Project 5: Application - Twitter data

Name: Jianxiong Wang, Yijun Wu, Yanzhao Wang,
Yutong Sun
Date: 2019.3.20

Question 1

Report the following statistics for each hashtag, i.e. each file:

- **Average number of tweets per hour**
- **Average number of followers of users posting the tweets per tweet (to make it simple, we average over the number of tweets; if a users posted twice, we count the user and the user's followers twice as well)**
- **Average number of retweets per tweet**

	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
Average tweets per hour	40.888695 652173915	292.09326 424870466	396.97103 918228277	750.63202 72572402	1275.5557 461406518	2067.824531 516184
Average followers per tweet	1427.2526 051635405	2217.9237 355281984	4662.3754 4523693	3280.4635 616550277	10374.160 292019487	8814.967994 24623
Average retweets per tweet	1.4081919 101697078	2.0132093 991319877	1.5344602 655543254	1.7852871 288476946	2.5271344 4111402	2.391189581 9207736

Table 1. Statistics for each hashtag

Question 2

Plot "number of tweets in hour" over time for #SuperBowl and #NFL (a histogram with 1-hour bins).

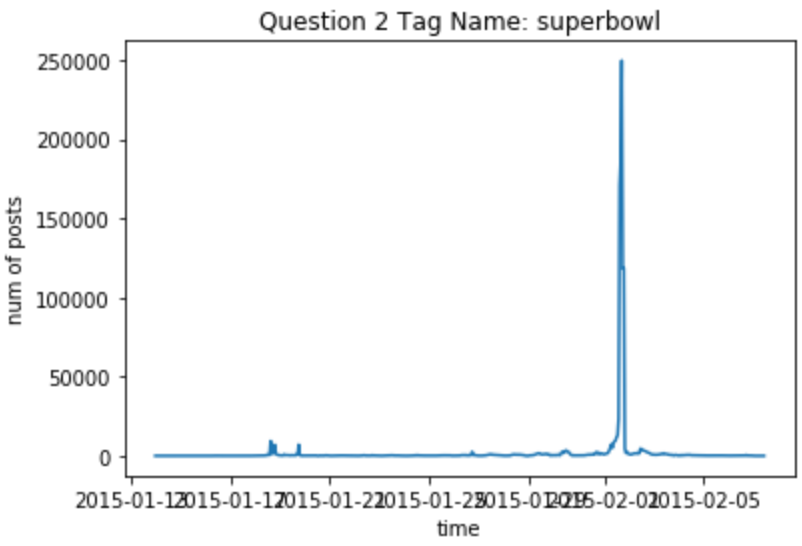


Figure 1. Number of tweets in hour vs. time for #SuperBowl

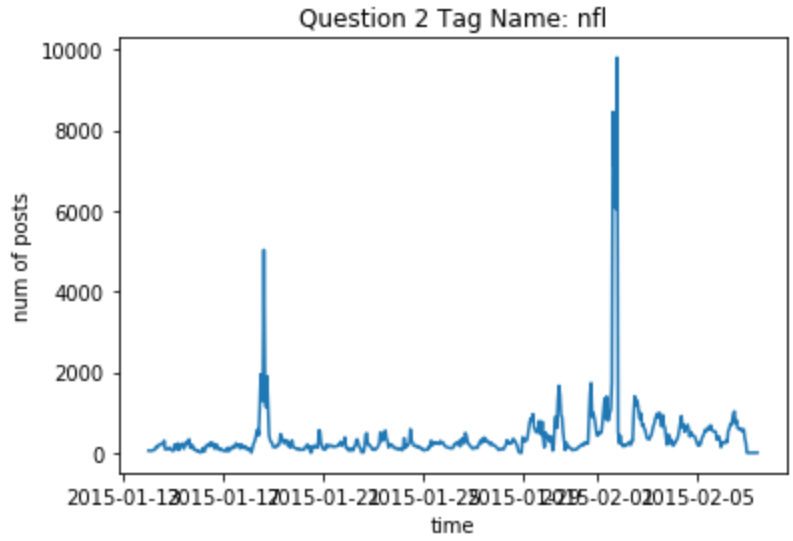


Figure 2. Number of tweets in hour vs. time for #NFL

Question 3

For each hashtag data file, fit a linear regression model using the following 5 features to predict number of tweets in the next hour, with features extracted from tweet data in the previous hour. The features you should use are:

- **Number of tweets**
- **Total number of retweets**
- **Sum of the number of followers of the users posting the hashtag**
- **Maximum number of followers of the users posting the hashtag**
- **Time of the day**

For each of your models, report your model's Mean Squared Error (MSE) and R-squared measure. Also, analyse the significance of each feature using the t-test and p-value.

	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
Mean squared error	27610.11	760963.89	273848.28	5189462.12	16199530.45	52488994.91
R squared value	0.6372256940098271	0.5041002071921129	0.6521006506177033	0.6794725841702596	0.8082352136527163	0.8030419980925358
P-value of number of tweets in the previous hour	0.27931161	2.15583400e-14	2.99044196e-06	7.59459794e-34	2.95505976e-34	9.67194039e-113
P-value of number of retweets	0.00941955	1.60947513e-03	4.79416843e-03	2.42607082e-01	4.00144711e-02	3.40873070e-008
P-value of sum of number of followers	0.5796288	1.52369053e-02	2.42865533e-05	6.60651423e-01	4.29712429e-01	6.71733985e-010
P-value of max followers	0.95397142	6.31486219e-01	2.48610844e-03	1.37065456e-01	2.27035922e-02	7.12296562e-007
P-value of time of the day	0.47142064	1.04619105e-02	1.30036264e-04	5.05783085e-01	8.12754589e-01	2.77906836e-001

Table 2. Mean Squared Error (MSE) and R-squared and p-value of each hashtag's model

P-value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be greater than or equal to the actual observed results. When a feature has lower p-value, it has a larger significance to the fitted model. Therefore, based on the p-value we got for all 5 features, we can conclude that for #gopatriots, number of retweets is the most significant feature for predicting the number of tweets in the next hour, and for #gohawks, #nfl, #patriots, #sb49 and #superbowl, number of tweets in the previous hour is the most significant feature.

Question 4

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

The model we designed is a linear regression model that takes 9 features, including time of the day, max followers, number of followers, number of retweets, happy emoji count, sad emoji count, user mentioned count, url count, number of tweets, and predicts the number of tweets in the next hour. And the model is fitted on data of all 6 hashtags. The MSE and p-values of the 6 fitted models we got are shown in Table 3 below.

	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
Mean squared error	9325.34	577818.60	225143.78	3694750.02	12311366.55	32528318.01
P-value of number of tweets in the previous hour	1.06749096e-01	3.47239965e-03	9.81328211e-02	3.35681622e-01	7.90197085e-06	1.34264138e-01
P-value of url count	1.43881749e-07	5.36889347e-17	3.63237350e-11	2.28000829e-01	5.68224321e-01	6.17142765e-01
P-value of user mentioned count	1.28988208e-73	2.82577279e-01	1.04115537e-04	6.73446215e-02	3.02801173e-02	5.67992456e-45
P-value of sad emoji count	3.45205616e-01	1.68160854e-12	2.16865080e-02	3.66906494e-03	1.20786429e-03	9.13382950e-07
P-value of happy emoji count	7.28216169e-81	1.22148788e-01	5.22271789e-15	3.92399082e-06	2.32953374e-16	8.40087510e-04
P-value of number of retweets	8.96991875e-06	6.83885610e-01	6.72989864e-01	1.11893084e-01	5.16399876e-02	9.69024473e-30

P-value of sum of number of followers	1.53634462e-05	4.00808352e-02	2.06030086e-02	1.55002302e-18	3.86919147e-02	1.26284992e-03
P-value of max followers	6.78204240e-07	5.69531034e-01	3.71828255e-02	1.46482278e-07	7.12492058e-03	7.09778878e-01
P-value of time of the day	7.01529382e-02	3.29245076e-01	7.37591942e-01	1.98390461e-01	2.37200940e-01	1.93197044e-02

Table 3. Mean Squared Error and p-value of each hashtag's model

Since we know that a feature is more significant if it has a lower p-value, we can get the significance ranking of the 9 features based on p-values in Table 3:

#gopatriots:
happy emoji count > user mentioned count > url count > max followers > number of retweets > sum of number of followers > time of the day > number of tweets > sad emoji count

#gohawks:
url count > sad emoji count > number of tweets > sum of number of followers > happy emoji count > user mentioned count > time of the day > max followers > number of retweets

#nfl:
happy emoji count > url count > user mentioned count > sum of number of followers > sad emoji count > max followers > number of tweets > number of retweets > time of the day

#patriots:
sum of number of followers > max followers > happy emoji count > sad emoji count > user mentioned count > number of retweets > time of the day > url count > number of tweets

#sb49:
happy emoji count > number of tweets > sad emoji count > max followers > user mentioned count > sum of number of followers > number of retweets > time of the day > url count

#superbowl:
user mentioned count > number of retweets > sad emoji count > happy emoji count > sum of number of followers > time of the day > number of tweets > url count > max followers

Question 5

For each of the top 3 features (i.e. with the smallest p-values) in your measurements, draw a scatter plot of predictant (number of tweets for next hour) versus value of that feature, using all the samples you have extracted, and analyze it. Do the regression coefficients agree with the trends in the plots? If not, why?

#gopatriots:

Top three features of #gopatriots:

1. Happy emoji count:
   p-value: 7.282161691952267e-81, coefficient: -174.48921096168107
2. User mentioned count:
   P-value: 1.2898820815369715e-73, coefficient: 8.044450240419753
3. Url count:
   P-value: 1.4388174876533967e-07, coefficient: 3.053911771619279

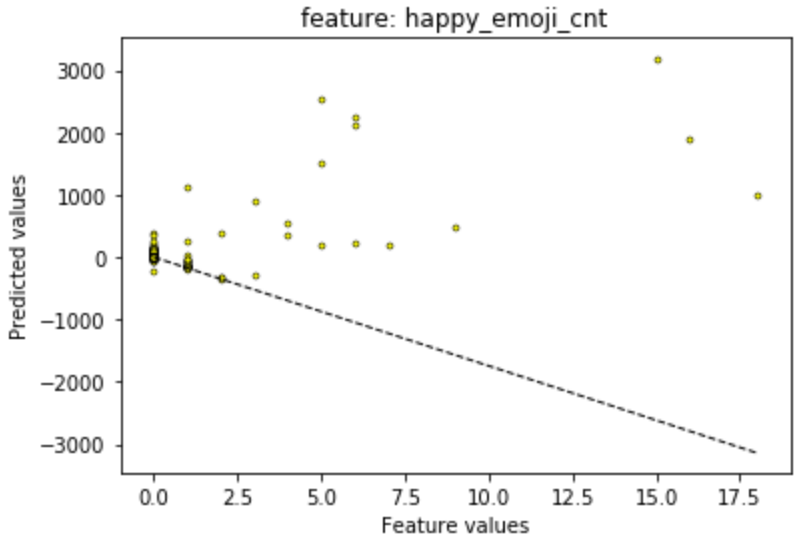


Figure 3. Scatter plot of predictant (number of tweets for next hour) vs. value of happy emoji count, #gopatriots

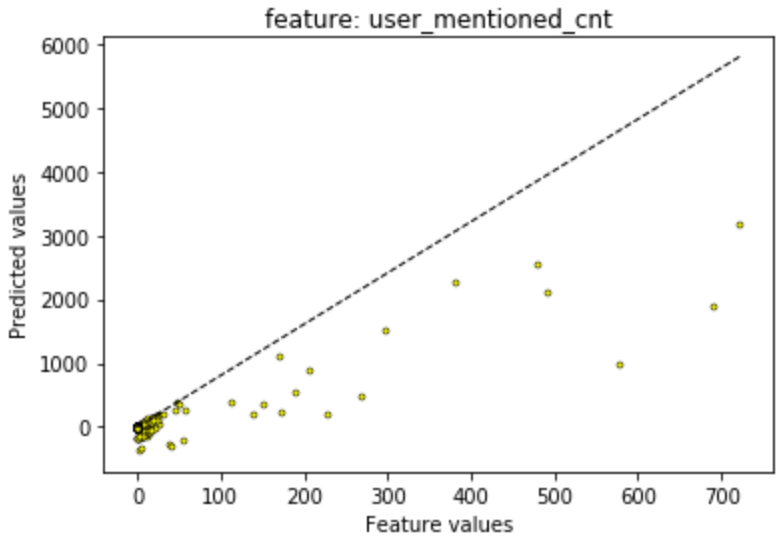


Figure 4. Scatter plot of predictant (number of tweets for next hour) vs. value of user mentioned count, #gopatriots

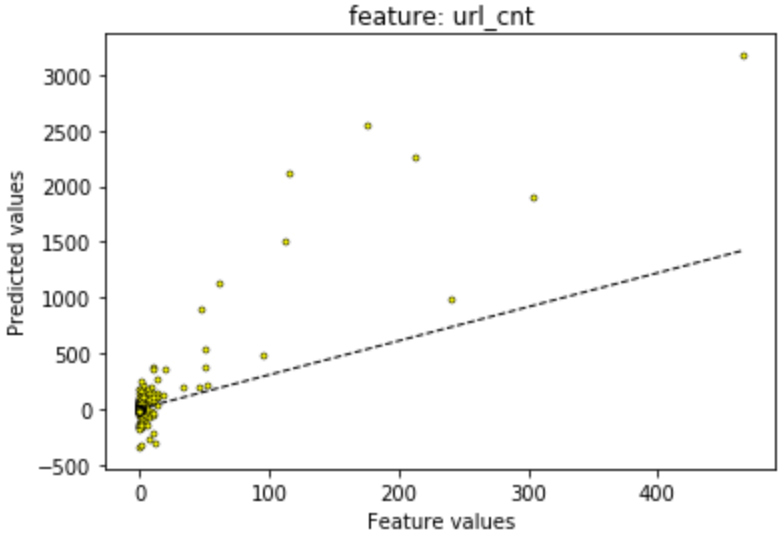


Figure 5. Scatter plot of predictant (number of tweets for next hour) vs. value of url count, #gopatriots

The dotted line in the scatter plot demonstrates the linear model fitted on the data in a scope of the feature. The coefficient of the feature in the fitted model should correspond to the slope of the fitted line, and if the line fit the data points better (has smaller loss) in the 2D space of the feature, this feature should be more significant for the model.

In the case of #gopatriots, the slopes of lines in Figure 3, 4, 5 agree with the coefficients -174.48921096168107, 8.044450240419753, 3.053911771619279. As happy emoji count and user mentioned count has similar p-value which is much smaller than the p-values of other

features, their significance should be close, which corresponding to similar fit of data points in 2D. And the fitted lines on these 2 features' plots fit the data points better than the linear model on url count plot. The scatter plots for happy emoji count, user mentioned count, and url count in Figure 3, 4, 5 agree with the coefficients of these 3 features in the model fitted with #gopatriots data.

#gohawks:

Top three features of #gohawks:

1. Url count:
   p-value: 5.368893471679995e-17, coefficient: 7.89154347035918
2. Sad emoji count:
   p-value: 1.681608538561575e-12, coefficient: 864.3932753249821
3. Number of tweets in the previous hour:
   p-value: 0.003472399650845431, coefficient: 0.644059944884301

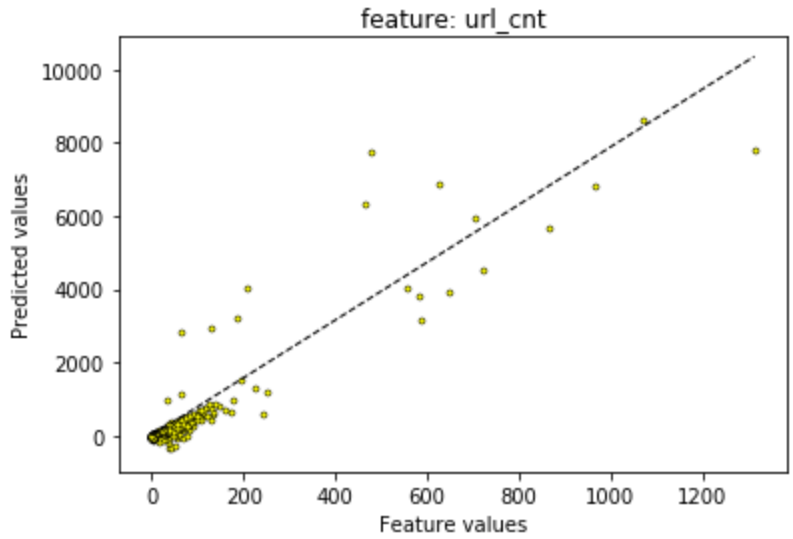


Figure 6. Scatter plot of predictant (number of tweets for next hour) vs. value of url count, #gohawks

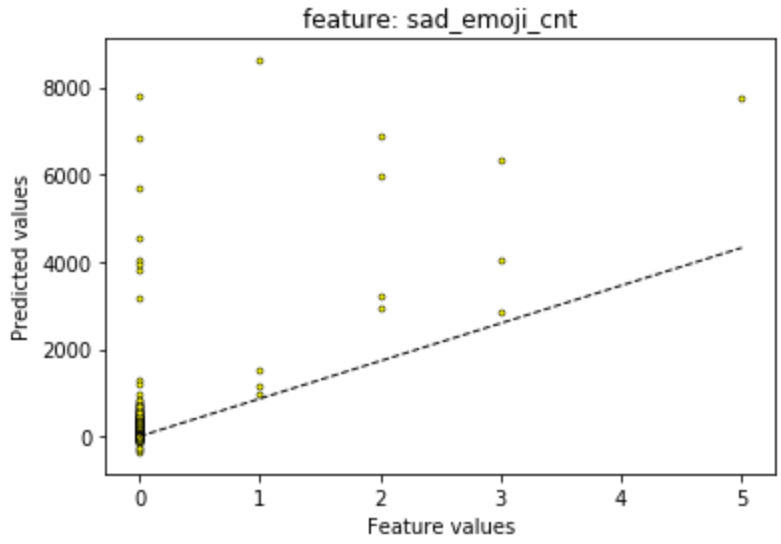


Figure 7. Scatter plot of predictant (number of tweets for next hour) vs. value of sad emoji count, #gohawks

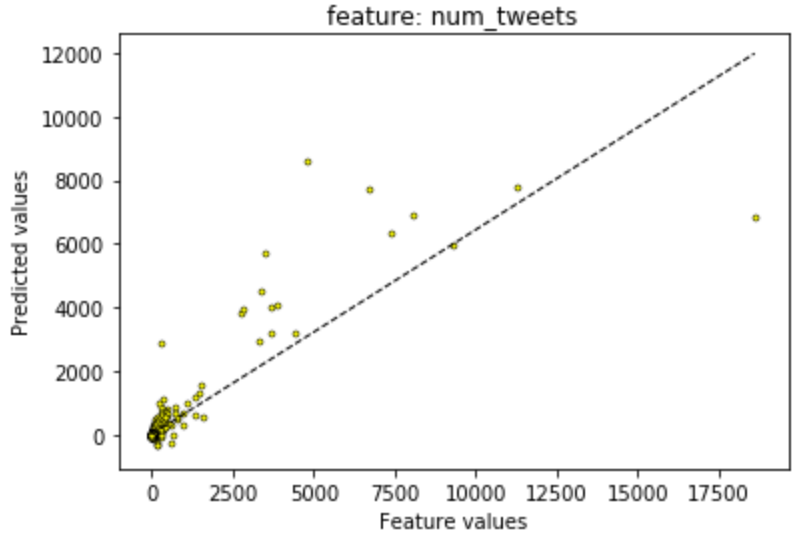


Figure 8. Scatter plot of predictant (number of tweets for next hour) vs. value of number of tweets, #gohawks

In the case of # gohawks, the slopes of lines in Figure 6, 7, 8 agree with the coefficients 7.89154347035918, 864.3932753249821, 0.644059944884301. The line on the url count plot fits the data points better than the line on the sad emoji count plot, and the line on the number of tweets plot, because the p-value of url count is smaller than the p-values of other features. Therefore, the scatter plots for url count, sad emoji count, number of tweets in Figure 6, 7, 8 agree with the coefficients of these 3 features in the model fitted with # gohawks data.

#nfl:

Top three features of #nfl:

1. Happy emoji count:
   p-value: 5.222717892875908e-15, coefficeint: 114.87343391680255
2. Url count:
   P value: 3.632373495197207e-11, coefficeint: 0.8674348079709744
3. User mentioned count:
   P value: 0.00010411553715715543, coefficeint: 2.6513348900536418

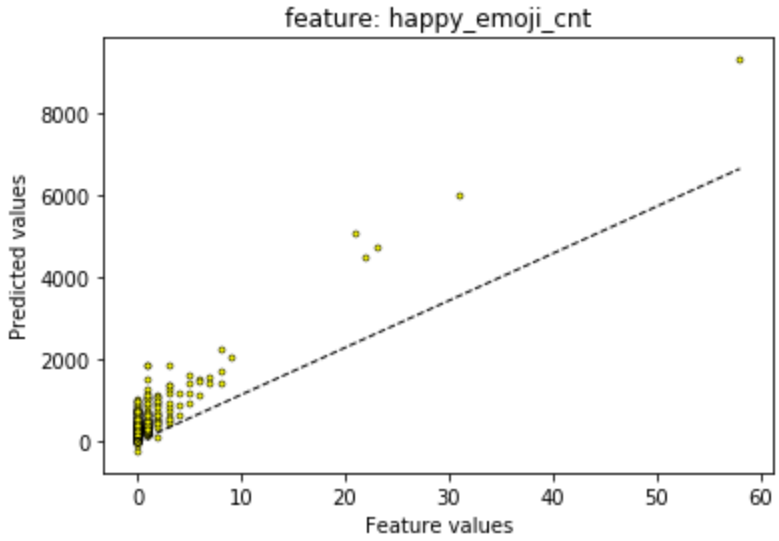


Figure 9. Scatter plot of predictant (number of tweets for next hour) vs. value of happy emoji count, #nfl

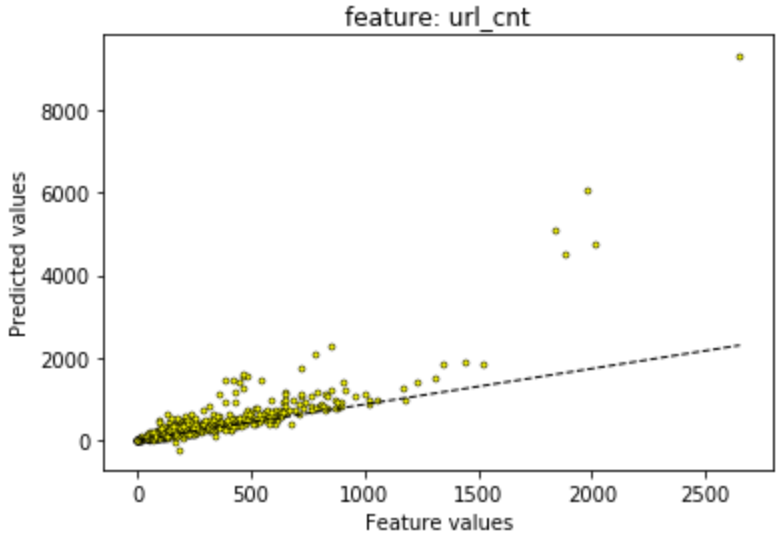


Figure 10. Scatter plot of predictant (number of tweets for next hour) vs. value of url count, #nfl

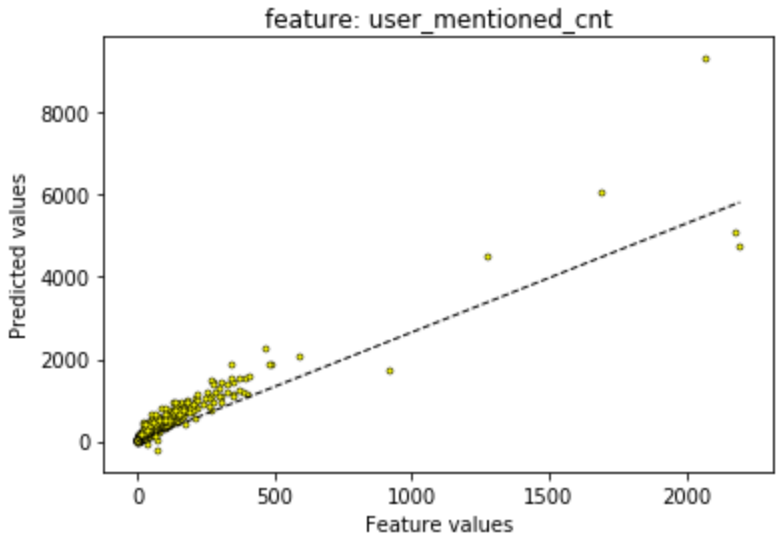


Figure 11. Scatter plot of predictant (number of tweets for next hour) vs. value of user mentioned count, #nfl

In the case of # nfl, the slopes of lines in Figure 9, 10, 11 agree with the coefficients 114.87343391680255, 0.8674348079709744, 2.6513348900536418. In Figure 9, data points on happy emoji count plot are better fitted than on the url count plot and user mentioned count plot, because the p-value of url count feature is smaller than the p-values of other features. Therefore, the scatter plots for feature happy emoji count, url count and user mentioned count in Figure 9, 10, 11 agree with the coefficients of these 3 features in the linear model fitted with # nfl data.

#patriots:

Top three features of #patriots:

1. Number of followers:
   P-value: 1.5500230198667302e-18, Coefficient: 0.0004314849349168283
2. Max followers:
   P-value: 1.4648227793658114e-07, Coefficient: -0.0005223201276815939
3. Happy emoji count:
   P-value: 3.923990817275451e-06, Coefficient: -191.0144245602798

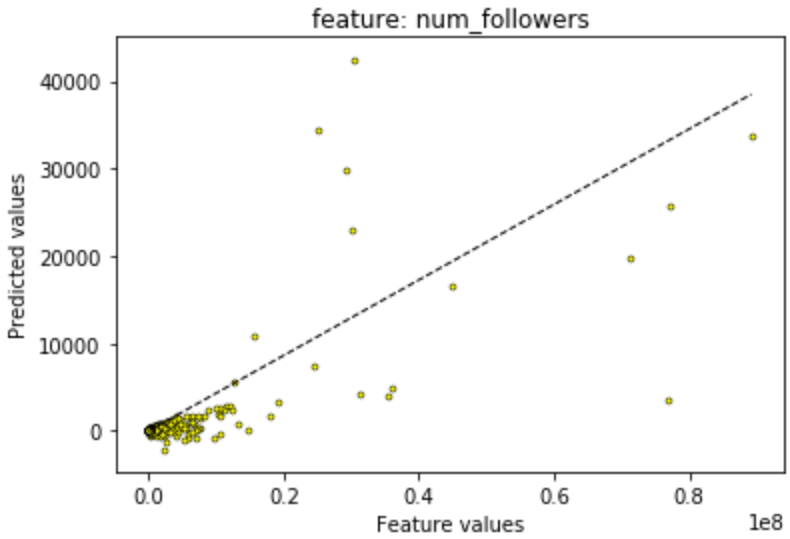


Figure 12. Scatter plot of predictant (number of tweets for next hour) vs. number of followers, #patriots

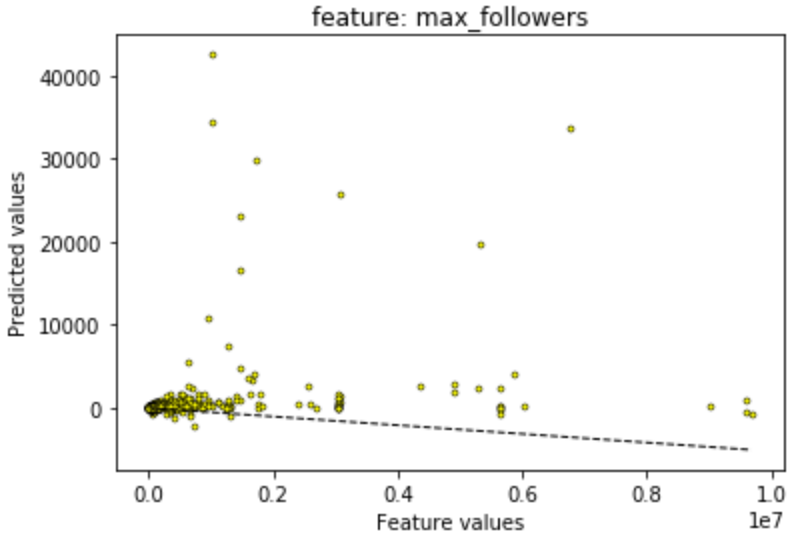


Figure 13. Scatter plot of predictant (number of tweets for next hour) vs. max followers, #patriots

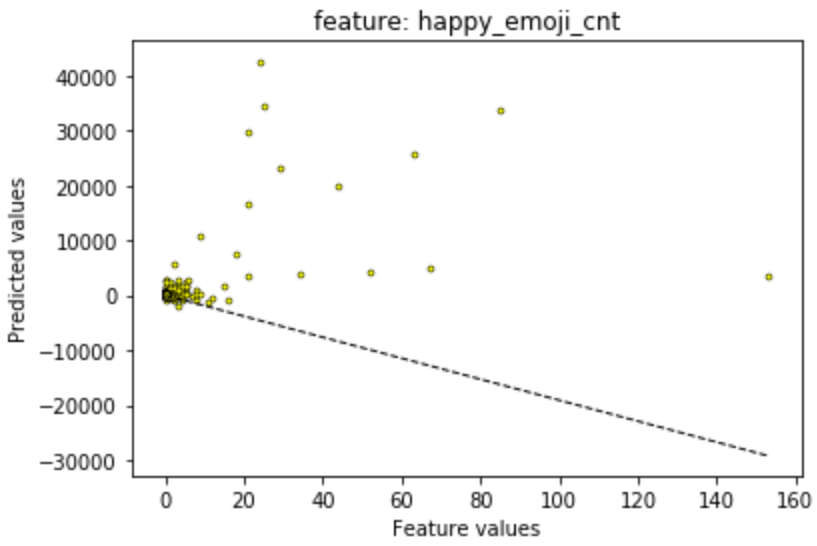


Figure 14. Scatter plot of predictant (number of tweets for next hour) vs. happy emoji count, #patriots

In the case of # patriots, the slopes of lines in Figure 12, 13, 14 agree with the coefficients 0.0004314849349168283, -0.0005223201276815939, -191.0144245602798. In Figure 12, data points on number of followers plot are better fitted than on the max follower plot and happy emoji count plot, because the p-value of number of followers feature is smaller than the p-values of other features. Therefore, the scatter plots for feature number of followers, max follower and happy emoji count in Figure 12, 13, 14 agree with the coefficients of these 3 features in the linear model fitted with #patriots data.

#sb49:

Top three features of #sb49:

1. Happy emoji count:
   P value: 2.329533744726656e-16, Coefficeint: 660.9137769895706
2. Number of tweets:
   P value: 7.901970851073756e-06, Coefficeint: -1.8340359143769183
3. Sad emoji count:
   P value: 0.001207864288965355, Coefficeint: -1857.2756778908556

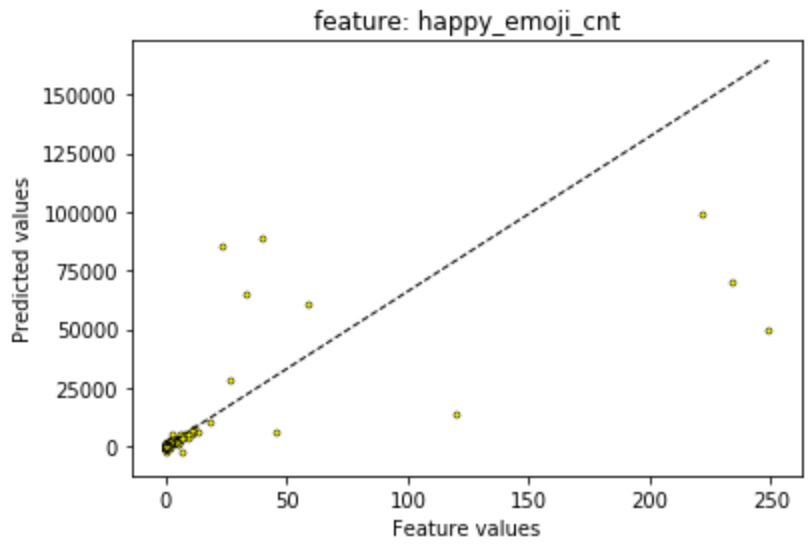


Figure 15. Scatter plot of predictant (number of tweets for next hour) vs. happy emoji count, #sb49

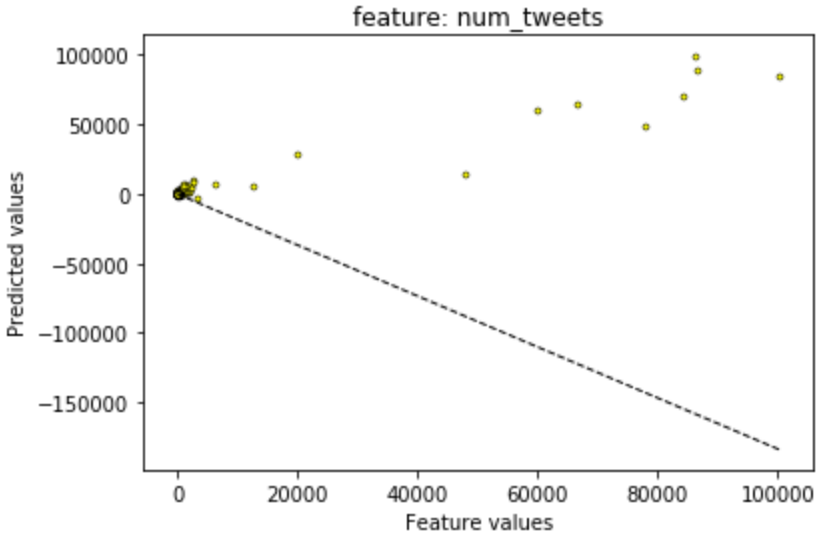


Figure 16. Scatter plot of predictant (number of tweets for next hour) vs. number of tweets, #sb49

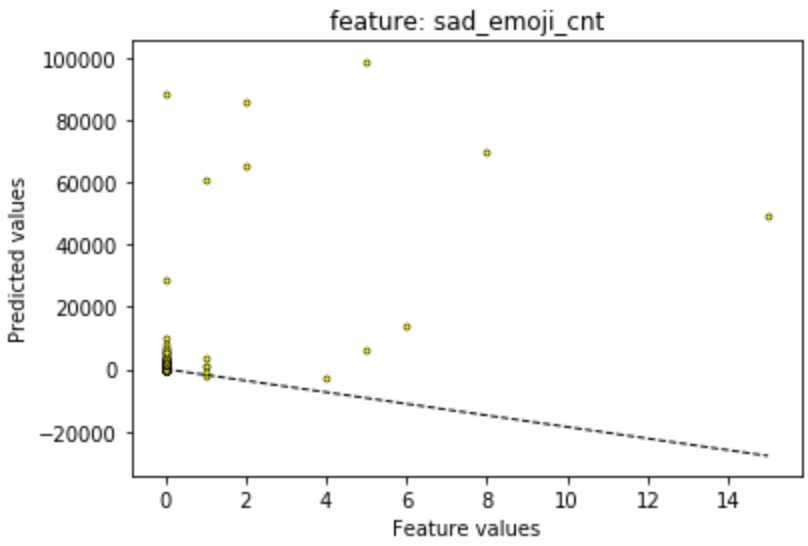


Figure 17. Scatter plot of predictant (number of tweets for next hour) vs. sad emoji count, #sb49

In the case of #sb49, the slopes of lines in Figure 15, 16, 17 agree with the coefficients 660.9137769895706, -1.8340359143769183, -1857.2756778908556. In Figure 15, data points on happy emoji count plot are better fitted than on the number of tweets plot and sad emoji count plot, because the p-value of happy emoji count feature is smaller than the p-values of other features. Therefore, the scatter plots for feature happy emoji count, number of tweets and sad emoji count in Figure 15, 16, 17 agree with the coefficients of these 3 features in the linear model fitted with #sb49 data.

#superbowl:

Top three features of #superbowl:

1. User mentioned count:
   P value: 5.679924559447648e-45, Coefficeint: 13.018558552627667
2. Number of retweets:
   P value: 9.690244729065033e-30, Coefficeint: -0.8415575922162639
3. Sad emoji count:
   P value: 9.133829498642921e-07, Coefficeint: -935.4871233612907

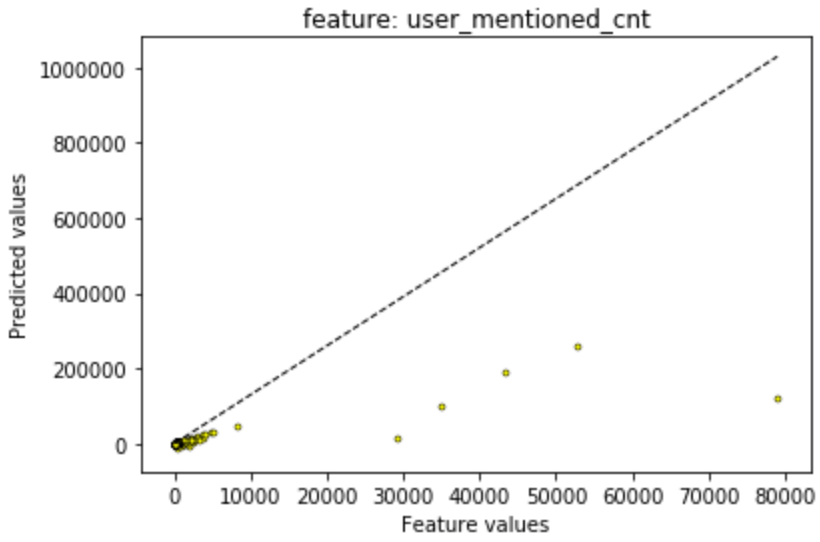


Figure 18. Scatter plot of predictant (number of tweets for next hour) vs. user mentioned count, #superbowl

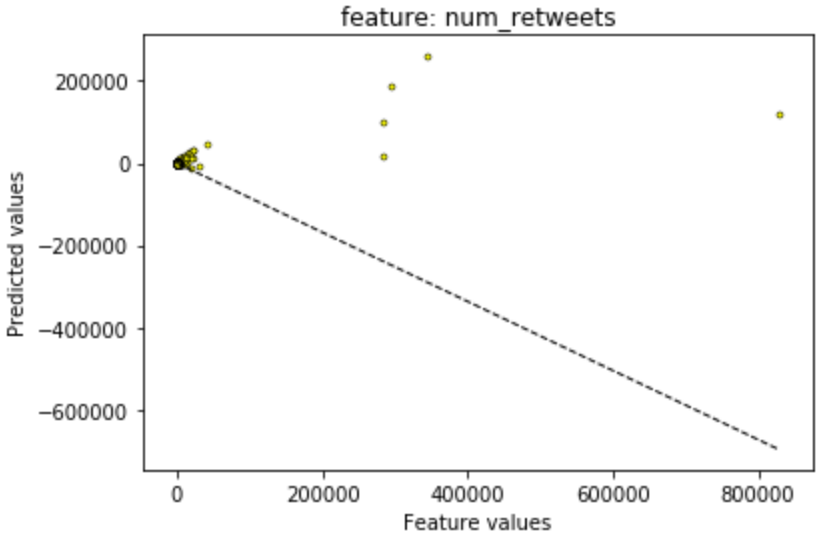


Figure 19. Scatter plot of predictant (number of tweets for next hour) vs. number of retweets, #superbowl

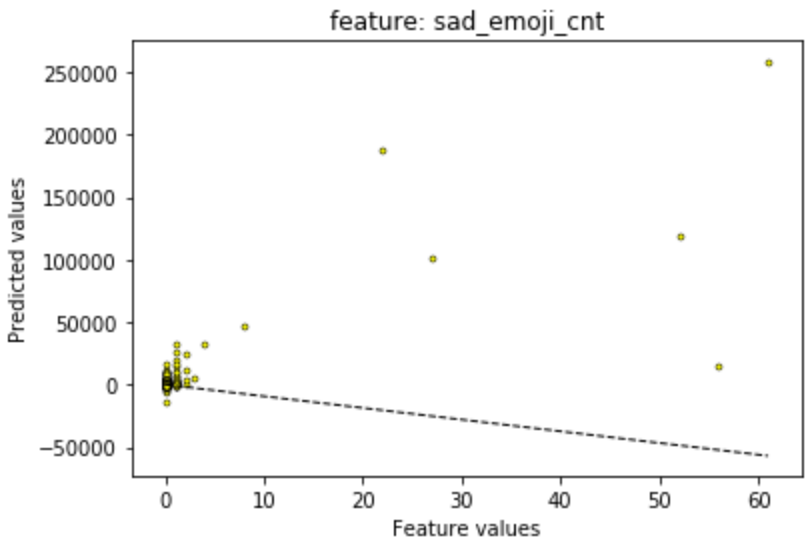


Figure 20. Scatter plot of predictant (number of tweets for next hour) vs. sad emoji count, #superbowl

In the case of #superbowl, the slopes of lines in Figure 18, 19, 20 agree with the coefficients 13.018558552627667, -1.8340359143769183, -1857.2756778908556. In Figure 18, data points on user mentioned count plot are better fitted than on the number of tweets plot and sad emoji count plot, because the p-value of user mentioned count feature is smaller than the p-values of other features. Therefore, the scatter plots for feature user mentioned count, number of tweets and sad emoji count in Figure 18, 19, 20 agree with the coefficients of these 3 features in the linear model fitted with #superbowl data.

Question 6

We define three time periods and their corresponding window length as follows:

1. **Before Feb. 1, 8:00 a.m.: 1-hour window**
2. **Between Feb. 1, 8:00 a.m. and 8:00 p.m.: 5-minute window**
3. **After Feb. 1, 8:00 p.m.: 1-hour window**

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.

The 3 periods correspond to period before superbowl, during superbowl, and after superbowl. In this part, we separated the original data into 3 json objects by time based on citation_date in the json object. MSE and R-squared score values for data of 6 hashtags are reported in Table 4 to Table 9.

#gopatriots:

	1st period	2nd period	3rd period
Mean squared error	1421.95	13591.79	17.13
R-squared score	0.668210980430 7836	0.654849434592 8143	0.9149492868667 216
P-value of number of tweets	1.31920334e-09	9.93656065e-01	1.27429091e-19
P-value of url count	8.33451432e-01	1.35692342e-01	1.83598996e-04
P-value of user mentioned count	1.62131766e-17	3.33502590e-02	6.47537115e-25
P-value of sad emoji count	3.23309345e-01	5.51928644e-01	3.20915063e-16
P-value of happy emoji count	2.23220087e-05	5.57074861e-01	5.96212834e-17
P-value of number of retweets	6.96352826e-04	1.32418146e-02	1.17337996e-02
P-value of sum of number of followers	2.29260308e-07	4.98855677e-01	5.51374278e-07
P-value of max followers	1.63628533e-07	6.42900740e-01	1.85111930e-04
P-value of time of the day	2.54966736e-03	7.08340220e-04	9.56741823e-01

Table 4. Mean Squared Error (MSE) and R-squared of gopatriots tag model for all 3 periods

#gohawks:

	1st period	2nd period	3rd period
Mean squared error	632589.52	68985.41	1415.48
R-squared score	0.4195209040988722	0.7711735772408022	0.885136180229501
P-value of number of tweets	2.04826459e-02	2.38011164e-04	2.85237715e-01
P-value of url count	3.85042709e-01	1.46938964e-02	8.29061192e-02
P-value of user mentioned count	6.97262520e-04	6.84535638e-02	4.71642247e-01
P-value of sad emoji count	2.64648416e-10	8.67441314e-01	2.90219650e-01
P-value of happy emoji count	8.24788800e-02	7.69613544e-01	1.34902811e-06
P-value of number of retweets	2.80434347e-02	6.81468517e-01	6.91941585e-02
P-value of sum of number of followers	5.40646966e-02	9.18635989e-02	1.34241059e-01
P-value of max followers	4.14109093e-01	1.76950978e-01	1.83076400e-03
P-value of time of the day	8.39941501e-01	4.45363139e-03	8.40585238e-01

Table 5. Mean Squared Error (MSE) and R-squared of gohawks tag model for all 3 periods

#nfl:

	1st period	2nd period	3rd period
Mean squared error	60824.06	19755.93	16884.89
R-squared score	0.7115216117056096	0.9113072989737111	0.9469024153469358
P-value of number of tweets	9.81518268e-01	6.74462897e-18	0.02891159
P-value of url count	2.02979397e-05	8.83563396e-01	0.52897557
P-value of user mentioned count	3.44756902e-02	7.79475035e-02	0.42488628
P-value of sad emoji count	7.60079414e-01	1.12142451e-02	0.33199689
P-value of happy emoji count	1.14724137e-07	6.02341449e-01	0.01722039
P-value of number of retweets	5.52556020e-01	1.44116936e-01	0.27589568
P-value of sum of number of followers	4.06592605e-01	1.46695618e-01	0.90644514
P-value of max followers	6.66618439e-01	7.17792494e-01	0.91496429
P-value of time of the day	7.09151729e-02	4.19999230e-03	0.34822884

Table 6. Mean Squared Error (MSE) and R-squared of nfl tag model for all 3 periods

#patriots:

	1st period	2nd period	3rd period
Mean squared error	314508.19	648968.27	8110.07
R-squared score	0.6412036336981739	0.8941854974883665	0.9309226730412096
P-value of number of tweets	1.02392471e-01	0.00171346	5.98209787e-15
P-value of url count	1.98101285e-02	0.9238926	6.37670773e-08
P-value of user mentioned count	9.72183636e-01	0.76774668	9.78315483e-01
P-value of sad emoji count	1.88071495e-01	0.43657724	4.55273158e-02
P-value of happy emoji count	3.19067021e-06	0.0816078	3.67107856e-05
P-value of number of retweets	7.23965213e-01	0.98952572	3.83478613e-22
P-value of sum of number of followers	8.43251227e-16	0.14373893	1.88759918e-01
P-value of max followers	3.73485383e-12	0.39148527	1.66755628e-01
P-value of time of the day	3.13037028e-03	0.00126445	6.21963751e-02

Table 7. Mean Squared Error (MSE) and R-squared of patriots tag model for all 3 periods

#sb49:

	1st period	2nd period	3rd period
Mean squared error	5993.60	1229086.76	51149.31
R-squared score	0.9022564613799776	0.9590600500051498	0.897978805132073
P-value of number of tweets	3.71478672e-52	0.00087893	2.25533463e-01
P-value of url count	2.36432874e-05	0.16391088	3.14763769e-06
P-value of user mentioned count	7.55301143e-01	0.06269751	9.68285585e-01
P-value of sad emoji count	2.94306263e-01	0.16941452	6.10000426e-01
P-value of happy emoji count	8.97166252e-01	0.70652296	8.22289381e-07
P-value of number of retweets	4.41844409e-03	0.4950804	6.12677206e-01
P-value of sum of number of followers	5.88328294e-01	0.1592681	6.14351182e-01
P-value of max followers	8.67581307e-01	0.60023582	8.48166999e-01
P-value of time of the day	2.05422767e-01	0.02852265	6.24671089e-01

Table 8. Mean Squared Error (MSE) and R-squared of sb49 tag model for all 3 periods

#superbowl:

	1st period	2nd period	3rd period
Mean squared error	482073.44	5396443.79	109651.39
R-squared score	0.5436752319218078	0.9481702055381944	0.9074163667096606
P-value of number of tweets	1.26598707e-01	2.36670554e-20	4.56764193e-04
P-value of url count	5.90176684e-01	5.43458274e-02	4.10071035e-02
P-value of user mentioned count	2.24010416e-05	1.78040054e-01	1.67593806e-01
P-value of sad emoji count	8.15399438e-01	2.08444025e-06	6.57265747e-01
P-value of happy emoji count	2.98159433e-03	1.68343200e-02	9.36262681e-01
P-value of number of retweets	5.63805663e-01	6.26434516e-03	6.86485400e-03
P-value of sum of number of followers	9.92018335e-01	6.84552648e-01	9.62435976e-01
P-value of max followers	9.95237593e-01	2.69132233e-01	8.52979006e-03
P-value of time of the day	4.29762293e-01	9.94280756e-01	7.61506958e-02

Table 9. Mean Squared Error (MSE) and R-squared of superbowl tag model for all 3 periods

Question 7

Also, aggregate the data of all hashtags, and train 3 models (for the intervals mentioned above) to predict the number of tweets in the next hour on the aggregated data. Perform the same evaluations on your combined model and compare with models you trained for individual hashtags.

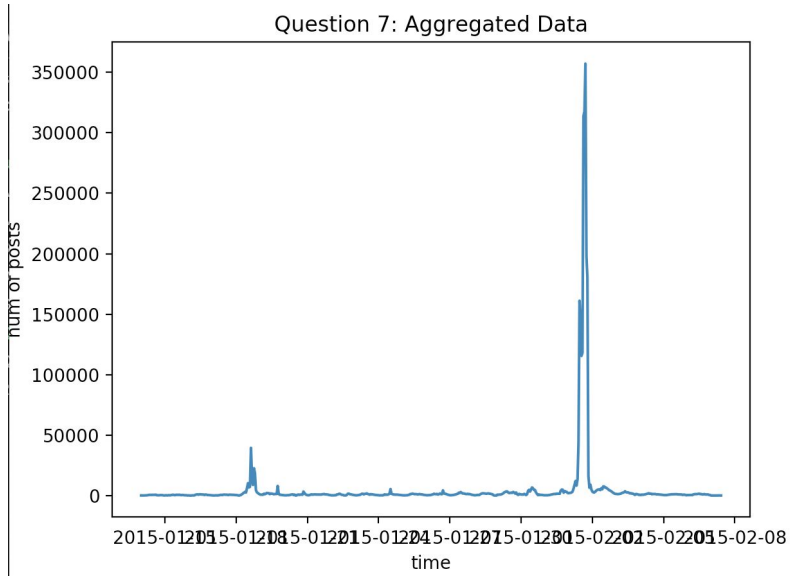


Figure 21. Number of tweets in each hour vs. time for all hashtags

Time Period	R2	MSE(Mean Squared Error)
Before Superbowl	0.5592890845076031	4137922.90
During Superbowl	0.9493834838422325	15363770.98
After Superbowl	0.9376224861600041	389601.33

Table 10. MSE and R-squared of the 3 models for 3 periods for the aggregated data

During Super Bowl, there is a spike in the number of tweets, which can result in very high R2 value. MSE is also the highest due to the fact the number of tweets changes a lot during the super bowl time.

For the first period, before superbowl, comparing the model with aggregated data with models with individual hashtag data, the model with aggregated data has lower R2 score than most of the models with individual hashtag data. Also, MSE is higher in the aggregated model.

For the second time period, during superbowl, the aggregated model has R2 score higher than most of the hashtags, and much higher than R2 of #gopatriots model and #gohawks model. It also has MSE much higher than MSE of models on all the hashtags.

For the third time period, after superbowl, the aggregated model has higher R2 score than most of the hashtags, and also has MSE much higher than MSE of models on all the hashtags.

Time Period	1st Significant Feature	2nd Significant Feature	3rd Significant Feature
Before Superbowl	Happy_emoji_cnt (p-value= 1.224538710691085 1e-05)	Sad_emoji_cnt (p-value = 0.000633678974600404)	User_mentioned_cnt (p-value = 0.0010297803385369878)
During Superbowl	Num_tweets (p-value = 2.987556627267579 7e-15)	Sad_emoji_cnt (p-value = 0.00012236459972569236)	Num_retweets (p-value = 0.028324334138776726)
After Superbowl	Happy_emoji_cnt (p-value = 2.792759313463979 8e-05)	Sad_emoji_cnt (p-value = 0.0008070372801387374)	Num_retweets (p-value = 0.019990806076962148)

Table 11. The top 3 features and p-value of each period for aggregated data

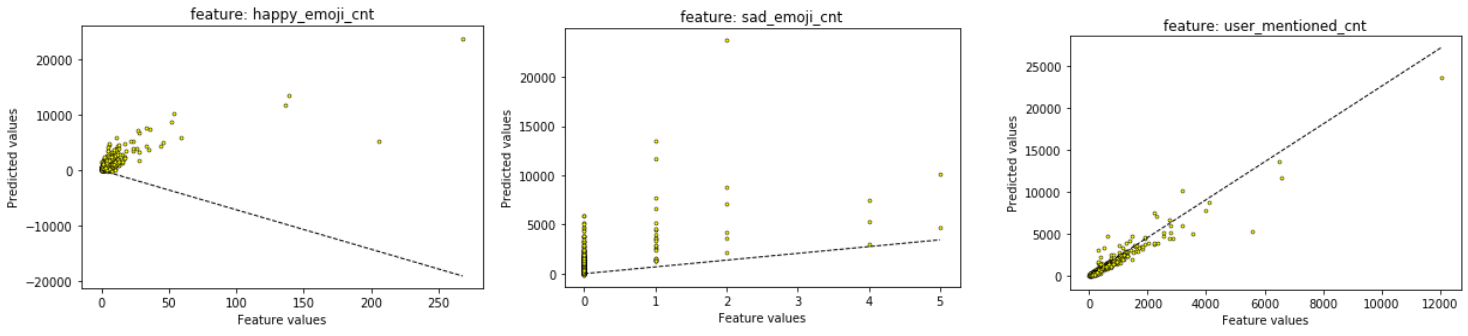


Figure 22. Scatter plot of the top 3 features of time before superbowl

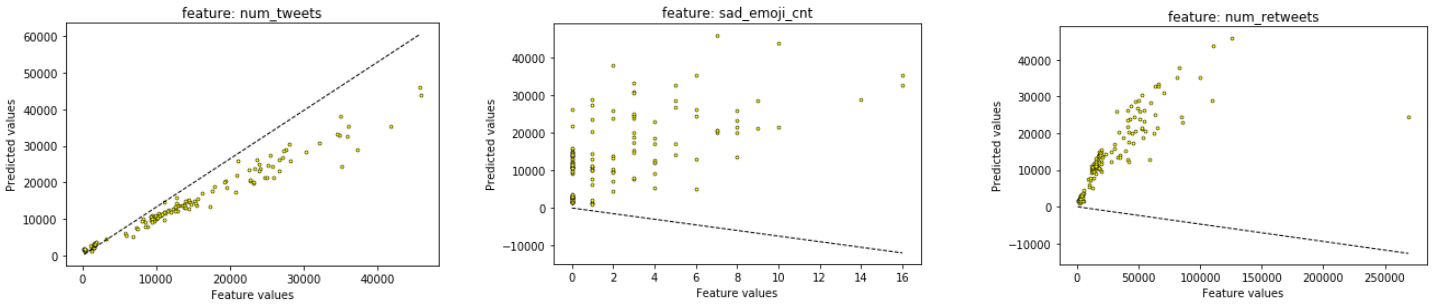


Figure 23. Scatter plot of the top 3 features of time during superbowl

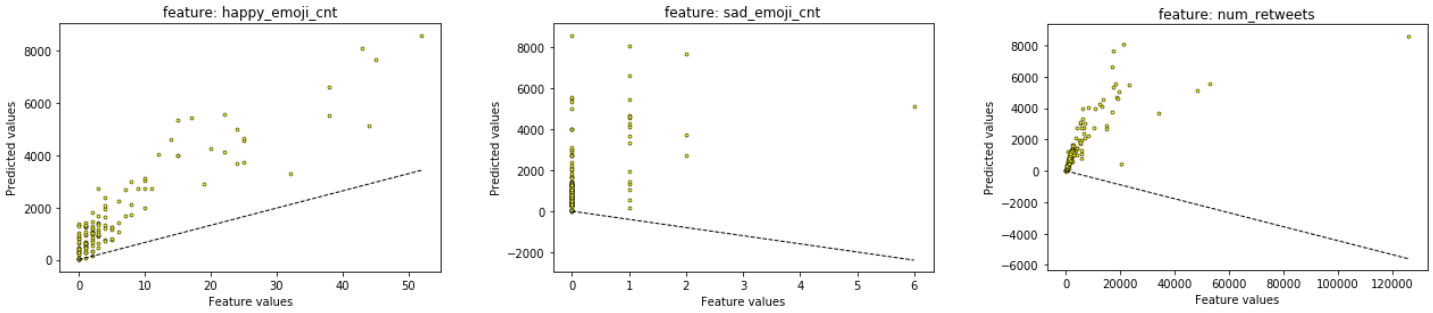


Figure 24. Scatter plot of the top 3 features of time after superbowl

From the plot of predicted value vs. feature values for the top 3 features of all three time periods, we find some of the coefficient agree with the trend of data points in the plot while some others not. The reason for it might be that the linear model doesn't fit well on the data divided into time periods (as the MSE value is high), and also some features are not so important in the linear model (as the p-value of some features is higher than 0.05, which means this feature is not very important in the linear model)

Question 8

Use grid search to find the best parameter set for RandomForestRegressor and GradientBoostingRegressor respectively. Use the following param_grid

Optimal Parameters for the Grid Search:

Types of classifiers	max features	n_estimators	min_samples_split	max_depth	min_samples_leaf
Random Forest	auto	200	2	None	1

Regressor					
Gradient Boost Regressor	sqrt	2000	10	80	4

Table 12. Optimal Parameters for the Grid Search for RandomForestRegressor and GradientBoostingRegressor

The test errors from cross-validation look good. Both the random forest regressor and the gradient boost regressor have a similar optimal cross-validation error. Below are the plots of the cross-validation errors of these two regressors. The x-axis corresponds to a particular iteration.

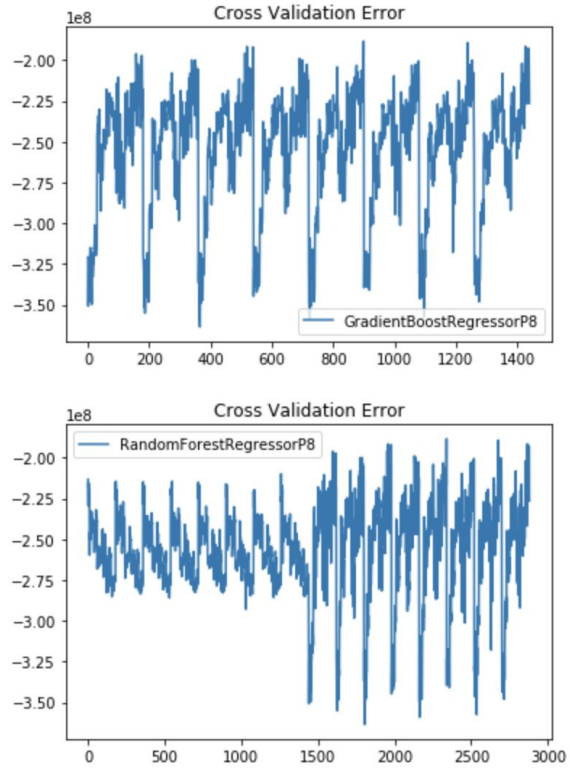


Figure 25. Cross validation error of random forest regressor and gradient boost regressor

Question 9

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

Type of Classifier	R2	MSE(Mean Squared Error)
Random Forest Regressor	0.96	32668118.25
Gradient Boost Regressor	1	0

OLS	0.856	118337780.3

Table 13. MSE and R-squared Comparison among models of Random Forest Regressor, Gradient Boost Regressor and OLS

The ensemble methods are better than the simple regression method. The gradient boost regressor achieves the perfect R^2 and MSE score, this is due to the fact training and testing are on the same set of data, and gradient boosting method tends to overfit.

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? Is the best parameter set you find in each period agree with those you found above?

Time Period	R2	MSE(Mean Squared Error)
Before Superbowl	1	0
During Superbowl	1	0
After Superbowl	1	6.80

Table 14. MSE and R-squared Comparison among GradientBoostingRegressor models of 3 periods, before superbowl, during superbowl, and after superbowl

Gradient boosting regressor tends to overfit the training data. Even though their R2 and MSE value are perfect. The cross-validation errors are different.

Time	Cross Validation Error
Before SuperBowl	-4953559
During SuperBowl	-24696085
After SuperBowl	-401288

Table 15. Cross validation error for GradientBoostingRegressor models of 3 periods, before superbowl, during superbowl, and after superbowl

The cross-validation after the super bowl is the lowest because the number of tweets during those two periods don't change very much. During the super bowl time, the number of tweets increased and decreased drastically, which results in larger cross-validation error. There is also some noise before the superbowl, so the cross validation error is the highest among the all three periods.

Time Period	max features	n_estimators	min_samples_split	max_depth	min_samples_leaf
Before Superbowl	sqrt	2000	10	80	1
During Superbowl	sqrt	1400	2	60	2
After Superbowl	auto	600	10	80	4
Aggregated Data	sqrt	2000	10	80	4

Table 16. Optimal parameters of GradientBoostingRegressor models of 3 periods, before superbowl, during superbowl, and after superbowl, and the aggregated data

The optimal parameters are different for each time period. The optimal parameters for "After Super Bowl" are the same the parameters for the aggregated data (Question 8). This makes sense because there is no change in the number of tweets most of the times (i.e. the aggregated data is not balanced). The number of tweets changes drastically only during a small time interval.

Question 11

Now try to regress the aggregated data with MLPRegressor. Try different architectures (i.e. the structure of the network) 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.

For the neural network we experimented with architectures as below. Since the input data size is not very large and there are only 9 features, we only tried architectures with at most 3 hidden layers and at most 50 neurons in each hidden layer. More experiments should be carried out if time permits.

Number of Neurons in First Hidden Layer	Number of Neurons in Second Hidden Layer	Number of Neurons in Third Hidden Layer	Mean Squared Error

5	N/A	N/A	4.25851950e+08 (Best Value)
6	N/A	N/A	9.10733996e+12
7	N/A	N/A	1.41093784e+09
8	N/A	N/A	2.28051061e+12
10	N/A	N/A	3.32433987e+13
20	N/A	N/A	1.24435722e+13
40	N/A	N/A	1.49941299e+13
50	N/A	N/A	1.52635133e+12
5	5	N/A	4.92155815e+13
5	6	N/A	8.41098297e+08
5	8	N/A	1.98773194e+13
5	10	N/A	1.69330523e+12
7	6	N/A	1.38742167e+10
7	8	N/A	1.46355156e+11
7	10	N/A	9.13752791e+08
7	20	N/A	6.89096012e+11
10	20	N/A	2.39689153e+12
10	40	N/A	8.68755400e+11
10	50	N/A	3.83809098e+12
20	40	N/A	8.93792109e+12
20	50	N/A	5.10923757e+12
40	50	N/A	1.95838725e+13

5	7	8	1.64190557e+13
5	8	10	4.32651027e+13
10	20	50	1.00275447e+11
20	40	50	3.49070922e+12

Table 17. Neural Network Architectures and Mean Squared Error for MLPRegressor on the aggregated data

From the table we find the best neural network architecture for fitting the aggregated data is one hidden layer with 5 neurons. The best Mean Squared Error is 425851950.13874

Question 12

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

Number of Neurons in First Hidden Layer	Number of Neurons in Second Hidden Layer	Number of Neurons in Third Hidden Layer	Mean Squared Error
5	N/A	N/A	2.40000859e+08
6	N/A	N/A	2.32474100e+08
7	N/A	N/A	2.41683765e+08
8	N/A	N/A	2.32949109e+08
10	N/A	N/A	2.36355757e+08
20	N/A	N/A	2.43370596e+08
40	N/A	N/A	2.37255421e+08
50	N/A	N/A	2.32231407e+08
5	5	N/A	2.35393686e+08
5	6	N/A	2.29907320e+08

5	8	N/A	2.33060694e+08
5	10	N/A	2.30415882e+08
7	6	N/A	2.32591684e+08
7	8	N/A	2.34731876e+08
7	10	N/A	2.38325046e+08
7	20	N/A	2.33651884e+08
10	20	N/A	2.54922542e+08
10	40	N/A	2.43585902e+08
10	50	N/A	2.27748669e+08
20	40	N/A	2.18399779e+08
20	50	N/A	2.25450987e+08
40	50	N/A	2.15162288e+08
5	7	8	2.35030355e+08
5	8	10	2.07600314e+08 (Best Value)
10	20	50	2.61876684e+08
20	40	50	2.19137093e+08

Table 18. Neural Network Architectures and Mean Squared Error for MLPRegressor on the aggregated data after standardization

From the table we find the best neural network architecture for fitting the aggregated data (with standardization) is three hidden layers with 5, 8, 10 neurons in each hidden layer. The best Mean Squared Error is 207600313.50296134. The performance has significantly increased after standardizing the data as the Mean Squared Error decreased.

Question 13

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

Period 1: Before Feb. 1, 8:00 a.m.: 1-hour window

Number of Neurons in First Hidden Layer	Number of Neurons in Second Hidden Layer	Number of Neurons in Third Hidden Layer	Mean Squared Error
5	N/A	N/A	4970872.68444338
6	N/A	N/A	5072705.01291805
7	N/A	N/A	4845616.51357766
8	N/A	N/A	5119032.55087788
10	N/A	N/A	5077172.20411519
20	N/A	N/A	5192048.34954916
40	N/A	N/A	5453524.16220057
50	N/A	N/A	5057915.96159057
5	5	N/A	4833482.11201255 (Best Value)
5	6	N/A	4868442.54789818
5	8	N/A	5321782.63824968
5	10	N/A	5013903.81086362
7	6	N/A	5256249.58811243
7	8	N/A	4884026.43910752
7	10	N/A	5048920.22274154

7	20	N/A	4904169.14717082
10	20	N/A	7408564.90611521
10	40	N/A	5032470.23319787
10	50	N/A	6607753.89589045
20	40	N/A	4900709.52617063
20	50	N/A	6346706.91280986
40	50	N/A	5167133.86634078
5	7	8	5092734.9682249
5	8	10	5050612.96588237
10	20	50	5177224.37224442
20	40	50	5167850.89439394

Table 19. Neural Network Architectures and Mean Squared Error for MLPRegressor on the 1st period of aggregated data after standardization

From the table we find the best neural network architecture for fitting the aggregated data (with standardization) is two hidden layers with 5, 5 neurons in each hidden layer. The best Mean Squared Error is 4833482.112012554

Period 2: Between Feb. 1, 8:00 a.m. and 8:00 p.m.: 5-minute window

Number of Neurons in First Hidden Layer	Number of Neurons in Second Hidden Layer	Number of Neurons in Third Hidden Layer	Mean Squared Error
5	N/A	N/A	2.75232032e+07
6	N/A	N/A	2.70112850e+07
7	N/A	N/A	2.58277012e+07
8	N/A	N/A	3.31880198e+07

10	N/A	N/A	2.28922377e+07
20	N/A	N/A	4.66904329e+07
40	N/A	N/A	7.85584160e+07
50	N/A	N/A	1.45513550e+08
5	5	N/A	4.15198439e+07
5	6	N/A	2.54795920e+07
5	8	N/A	2.94116560e+07
5	10	N/A	2.68271700e+07
7	6	N/A	2.80636743e+07
7	8	N/A	2.25750987e+07 (Best Value)
7	10	N/A	3.19055947e+07
7	20	N/A	3.13404047e+07
10	20	N/A	3.18102954e+07
10	40	N/A	3.36575301e+07
10	50	N/A	4.19778328e+07
20	40	N/A	4.10268743e+07
20	50	N/A	4.05757062e+07
40	50	N/A	4.15812613e+07
5	7	8	3.21445178e+07
5	8	10	2.59434703e+07
10	20	50	6.56831820e+07
20	40	50	6.54598584e+07

Table 20. Neural Network Architectures and Mean Squared Error for MLPRegressor on the 2nd period of aggregated data after standardization

From the table we find the best neural network architecture for fitting the aggregated data (with standardization) is two hidden layers with 7, 8 neurons in each hidden layer. The best Mean Squared Error is 22575098.69910023

Period 3: After Feb. 1, 8:00 p.m.: 1-hour window

Number of Neurons in First Hidden Layer	Number of Neurons in Second Hidden Layer	Number of Neurons in Third Hidden Layer	Mean Squared Error
5	N/A	N/A	2088250.03372647
6	N/A	N/A	2031554.25383256
7	N/A	N/A	1554066.35735982
8	N/A	N/A	2362468.42625443
10	N/A	N/A	1882676.3000007
20	N/A	N/A	1692744.38175293
40	N/A	N/A	973297.30589679
50	N/A	N/A	1541992.12444708
5	5	N/A	1733754.99312992
5	6	N/A	2135059.03867981
5	8	N/A	1693714.7195382
5	10	N/A	1984802.38637881
7	6	N/A	2189950.12782359
7	8	N/A	2096848.48120062
7	10	N/A	1974696.18036853

7	20	N/A	1239628.72785991
10	20	N/A	2074859.67583394
10	40	N/A	2055255.33265866
10	50	N/A	812411.021172 (Best Value)
20	40	N/A	1436292.9054873
20	50	N/A	1900205.43970267
40	50	N/A	1685755.72079013
5	7	8	1894577.89980018
5	8	10	2044836.14196496
10	20	50	2259447.80566032
20	40	50	2102983.84536623

Table 21. Neural Network Architectures and Mean Squared Error for MLPRegressor on the 3rd period of aggregated data after standardization

From the table we find the best neural network architecture for fitting the aggregated data (with standardization) is two hidden layers with 10, 50 neurons in each hidden layer. The best Mean Squared Error is 812411.0211719986

Question 14

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

Training Data Preparation

Since in the question, the test data are the tweets collected in a 6-hour (or 6x5 = 30 minutes) window, we are going to use the data in 6x window to predict the number of tweets in the next window (next hour or next 5 minutes). Therefore, in training data preparation stage, for each timestamp (every hour or every 5 minute from 8 am to 8 pm on 02/01/2015), we record the

features of tweets in a 6x window before it as the training features. Then we record the tweet numbers in the next time window as the training labels.

Model Selection

We choose Random Forest Model for this question as the Random Forest Model gives the best overall performance in above question. We train the Random Forest Model on the entire aggregated data with the data preprocessing methods above. The Mean Squared Error for the training is 1001680.98 and R2 score is 0.98.

Prediction

The predicted results of tweet numbers in next window is as in table below.

Test time period	Predicted number of tweets in next window (rounded to integer)
sample0_period1	581
sample0_period2	2015
sample0_period3	81
sample1_period1	861
sample1_period2	1919
sample1_period3	307
sample2_period1	277
sample2_period2	82
sample2_period3	406

Table 22. Predicted number of tweets in next window using data from a 6x window using Random Forest Regressor

Question 15

1. Explain the method you use to determine whether the location is in Washington, Massachusetts or neither. Only use the tweets whose authors belong to either Washington or Massachusetts for the next part.

2. 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 (using the techniques you learnt in project 1). Try different classification algorithms (at least 3). For each, plot ROC curve, report confusion matrix, and calculate accuracy, recall and precision.

Location Determination

To determine the location of the user of each tweet, we check its "location" attribute. We try to determine the location from either the city name or the state name. In particular, in order to be classified as "Massachusetts", the location attribute has to contain either the string "Boston" (patriots team's city) or the string "Massachusetts", "MA". Similarly, in order to be classified as "Washington", the location attribute has to contain either the string "Seattle"/ "Kirkland", or the string "Washington"/"WA". Additionally, the string should NOT contain "DC"/"D.C." because locations with that certainly refers to the capital, not the Washington state (e.g. "Washington, D.C.").

Binary Classification

Meta Data

After applying such criteria to all tweets in #superbowl, 38124 tweets were selected, 18337 of which are from Washington, the rest 19787 are from Massachusetts. So the two classes are well balanced.

Data Preprocessing

The next step is to tokenize all tweets with two filters. The first is English stopwords. The second is keeping only words with letters and numbers i.e. no special characters allowed, with only one exception, hashtag. Keeping hashtag would have a huge benefit in the following classification task, improving the accuracy by more than 10%. One possible reason for such significant improvement may be that hashtag is usually a good summary of the tweet, capable of representing the user inclination very well. After all tweets are converted to word vectors, LSI is performed to extract high level implicit topics. Here the number of components is 100.

Training Method

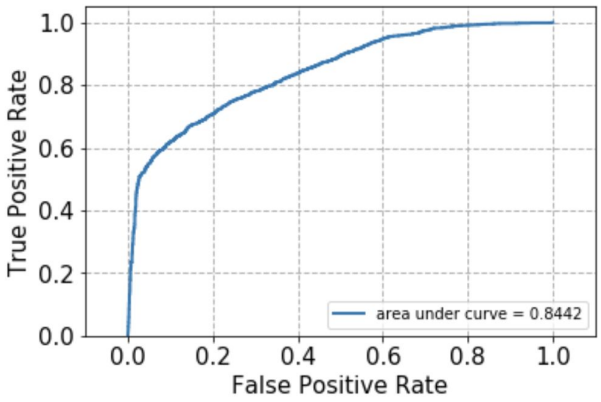
We split the data into two parts. 90% becomes training data and 10% becomes testing data.

1. The first algorithm is linear SVM with high loss, C = 0.025

SVM, C= 0.025	
accuracy	0.7416732231838448

precision	0.94655704008222
recall	0.4967637540453074
F1 score	0.6515741068270251

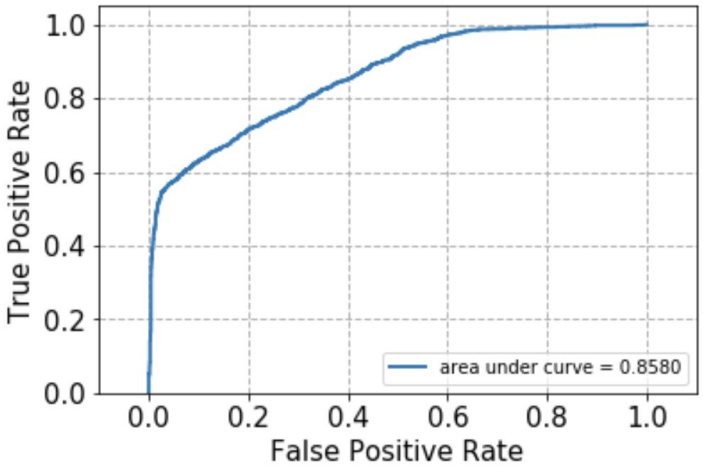
SVM, C= 0.025	Predicted Massachusetts	Predicted Washington
Actual Massachusetts	1907	52
Actual Washington	933	921



2. The second model is Logistic Regression with l2 norm and C = 10

Logistic Regression C=10	
accuracy	0.7684238132703908
precision	0.8963265306122449
recall	0.5922330097087378
F1 score	0.7132185774602143

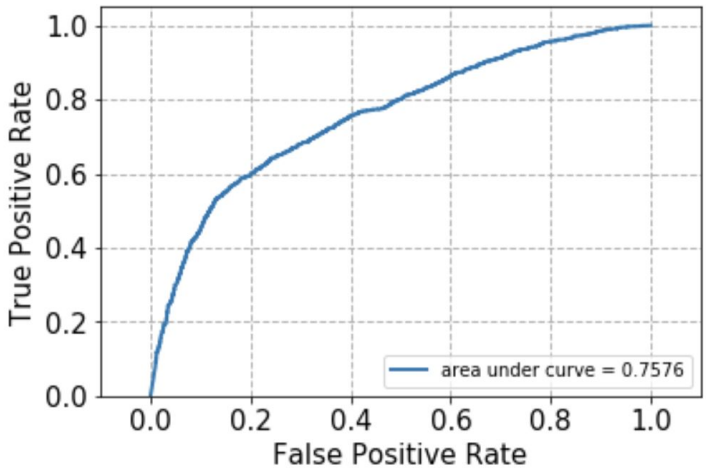
Logistic Regression C=10	Predicted Massachusetts	Predicted Washington
Actual Massachusetts	1832	127
Actual Washington	756	1098



3. The third model is Gaussian Naive Bayes

Gaussian Naive Bayes	
accuracy	0.7060057697351168
precision	0.7631012203876526
recall	0.5733549083063646
F1 score	0.6547582383738836

Gaussian Naive Bayes	Predicted Massachusetts	Predicted Washington
Actual Massachusetts	1629	330
Actual Washington	791	1063



Among the three models, **logistic regression performs the best,** with SVM being a close second and GNB is notably worse.

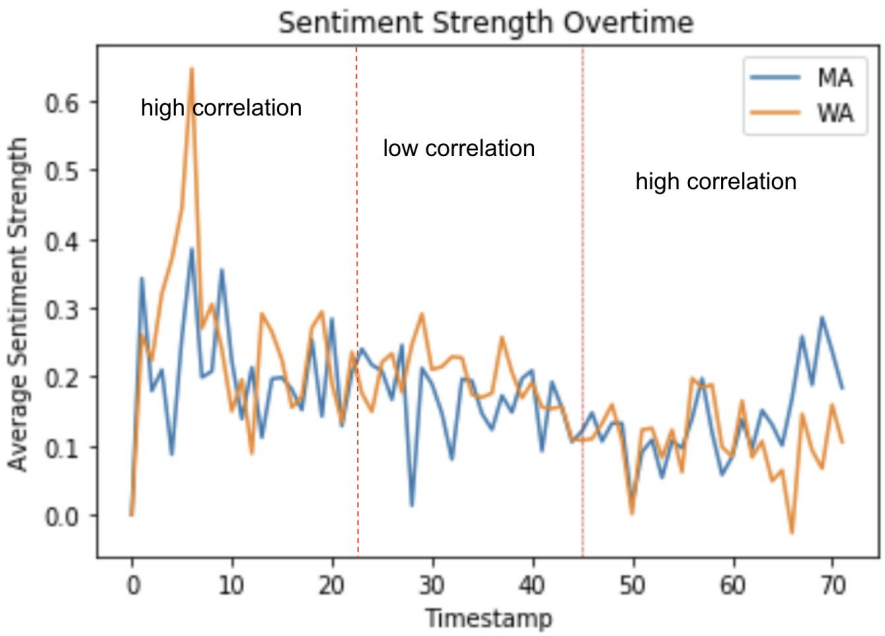
Question 16

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.

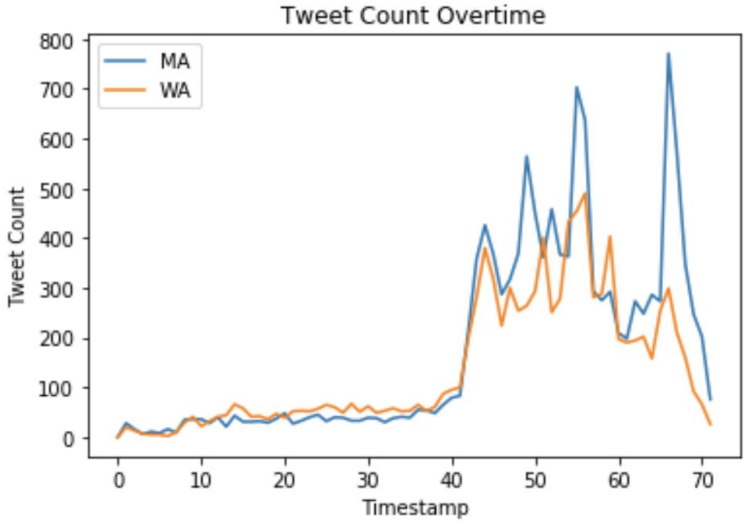
In this part, we would like to investigate how the sentiment of tweets sent from both Massachusetts and Washington changes over time. We assume that tweets from Massachusetts are good representation of patriots' fans and those from Washington are good representation of hawk's fans. We would also like to see what could be inferred from that.

We focus on the time between 8am to 8pm on the day of superbowl as that is the time when the game happened. We separated the tweets into 10-minute intervals. To understand the sentiment of each tweet, we used the VADER sentiment analysis tool which was a parsimonious rule-based model for sentiment analysis of social media text. It provides a method that could take a tweet as input and output its sentiment strength score, positive values are positive valence, negative values are negative valence. The average sentiment strength of tweets over every 10-minute period is calculated and plotted.

As we can see from the plot, before the game started (t=45, which is 3:30pm), the sentiment score of both teams are about the same, with hawks being slightly higher. After the game started, both scores remained very close up to about t=62. Afterwards, there was a sharp drop in sentiment strength in hawk's fans and the gap remained until the end. This is a good reflection of the fact that Patriots ended up winning the game.



Another interesting observation from the plot is that in most of the time, the sentiment strength of the two teams are highly correlated. Most of the ups and downs happened to both of them at the same time. The only two exception is between t=22 to 45 where patriots suffered from a significant drop and t=60 to 65 where hawks suffered because of the loss.



The above is a plot of the total number of tweets in each interval. As we can see, there are several peaks in the graph. The first happens at t=45, which is when the game started. The last happened at t = 65 which coincided with when there was a big drop in hawk fans' sentiment strength. We can inferred that it was when the hawk lost the game. Such observation leads to

another aspect of studying tweets, looking at tweet peaks because often time that is a good indicator that something worth noting has happened.