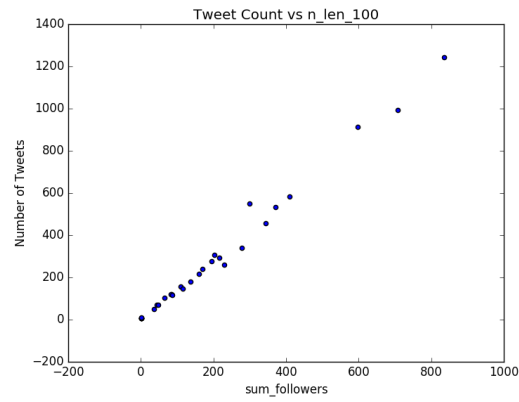(c) Tweet Count vs Number of Tweets with 100 or More Characters

Figure 4: #nfl: Scatter plots for top 3 features in measurements

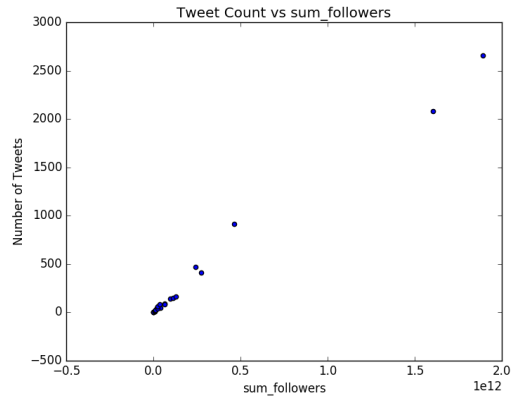(a) Tweet Count vs Sum of Total Number of Followers

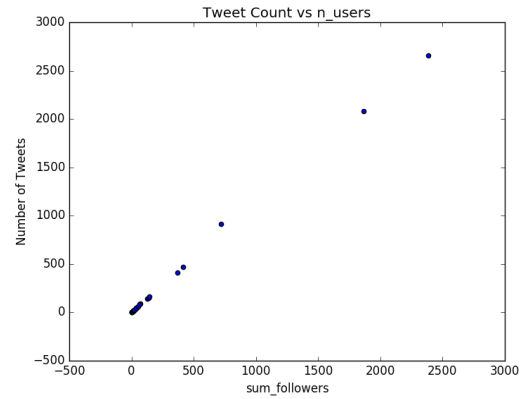(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour

(c) Tweet Count vs Number of Tweets with 100 or More Characters
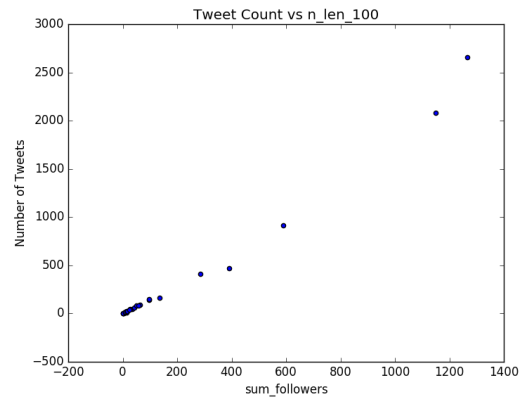
Figure 5: #patriots: Scatter plots for top 3 features in measurements

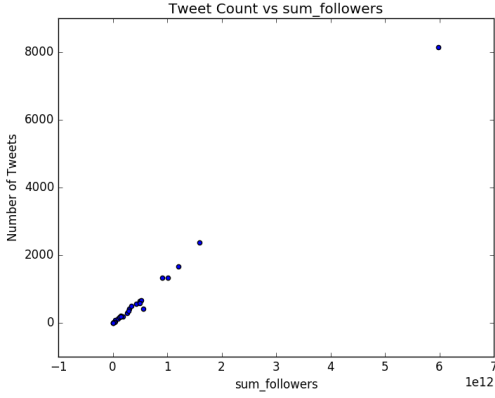(a) Tweet Count vs Sum of Total Number of Followers

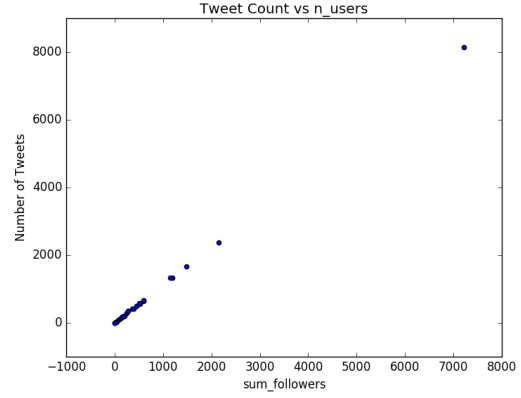(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour

(c) Tweet Count vs Number of Tweets with 100 or More Characters

Figure 6: #sb49: Scatter plots for top 3 features in measurements

17

(a) Tweet Count vs Sum of Total Number of Followers

(b) Tweet Count vs Number of Users Who Tweeted 3 or More Times Per Hour

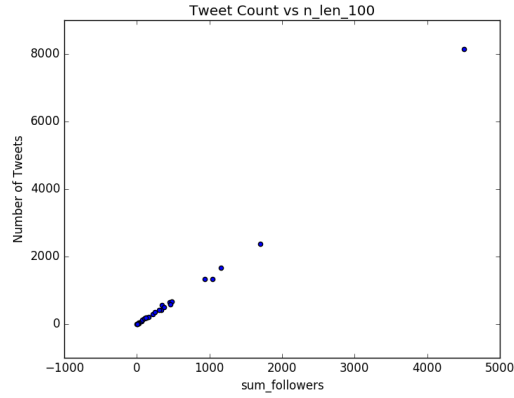(c) Tweet Count vs Number of Tweets with 100 or More Characters

Figure 7: #superbowl: Scatter plots for top 3 features in measurements

1.4   Part 4:

In part 4 we split our feature data into 10 parts for cross-validation. 10 tests are run and each time the model is fitted on 9 parts and the number of tweets is predicted for the 1 remaining part. The average prediction error is calculated over samples in the remaining part and then these values are averaged over the 10 tests.

Different regression models are then created according to the given three time periods (Period1: Before Feb 1, 8AM, Period 2: Between Feb 1, 8AM and Period 3: 8PM, and After Feb 1 8PM). These times reflect a period of time before the Super Bowl, the day including the actual Super Bowl game and a period of time after the Super Bowl. These periods of time are expected to follow respective trends of inactivity, then high bursting activity and then inactivity. Cross validation errors for the 3 models are reported in the tables below.

Period	Average Prediction Error
1	27.448149
2	36.383078
3	16.337346

Table 31: #gohawks: Average prediction error for different periods (before, during, after)

Period	Average Prediction Error
1	61.608437
2	65.836415
3	10.785898

Table 32: #gopatriots: Average prediction error for different periods (before, during, after)

Period	Average Prediction Error
1	51.328586
2	170.605770
3	27.154346

Table 33: #nfl: Average prediction error for different periods (before, during, after)

Period	Average Prediction Error
1	84.086567
2	65.676033
3	16.494728

Table 34: #patriots: Average prediction error for different periods (before, during, after)

Period	Average Prediction Error
1	3436.743379
2	1964.359263
3	40.565964

Table 35: #sb49: Average prediction error for different periods (before, during, after)

Period	Average Prediction Error
1	542.311391
2	1533.897981
3	212.234838

Table 36: #superbowl: Average prediction error for different periods (before, during, after)

1.5   Part 5:

The test data set was downloaded and our model was used to make predictions for the next hour in each test case. The predicted number of tweets for the next hour in each sample set are shown below.

Sample File	Predicted Tweets
1	60.125156
2	99.887770
3	78000.896189
4	603.955758
5	276.738015
6	287.107220
7	41118.362584
8	57.667595
9	44.229527
10	1551.637766

Table 37: #superbowl: Average prediction error for different periods (before, during, after)

1.6   Part 6:

In this portion we used bag of words representation of all the tweets regarding gopatriots and gohawks to formulate a sparse model to categorize future tweets based on the context of the tweets themselves. To start, we created a JSON parsing function to detect tweet bodies from the JSON structures. We then output each tweet into their own .txt file for sklearn to import and fit our model. For all .txt files related to gopatriots and gohawks we produced a sparse feature vector of words which excluded the words gopatriots and gohawks to discard any bias terms the sklearn?s Logistic Regression would be skewed upon. After using the TFIDF Vectorizer from the sklearn library, we were able to create a numerical representation of our tweets. As mentioned previously, we used the Logistic Regression algorithm in a 10-fold cross-validation accuracy calculation. We observed the following results of classifying tweets between gopatriots and gohawks.

Iteration	Cross Validation Score
1	91.11%
2	91.70%
3	91.54%
4	91.53%
5	91.51%
6	91.44%
7	91.48%
8	91.45%
9	91.59%
10	91.40%
Mean Accuracy	91.47%

Table 38: 10-Fold Cross Validation Scores