March 22nd, 2020

ECE 219 Project 5: Twitter Data Application

Bryan Bednarski 005428092
KeYu Chen 105331424
Hao-Jen Chien 405219534
Anthony Wang 404483002

Introduction

In this report, we detail the process of analyzing tweet metadata for six different hashtags that were used in abundance in the two weeks before and 1 week after the 2015 Superbowl game between the Seattle Seahawks and New England Patriots. These hashtags pertain to both the individual teams (and the game itself) and demonstrate the sheer magnitude of tweets posted during this time frame.

In Part 1.1 of this report, we analyze the data set, presenting some basic statistics for the data as a whole, and parse some baseline metrics for analysis. We fit a linear OLS model to these preliminary statistics and analyze the viability of this model as a predictor for the number of tweets containing each hashtag in the following hours. Next we parse some additional, nontrivial statistics from the data set, referencing the paper from Kong et al. [2014] that was provided as a reference with the assignment. We again fit a linear model to these features and analyze the success of these new predictors in a linear model. Furthermore, we perform additional time-based analysis on the hashtags by windowing them according to the number of incoming tweets, and by aggregating the data from all hashtags together.

In Part 1.2, we use the features accumulated from above to fit our models to Random Forest Regressors and Gradient Boosting Regression models from the sklearn API. We use a randomized grid search to find the best parameters from each of these methods and compare our results to the previous linear regression. Finally, we perform the same time-windowed analysis as in Part 1.1, this time using non-linear regressors to predict the number of tweets in the next time window.

Part 1.3 continues nonlinear regression, using a neural network (MLPRegressor) to predict the number of tweets in the next hour. We try both a standard scaling method and grid search to optimize our results.

In Part 2 of this project, we built a binary classifer to predict the location of fans posting tweets from the location field of the JSON metadata. This binary classifier is built to determine whether the tweet was sent from Washington State or Massachusetts. We plot the ROC curve, confusion matrix, and calculate model success metrics to analyze our results.

In Part 3 we create our own project based on classifying the tweets based on locations. We scrape through every tweet, save the GPS coordinates if available, and then use a random forest classifier (with cross validation and a randomized grid search). We do this for the ”gopatriots” versus the ”gohawks” as a test, which achieves a great accuracy. Then we do it for all 6 hashtags.

Part 1: Popularity Prediction

Question 1: Basic Statistics: Report the following statistics for each hashtag. Average number of tweets per hour, average number of followers of users posting each tweet*, average number of retweets per tweet.

** Note: we average over the number of tweets; if a users posted twice, we count the user and the user's followers twice as well*

Results for basic statistics parsing and analysis below. From this data, you can see that the most popular hashtags tend to be the most generic as they play into much larger pools of metadata. For example, #superbowl and #sb49 represent both the largest overall number of tweets, but also the largest follower per tweet and retweet ratios. We presume that these statistics are likely a result of the large follower bases that major media outlets tend to have, and their tendency to promote the game more generally than individual fans with fewer followers who are more likely to promote a particular team like #gopatriots and #gohawks.

Hashtag	Number of Tweets	Tweets / Hour (avg.)	Follower / Tweet (avg.)	Retweets / Tweet (avg.)
#patriots	440,521	750.89	3,280.46	1.79
#superbowl	1,213,812	2,072.12	8,814.97	2.39
#nfl	223,022	397.02	4,662.38	1.53
#gopatriots	23,511	40.96	1427.25	1.41
#sb49	743,649	1,276.86	10,374.16	2.53
#gohawks	169122	292.49	2,217.92	2.01

Question 2: Visualize Tweet Rate: Plot "number of tweets in hour" over time for #superbowl and #NFL as a bar plot with 1-hour bins.

The result below visualizes the number of tweets per hour that were submitted in the two weeks leading up to, and one week after the Superbowl game. In this section, we only plot the data for two hashtags, #superbowl and #NFL. For each of these hashtags, we present plots of both linear and logarithmic scale on the y-axis to provide different interpretations of the resulting data.

For the #superbowl data on a log scale, we can see the number of tweets per hour rising steady on an exponential scale from an order of 10e1 at hour zero, to around 10e3 around hour 400 (the day or two before the game). However, when game day comes around the order of magnitude of references jumps another two orders and for an hour, and almost a quarter million tweets are being posted containing #superbowl. After the game, the number drops at a faster rate than it was rising in the days before, over the next week. From these plots we can also see a clear trend in the number of posts during the North American days and nights.

A similar trend can be observed for tweets containing #NFL. However, these tweets occur with an order of magnitude fewer than #superbowl, capping at just above 11,000 at its peak. Additionally, there is a rise in the number of tweets per hour around hour 100 from the beginning of the time period for both categories, which would likely correspond to some news about the game or interview with players.

#superbowl

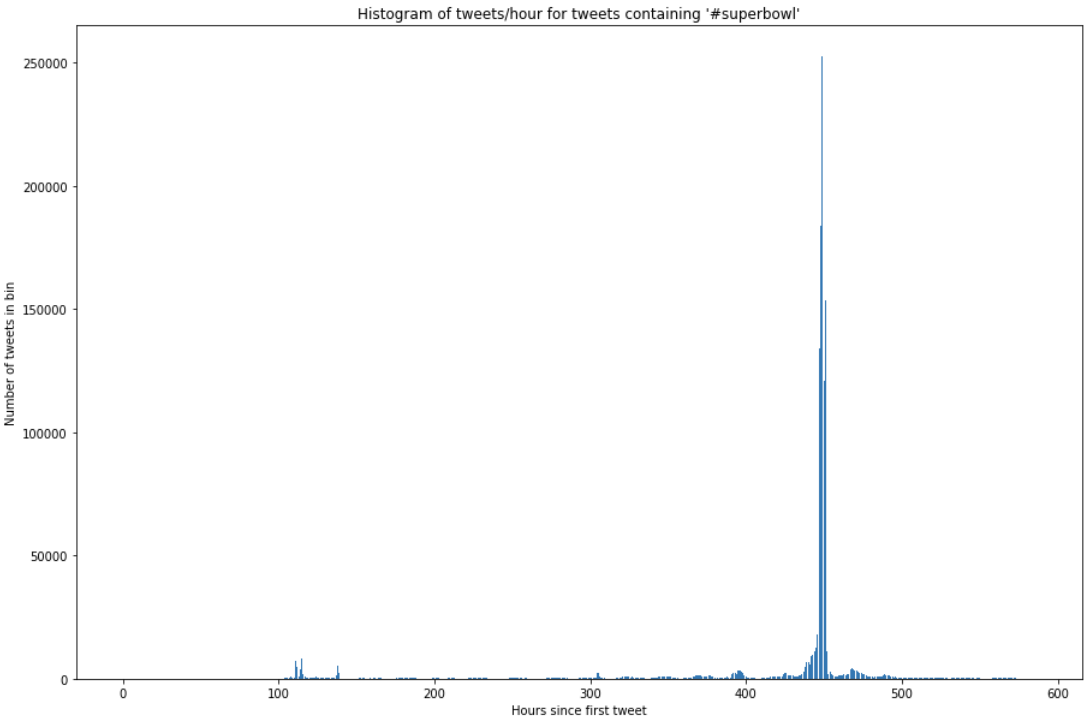


Figure 1: Tweets per hour containing #superbowl, in 1 hour bins. Y-axis on a linear scale.

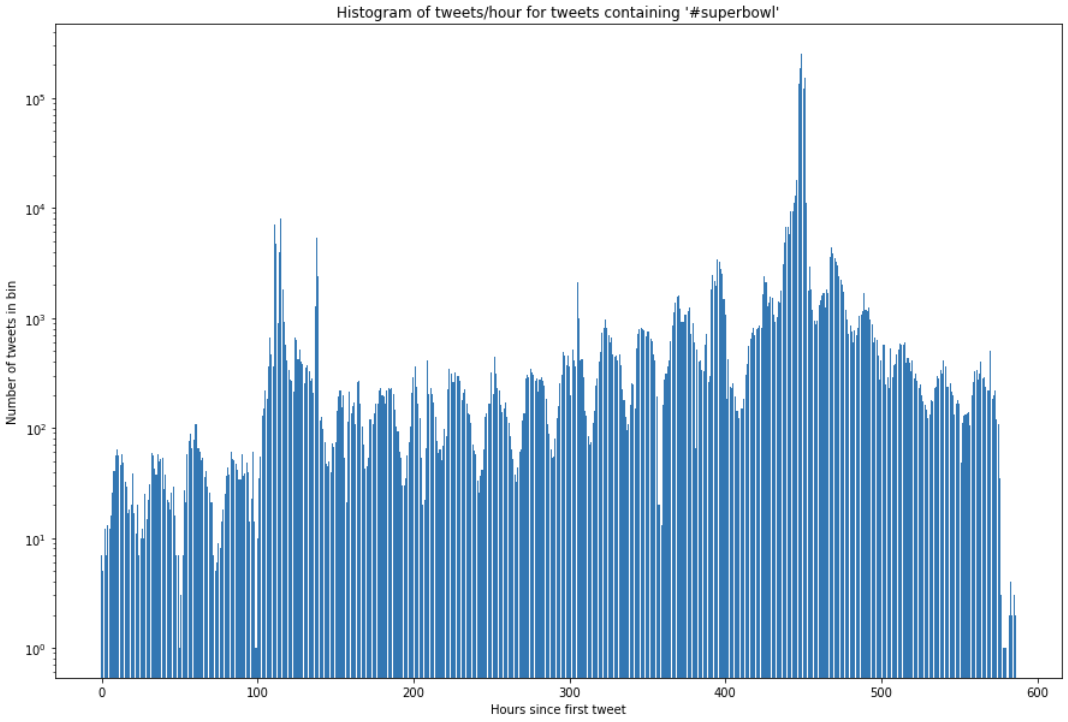


Figure 2: Tweets per hour containing #superbowl, in 1 hour bins. Y-axis on a logarithmic scale.

#NFL

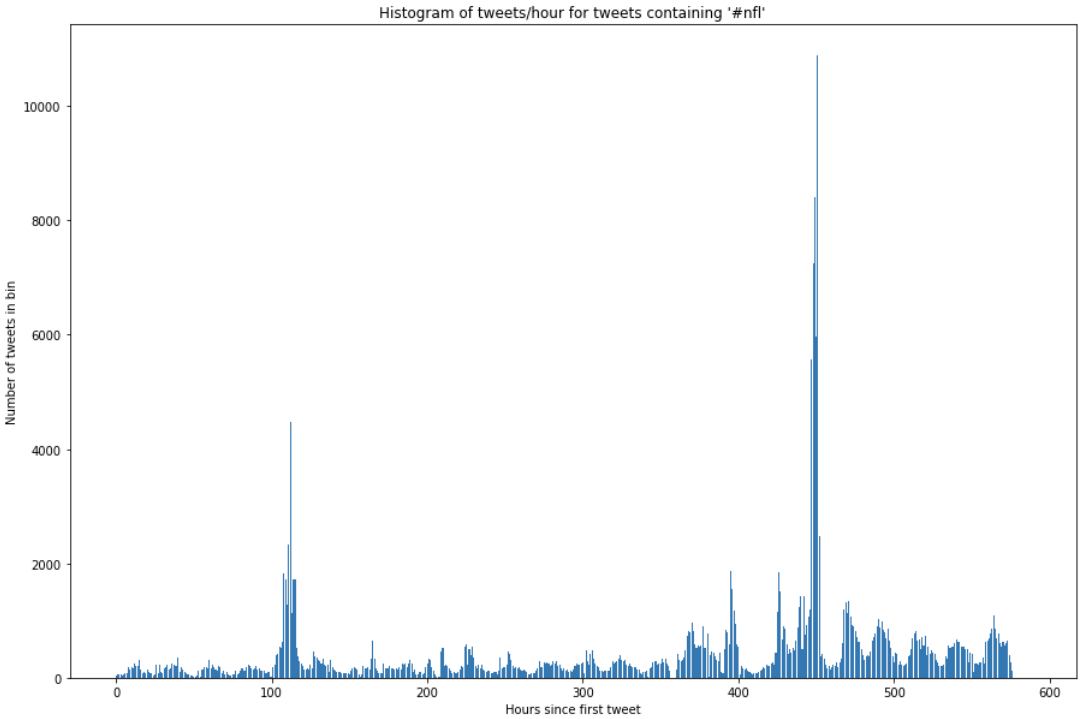


Figure 3: Tweets per hour containing #NFL, in 1 hour bins. Y-axis on a linear scale.

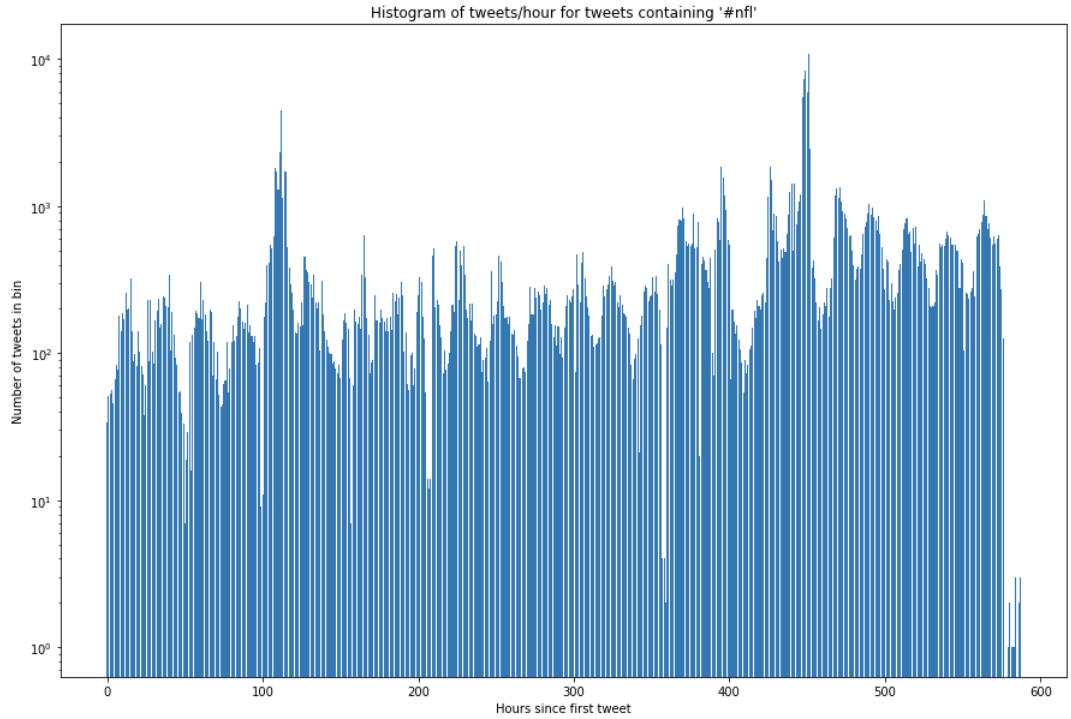


Figure 4: Tweets per hour containing #NFL, in 1 hour bins. Y-axis on a logarithmic scale.

Question 3: Linear Regression: Extract 5 basic features from the data and fit a linear regression model for each hashtag in dataset. Report MSE and R-Squared measure. Analyze the significance of t-test and p-value.

In this section, we apply a linear regression model to each of the 6 hashtags after extracting 5 basic features from the metadata. These features are totaled into one hour bins so that the linear model is capable of predicting the total number of tweets in the following hour. The features that we use for this analysis are as follows:

- Number of tweets
- Total number of retweets
- Sum of the number of followers of the users posting the hashtags
- Max number of followers of the users posting the hashtags
- Time of day (0-23) in a specific timezone

Results for each hashtag, fit to a linear regression model (using the OLS model from the StatsModels API) follow. For each model, we will consider a number of analytic results. First, we summarize the success of the model by the RMSE of the predicted number of tweets in the next hour compared to the actual count. Additionally, we consider the R-squared metric, which describes the total fraction of each truth value that can be accounted for by the model's prediction while considering the features that the model was provided. Finally, we describe the t-test, and two-tailed p-test results that are summarized by the StatsModels API, describing the statistical significance of each model. In a linear model, each feature should represent a positive or negative coefficient to the line of best fit. Therefore, if the 95% confidence interval contains 0 (bounded by both positive and negative values), we can determine the feature to be generally insignificant.

OLS Fit Results Considering 5 Basic Features

Hashtag	RMSE	Adjusted R-Squared Value
#patriots	2,276.157	0.666
#superbowl	7,244.548	0.798
#nfl	519.579	0.567
#gopatriots	166.095	0.626
#sb49	4,023.484	0.803
#gohawks	870.95	0.472

#patriots

```
Regression analysis on '#patriots' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.668
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     233.8
Date:                Fri, 20 Mar 2020   Prob (F-statistic):          1.91e-136
Time:                        17:32:15   Log-Likelihood:                -5361.4
No. Observations:                 586   AIC:                         1.073e+04
Df Residuals:                     580   BIC:                         1.076e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         180.1751    183.925      0.980      0.328    -181.066     541.416
num_tweets      0.9145      0.071     12.937      0.000       0.776       1.053
num_retweets   -0.0681      0.058     -1.178      0.239      -0.181       0.045
num_followers -1.098e-05   2.63e-05     -0.417      0.677   -6.27e-05    4.07e-05
max_followers   0.0001    9.17e-05      1.340      0.181   -5.72e-05       0.000
hour_of_day    -5.8597     13.765     -0.426      0.670     -32.896      21.176
==============================================================================
Omnibus:                      887.682   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           690539.222
Skew:                           7.937   Prob(JB):                         0.00
Kurtosis:                     170.420   Cond. No.                     1.60e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.6e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  5180890.103
Model RMSE:  2276.157
Model R-squared:  0.668
```

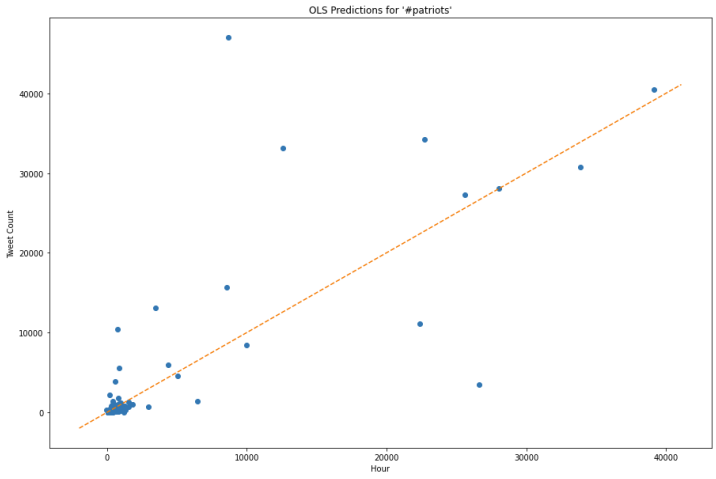


Figure 5: Regression results predicting number of tweets in next hour window for #patriots and OLS fit predicting number of tweets in next hour window for #patriots

From the results for #patriots fit to an OLS model, we determine an R^2 value of 0.668 and RMSE of 2276. The R^2 value seen here is significantly high for a simple model and only 5 features. The RMSE represents error only in tweets containing #patiots in the content. From the statistical analysis of the features, we determine that 'num_tweets', the number of tweets in the previous hour, to be the best predictor for this hashtag. 'num_tweets' in this case has a two-tailed p-test value of 0, which is good and a t-value of 12.937, showing that this feature is statistically significant. Other features considered here: 'num_retweets', 'num_followers', 'max_followers' and 'hour_of_day' in this case all have p-values greater than zero, and fairly low t-test values, showing that in this case they are not very statistically significant.

#superbowl

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Regression analysis on '#superbowl' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     463.5
Date:                Fri, 20 Mar 2020   Prob (F-statistic):          6.72e-200
Time:                        17:32:15   Log-Likelihood:                -6039.9
No. Observations:                 586   AIC:                         1.209e+04
Df Residuals:                     580   BIC:                         1.212e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -149.5572    605.382     -0.247      0.805   -1338.565    1039.451
num_tweets     2.2766      0.080     28.537      0.000       2.120       2.433
num_retweets  -0.2543      0.046     -5.544      0.000      -0.344      -0.164
num_followers -0.0001   2.2e-05     -6.265      0.000      -0.000   -9.47e-05
max_followers  0.0007      0.000      4.889      0.000       0.000       0.001
hour_of_day  -20.4965     43.624     -0.470      0.639    -106.177      65.184
==============================================================================
Omnibus:                      973.862   Durbin-Watson:                   2.283
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1787388.254
Skew:                           9.272   Prob(JB):                         0.00
Kurtosis:                     272.925   Cond. No.                     2.21e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.21e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  52483472.229
Model RMSE:  7244.548
Model R-squared:  0.8
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

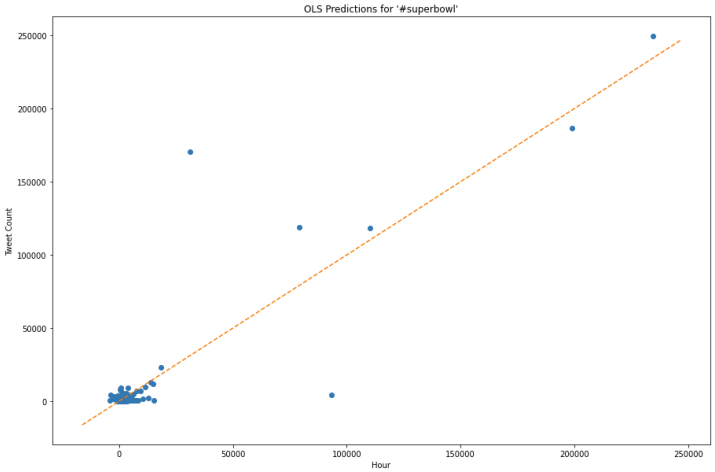


Figure 6: Regression results predicting number of tweets in next hour window for #superbowl and OLS fit predicting number of tweets in next hour window for #superbowl

The results of an OLS fit for #superbowl, show an R-squared value of 0.800 and RMSE of 7244.548. Predicting tweets for #superbowl in this case has a higher R-squared coefficient than for #patriots, showing that in general, the features we are considering are more significant predictors. However, the RMSE is higher because the scale of tweets per hour that we are working with is significantly larger. Referring to our results in question 2, we see that on average #superbowl has almost 3 times the number of tweets per hour than #patriots.

Additionally, when looking at the statistical significance of each feature through the two-tailed p-value results, we see that 4 of the five basic features have a value of 0.0. Only 'hour_of_day' has a non-zero p-value, deeming it insignificant. Additionally, the t-test values for the other four features all have significant magnitudes, specifically: 28.537, -5.544, -6.265, 4.889. None of these features have 95% confidence intervals that cross zero. 'max_followers' 95% confidence interval contains zero, and cohesively has the lowest magnitude of t-test values.

#nfl

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Regression analysis on '#nfl' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.571
Model:                            OLS   Adj. R-squared:                  0.567
Method:                 Least Squares   F-statistic:                     154.3
Date:                Fri, 20 Mar 2020   Prob (F-statistic):          4.56e-104
Time:                        17:32:15   Log-Likelihood:                -4495.8
No. Observations:                 586   AIC:                             9004.
Df Residuals:                     580   BIC:                             9030.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        123.9278     42.889      2.889      0.004      39.691     208.165
num_tweets     0.5671      0.135      4.203      0.000       0.302       0.832
num_retweets  -0.1653      0.064     -2.592      0.010      -0.291      -0.040
num_followers  0.0001   2.5e-05      4.578      0.000    6.53e-05       0.000
max_followers -0.0001   3.31e-05     -3.524      0.000      -0.000   -5.16e-05
hour_of_day    0.4058      3.155      0.129      0.898      -5.791       6.603
==============================================================================
Omnibus:                      670.042   Durbin-Watson:                   2.373
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           350954.169
Skew:                           4.595   Prob(JB):                         0.00
Kurtosis:                     122.537   Cond. No.                     8.60e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.6e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  269962.153
Model RMSE:  519.579
Model R-squared:  0.571
```

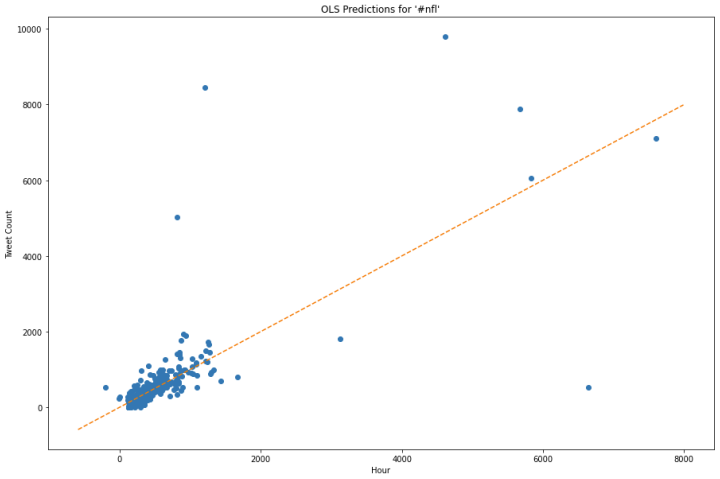


Figure 7: Regression results predicting number of tweets in next hour window for #nfl and OLS fit predicting number of tweets in next hour window for #nfl

#nfl has a lower number of tweets per hour than both of the previous hashtagged tweets discussed in this section. As a result, this hashtag also has a very low RMSE value of 519. However, the R-squared value is 0.567, which is lower than the two previous classes that we have analyzed. For this hashtag, there are three statistically significant features: 'num_tweets', 'num_followers' and 'max_followers'. These features have corresponding significant magnitudes for t-test results of 4.203, -2.592 and -3.53 and confidence intervals that do not contain zero. The remaining two features, 'num_followers' and 'hours_of_day' are shown to not be statistically significant.

#gopatriots

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Regression analysis on '#gopatriots' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.629
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     192.9
Date:                Fri, 20 Mar 2020   Prob (F-statistic):          6.97e-120
Time:                        17:32:15   Log-Likelihood:                -3749.1
No. Observations:                 574   AIC:                             7510.
Df Residuals:                     568   BIC:                             7536.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.2632     13.562      0.683      0.495     -17.375      35.901
num_tweets     0.3054      0.285      1.073      0.284      -0.254       0.865
num_retweets   0.4947      0.191      2.590      0.010       0.120       0.870
num_followers -0.0001      0.000     -0.525      0.599      -0.001       0.000
max_followers -1.937e-05   0.000     -0.089      0.929      -0.000       0.000
hour_of_day    -0.1991     1.017     -0.196      0.845      -2.197       1.799
==============================================================================
Omnibus:                      486.048   Durbin-Watson:                   1.909
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            29105.311
Skew:                           2.526   Prob(JB):                         0.00
Kurtosis:                     113.192   Cond. No.                     7.48e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  27587.448
Model RMSE:  166.095
Model R-squared:  0.629
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

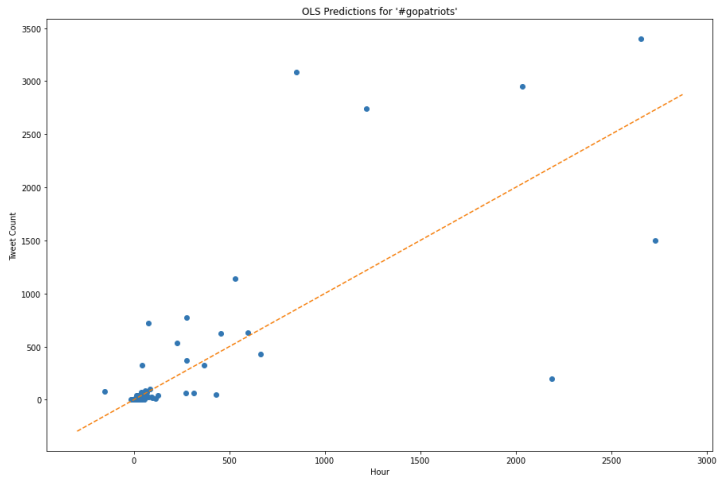


Figure 8: Regression results predicting number of tweets in next hour window for #gopatriots and OLS fit predicting number of tweets in next hour window for #gopatriots

Results for OLS fit on tweets containing #gopatriots demonstrates the lowest RMSE values that we had observed in the previous 3 tweets with a value of 166.095. These results show a average-level R-squared value of 0.626 and no features of particular statistical importance. Of the 5 features being considered, none of them demonstrate a two-trailed p-test that is low enough (less than alpha) to be considered important. Additionally, all of them have fairly low t-test results and poor 95% confidence interval results. As a result, I would not recommend this model with this data set, as perturbations in any feature would not be valuable for predicting the resulting number of tweets.

#sb49

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Regression analysis on '#sb49' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     474.3
Date:                Fri, 20 Mar 2020   Prob (F-statistic):          1.66e-201
Time:                        17:32:15   Log-Likelihood:                -5656.4
No. Observations:                 582   AIC:                         1.132e+04
Df Residuals:                     576   BIC:                         1.135e+04
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          206.5402    328.558      0.629      0.530    -438.777     851.857
num_tweets       1.1363      0.087     13.020      0.000       0.965       1.308
num_retweets    -0.1605      0.079     -2.039      0.042      -0.315      -0.006
num_followers  9.719e-06   1.25e-05      0.777      0.438   -1.49e-05    3.43e-05
max_followers  9.449e-05   4.36e-05      2.166      0.031    8.79e-06       0.000
hour_of_day     -15.9597     24.434     -0.653      0.514     -63.950      32.031
==============================================================================
Omnibus:                     1179.269   Durbin-Watson:                   1.674
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2205207.514
Skew:                          14.582   Prob(JB):                         0.00
Kurtosis:                     303.143   Cond. No.                     1.57e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.57e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  16188424.182
Model RMSE:  4023.484
Model R-squared:  0.805
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

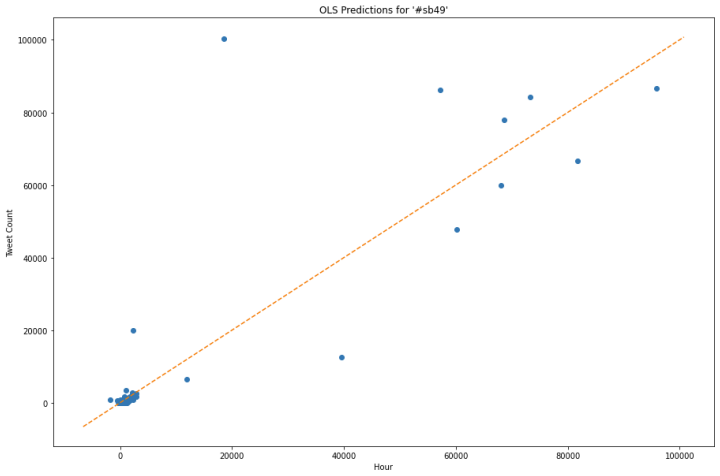


Figure 9: Regression results predicting number of tweets in next hour window for #sb49 and OLS fit predicting number of tweets in next hour window for #sb49

Result for tweets containing #sb49 are more significant than the previous model, demonstrating a R-squared value of 0.805, which is nearly the highest that we have seen so far with an OLS fit and only five features being considered. Additionally, with an RMSE of 4,023.48 and the second highest number of tweets per hour, we believe that the OLS does a pretty good job of fitting predictions for #sb49. Interestingly, so far we have seen the highest R-squared values with the classes that are the most general in their description, meaning that that they do not cater to a particular audience like the team-based tweets. This sort of description may lead to qualities in the features that are simpler and easier to fit with a linear model.

For this model, we only see one feature with a two-tailed p-value result of 0.0, showing statistical significance. This feature is 'num_tweets', and it additionally has a very high t-test magnitude and 95% confidence interval that does not contain zero. The other features have positive p-value results and low t-test results.

#gohawks

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Regression analysis on '#gohawks' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.476
Model:                            OLS   Adj. R-squared:                  0.472
Method:                 Least Squares   F-statistic:                     104.1
Date:                Fri, 20 Mar 2020   Prob (F-statistic):           5.01e-78
Time:                        17:32:15   Log-Likelihood:                -4733.0
No. Observations:                 578   AIC:                             9478.
Df Residuals:                     572   BIC:                             9504.
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           95.0899     70.543      1.348      0.178     -43.465     233.645
num_tweets       1.2827      0.164      7.831      0.000       0.961       1.604
num_retweets    -0.1364      0.043     -3.138      0.002      -0.222      -0.051
num_followers   -0.0002      8e-05     -2.429      0.015      -0.000   -3.72e-05
max_followers  6.154e-05      0.000      0.413      0.680      -0.000       0.000
hour_of_day       1.6195      5.325      0.304      0.761      -8.839      12.078
==============================================================================
Omnibus:                      916.585   Durbin-Watson:                   2.216
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           783084.769
Skew:                           8.690   Prob(JB):                         0.00
Kurtosis:                     182.481   Cond. No.                     5.13e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.13e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  758554.248
Model RMSE:  870.95
Model R-squared:  0.476
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

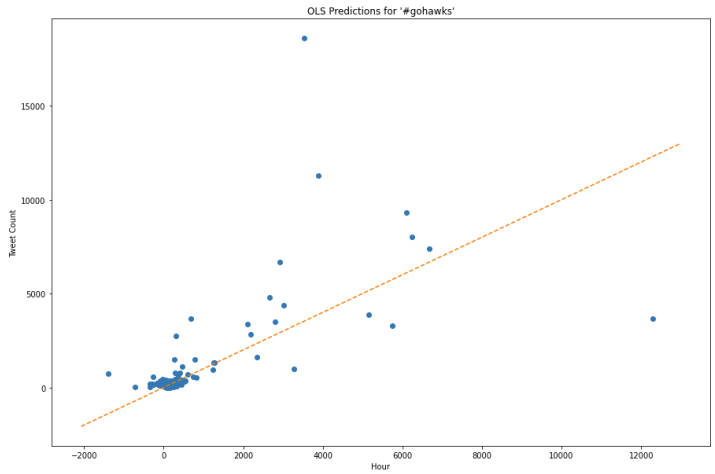


Figure 10: Regression results predicting number of tweets in next hour window for #gohawks and OLS fit predicting number of tweets in next hour window for #gohawks

Lastly, the #gohawks tweets were fit to an OLS linear model and demonstrate a very low R-squared value of 0.472 with an RMSE of 870.95. This low R-squared value indicates that none of the features are great predictors for this class of hashtag tweets and fits our previous hypothesis that team-based tweets are generally more complex classes to predict than the general ones that just pertain to the game. From this OLS fit, we saw only one feature, just like the other models, that had statistical significance. This feature is 'num_tweets' which is understandable as we are predicting the number of tweets in the following hour. We see a correspondingly high t-test value for this feature and the opposite for the other features. In this case, we do not have evidence that any of the other features are statistically significant for not only #gohawks, but also the other hashtags. Across the board we have seen statistical significance with 'num_tweets' in the previous hour, as a predictor for the number of tweets in the following hour.

Question 4: Design a regression model using any additional features from the reference paper.

Once parsed, fit a linear OLS model to these features and the 5 basic features used in the previous question After analyzing

the paper provided as reference from Kong et al. [2014] (https://arxiv.org/abs/1401.2018), we determined 8 additional features to help us predict the number of tweets containing respective hashtags in the following models. These additional features are as follows:

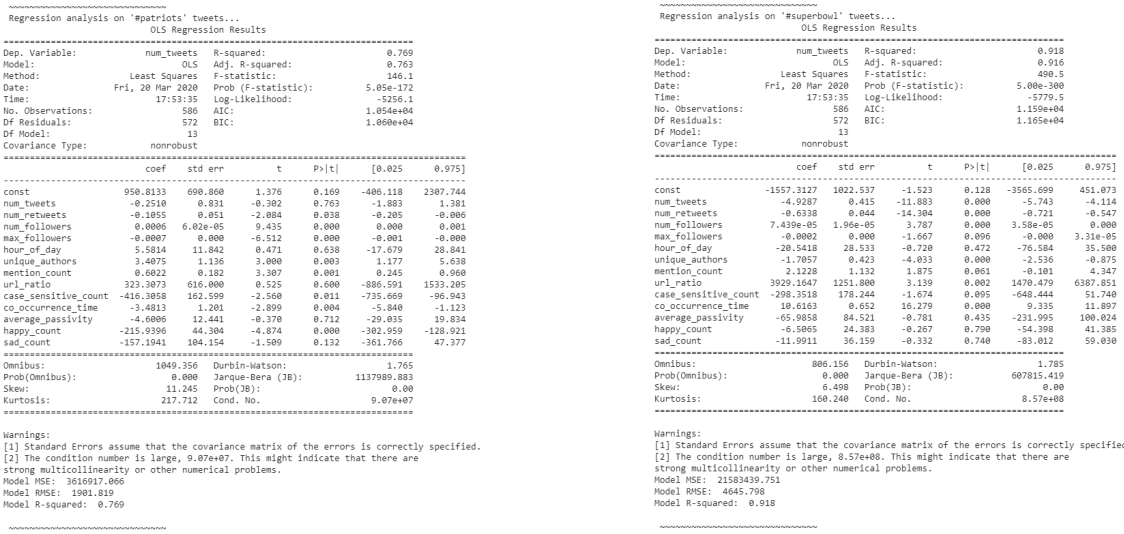
- **Author Count: unique_authors** - describes the number of unique authors who has posted tweets containing this hashtag. Excludes counting tweets from authors that have already posted a tweet containing this hashtag.
- **Mention Count: 'mention_count'** - Counts the total number of mentions across tweets for a hashtag in a time period. 'Mentions are a directional sharing behavior' within the twitter platform, and may be considered a good form of user engagement.
- **URL Ratio: 'url_ratio'** - The ratio of tweets that contain a URL. Tweets containing URLs are thought to be more specifically referring to events or content that may have more likelihood to burst.
- **Case Sensitive Count: 'case_sensitive_count'** - Popular hashtags tend to have more case sensitive occurrences, meaning that the same hashtag (in different case combinations) occurs regularly.
- **Co-occurrence Times: 'co_occurrence_times'** - It is thought that popular events typically contain more than one hashtag in occurrence. The co-occurrence time feature counts the number of hashtags used in combination with the one under analysis.
- **Average Passivity: 'average_passivity':** - Describes the average number of tweets posted by a user in a day, in order to capture that tendency for a user to post tweets following the generation of other popular hashtags. $P_{sv()u_i} = \frac{N_{d(u_i)}}{1.0+N_{t(u_i)}}$ describes this feature where N_d is the number of days since the account was created and N_t describes the total number of tweets from that user.
- **Happy Count and Sad Count: 'happy_count'/'sad_count'** - This feature describes the number emoticons within a tweet that are generally thought of as representing happy vs. sad emotions.

Using these additional features, we fit new linear OLS models to our data which resulted in the following model summaries.

OLS Fit Results Considering 13 Extended Features

Hashtag	RMSE	Adjusted R-Squared Value
#patriots	1,901.819	0.763
#superbowl	4,645.80	0.916
#nfl	406.633	0.731
#gopatriots	73.966	0.925
#sb49	3,504.86	0.848
#gohawks	755.163	0.597

#patriots & #superbowl



```
Regression analysis on '#patriots' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:              num_tweets   R-squared:                       0.769
Model:                             OLS   Adj. R-squared:                  0.763
Method:                  Least Squares   F-statistic:                     146.1
Date:                 Fri, 20 Mar 2020   Prob (F-statistic):          5.05e-172
Time:                         17:53:35   Log-Likelihood:                -5256.1
No. Observations:                  586   AIC:                         1.054e+04
Df Residuals:                      572   BIC:                         1.060e+04
Df Model:                           13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const            950.8133    690.860      1.376      0.169    -406.118    2307.744
num_tweets        -0.2510      0.831     -0.302      0.763      -1.883       1.381
num_retweets      -0.1055      0.051     -2.084      0.038      -0.205      -0.006
num_followers      0.0006   6.02e-05      9.435      0.000       0.000       0.001
max_followers     -0.0007      0.000     -6.512      0.000      -0.001      -0.000
hour_of_day        5.5814     11.842      0.471      0.638     -17.679      28.841
unique_authors     3.4075      1.136      3.000      0.003       1.177       5.638
mention_count      0.6022      0.182      3.307      0.001       0.245       0.960
url_ratio        323.3073    616.000      0.525      0.600    -886.591    1533.205
case_sensitive_count -416.3058  162.599     -2.560      0.011    -735.669     -96.943
co_occurrence_time    -3.4813    1.201     -2.899      0.004      -5.840      -1.123
average_passivity     -4.6006   12.441     -0.370      0.712     -29.035      19.834
happy_count         -215.9396   44.304     -4.874      0.000    -302.959    -128.921
sad_count           -157.1941  104.154     -1.509      0.132    -361.766      47.377
==============================================================================
Omnibus:                     1049.356   Durbin-Watson:                   1.765
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1137989.883
Skew:                          11.245   Prob(JB):                         0.00
Kurtosis:                     217.712   Cond. No.                     9.07e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.07e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  3616917.066
Model RMSE:  1901.819
Model R-squared:  0.769
```

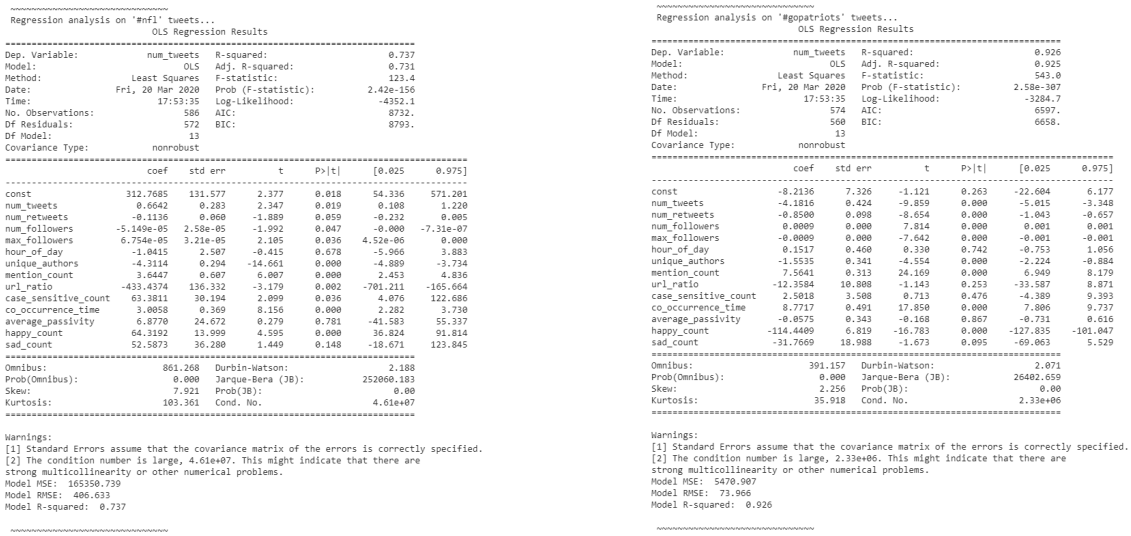
```
Regression analysis on '#superbowl' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:              num_tweets   R-squared:                       0.918
Model:                             OLS   Adj. R-squared:                  0.916
Method:                  Least Squares   F-statistic:                     490.5
Date:                 Fri, 20 Mar 2020   Prob (F-statistic):          5.00e-300
Time:                         17:53:35   Log-Likelihood:                -5779.5
No. Observations:                  586   AIC:                         1.159e+04
Df Residuals:                      572   BIC:                         1.165e+04
Df Model:                           13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          -1557.3127  1022.537     -1.523      0.128   -3565.699     451.073
num_tweets        -4.9287     0.415    -11.883      0.000      -5.743      -4.114
num_retweets      -0.6338     0.044    -14.304      0.000      -0.721      -0.547
num_followers   7.439e-05  1.96e-05      3.787      0.000    3.58e-05       0.000
max_followers     -0.0002     0.000     -1.667      0.096      -0.000    3.31e-05
hour_of_day      -20.5418    28.533     -0.720      0.472     -76.584      35.500
unique_authors    -1.7057     0.423     -4.033      0.000      -2.536      -0.875
mention_count      2.1228     1.132      1.875      0.061      -0.101       4.347
url_ratio       3929.1647  1251.800      3.139      0.002    1470.479    6387.851
case_sensitive_count -298.3518  178.244     -1.674      0.095    -648.444      51.740
co_occurrence_time    10.6163     0.652     16.279      0.000       9.335      11.897
average_passivity    -65.9858    84.521     -0.781      0.435    -231.995     100.024
happy_count           -6.5065    24.383     -0.267      0.790     -54.398      41.385
sad_count            -11.9911    36.159     -0.332      0.740     -83.012      59.030
==============================================================================
Omnibus:                      806.156   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           607815.419
Skew:                           6.498   Prob(JB):                         0.00
Kurtosis:                     160.240   Cond. No.                     8.57e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.57e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  21583439.751
Model RMSE:  4645.798
Model R-squared:  0.918
```

Figure 11: Regression results predicting number of tweets in next hour window for #patriots and #superbowl

(#patriots) From the results above, we can see that the R-squared value has improved significantly from 0.666 with only 5 basic features to 0.763 with 13 extended features. Additionally, the RMSE has dropped from 2,276.16 to 1,901.82. Considering all 13 features, we now are able to determine more accurately which features are significant predictors or the number of tweets in the coming hours. For #patriots, 'num_followers', 'max_followers', and 'happy_count' all have two-trailed p-values of zero, and significantly high t-test values of 9.435, -6.512 and -4.874. These features will be analyzed further in Question 5. For this case, the worst predictors are shown to be 'case_sensitive_count', 'url_count' and 'sad_count'.

(#superbowl) The results above show an improvement for adjusted R-squared from 0.798 to 0.918 for tweets containing #superbowl. This improvement accounts for the massive improvement in RMSE from 7,244.55 to 4,645.8, for the most popular tweet over this three week period. For this tweet there were 5 features that were determined significant, those being: 'num_retweets', 'num_followers', 'num_tweets', 'unique_authors' and 'co_occurence_times'.

#nfl & #gopatriots



```
Regression analysis on '#nfl' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:              num_tweets   R-squared:                       0.737
Model:                             OLS   Adj. R-squared:                  0.731
Method:                  Least Squares   F-statistic:                     123.4
Date:                 Fri, 20 Mar 2020   Prob (F-statistic):          2.42e-156
Time:                         17:53:35   Log-Likelihood:                -4352.1
No. Observations:                  586   AIC:                             8732.
Df Residuals:                      572   BIC:                             8793.
Df Model:                           13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           312.7685    131.577      2.377      0.018      54.336     571.201
num_tweets        0.6642      0.283      2.347      0.019       0.108       1.220
num_retweets     -0.1136      0.060     -1.889      0.059      -0.232       0.005
num_followers -5.149e-05   2.58e-05     -1.992      0.047      -0.000   -7.31e-07
max_followers  6.754e-05   3.21e-05      2.105      0.036    4.52e-06       0.000
hour_of_day       -1.0415      2.507     -0.415      0.678      -5.966       3.883
unique_authors    -4.3114      0.294    -14.661      0.000      -4.889      -3.734
mention_count      3.6447      0.607      6.007      0.000       2.453       4.836
url_ratio       -433.4374   136.332     -3.179      0.002    -701.211    -165.664
case_sensitive_count    63.3811     30.194      2.099      0.036       4.076     122.686
co_occurrence_time       3.0058      0.369      8.156      0.000       2.282       3.730
average_passivity        6.8770     24.672      0.279      0.781     -41.583      55.337
happy_count             64.3192     13.999      4.595      0.000      36.824      91.814
sad_count               52.5873     36.280      1.449      0.148     -18.671     123.845
==============================================================================
Omnibus:                      861.268   Durbin-Watson:                   2.188
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           252060.183
Skew:                           7.921   Prob(JB):                         0.00
Kurtosis:                     103.361   Cond. No.                     4.61e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.61e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  165350.739
Model RMSE:  406.633
Model R-squared:  0.737
```

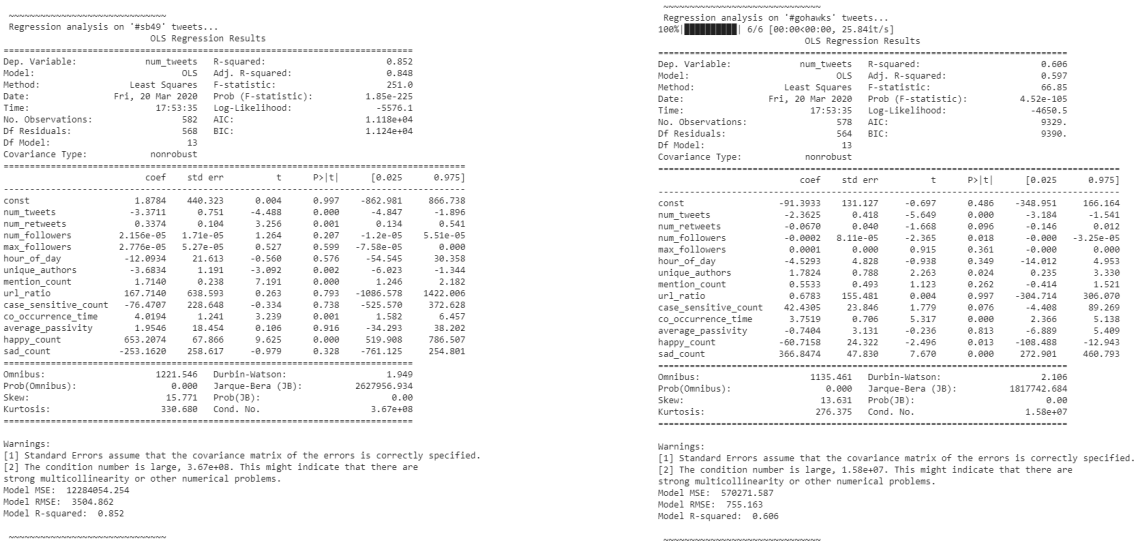
```
Regression analysis on '#gopatriots' tweets...
                            OLS Regression Results
==============================================================================
Dep. Variable:              num_tweets   R-squared:                       0.926
Model:                             OLS   Adj. R-squared:                  0.925
Method:                  Least Squares   F-statistic:                     543.0
Date:                 Fri, 20 Mar 2020   Prob (F-statistic):          2.58e-307
Time:                         17:53:35   Log-Likelihood:                -3284.7
No. Observations:                  574   AIC:                             6597.
Df Residuals:                      560   BIC:                             6658.
Df Model:                           13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           -8.2136      7.326     -1.121      0.263     -22.604       6.177
num_tweets      -4.1816      0.424     -9.859      0.000      -5.015      -3.348
num_retweets    -0.8500      0.098     -8.654      0.000      -1.043      -0.657
num_followers    0.0009      0.000      7.814      0.000       0.001       0.001
max_followers   -0.0009      0.000     -7.642      0.000      -0.001      -0.001
hour_of_day      0.1517      0.460      0.330      0.742      -0.753       1.056
unique_authors  -1.5535      0.341     -4.554      0.000      -2.224      -0.884
mention_count    7.5641      0.313     24.169      0.000       6.949       8.179
url_ratio      -12.3584     10.808     -1.143      0.253     -33.587       8.871
case_sensitive_count     2.5018      3.508      0.713      0.476      -4.389       9.393
co_occurrence_time       8.7717      0.491     17.850      0.000       7.806       9.737
average_passivity       -0.0575      0.343     -0.168      0.867      -0.731       0.616
happy_count           -114.4409      6.819    -16.783      0.000    -127.835    -101.047
sad_count              -31.7669     18.988     -1.673      0.095     -69.063       5.529
==============================================================================
Omnibus:                      391.157   Durbin-Watson:                   2.071
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            26402.659
Skew:                           2.256   Prob(JB):                         0.00
Kurtosis:                      35.918   Cond. No.                     2.33e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.33e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE:  5470.907
Model RMSE:  73.966
Model R-squared:  0.926
```

Figure 12: Regression results predicting number of tweets in next hour window for #nfl and #gopatriots

(#nfl) Again, an improvement is seen when considering 13 features instead of 5 for tweets containing #nfl with Adjusted R-Squared values improving from 0.567 to 0.731 and RMSE decreasing from 519.579 to 406.633 across this class. Features deemed significant from this dataset are as follows: 'unique_authors', 'mention_count', and 'co_occurence_time'.

(#gopatriots) OLS regression for tweets containing #go_patriots improved from when considering 13 features instead of just 5. The adjusted R-squared ratio jumped from 0.626 to 0.925 and the RMSE fell from 166.1 to 73.966. This dramatic improvement can be attributed to the number of features that are statistically significant to this linear classifier. The following features all had two-tailed p-values of value zero with significant t-test values: 'num_tweets', 'num_retweets', 'num_followers', 'max_followers', 'mention_count', 'co_occurrence_time' and 'happy_count'.

#sb49 & #gohawks



Regression analysis on '#sb49' tweets...

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.848
Method:	Least Squares	F-statistic:	251.0
Date:	Fri, 20 Mar 2020	Prob (F-statistic):	1.85e-225
Time:	17:53:35	Log-Likelihood:	-5576.1
No. Observations:	582	AIC:	1.118e+04
Df Residuals:	568	BIC:	1.124e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8784	440.323	0.004	0.997	-862.981	866.738
num_tweets	-3.3711	0.751	-4.488	0.000	-4.847	-1.896
num_retweets	0.3374	0.104	3.256	0.001	0.134	0.541
num_followers	2.156e-05	1.71e-05	1.264	0.207	-1.2e-05	5.51e-05
max_followers	2.776e-05	5.27e-05	0.527	0.599	-7.58e-05	0.000
hour_of_day	-12.0934	21.613	-0.560	0.576	-54.545	30.358
unique_authors	-3.6834	1.191	-3.092	0.002	-6.023	-1.344
mention_count	1.7140	0.238	7.191	0.000	1.246	2.182
url_ratio	167.7140	638.593	0.263	0.793	-1086.578	1422.006
case_sensitive_count	-76.4707	228.648	-0.334	0.738	-525.570	372.628
co_occurrence_time	4.0104	1.241	3.239	0.001	1.582	6.457
average_passivity	1.9546	18.454	0.106	0.916	-34.293	38.202
happy_count	653.2074	67.866	9.625	0.000	519.908	786.507
sad_count	-253.1620	258.617	-0.979	0.328	-761.125	254.801

Omnibus:	1221.546	Durbin-Watson:	1.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2627956.934
Skew:	15.771	Prob(JB):	0.00
Kurtosis:	330.680	Cond. No.	3.67e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.67e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE: 12284054.254
Model RMSE: 3504.862
Model R-squared: 0.852

Regression analysis on '#gohawks' tweets...
100%|██████████| 6/6 [00:00<00:00, 25.84it/s]

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.606
Model:	OLS	Adj. R-squared:	0.597
Method:	Least Squares	F-statistic:	66.85
Date:	Fri, 20 Mar 2020	Prob (F-statistic):	4.52e-105
Time:	17:53:35	Log-Likelihood:	-4650.5
No. Observations:	578	AIC:	9329.
Df Residuals:	564	BIC:	9390.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-91.3933	131.127	-0.697	0.486	-348.951	166.164
num_tweets	-2.3625	0.418	-5.649	0.000	-3.184	-1.541
num_retweets	-0.0670	0.040	-1.668	0.096	-0.146	0.012
num_followers	-0.0002	8.11e-05	-2.365	0.018	-0.000	-3.25e-05
max_followers	0.0001	0.000	0.915	0.361	-0.000	0.000
hour_of_day	-4.5293	4.828	-0.938	0.349	-14.012	4.953
unique_authors	1.7824	0.788	2.263	0.024	0.235	3.330
mention_count	0.5533	0.493	1.123	0.262	-0.414	1.521
url_ratio	0.6783	155.481	0.004	0.997	-304.714	306.070
case_sensitive_count	42.4305	23.846	1.779	0.076	-4.408	89.269
co_occurrence_time	3.7519	0.706	5.317	0.000	2.366	5.138
average_passivity	-0.7404	3.131	-0.236	0.813	-6.889	5.409
happy_count	-60.7158	24.322	-2.496	0.013	-108.488	-12.943
sad_count	366.8474	47.830	7.670	0.000	272.901	460.793

Omnibus:	1135.461	Durbin-Watson:	2.106
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1817742.684
Skew:	13.631	Prob(JB):	0.00
Kurtosis:	276.375	Cond. No.	1.58e+07

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
Model MSE: 570271.587
Model RMSE: 755.163
Model R-squared: 0.606

Figure 13: Regression results predicting number of tweets in next hour window for #sb49 and #gohawks

(#sb49) Results for hashtags containing #sb49 also improved using this additional features. Adjusted R-Squared inched upwards from 0.803 to 0.848, while the RMSE dropped from 4,023 to 3,504. A number of significant features were also observed with this hashtag including 'num_tweets', and 'mention_count'. There were additionally two features with two-trailed p-value results less than alpha, including 'num_retweets',and 'co_occurence_times.'

(#gohawks) Lastly, we ran another OLS linear fit with tweets containing #gohawks. This hashtag is another that represents a specific reference to a team. We see a corresponding high adjusted R-squared value with this class, and a number of features that correspond to the success of the classifier. The R-squared value improved from 0.472 to 0.567 and the RMSE improved from 870.95 to 755. From the p-values and t-test results, we have determined that for this linear model, 'num_tweets', 'co_occurence_times', and 'sad_count' have been determined to be good classifiers.

To conclude this section, it is worth noting that with an understanding of the context that these tweets were published, we can see some clear trends in the data and feature analysis. For example, we see that the basic features (five original) are adequate predictors of frequency of hashtag use for hashtags that are more general in their scope. For example, we saw an adjusted R-squared value of 0.472 for #gohawks, which is very specific for seahawks fans. However, for something with national traction like the superbowl game itself, we observed an adjusted R-squared value of 0.803 for #sb49 tweets, referring to the game in general. Additionally, we can see some direct correlation between the significance of a feature and its relevance to the hashtag itself. For example, for tweets containing #gohawks, we see significance for the sad emoticon usage, while for #gopatriots we see significance for the happy emoticon occurrence; it seems like more than a coincidence that the Seattle Seahawks lost this game very dramatically to the Patriots.

Question 5: For each of the top three features for each hashtag, plot the predictant versus the value of the feature and analyze it

Do regression coefficients agree with the trends in the plots?

In general, the regression coefficients agree with the trends that we observe in the following plots of feature values versus predictant values. As you can observe in the following plot of features that have been determined as statistically significant, there is a definitive cluster of values for low values in each plot. This shows that our predictions are generally more accurate when we are predicting low, non-bursting values. And this trend holds when we observe the RMSE of each hashtag during different time windows. When the hashtags are more likely to be bursting, we actually see a much higher RMSE, out of scale with the number of hashtags. This shows that the models have a hard time predicting bursting hashtag counts. Additionally, we can see in the following plots throughout this section that as we increase the number of tweets in the following hour, we additionally start to see a lot more spread in the feature count as well.

By observing each of the following plots, and comparing to the two-tailed p-value and t-test results described in the section above, you can begin to see some clear trends in the plots and the error that increases that the number of total tweets in the next hour increases.

#patriots top features

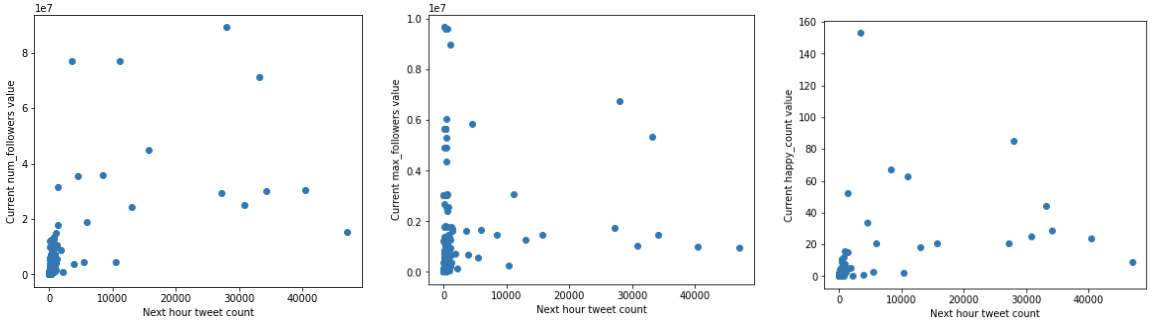


Figure 14: #patriots number of followers vs. predictant values (top left); max followers vs. predictant values (top right); happy emoticon count vs. predictant values

#superbowl top features

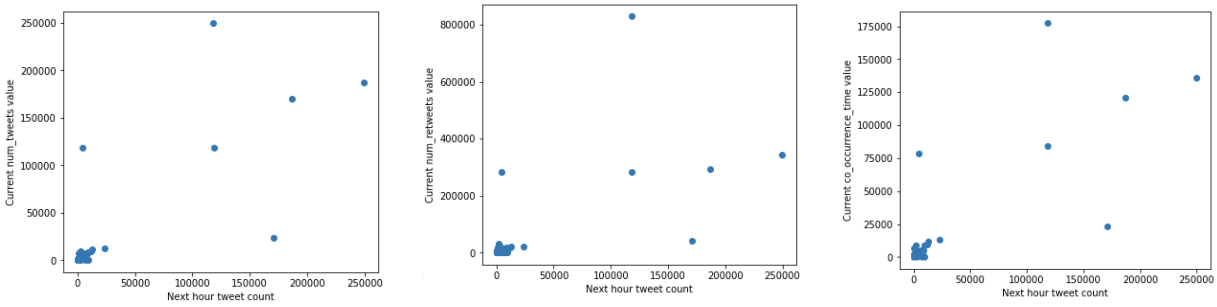


Figure 15: #superbowl number of tweets vs. predictant values (top left); number of retweets vs. predictant values (top right); co-occurrence time vs. predictant values

#nfl top features

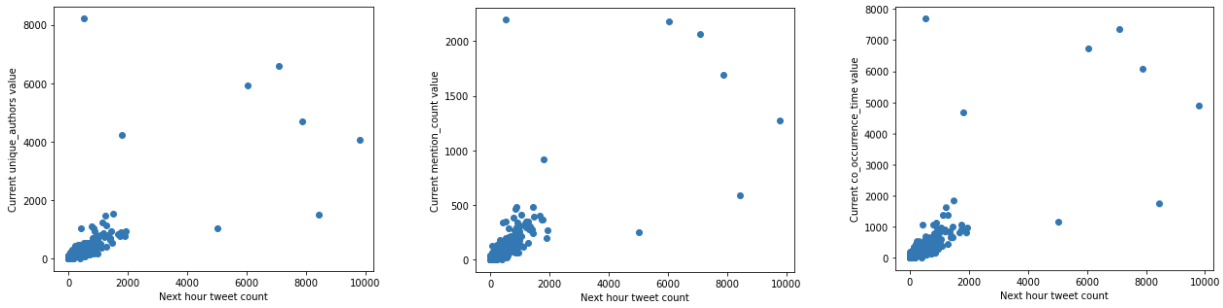


Figure 16: #nfl unique authors vs. predictant valuess (top left); mention count vs. predictant values (top right); co-occurrence time vs. predictant values

#gopatriots top features

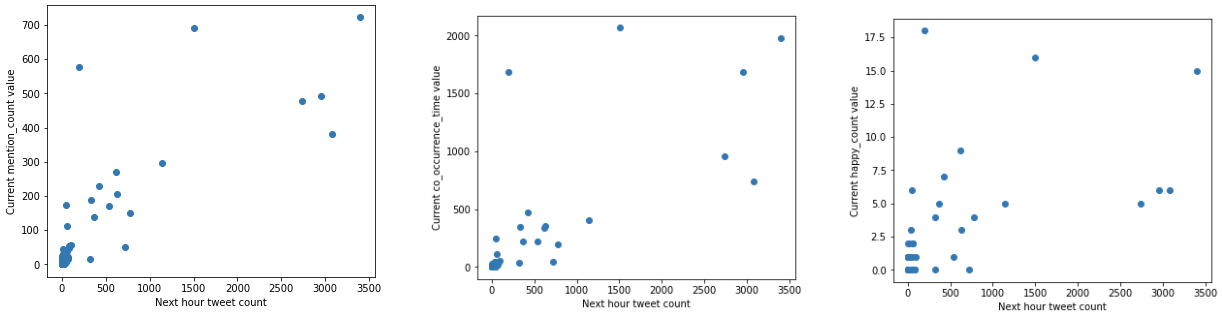


Figure 17: #gopatriots mention count vs. predictant values (top left); co-occurrence time vs. predictant values (top right); happy count vs. predictant values

#sb49 top features

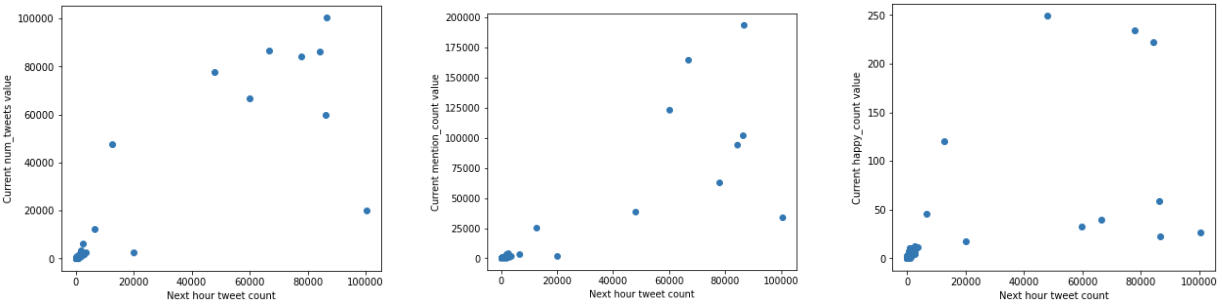


Figure 18: #sb49 number of tweets vs. predictant values (top left); Mention Count Time vs. predictant values (top right); Happy Count vs. predictant values

#gohawks top features

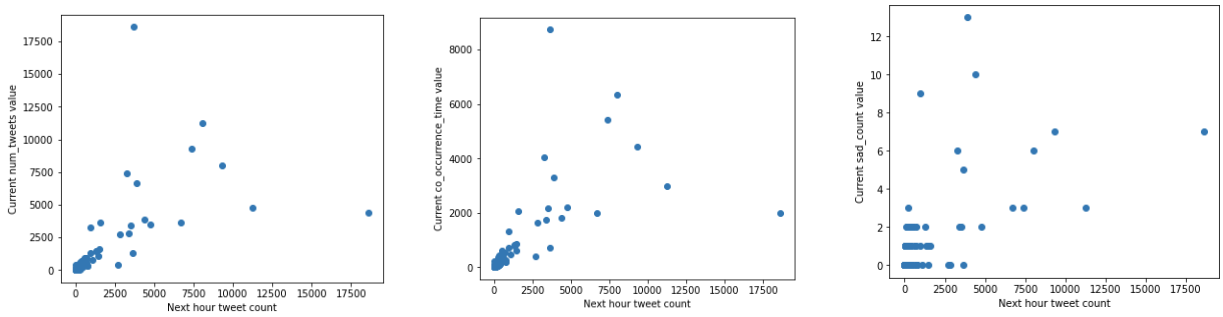


Figure 19: #gohawks Number of Tweets vs. predictant values (top left); Co-occurrence times vs. predictant values (top right); Sad Count vs. predictant values

Question 6: For each hashtag, train 3 regression models, one for each of these time periods (the times are all in PST). Report the MSE and R-squared score for each case.

We used the same features as Question 4. We separated the data with their time into three periods to do linear regression.

A. Time before 02/01/2015 8am PST – one-hour window

OLS Fit Results before 8am

Hashtag	RMSE	R-Squared Value
#patriots	631.242	0.554
#superbowl	692.287	0.351
#nfl	244.735	0.548
#gopatriots	29.301	0.807
#sb49	65.343	0.863
#gohawks	732.511	0.488

#patriots & #superbowl

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.554
Model:	OLS	Adj. R-squared:	0.542
Method:	Least Squares	F-statistic:	43.23
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	1.20e-65
Time:	10:50:39	Log-Likelihood:	-3382.7
No. Observations:	430	AIC:	6791.
Df Residuals:	417	BIC:	6844.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	269.7189	279.529	0.965	0.335	-279.743	819.181
num_retweets	0.0576	0.080	0.719	0.473	-0.100	0.215
num_followers	0.0004	3.7e-05	10.428	0.000	0.000	0.000
max_followers	-0.0005	5.68e-05	-9.160	0.000	-0.001	-0.000
hour_of_day	3.6403	4.669	0.780	0.436	-5.538	12.819
unique_authors	2.2600	0.601	3.762	0.000	1.079	3.441
mention_count	-1.4708	0.881	-1.669	0.096	-3.203	0.261
url_ratio	-382.5055	268.057	-1.427	0.154	-909.417	144.406
case_sensitive_count	37.3989	73.635	0.508	0.612	-107.342	182.140
co_occurrence_time	-3.1641	0.778	-4.066	0.000	-4.694	-1.634
average_passivity	28.6309	22.631	1.265	0.207	-15.855	73.117
happy_count	-4.7795	26.210	-0.182	0.855	-56.299	46.740
sad_count	-143.0173	47.950	-2.983	0.003	-237.271	-48.764

Omnibus:	616.824	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	180547.849
Skew:	7.149	Prob(JB):	0.00
Kurtosis:	102.361	Cond. No.	4.96e+07

```
Model MSE:  398466.865
Model RMSE:  631.242
Model R-squared:  0.554
```

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.351
Model:	OLS	Adj. R-squared:	0.332
Method:	Least Squares	F-statistic:	18.77
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	1.54e-32
Time:	10:50:39	Log-Likelihood:	-3422.3
No. Observations:	430	AIC:	6871.
Df Residuals:	417	BIC:	6924.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	474.4654	207.323	2.289	0.023	66.938	881.993
num_retweets	-0.1387	0.079	-1.758	0.079	-0.294	0.016
num_followers	-1.007e-07	9.98e-06	-0.010	0.992	-1.97e-05	1.95e-05
max_followers	2.963e-06	3.05e-05	0.097	0.923	-5.7e-05	6.29e-05
hour_of_day	-0.2353	5.053	-0.047	0.963	-10.169	9.698
unique_authors	0.5364	0.446	1.202	0.230	-0.341	1.414
mention_count	1.5813	0.334	4.733	0.000	0.925	2.238
url_ratio	-935.8397	265.883	-3.520	0.000	-1458.478	-413.201
case_sensitive_count	64.0214	35.955	1.781	0.076	-6.655	134.697
co_occurrence_time	-0.4261	0.437	-0.976	0.330	-1.284	0.432
average_passivity	-5.8280	15.954	-0.365	0.715	-37.188	25.532
happy_count	-13.7454	5.788	-2.375	0.018	-25.122	-2.369
sad_count	185.4844	74.739	2.482	0.013	38.571	332.397

Omnibus:	695.511	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	188128.253
Skew:	9.221	Prob(JB):	0.00
Kurtosis:	103.797	Cond. No.	1.28e+08

```
Model MSE:  479261.262
Model RMSE:  692.287
Model R-squared:  0.351
```

Figure 20: (left) Regression results for #patriots; (right) Regression results for #superbowl

#nfl & #gopatriots

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	42.11
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.32e-64
Time:	10:50:39	Log-Likelihood:	-2975.2
No. Observations:	430	AIC:	5976.
Df Residuals:	417	BIC:	6029.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	314.5148	123.784	2.541	0.011	71.196	557.834
num_retweets	0.0251	0.047	0.530	0.596	-0.068	0.118
num_followers	6.066e-05	2.35e-05	2.581	0.010	1.45e-05	0.000
max_followers	-5.846e-05	2.73e-05	-2.142	0.033	-0.000	-4.81e-06
hour_of_day	-0.0007	1.766	-0.000	1.000	-3.473	3.472
unique_authors	-0.1397	0.428	-0.326	0.744	-0.981	0.702
mention_count	1.6483	0.498	3.308	0.001	0.669	2.628
url_ratio	-287.9831	124.377	-2.315	0.021	-532.468	-43.498
case_sensitive_count	19.5172	25.295	0.772	0.441	-30.205	69.239
co_occurrence_time	-0.0824	0.367	-0.224	0.823	-0.804	0.639
average_passivity	0.7802	16.716	0.047	0.963	-32.077	33.637
happy_count	67.2739	19.231	3.498	0.001	29.472	105.076
sad_count	114.1008	45.968	2.482	0.013	23.743	204.459

Omnibus:	722.735	Durbin-Watson:	2.236
Prob(Omnibus):	0.000	Jarque-Bera (JB):	407086.814
Skew:	9.569	Prob(JB):	0.00
Kurtosis:	152.515	Cond. No.	3.96e+07

Model MSE: 59895.428
Model RMSE: 244.735
Model R-squared: 0.548

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.807
Model:	OLS	Adj. R-squared:	0.802
Method:	Least Squares	F-statistic:	145.2
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.32e-140
Time:	10:50:39	Log-Likelihood:	-2057.7
No. Observations:	429	AIC:	4141.
Df Residuals:	416	BIC:	4194.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.3941	3.638	-0.933	0.351	-10.546	3.758
num_retweets	-0.1228	0.092	-1.331	0.184	-0.304	0.059
num_followers	-0.0032	0.000	-18.779	0.000	-0.004	-0.003
max_followers	0.0031	0.000	18.658	0.000	0.003	0.003
hour_of_day	-0.2254	0.212	-1.061	0.289	-0.643	0.192
unique_authors	3.9903	0.229	17.416	0.000	3.540	4.441
mention_count	0.4538	0.376	1.207	0.228	-0.286	1.193
url_ratio	-0.5758	4.934	-0.117	0.907	-10.275	9.123
case_sensitive_count	-1.9408	1.924	-1.009	0.314	-5.723	1.841
co_occurrence_time	1.0983	0.370	2.972	0.003	0.372	1.825
average_passivity	-0.4230	0.208	-2.034	0.043	-0.832	-0.014
happy_count	-18.4131	3.805	-4.839	0.000	-25.893	-10.934
sad_count	-5.4374	16.938	-0.321	0.748	-38.732	27.857

Omnibus:	497.014	Durbin-Watson:	1.718
Prob(Omnibus):	0.000	Jarque-Bera (JB):	65588.564
Skew:	5.021	Prob(JB):	0.00
Kurtosis:	62.736	Cond. No.	2.97e+06

Model MSE: 858.544
Model RMSE: 29.301
Model R-squared: 0.807

Figure 21: (left) Regression results for #nfl; (right) Regression results for #gopatriots

#sb49 & #gohawks

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.859
Method:	Least Squares	F-statistic:	216.9
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	8.20e-170
Time:	10:50:39	Log-Likelihood:	-2385.0
No. Observations:	426	AIC:	4796.
Df Residuals:	413	BIC:	4849.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.2608	9.286	0.997	0.319	-8.992	27.514
num_retweets	-0.0029	0.005	-0.640	0.523	-0.012	0.006
num_followers	7.168e-06	1.47e-06	4.882	0.000	4.28e-06	1.01e-05
max_followers	-1.077e-05	3.22e-06	-3.348	0.001	-1.71e-05	-4.45e-06
hour_of_day	-0.8219	0.478	-1.718	0.086	-1.762	0.118
unique_authors	0.7073	0.226	3.134	0.002	0.264	1.151
mention_count	-0.0078	0.123	-0.063	0.950	-0.251	0.235
url_ratio	3.5984	12.411	0.290	0.772	-20.799	27.995
case_sensitive_count	5.8181	5.325	1.093	0.275	-4.650	16.286
co_occurrence_time	0.2792	0.210	1.327	0.185	-0.134	0.693
average_passivity	-0.2680	0.465	-0.577	0.564	-1.181	0.645
happy_count	-5.5708	4.692	-1.187	0.236	-14.794	3.652
sad_count	-61.8747	16.228	-3.813	0.000	-93.775	-29.975

Omnibus:	212.538	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6696.717
Skew:	1.496	Prob(JB):	0.00
Kurtosis:	22.192	Cond. No.	4.72e+07

Model MSE: 4269.771
Model RMSE: 65.343
Model R-squared: 0.863

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.488
Model:	OLS	Adj. R-squared:	0.473
Method:	Least Squares	F-statistic:	33.14
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.26e-53
Time:	10:50:39	Log-Likelihood:	-3446.6
No. Observations:	430	AIC:	6919.
Df Residuals:	417	BIC:	6972.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-330.9767	187.553	-1.765	0.078	-699.643	37.690
num_retweets	-0.2448	0.106	-2.311	0.021	-0.453	-0.037
num_followers	-3.432e-07	0.000	-0.002	0.998	-0.000	0.000
max_followers	-0.0001	0.000	-0.695	0.488	-0.001	0.000
hour_of_day	-6.2877	5.681	-1.107	0.269	-17.454	4.879
unique_authors	7.3857	1.075	6.869	0.000	5.272	9.499
mention_count	0.2602	0.614	0.424	0.672	-0.947	1.467
url_ratio	257.6041	251.451	1.024	0.306	-236.666	751.874
case_sensitive_count	108.5325	31.243	3.474	0.001	47.119	169.946
co_occurrence_time	-10.9442	1.717	-6.373	0.000	-14.320	-7.568
average_passivity	-1.7142	9.105	-0.188	0.851	-19.612	16.184
happy_count	45.2141	27.554	1.641	0.102	-8.948	99.376
sad_count	555.4616	56.758	9.787	0.000	443.895	667.028

Omnibus:	800.174	Durbin-Watson:	2.192
Prob(Omnibus):	0.000	Jarque-Bera (JB):	751516.948
Skew:	11.669	Prob(JB):	0.00
Kurtosis:	206.471	Cond. No.	1.88e+07

Model MSE: 536572.366
Model RMSE: 732.511
Model R-squared: 0.488

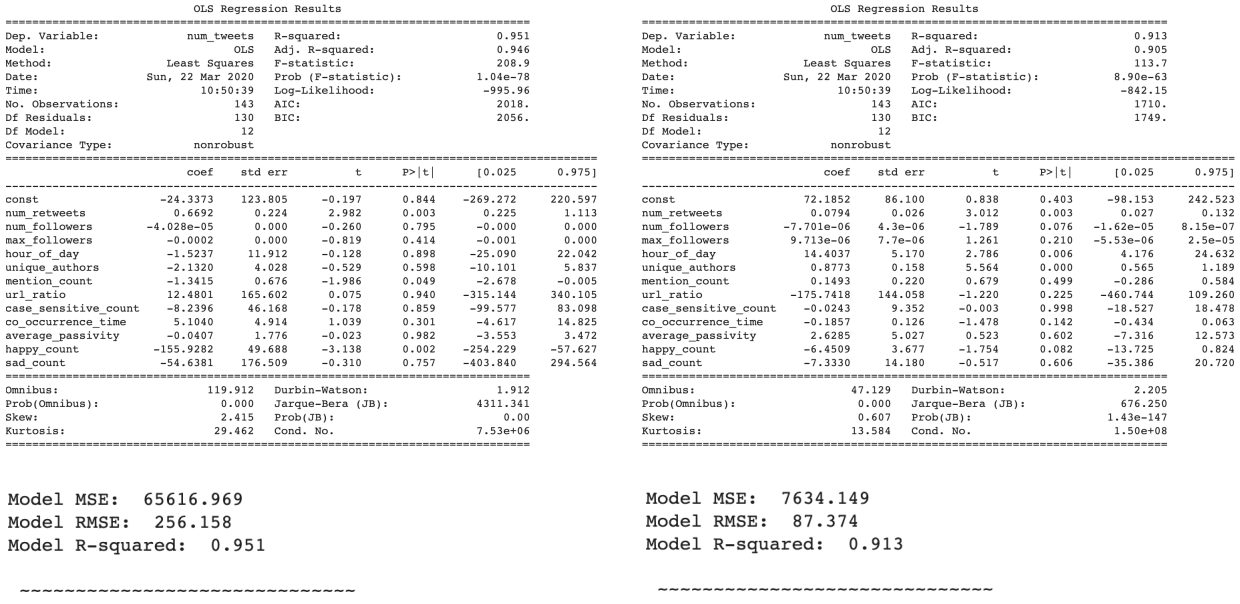
Figure 22: (left) Regression results for #sb49; (right) Regression results for #gohawks

B. Time between 02/01/2015 8am PST and 02/01/2015 8pm PST – five-minute window

OLS Fit Results between 8am and 8pm

Hashtag	RMSE	R-Squared Value
#patriots	256.158	0.951
#superbowl	87.374	0.913
#nfl	21.449	0.705
#gopatriots	3.332	0.898
#sb49	532.153	0.952
#gohawks	18.502	0.949

#patriots & #superbowl



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.951
Model:                            OLS   Adj. R-squared:                  0.946
Method:                 Least Squares   F-statistic:                     208.9
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           1.04e-78
Time:                        10:50:39   Log-Likelihood:                -995.96
No. Observations:                 143   AIC:                             2018.
Df Residuals:                     130   BIC:                             2056.
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               -24.3373    123.805     -0.197      0.844    -269.272     220.597
num_retweets          0.6692      0.224      2.982      0.003       0.225       1.113
num_followers    -4.028e-05      0.000     -0.260      0.795      -0.000       0.000
max_followers        -0.0002      0.000     -0.819      0.414      -0.001       0.000
hour_of_day          -1.5237     11.912     -0.128      0.898     -25.090      22.042
unique_authors       -2.1320      4.028     -0.529      0.598     -10.101       5.837
mention_count        -1.3415      0.676     -1.986      0.049      -2.678      -0.005
url_ratio            12.4801    165.602      0.075      0.940    -315.144     340.105
case_sensitive_count  -8.2396     46.168     -0.178      0.859     -99.577      83.098
co_occurrence_time    5.1040      4.914      1.039      0.301      -4.617      14.825
average_passivity    -0.0407      1.776     -0.023      0.982      -3.553       3.472
happy_count        -155.9282     49.688     -3.138      0.002    -254.229     -57.627
sad_count           -54.6381    176.509     -0.310      0.757    -403.840     294.564
==============================================================================
Omnibus:                      119.912   Durbin-Watson:                   1.912
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4311.341
Skew:                           2.415   Prob(JB):                         0.00
Kurtosis:                      29.462   Cond. No.                     7.53e+06
==============================================================================
```

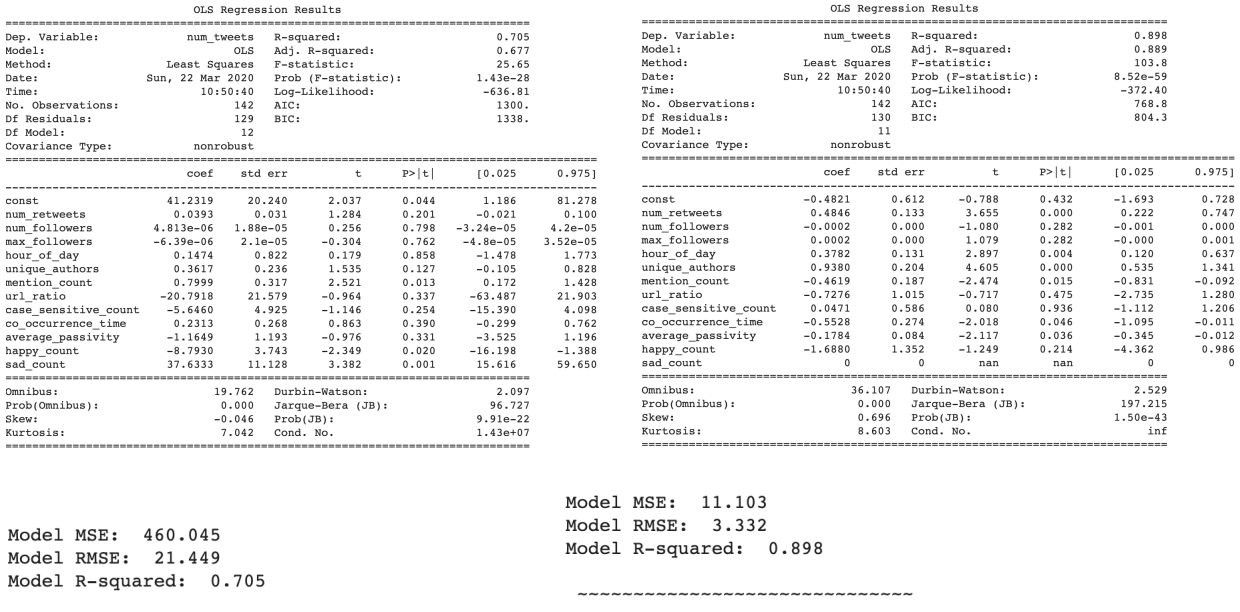
```
Model MSE:  65616.969
Model RMSE:  256.158
Model R-squared:  0.951
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.913
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                     113.7
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           8.90e-63
Time:                        10:50:39   Log-Likelihood:                -842.15
No. Observations:                 143   AIC:                             1710.
Df Residuals:                     130   BIC:                             1749.
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               72.1852     86.100      0.838      0.403     -98.153     242.523
num_retweets         0.0794      0.026      3.012      0.003       0.027       0.132
num_followers   -7.701e-06    4.3e-06     -1.789      0.076   -1.62e-05    8.15e-07
max_followers    9.713e-06    7.7e-06      1.261      0.210   -5.53e-06     2.5e-05
hour_of_day         14.4037      5.170      2.786      0.006       4.176      24.632
unique_authors       0.8773      0.158      5.564      0.000       0.565       1.189
mention_count        0.1493      0.220      0.679      0.499      -0.286       0.584
url_ratio         -175.7418    144.058     -1.220      0.225    -460.744     109.260
case_sensitive_count  -0.0243      9.352     -0.003      0.998     -18.527      18.478
co_occurrence_time    -0.1857      0.126     -1.478      0.142      -0.434       0.063
average_passivity     2.6285      5.027      0.523      0.602      -7.316      12.573
happy_count          -6.4509      3.677     -1.754      0.082     -13.725       0.824
sad_count            -7.3330     14.180     -0.517      0.606     -35.386      20.720
==============================================================================
Omnibus:                       47.129   Durbin-Watson:                   2.205
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              676.250
Skew:                           0.607   Prob(JB):                     1.43e-147
Kurtosis:                      13.584   Cond. No.                     1.50e+08
==============================================================================
```

```
Model MSE:  7634.149
Model RMSE:  87.374
Model R-squared:  0.913
```

Figure 23: (left) Regression results for #patriots; (right) Regression results for #superbowl

#nfl & #gopatriots



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.705
Model:                            OLS   Adj. R-squared:                  0.677
Method:                 Least Squares   F-statistic:                     25.65
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           1.43e-28
Time:                        10:50:40   Log-Likelihood:                -636.81
No. Observations:                 142   AIC:                             1300.
Df Residuals:                     129   BIC:                             1338.
Df Model:                          12
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               41.2319     20.240      2.037      0.044       1.186      81.278
num_retweets         0.0393      0.031      1.284      0.201      -0.021       0.100
num_followers    4.813e-06    1.88e-05      0.256      0.798   -3.24e-05     4.2e-05
max_followers   -6.39e-06     2.1e-05     -0.304      0.762    -4.8e-05    3.52e-05
hour_of_day          0.1474      0.822      0.179      0.858      -1.478       1.773
unique_authors       0.3617      0.236      1.535      0.127      -0.105       0.828
mention_count        0.7999      0.317      2.521      0.013       0.172       1.428
url_ratio          -20.7918     21.579     -0.964      0.337     -63.487      21.903
case_sensitive_count  -5.6460      4.925     -1.146      0.254     -15.390       4.098
co_occurrence_time    0.2313      0.268      0.863      0.390      -0.299       0.762
average_passivity    -1.1649      1.193     -0.976      0.331      -3.525       1.196
happy_count          -8.7930      3.743     -2.349      0.020     -16.198      -1.388
sad_count            37.6333     11.128      3.382      0.001      15.616      59.650
==============================================================================
Omnibus:                       19.762   Durbin-Watson:                   2.097
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               96.727
Skew:                          -0.046   Prob(JB):                     9.91e-22
Kurtosis:                       7.042   Cond. No.                     1.43e+07
==============================================================================
```

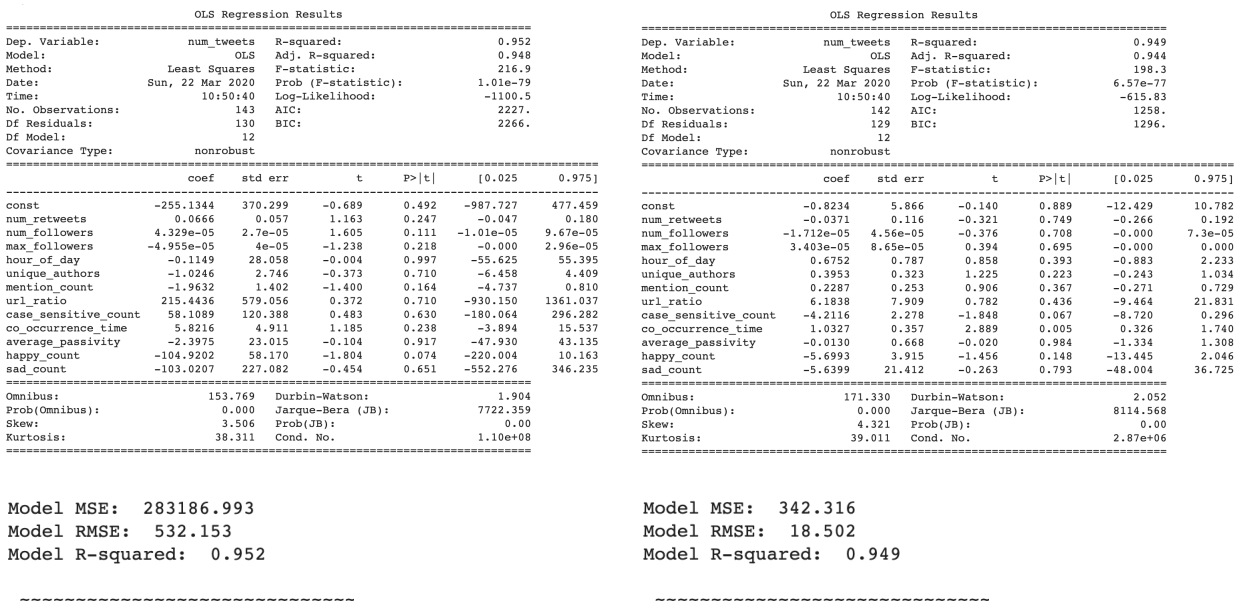
```
Model MSE:  460.045
Model RMSE:  21.449
Model R-squared:  0.705
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.889
Method:                 Least Squares   F-statistic:                     103.8
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           8.52e-59
Time:                        10:50:40   Log-Likelihood:                -372.40
No. Observations:                 142   AIC:                             768.8
Df Residuals:                     130   BIC:                             804.3
Df Model:                          11
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const              -0.4821      0.612     -0.788      0.432      -1.693       0.728
num_retweets        0.4846      0.133      3.655      0.000       0.222       0.747
num_followers      -0.0002      0.000     -1.080      0.282      -0.001       0.000
max_followers       0.0002      0.000      1.079      0.282      -0.000       0.001
hour_of_day         0.3782      0.131      2.897      0.004       0.120       0.637
unique_authors      0.9380      0.204      4.605      0.000       0.535       1.341
mention_count      -0.4619      0.187     -2.474      0.015      -0.831      -0.092
url_ratio          -0.7276      1.015     -0.717      0.475      -2.735       1.280
case_sensitive_count   0.0471      0.586      0.080      0.936      -1.112       1.206
co_occurrence_time    -0.5528      0.274     -2.018      0.046      -1.095      -0.011
average_passivity     -0.1784      0.084     -2.117      0.036      -0.345      -0.012
happy_count           -1.6880      1.352     -1.249      0.214      -4.362       0.986
sad_count                  0          0        nan        nan           0           0
==============================================================================
Omnibus:                       36.107   Durbin-Watson:                   2.529
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              197.215
Skew:                           0.696   Prob(JB):                     1.50e-43
Kurtosis:                       8.603   Cond. No.                          inf
==============================================================================
```

```
Model MSE:  11.103
Model RMSE:  3.332
Model R-squared:  0.898
```

Figure 24: (left) Regression results for #nfl; (right) Regression results for #gopatriots

#sb49 & #gohawks



OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.952
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	216.9
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	1.01e-79
Time:	10:50:40	Log-Likelihood:	-1100.5
No. Observations:	143	AIC:	2227.
Df Residuals:	130	BIC:	2266.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-255.1344	370.299	-0.689	0.492	-987.727	477.459
num_retweets	0.0666	0.057	1.163	0.247	-0.047	0.180
num_followers	4.329e-05	2.7e-05	1.605	0.111	-1.01e-05	9.67e-05
max_followers	-4.955e-05	4e-05	-1.238	0.218	-0.000	2.96e-05
hour_of_day	-0.1149	28.058	-0.004	0.997	-55.625	55.395
unique_authors	-1.0246	2.746	-0.373	0.710	-6.458	4.409
mention_count	-1.9632	1.402	-1.400	0.164	-4.737	0.810
url_ratio	215.4436	579.056	0.372	0.710	-930.150	1361.037
case_sensitive_count	58.1089	120.388	0.483	0.630	-180.064	296.282
co_occurrence_time	5.8216	4.911	1.185	0.238	-3.894	15.537
average_passivity	-2.3975	23.015	-0.104	0.917	-47.930	43.135
happy_count	-104.9202	58.170	-1.804	0.074	-220.004	10.163
sad_count	-103.0207	227.082	-0.454	0.651	-552.276	346.235

Omnibus:	153.769	Durbin-Watson:	1.904
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7722.359
Skew:	3.506	Prob(JB):	0.00
Kurtosis:	38.311	Cond. No.	1.10e+08

Model MSE: 283186.993
Model RMSE: 532.153
Model R-squared: 0.952

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.949
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	198.3
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	6.57e-77
Time:	10:50:40	Log-Likelihood:	-615.83
No. Observations:	142	AIC:	1258.
Df Residuals:	129	BIC:	1296.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.8234	5.866	-0.140	0.889	-12.429	10.782
num_retweets	-0.0371	0.116	-0.321	0.749	-0.266	0.192
num_followers	-1.712e-05	4.56e-05	-0.376	0.708	-0.000	7.3e-05
max_followers	3.403e-05	8.65e-05	0.394	0.695	-0.000	0.000
hour_of_day	0.6752	0.787	0.858	0.393	-0.883	2.233
unique_authors	0.3953	0.323	1.225	0.223	-0.243	1.034
mention_count	0.2287	0.253	0.906	0.367	-0.271	0.729
url_ratio	6.1838	7.909	0.782	0.436	-9.464	21.831
case_sensitive_count	-4.2116	2.278	-1.848	0.067	-8.720	0.296
co_occurrence_time	1.0327	0.357	2.889	0.005	0.326	1.740
average_passivity	-0.0130	0.668	-0.020	0.984	-1.334	1.308
happy_count	-5.6993	3.915	-1.456	0.148	-13.445	2.046
sad_count	-5.6399	21.412	-0.263	0.793	-48.004	36.725

Omnibus:	171.330	Durbin-Watson:	2.052
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8114.568
Skew:	4.321	Prob(JB):	0.00
Kurtosis:	39.011	Cond. No.	2.87e+06

Model MSE: 342.316
Model RMSE: 18.502
Model R-squared: 0.949

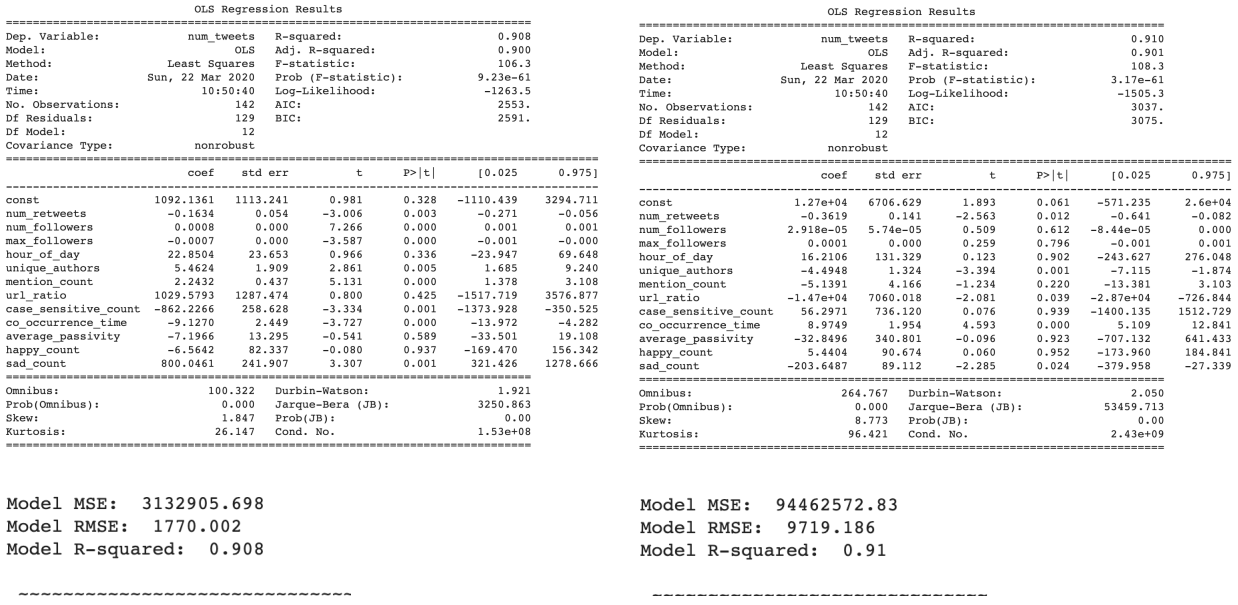
Figure 25: (left) Regression results for #sb49; (right) Regression results for #gohawks

C. Time after 02/01/2015 8pm PST – one-hour window

OLS Fit Results after 8pm

Hashtag	RMSE	R-Squared Value
#patriots	1770.002	0.908
#superbowl	9719.186	0.91
#nfl	611.186	0.812
#gopatriots	87.54	0.974
#sb49	2699.365	0.965
#gohawks	405.204	0.939

#patriots & #superbowl



OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.908
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	106.3
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	9.23e-61
Time:	10:50:40	Log-Likelihood:	-1263.5
No. Observations:	142	AIC:	2553.
Df Residuals:	129	BIC:	2591.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1092.1361	1113.241	0.981	0.328	-1110.439	3294.711
num_retweets	-0.1634	0.054	-3.006	0.003	-0.271	-0.056
num_followers	0.0008	0.000	7.266	0.000	0.001	0.001
max_followers	-0.0007	0.000	-3.587	0.000	-0.001	-0.000
hour_of_day	22.8504	23.653	0.966	0.336	-23.947	69.648
unique_authors	5.4624	1.909	2.861	0.005	1.685	9.240
mention_count	2.2432	0.437	5.131	0.000	1.378	3.108
url_ratio	1029.5793	1287.474	0.800	0.425	-1517.719	3576.877
case_sensitive_count	-862.2266	258.628	-3.334	0.001	-1373.928	-350.525
co_occurrence_time	-9.1270	2.449	-3.727	0.000	-13.972	-4.282
average_passivity	-7.1966	13.295	-0.541	0.589	-33.501	19.108
happy_count	-6.5642	82.337	-0.080	0.937	-169.470	156.342
sad_count	800.0461	241.907	3.307	0.001	321.426	1278.666

Omnibus:	100.322	Durbin-Watson:	1.921
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3250.863
Skew:	1.847	Prob(JB):	0.00
Kurtosis:	26.147	Cond. No.	1.53e+08

Model MSE: 3132905.698
Model RMSE: 1770.002
Model R-squared: 0.908

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.910
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	108.3
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	3.17e-61
Time:	10:50:40	Log-Likelihood:	-1505.3
No. Observations:	142	AIC:	3037.
Df Residuals:	129	BIC:	3075.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.27e+04	6706.629	1.893	0.061	-571.235	2.6e+04
num_retweets	-0.3619	0.141	-2.563	0.012	-0.641	-0.082
num_followers	2.918e-05	5.74e-05	0.509	0.612	-8.44e-05	0.000
max_followers	0.0001	0.000	0.259	0.796	-0.001	0.001
hour_of_day	16.2106	131.329	0.123	0.902	-243.627	276.048
unique_authors	-4.4948	1.324	-3.394	0.001	-7.115	-1.874
mention_count	-5.1391	4.166	-1.234	0.220	-13.381	3.103
url_ratio	-1.47e+04	7060.018	-2.081	0.039	-2.87e+04	-726.844
case_sensitive_count	56.2971	736.120	0.076	0.939	-1400.135	1512.729
co_occurrence_time	8.9749	1.954	4.593	0.000	5.109	12.841
average_passivity	-32.8496	340.801	-0.096	0.923	-707.132	641.433
happy_count	5.4404	90.674	0.060	0.952	-173.960	184.841
sad_count	-203.6487	89.112	-2.285	0.024	-379.958	-27.339

Omnibus:	264.767	Durbin-Watson:	2.050
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53459.713
Skew:	8.773	Prob(JB):	0.00
Kurtosis:	96.421	Cond. No.	2.43e+09

Model MSE: 94462572.83
Model RMSE: 9719.186
Model R-squared: 0.91

Figure 26: (left) Regression results for #patriots; (right) Regression results for #superbowl

#nfl & #gopatriots

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	46.56
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	5.52e-41
Time:	10:50:40	Log-Likelihood:	-1112.5
No. Observations:	142	AIC:	2251.
Df Residuals:	129	BIC:	2289.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	99.6973	283.294	0.352	0.725	-460.807	660.202
num_retweets	-0.3878	0.179	-2.167	0.032	-0.742	-0.034
num_followers	-0.0002	8.21e-05	-2.728	0.007	-0.000	-6.15e-05
max_followers	0.0003	0.000	2.229	0.028	3.19e-05	0.001
hour_of_day	0.2468	8.385	0.029	0.977	-16.343	16.837
unique_authors	-5.7526	0.621	-9.267	0.000	-6.981	-4.524
mention_count	8.1991	1.852	4.427	0.000	4.535	11.863
url_ratio	-24.0971	342.357	-0.070	0.944	-701.459	653.264
case_sensitive_count	30.9277	80.843	0.383	0.703	-129.023	190.878
co_occurrence_time	4.8557	0.670	7.246	0.000	3.530	6.182
average_passivity	-24.2206	98.495	-0.246	0.806	-219.096	170.655
happy_count	84.7599	39.718	2.134	0.035	6.177	163.343
sad_count	201.7295	112.044	1.800	0.074	-19.953	423.412

Omnibus:	138.486	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3349.109
Skew:	3.294	Prob(JB):	0.00
Kurtosis:	25.861	Cond. No.	5.57e+07

```
Model MSE:  373547.851
Model RMSE:  611.186
Model R-squared:  0.812
```

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.974
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	374.7
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	7.65e-88
Time:	10:50:40	Log-Likelihood:	-771.72
No. Observations:	131	AIC:	1569.
Df Residuals:	118	BIC:	1607.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-17.1923	18.232	-0.943	0.348	-53.296	18.912
num_retweets	-1.3549	0.195	-6.942	0.000	-1.741	-0.968
num_followers	0.0019	0.000	7.537	0.000	0.001	0.002
max_followers	-0.0037	0.000	-7.479	0.000	-0.005	-0.003
hour_of_day	1.2995	1.237	1.051	0.296	-1.150	3.748
unique_authors	-2.3360	0.540	-4.327	0.000	-3.405	-1.267
mention_count	11.0785	0.623	17.774	0.000	9.844	12.313
url_ratio	-24.9521	30.197	-0.826	0.410	-84.750	34.846
case_sensitive_count	22.4514	10.671	2.104	0.037	1.321	43.582
co_occurrence_time	1.8624	1.194	1.560	0.121	-0.502	4.227
average_passivity	-0.2949	0.577	-0.511	0.610	-1.437	0.847
happy_count	-144.2895	20.999	-6.871	0.000	-185.873	-102.706
sad_count	10.1711	43.082	0.236	0.814	-75.143	95.485

Omnibus:	82.110	Durbin-Watson:	2.953
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4113.860
Skew:	1.279	Prob(JB):	0.00
Kurtosis:	30.334	Cond. No.	3.72e+06

```
Model MSE:  7663.18
Model RMSE:  87.54
Model R-squared:  0.974
```

Figure 27: (left) Regression results for #nfl; (right) Regression results for #gopatriots

#patriots & #gohawks

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.965
Model:	OLS	Adj. R-squared:	0.962
Method:	Least Squares	F-statistic:	295.2
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	1.54e-87
Time:	10:50:40	Log-Likelihood:	-1323.4
No. Observations:	142	AIC:	2673.
Df Residuals:	129	BIC:	2711.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	249.6573	2794.584	0.089	0.929	-5279.495	5778.810
num_retweets	-0.2918	0.087	-3.365	0.001	-0.463	-0.120
num_followers	6.488e-05	1.43e-05	4.547	0.000	3.66e-05	9.31e-05
max_followers	-2.529e-05	4.96e-05	-0.510	0.611	-0.000	7.29e-05
hour_of_day	26.9530	35.947	0.750	0.455	-44.169	98.075
unique_authors	-0.9732	1.113	-0.874	0.384	-3.176	1.230
mention_count	1.7924	0.192	9.324	0.000	1.412	2.173
url_ratio	108.1482	3130.283	0.035	0.972	-6085.193	6301.490
case_sensitive_count	-273.4415	494.858	-0.553	0.582	-1252.530	705.647
co_occurrence_time	-1.4508	1.112	-1.305	0.194	-3.651	0.749
average_passivity	4.8358	33.665	0.144	0.886	-61.771	71.443
happy_count	277.6340	64.227	4.323	0.000	150.559	404.709
sad_count	578.4933	224.425	2.578	0.011	134.464	1022.523

Omnibus:	196.661	Durbin-Watson:	2.206
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16387.888
Skew:	5.220	Prob(JB):	0.00
Kurtosis:	54.583	Cond. No.	2.74e+09

```
Model MSE:  7286573.874
Model RMSE:  2699.365
Model R-squared:  0.965
```

OLS Regression Results

Dep. Variable:	num_tweets	R-squared:	0.939
Model:	OLS	Adj. R-squared:	0.933
Method:	Least Squares	F-statistic:	155.1
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.15e-67
Time:	10:50:40	Log-Likelihood:	-994.73
No. Observations:	134	AIC:	2015.
Df Residuals:	121	BIC:	2053.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.1221	112.306	-0.072	0.942	-230.461	214.217
num_retweets	-0.1785	0.034	-5.316	0.000	-0.245	-0.112
num_followers	-0.0009	0.000	-9.039	0.000	-0.001	-0.001
max_followers	0.0028	0.000	7.605	0.000	0.002	0.003
hour_of_day	-1.4744	5.641	-0.261	0.794	-12.643	9.694
unique_authors	1.2077	0.907	1.332	0.185	-0.587	3.003
mention_count	5.2870	0.634	8.345	0.000	4.033	6.541
url_ratio	-5.2987	123.938	-0.043	0.966	-250.667	240.070
case_sensitive_count	2.0645	43.531	0.047	0.962	-84.116	88.245
co_occurrence_time	3.4325	0.905	3.794	0.000	1.641	5.224
average_passivity	0.2190	1.869	0.117	0.907	-3.480	3.918
happy_count	-422.0944	35.556	-11.871	0.000	-492.487	-351.702
sad_count	-316.3773	132.723	-2.384	0.019	-579.138	-53.617

Omnibus:	67.725	Durbin-Watson:	2.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1678.480
Skew:	-1.076	Prob(JB):	0.00
Kurtosis:	20.204	Cond. No.	1.53e+07

```
Model MSE:  164190.095
Model RMSE:  405.204
Model R-squared:  0.939
```

Figure 28: (left) Regression results for #sb49; (right) Regression results for #gohawks

Question 7: Aggregate the data of all hashtags, and train 3 models (for the intervals mentioned above) to predict the number of tweets in the next time window on the aggregated data.

We concatenated data of different hashtags to get the aggregated data with a hashtag #all. As in question 6, we did the same linear regression for three different time periods.

A. Time before 02/01/2015 8am PST – one-hour window
OLS Fit Results before 8am

Hashtag	RMSE	R-Squared Value
#all	1863.146	0.517

```
Model MSE:  3471311.742
Model RMSE:  1863.146
Model R-squared:  0.517

 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

hashtag #all

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.517
Model:                            OLS   Adj. R-squared:                  0.503
Method:                 Least Squares   F-statistic:                     37.24
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           1.48e-58
Time:                        11:01:32   Log-Likelihood:                -3848.1
No. Observations:                 430   AIC:                             7722.
Df Residuals:                     417   BIC:                             7775.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                   1322.6082    932.992      1.418      0.157    -511.346    3156.562
num_retweets               0.0579      0.071      0.819      0.413      -0.081       0.197
num_followers         -5.372e-07   1.18e-05     -0.045      0.964   -2.38e-05    2.27e-05
max_followers         -4.732e-05   5.45e-05     -0.868      0.386      -0.000    5.99e-05
hour_of_day               -8.4938     14.331     -0.593      0.554     -36.664      19.676
unique_authors             2.0963      0.428      4.894      0.000       1.254       2.938
mention_count              1.1589      0.549      2.111      0.035       0.080       2.238
url_ratio              -2354.8055    889.949     -2.646      0.008   -4104.151    -605.461
case_sensitive_count      71.8208     41.156      1.745      0.082      -9.079     152.721
co_occurrence_time        -1.4982      0.405     -3.703      0.000      -2.294      -0.703
average_passivity       -205.2421     97.056     -2.115      0.035    -396.022     -14.462
happy_count              -92.5488     16.265     -5.690      0.000    -124.520     -60.577
sad_count                418.7690     72.147      5.804      0.000     276.953     560.585
==============================================================================
Omnibus:                      686.731   Durbin-Watson:                   2.234
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           268782.834
Skew:                           8.765   Prob(JB):                         0.00
Kurtosis:                     124.221   Cond. No.                     3.43e+08
==============================================================================
```

Figure 29: Regression results predicting number of tweets in next hour window for before 02/01/2015 8am PST #patriots

B. Time between 02/01/2015 8am and 8pm PST – five-minute window
OLS Fit Results between 8am and 8pm

Hashtag	RMSE	R-Squared Value
#all	784.751	0.958

```
Model MSE:  3471311.742
Model RMSE:  1863.146
Model R-squared:  0.517

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

hashtag #all

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              num_tweets   R-squared:                       0.958
Model:                            OLS   Adj. R-squared:                  0.954
Method:                 Least Squares   F-statistic:                     248.2
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           2.44e-83
Time:                        11:01:32   Log-Likelihood:                -1156.1
No. Observations:                 143   AIC:                             2338.
Df Residuals:                     130   BIC:                             2377.
Df Model:                          12
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                    -82.9938    943.806     -0.088      0.930   -1950.202    1784.214
num_retweets               0.1530      0.072      2.126      0.035       0.011       0.295
num_followers           8.179e-06   1.85e-05      0.443      0.658   -2.83e-05    4.47e-05
max_followers          -3.589e-05   4.16e-05     -0.864      0.389      -0.000    4.63e-05
hour_of_day                61.7650     45.334      1.362      0.175     -27.922     151.453
unique_authors              0.5508      1.033      0.533      0.595      -1.493       2.595
mention_count              -0.7475      0.308     -2.427      0.017      -1.357      -0.138
url_ratio                 514.5298   1309.970      0.393      0.695   -2077.090    3106.149
case_sensitive_count      -26.8853     43.524     -0.618      0.538    -112.992      59.222
co_occurrence_time          1.8971      1.079      1.759      0.081      -0.237       4.031
average_passivity          -6.5365     46.578     -0.140      0.889     -98.686      85.613
happy_count              -112.2505     26.631     -4.215      0.000    -164.937     -59.564
sad_count                -171.0700    107.625     -1.590      0.114    -383.993      41.853
==============================================================================
Omnibus:                      114.266   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3216.892
Skew:                           2.331   Prob(JB):                         0.00
Kurtosis:                      25.763   Cond. No.                     3.12e+08
==============================================================================
                            Model MSE:  615834.403
                            Model RMSE:  784.751
                            Model R-squared:  0.958

                             ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Figure 30: Regression results predicting number of tweets in next hour window for between 02/01/2015 8am – 8pm PST #all

C. Time after 02/01/2015 8pm PST – one-hour window
OLS Fit Results after 8pm

Hashtag	RMSE	R-Squared Value
#all	14070.373	0.93

```
Model MSE:  197975393.677
Model RMSE:  14070.373
Model R-squared:  0.93

 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

hashtag #all

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             num_tweets   R-squared:                       0.930
Model:                            OLS   Adj. R-squared:                  0.924
Method:                 Least Squares   F-statistic:                     143.4
Date:                Sun, 22 Mar 2020   Prob (F-statistic):           2.10e-68
Time:                        11:01:32   Log-Likelihood:                -1557.8
No. Observations:                 142   AIC:                             3142.
Df Residuals:                     129   BIC:                             3180.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                  -4205.8907   1.55e+04     -0.271      0.787   -3.49e+04    2.65e+04
num_retweets              -0.3371      0.067     -5.058      0.000      -0.469      -0.205
num_followers              0.0002   3.45e-05      5.733      0.000       0.000       0.000
max_followers             -0.0006      0.000     -1.918      0.057      -0.001    1.74e-05
hour_of_day               36.4803    191.063      0.191      0.849    -341.543     414.503
unique_authors            -9.0435      0.909     -9.951      0.000     -10.842      -7.246
mention_count              0.2181      0.408      0.534      0.594      -0.589       1.025
url_ratio               1449.2068   1.51e+04      0.096      0.924   -2.84e+04    3.13e+04
case_sensitive_count     308.6667    370.646      0.833      0.407    -424.665    1041.999
co_occurrence_time         7.0499      0.918      7.678      0.000       5.233       8.867
average_passivity         66.8011    251.821      0.265      0.791    -431.433     565.035
happy_count              215.6794     65.689      3.283      0.001      85.712     345.646
sad_count               -134.8946    132.876     -1.015      0.312    -397.792     128.003
==============================================================================
Omnibus:                      234.273   Durbin-Watson:                   1.908
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            39971.620
Skew:                           6.883   Prob(JB):                         0.00
Kurtosis:                      84.032   Cond. No.                     6.78e+09
==============================================================================
                            Model MSE:  197975393.677
                            Model RMSE:  14070.373
                            Model R-squared:  0.93

                             ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Figure 31: Regression results predicting number of tweets in next hour window for after 02/01/2015 8pm PST #all

Question 8: Use grid search to find the best parameter set for RandomForestRegressor and GradientBoostingRegressor respectively.

We did a randomized grid search to cut down on the computational cost. These are the best parameters we got:

```
[Parallel(n_jobs=-1)]: Done 2500 out of 2500 | elapsed: 47.9min finished
{'max_depth': 10,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 400}
```

Figure 32: Best parameters for the RandomForestRegressor.

```
[Parallel(n_jobs=-1)]: Done 2500 out of 2500 | elapsed: 47.9min finished
{'max_depth': 100,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 2000}
```

Figure 33: Best parameters for the GradientBoostingRegressor.

The test error from cross-validation doesn't look good. For the RandomForestRegressor we got a RMSE of [7134.99896526, 1986.63730871, 639.15572004, 60367.93969278, 1541.9491199] and for the GradientBoostingRegressor we got a RMSE of [3594.82938006, 3422.89157611, 833.85500051, 61578.83961979, 510.93299698]. The test RMSE has a very high variance; this is probably because we're overfitting when using these models. We might need to make more shallow trees (i.e. reduce the max depth parameter) and/or change other parameters, such as max features.

Question 9: Compare the best estimator you found in the grid search with OLS on the entire dataset.

The RMSE for the RandomForestRegressor was 27210 and for the GradientBoostingRegressor it was 27631. This performs worse than linear regression, but when comparing to a neural network (question 11 below) it does about the same. (The neural network achieves a RMSE in the mid-20,000s). As in question 8, this shows that the random forests are overfitting with these parameters.

Question 10: For each time period described in Question 6, perform the same grid search above for GradientBoostingRegressor (with corresponding time window length). Does the cross-validation test error change? Are the best parameter set you find in each period agree with those you found above?

We did the same randomized grid search to cut down on the computational cost. For each of the 3 periods, we found these parameters to be the best:

```
[Parallel(n_jobs=-1)]: Done 5000 out of 5000 | elapsed: 64.2min finished
{'max_depth': 200,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 1000}
```

Figure 34: Best parameters for the GradientBoostingRegressor for the "Before" period.

```
[Parallel(n_jobs=-1)]: Done 5000 out of 5000 | elapsed: 15.5min finished
{'max_depth': 30,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 1000}
```

Figure 35: Best parameters for the GradientBoostingRegressor for the ”Between” period..

```
[Parallel(n_jobs=-1)]: Done 5000 out of 5000 | elapsed: 20.4min finished
{'max_depth': 10,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 1800}
```

Figure 36: Best parameters for the GradientBoostingRegressor for the ”After” period.

Yes, the cross-validation test error changes as well. Here, for the ”Before” period we get an RMSE of 10422 with a cross-validation of [366.64111204, 15733.8074463, 578.98207621, 919.97182442, 17155.65783085]. ”Between” has an RMSE of 3766 with with a cross-validation of [156.72070234, 59.36458139, 178.83880908, 344.60928508, 8412.50336415]. ”After” has an RMSE of 53761 with a cross-validation of [120180.50971023, 2769.03214769, 271.9940225, 245.12402345, 505.55402681]. As with before, we are definitely overfitting.

No, the parameters in question 11 for either of the periods isn't the same as in question 8. The first reason is that we're using a randomized grid search. Since it looks at a random set of parameters, there's no guarantee that we will even test the same set. But even if we did use a full grid search (no randomization), it makes sense that different periods have different best parameters. This is because each period has different characteristics, much like how piece-wise linear regression is different from linear regression.

Question 11: Now try to regress the aggregated data with MLPRegressor by adjusting hidden layer sizes. You should try at least 5 architectures with various numbers of layers and layer sizes. Report the architectures you tried, as well as its MSE of fitting the entire aggregated data.

We aggregated the data as we did before and tried 7 different architectures by adjusting hidden_layer_sizes. We use cross validation to evaluate our models with split = 10.

1/7 : Evaluate Multilayer Perceptron with 2 hidden layers

```
hidden_layer_sizes = (5,5,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 37: hidden_layer_sizes = (5,5,)

Result of hidden_layer_sizes = (5,5,):

	Training	**Test**
Average_MSE	968042477202.858	113031662507.279

```
Average MSE from training:  968042477202.8578
Average MSE from testing:  113031662507.27873
```

2/7 : Evaluate Multilayer Perceptron with 4 hidden layers

```
hidden_layer_sizes = (50,50,50,50,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 38: hidden_layer_sizes = (50,50,50,50,)

Result of hidden_layer_sizes = (50,50,50,50,):

	Training	Test
Average_MSE	186039106812.395	282942876639.491

```
Average MSE from training:  186039106812.39508
Average MSE from testing:  282942876639.49054
```

3/7 : Evaluate Multilayer Perceptron with 8 hidden layers

```
hidden_layer_sizes = (100,100,100,100,100,100,100,100,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 39: hidden_layer_sizes = (100,100,100,100,100,100,100,100,)

Result of hidden_layer_sizes = (100,100,100,100,100,100,100,100,):

	Training	Test
Average_MSE	16715923864.696	2477596852.691

```
Average MSE from training:  16715923864.69574
Average MSE from testing:  2477596852.691178
```

4/7 : Evaluate Multilayer Perceptron with 16 hidden layers

```
hidden_layer_sizes = (500,500,500,500,500,500,500,500,500,500,500,500,500,500,500,500,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 40: hidden_layer_sizes = (500,500,500,500,500,500,500,500,500,500,500,500,500,500,500,500,)

Result of hidden_layer_sizes:

	Training	Test
Average_MSE	336611289.838	340447558.990

```
Average MSE from training:  336611289.8377611
Average MSE from testing:  340447558.9899054
```

5/7 : Evaluate Multilayer Perceptron with 7 hidden layers

```
hidden_layer_sizes = (50,100,500,1000,500,100,50,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 41: hidden_layer_sizes = (50,100,500,1000,500,100,50,)

Result of hidden_layer_sizes:

	Training	Test
Average_MSE	449856563786.001	117612906092.342

```
Average MSE from training:  449856563786.0007
Average MSE from testing:  117612906092.34216
```

6/7 : Evaluate Multilayer Perceptron with 30 hidden layers

```
hidden_layer_sizes = (100,100,100,100,100,100,100,100,100,100,
                      100,100,100,100,100,100,100,100,100,100,
                      100,100,100,100,100,100,100,100,100,100,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 42: 30 hidden layers

Result of hidden_layer_sizes:

	Training	Test
Average_MSE	362414196.807	614592860.794

```
Average MSE from training:  362414196.80677474
Average MSE from testing:  614592860.7938277
```

7/7 : Evaluate Multilayer Perceptron with 3 hidden layers

```
hidden_layer_sizes = (1000,1000,1000,)
network_model = MLPRegressor(hidden_layer_sizes=hidden_layer_sizes, activation='relu', solver='adam',
                             alpha=0.0001, batch_size='auto', learning_rate='constant',
                             learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,
                             random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9,
                             nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1,
                             beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10, max_fun=15000)
```

Figure 43: hidden_layer_sizes = (1000, 1000, 1000,)

Result of hidden_layer_sizes:

	Training	**Test**
Average_MSE	543647341879.960	135576185716.065

```
Average MSE from training:  543647341879.96014
Average MSE from testing:  135576185716.06453
```

Question 12: Use StandardScaler to scale the features before feeding it to MLPRegressor (with the best architecture you got above). Does its performance increase?

Among the above architectures, we found the one with 16 hidden layers performs best. This time we use the StandardScaler to scale the features before training. The result is shown below, demonstrating that the performance decreases.

	Training	**Test**
Average_RMSE_Scaled	28940.513	23138.667
Average_RMSE_Non_Scaled	17544.731	15938.11

```
Average RMSE from training:  28940.513212999143
Average RMSE from testing:  23138.666517141683

Average RMSE from training:  17544.731083587052
Average RMSE from testing:  15938.109975420954
```

Figure 44: RMSE : scaled data (above) vs non_scaled data (below)

Question 13: Using grid search, find the best architecture (for scaled data) for each period (with corresponding window length) described in Question 6.

We separated the data into three time periods as we did above. Following Question 11, we use grid search on different hidden_layer_sizes. Below is the best architecture we got for the three time periods.
We only adjust the architectures by adjusting the hidden_layer_sizes as required by Question 11; the other parameters are:

- 'activation': ['relu'],
- 'alpha': [0.001],
- 'learning_rate_init': [0.005]

Before 8am	'hidden_layer_sizes': (100, 100, 100, 100, 100, 100, 100, 100)
Between 8am-8pm	'hidden_layer_sizes': (100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100)
After 8pm	'hidden_layer_sizes': (500, 500, 500, 500, 500, 500, 500, 500)

```
Best MLP Paramters Determined from Grid Search:(parameters)
{'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (100, 100, 100, 100, 100, 100, 100, 100), 'learning_rate_init': 0.005}
```

Figure 45: Best Parameter before 8am

```
Best MLP Paramters Determined from Grid Search:(parameters)
{'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100), 'learning_rate_init': 0.005}
```

Figure 46: Best Parameter between 8am and 8pm

```
Best MLP Paramters Determined from Grid Search:(parameters)
{'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (500, 500, 500, 500, 500, 500, 500, 500), 'learning_rate_init': 0.005}
```

Figure 47: Best Parameter after 8pm

Question 14: Report the model you use. For each test file, provide your predictions on the number of tweets in the next time window

We obtained the windowed aggregated data as we did in the previous questions. Based on the question description, we created a 6x window to accumulate the data. We used the 1x data after the 6x window to be the truth that we used for the model training. As we got the training data, we do the same process on the test files.

We used RandomForestRegressor and trained it with the training data to predict on the test data. Among the models, we chose to use the random forest regression to do the work.

Table 1: Prediction on different samples in different time periods

Sample	Time period	Prediction
Sample0_period1	Before 8am	6745.3
Sample1_period1	Before 8am	6884.936
Sample2_period1	Before 8am	6750.109
Sample0_period2	Between 8am and 8pm	12372.535
Sample1_period2	Between 8am and 8pm	12319.337
Sample2_period2	Between 8am and 8pm	12309.155
Sample0_period3	After 8 pm	94486.706
Sample1_period3	After 8 pm	94450.083
Sample2_period3	After 8 pm	94450.083

```
sample0_period1 : sample0_period2 : sample0_period3 :
[6745.29999338]   [12372.53506095]  [94486.70634794]

sample1_period1 : sample1_period2 : sample1_period3 :
[6884.9360648]    [12319.33765626]  [94450.08332636]

sample2_period1 : sample2_period2 : sample2_period3 :
[6750.10916004]   [12309.15502094]  [94450.08332636]
```

Figure 48: Prediction on test data

Part 2: Fan Base Prediction

Question 15: Explain the method you use to determine the location. Train a binary classifier to predict the location of the author of a tweet (Washington or Massachusetts), given only the textual content of the tweet. Try different classification algorithms (at least 3). For each, plot ROC curve, report confusion matrix, and calculate accuracy, recall and precision.

The method of choosing the tweets from Washington and Massachusetts is that first we get all the unique locations of the dataset, and then scan through them to learn the keywords that are used by people in Washington and Massachusetts.

For Washington, we only use keywords such as "WA" and "Washington"; the results including "Washington D.C." are not desired. Therefore, after selecting locations including "WA" or "Washington", we run another loop to exclude those including "dc", "D.C", or "d.c". We do not need to include the last dot in "D.C." as it will exclude both "D.C" and "D.C.". Sometimes, users only have their city names as their locations; however, other states might have cities that have the exact same names. Fortunately, those common city names do not have a huge population, and the first few top big cities usually have unique names. Therefore, we also include the name "Seattle" or "Kirkland".

For Massachusetts, the case is easier because there are not states that have a similar name. We use "MA", "Massachusetts" to search the desired locations in the unique location lists. We also look up a few large cities in Massachusetts, such as "Worcester" or "Boston", as keywords to include more data points.

SVM Results

Accuracy	0.7471897224133975
Precision	0.6718799872326844
Recall	0.966039467645709

Table 2: SVM metrics

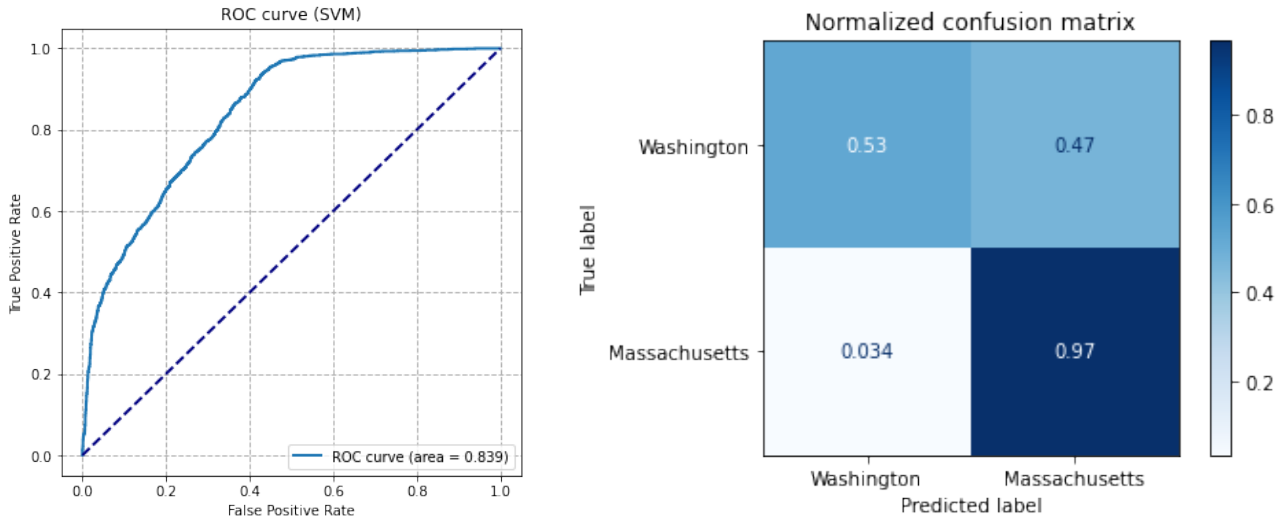


Figure 49: SVM confusion matrix and ROC curve

Logistic Regression Results

Accuracy	0.7531543932094517
Precision	0.7046382189239332
Recall	0.8715006883891694

Table 3: Logistic Regression metrics

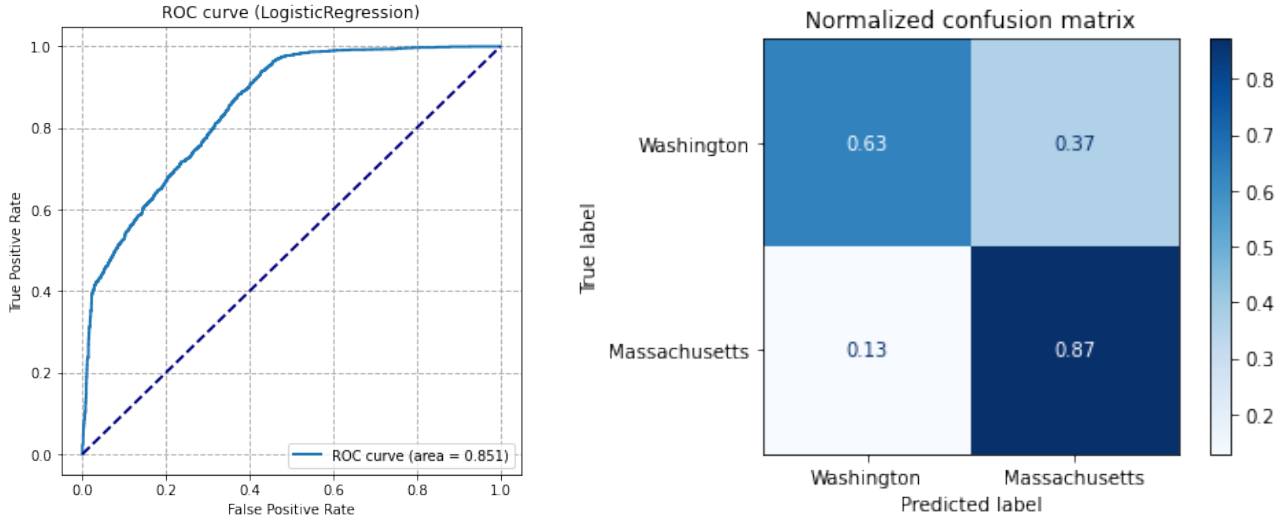


Figure 50: Logistic Regression confusion matrix and ROC curve

GuassianNB Results

Accuracy	0.676990135352145
Precision	0.6984045290787442
Recall	0.6227627351996329

Table 4: GuassianNB metrics

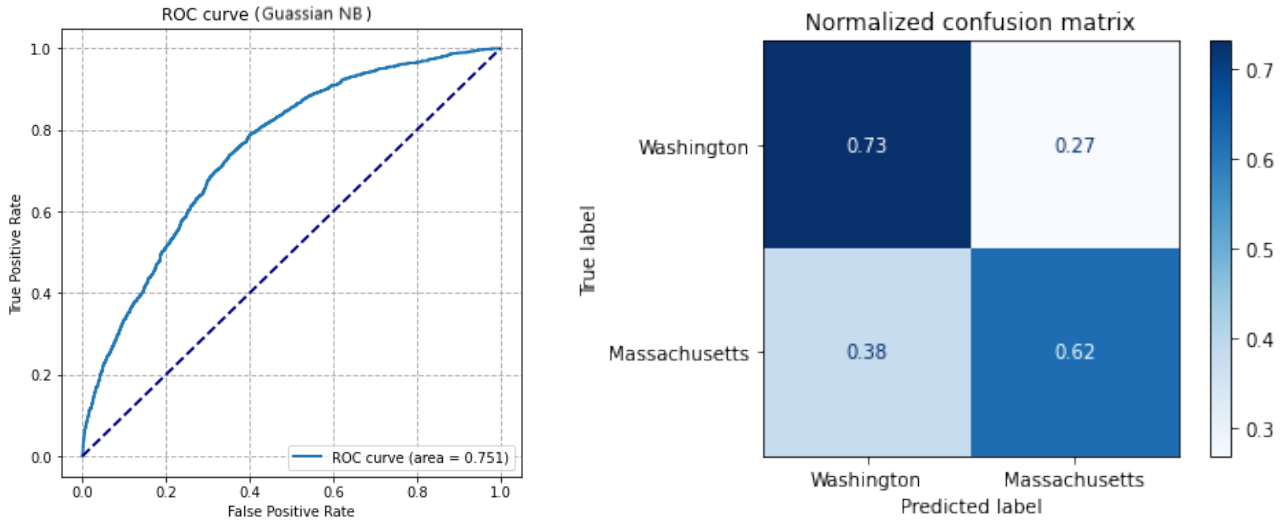


Figure 51: GuassianNB confusion matrix and ROC curve

Part 3: Define Your Own Project

Question 16: The dataset in hands is rich as there is a lot of metadata to each tweet. Be creative and propose a new problem (something interesting that can be inferred from this dataset) other than the previous parts. You can look into the literature of Twitter data analysis to get some ideas. Implement your idea and show that it works. As a suggestion, you might provide some analysis based on changes of tweet sentiments for fans of the opponent teams participating in the match. You get full credit for briniging in novelty and full or partial implementation of your new ideas.

In this section, we preprocess every tweet to look for the GPS location, which gives us 2 numbers: latitude and longitude. Note that not every tweet has this. For those tweets, we simply ignore them. Below is a figure of the tweet locations for the ”superbowl” tweets. It seems like people are tweeting about it all around the world!

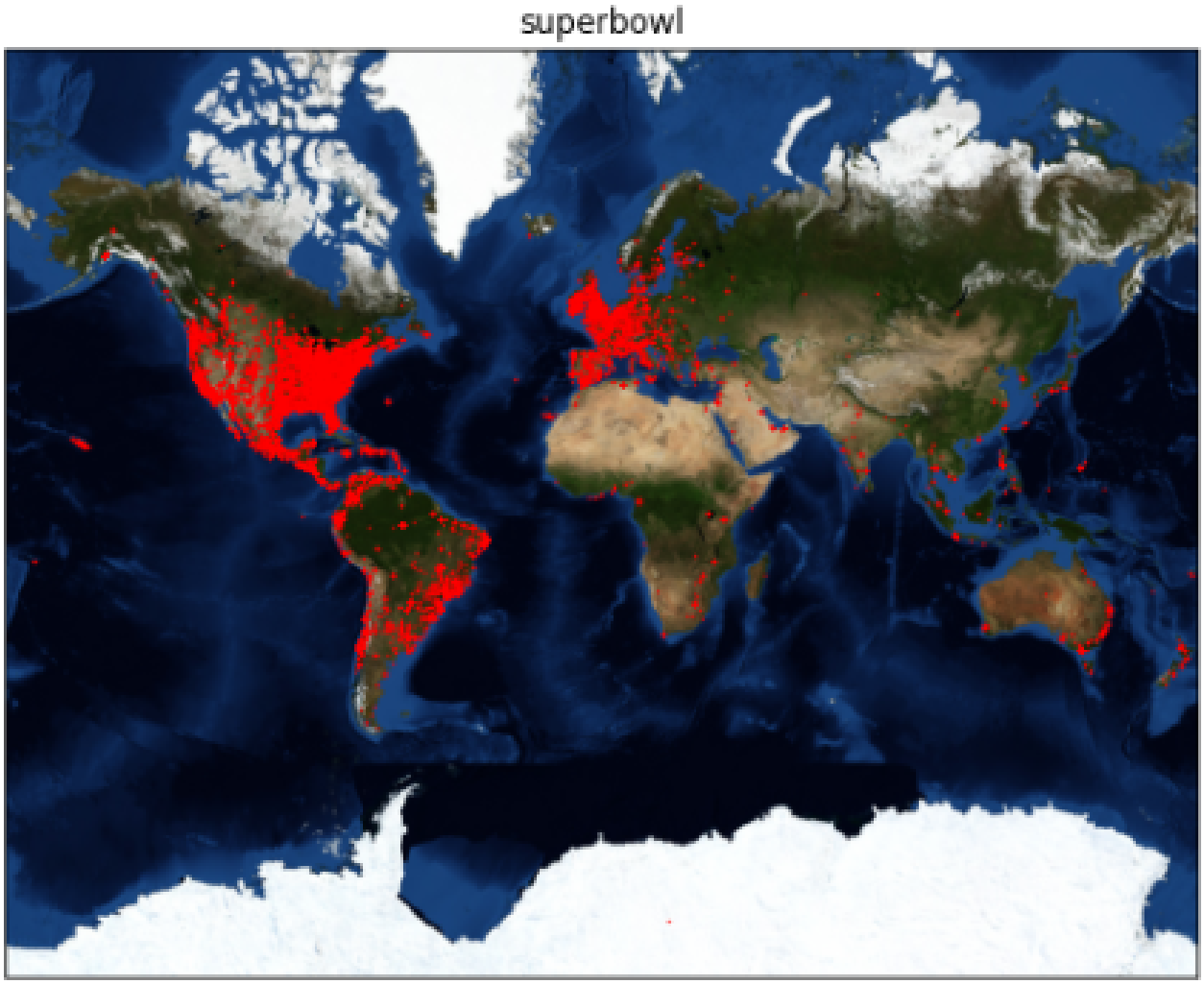


Figure 52: Locations for ”superbowl” tweets.

Here are the maps for each individual tweet. For clarity, each map is zoomed in on the continental US (the lower 48 states). **Note that for the ”gopatriots” hashtag there's a very clear cluster around Boston and for the ”gohawks” hashtag there's a very clear cluster around Seattle.** This makes sense, since that's where the teams are. Also, there's a sizeable cluster near Phoenix, Arizona (since that's where the Superbowl was held).

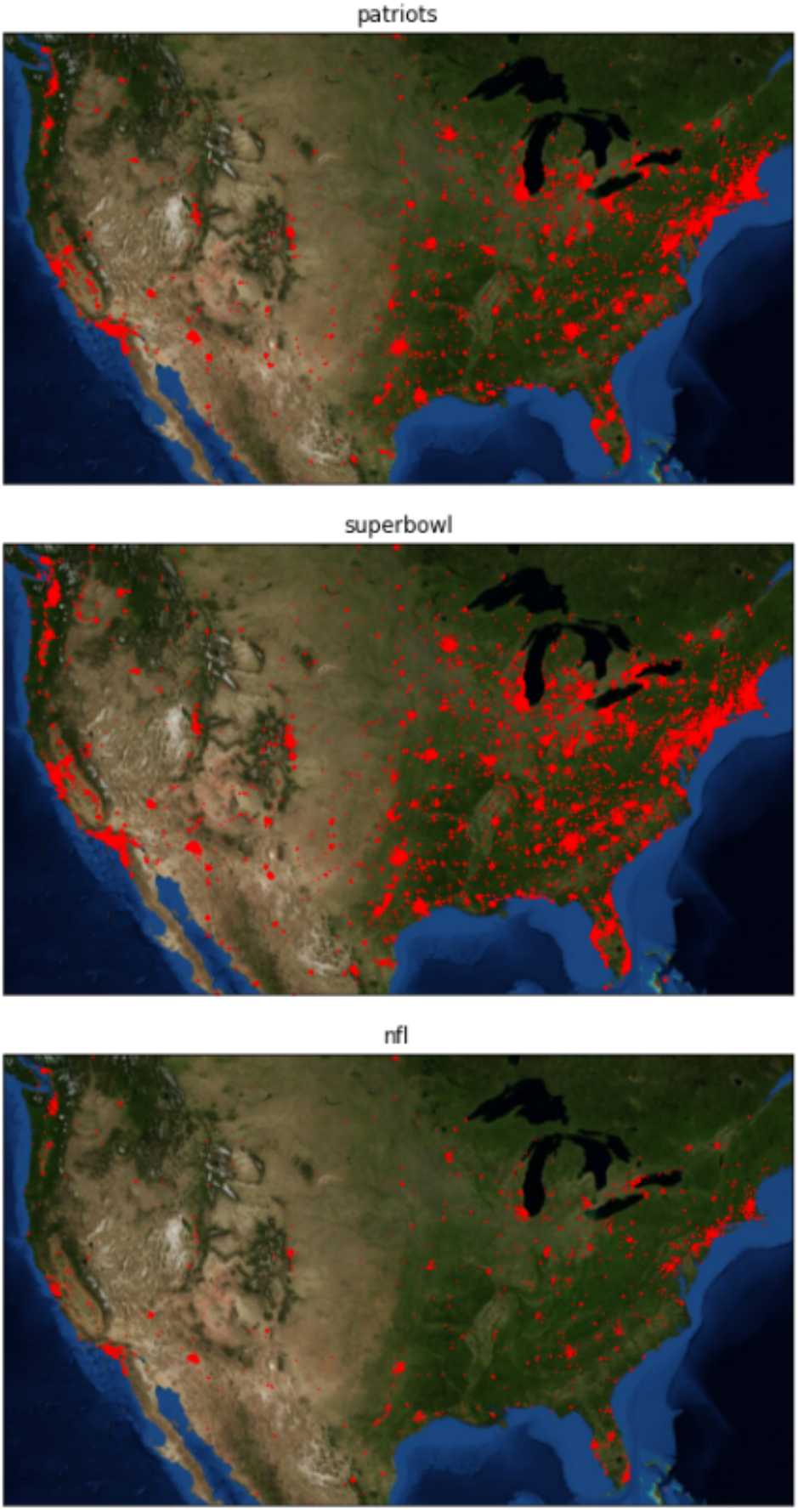


Figure 53: Locations for the ”patriots”, ”superbowl”, and ”nfl” tweets for the continental US.

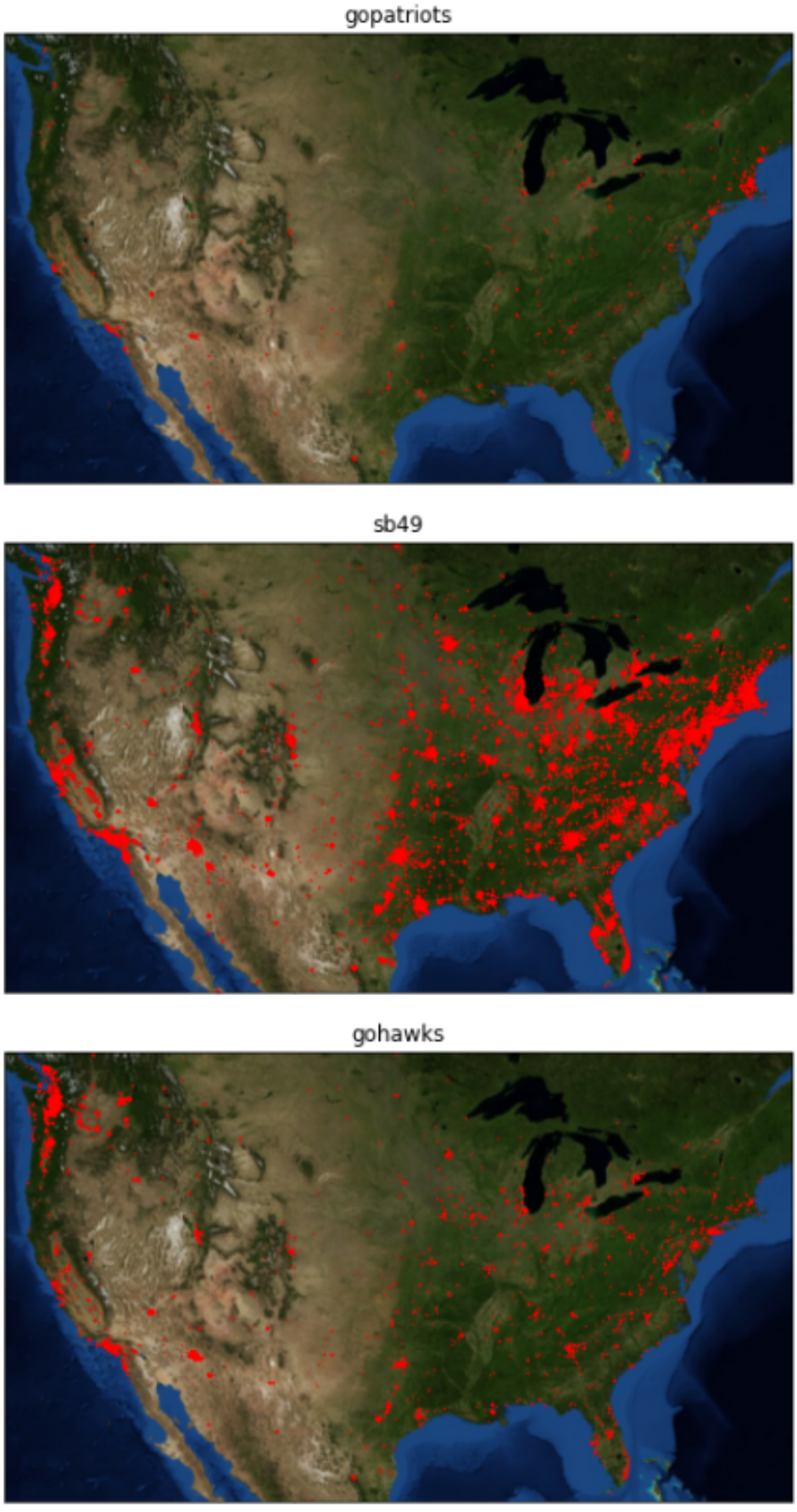


Figure 54: Locations for the ”gopatriots”, ”sb49”, and ”gohawks” tweets for the continental US. Note the geographic location for the ”gopatriots” versus ”gohawks” tweets.

Now, we try to predict the hashtag of the tweet based on this location data. For now, we perform binary classificaton (i.e. either ”gopatriots” or ”gohawks”). Later we will predict on all 6. For this binary classification, we used a random forest classifier. We used a randomized grid search with 5-fold cross validation with these parameters:

Hyperparameters for Binary Classification	**Values Tested**
Minimum Samples per Leaf	[1, 2, 3, 4]
Minimum Samples per Split	[2, 5, 10]
Number of Trees	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
Depth of Each Tree	[10, 30, 50, 70, 100, 200, None]

We ended up using these values:

Hyperparameters for Binary Classification	**Values Used**
Minimum Samples per Leaf	4
Minimum Samples per Split	10
Number of Trees	1600
Depth of Each Tree	70

Here’s the precision, recall, F1, and accuracy for the binary classification. Overall, we do pretty well and achieve a 95 percent accuracy.

	precision	recall	f1-score	support
gopatriots	0.89	0.68	0.77	1599
gohawks	0.95	0.99	0.97	10774
accuracy			0.95	12373
macro avg	0.92	0.83	0.87	12373
weighted avg	0.95	0.95	0.94	12373

Figure 55: Statistics for the ”gopatriots” versus ”gohawks” binary classification.

Now we consider the harder problem of classifying all 6 hashtags. This problem is harder because there are more categories but also the hashtags start losing their geographical association. For instance, it’s not clear where most ”superbowl” hashtag users are compared to ”sb49” hashtag users. We use the same randomized grid search over the parameters above. (Also using a random forest classifier and 5-fold cross validation.) We ended up using the parameters below.

Hyperparameters for Classifying 6 Classes	**Values Used**
Minimum Samples per Leaf	2
Minimum Samples per Split	5
Number of Trees	200
Depth of Each Tree	50

As the results below show, we do a lot worse than before althoug we still get a 76 percent accuracy. This shows that for these hashtags it’s harder to classify based on GPS data alone.

	precision	recall	f1-score	support
gopatriots	0.78	0.28	0.41	1599
gohawks	0.81	0.76	0.78	10774
patriots	0.67	0.40	0.50	19732
superbowl	0.80	0.92	0.85	58663
nfl	0.68	0.28	0.39	5253
sb49	0.73	0.80	0.76	37322
accuracy			0.76	133343
macro avg	0.74	0.57	0.62	133343
weighted avg	0.75	0.76	0.75	133343

Figure 56: Statistics for all 6 classifications.