Project 4 Popularity Prediction on Twitter

Huang xinxin 104589081

Lu qiujing 704617222

Niu longjia 304590762

Problem 1:

(1) Calculate these statistics for each hashtag:

average number of tweets per hour

average number of followers of users posting the tweets

average number of retweets

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

Solution:

Note:

The definition of average number of followers of users posting the tweets is not clear, we include two possible interpretations in our table. The first interpretation is expressed as followers of users posting the tweets/per user, and second interpretation is expressed as followers of users posting the tweets/per unique user.

Table 1-1. Statistics information for each hashtag (average number)

Hashtags\labels	Tweets/hour	followers of users posting the tweets/per unique user	followers of users posting the tweets/per user	Retweets /per user
#gohawks	272.50	1724.91	2180.76	2.04
#gopatriots	49.90	1100.23	1272.57	1.38
#nfl	463.24	4510.41	4792.42	1.54
#patriots	934.60	1646.80	3209.21	1.75
#sb49	1639.63	2231.62	10268.89	2.51
#superbowl	2660.75	3590.97	8872.25	2.39

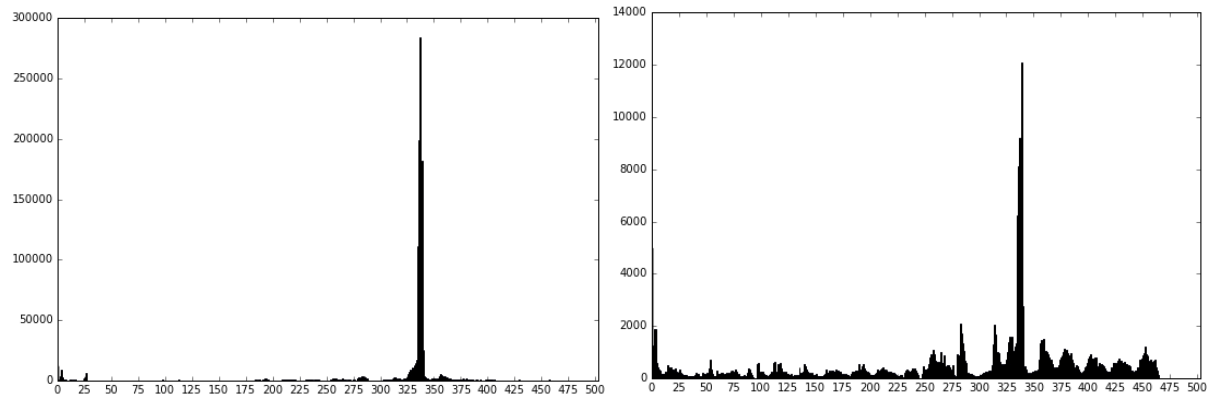


Fig 1-1. Number of tweets in hour over time for #superbowl (left) and #nfl (right)

Problem 2:

(1) Fit a Linear Regression model using 5 features to predict number of tweets in the next hour:

X1: number of tweets

X2: total number of retweets

X3: sum of the number of followers of the users posting the hashtag

X4: maximum number of followers of the users posting the hashtag

X5: time of the day

(2) Explain your model's training accuracy and the significance of each feature using the t-test and P-value results of fitting the model.

Solution：

Table 2-1. R squared value of linear regression model for different hashtags

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
R^2	0.665	0.593	0.721	0.720	0.843	0.768

Table 2-2. Significance of different features (x1~x5 indicate features mentioned in problem 2) in different hashtags

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
x1	0.1682	-1.3334	0.1787	1.8487	1.2272	1.1806
x2	-0.1463	3.2036	-0.0871	-0.8281	-0.2284	0.5485
x3	0.0004	-0.0018	0.0002	8.006e-05	1.781e-05	-0.0003
x4	-0.0005	0.0010	-0.0003	3.4e-05	0.0002	0.0016
x5	7.9487	-0.6286	10.3542	18.5189	-18.1702	-118.4291

Table 2-2 indicates the significance of each feature from x1~x5. Obviously, the values with larger absolute value may have larger effects on the model. We can conclude that :

For **#gohawks:**

x5 (time of the day) contributes the most, x1(number of tweets) and x2(total number of retweets) contribute a little;

For **#gopatriots:**

x2(total number of retweets) contribute the most, x1(number of tweets) and x5(time of the day) contribute a little;

For **#nfl:**

x5(time of the day) contributes the most, x1(number of tweets) and x2(total number of retweets) contribute a little;

For **#patriots:**

x5(time of the day) contributes the most, x1(number of tweets) and x2(total number of retweets) contribute a little;

For **#sb49:**

x5(time of the day) contributes the most, x1(number of tweets) contributes a little and x2(total number of retweets) contribute a little;

For **#superbowl**:

x5(time of the day) contributes the most, x1(number of tweets) and x2(total number of retweets) contribute a little.

Overall, x5(time of the day) contributes most to almost each of the hashtag, x1(number of tweets) and x2(total number of retweets) have minor contributions to the model.

Table 2-3. P-value of different features in different hashtags

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
x1	0.187	0.000	0.231	0.000	0.000	0.007
x2	0.000	0.000	0.260	0.000	0.026	0.016
x3	0.000	0.000	0.000	0.034	0.268	0.000
x4	0.000	0.000	0.000	0.760	0.000	0.000
x5	0.000	0.411	0.000	0.053	0.261	0.002

P-value, expresses the results of the hypothesis test as a significance level. Conventionally, P-values smaller than 0.05 are taken as evidence that the population coefficient is nonzero.

For **#gohawks**, x2~x5 have better training accuracy than x1.

For **#gopatriots**, x1~x4 have better training accuracy than x5.

For **#nfl**, x3~x5 have better training accuracy than x1 and x2.

For **#patriots**, x1 and x2 have better training accuracy than the rest.

For **#sb49**, x1 and x4 have better training accuracy than the rest.

For **#superbowl**, x3 and x4 have better training accuracy than the rest.

Problem 3:

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 and report fitting accuracy and significance of variables. For the top 3 features in

your measurements, draw a scatter plot of predictant (number of tweets for next hour) versus feature value, using all the samples you have extracted and analyze it.

Solution:

Besides the features in last two problems, we extract 3 new features in this problem according to the paper [1] and also our observation. So in our model, the total input for each sample is 8. These features are:

- Feature in problem 2：

X1：number of tweets,

X2: total number of retweets, counting by ["metrics"]["citations"]["total"]

X3: sum of the number of followers of the users' posting the hashtag,

X4: maximum number of followers of the users' posting the hashtag,

X5: time of the day

- Features in this part:

X1：number of tweets,

X2: total number of retweets counting by ["tweet"]["retweet_count"],

X3: sum of the number of followers of the users' posting the hashtag,

X4: maximum number of followers of the users' posting the hashtag,

X5: time of the day

X6: number of celebrity

We suppose that if a user's followers number is larger than a threshold, we consider the user as a celebrity. We first test the different numbers of the threshold, as 100,1000 10000 and 100000. According to our experiment, the threshold of defining the celebrity will influence the importance of other attributes, especially author count. For dataset gopatriots, when the threshold is larger than 1000, the p-value of author becomes larger than 0. According to our carefully test, when the threshold equals to 1000, the models has the smallest AIC, which implies the model perform best. So we choose 1000 as threshold.

X7: author count: total number of unique users post tweets in one hour.

X8: mention count : total number of mention per hour(done)

In this part we mainly use R-square as a judgment of our models' performance.

To save time, we use the gopatriots set for testing. We test the model performance of different kinds of input. And the results are shown below:

Table 3-1 For gopatriots, model performance of different input

Input (number of input)	R	AIC	Top 3 important features
X1-X5 (5)	0.653	6782	X2,X1,X4
X1-X5,X6 (6)	0.671	6756	X2,X6,X1

X1-X5,X7 (6)	0.678	6746	X2,X7,X1
X1-X5,X8 (6)	0.689	6728	X2,X8,X1
X1-X5,X6,X7 (7)	0.715	6686	X7,X1,X6
X1-X5,X6,X8 (7)	0.698	6715	X2,X8,X6
X1-X5,X7,X8 (7)	0.690	6757	X2,X8,X3
X1-X8 (8)	0.715	6688	X6,X1,X7

As we can see from the table above, when we use X1-X7 as input and X1-X8 as input, the models performs best. Most important features now are X1, X6, and X7. So for other dataset, it is reliable to assume that the 8 inputs perform best.

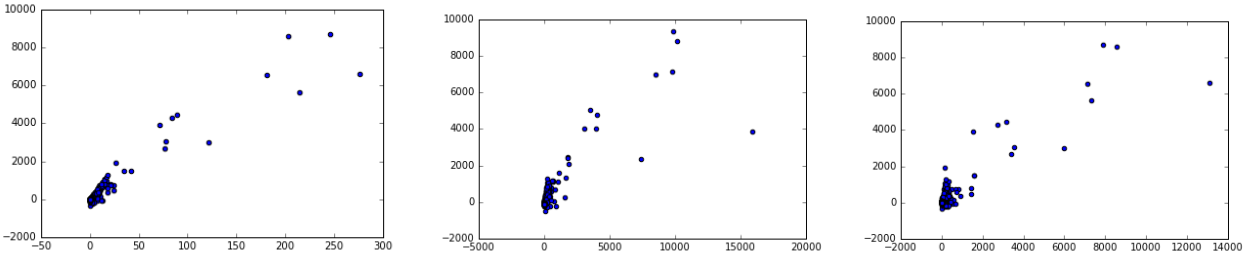
Table 3-2. Performance and top 3 features for different models and datasets

Data Set	R-square in part 2	R-square with 8 input	Top3 features (8 inputs)
gohawks	0.665	0.878	X6,X1,X7
gopatriots	0.693	0.715	X6,X1,X7
nfl	0.721	0.805(X1-X6,X8)	X2,X8,X3
patriots	0.720	0.792	X8,X6,X1
sb49	0.843	0.915	X8,X1,X6
superbowl	0.768	0.899	X7,X3,X1

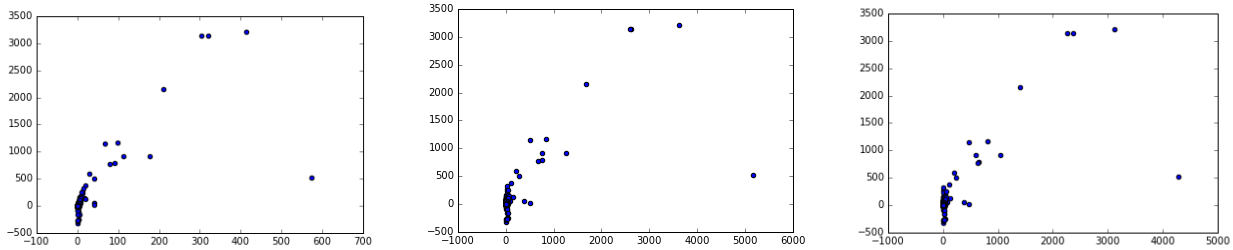
According to our experiment, with 8 inputs for nfl, the R-square becomes to 0.777. The performance becomes worse. So we implement different kinds of inputs and find out that when we use X1-X6 and X8 as inputs, the R-square rises to 0.805, which is the best of all feature combinations. So we use this one instead in the table. As we can see, the performance of linear model become better for all data sets after adding our new features. Feature X6,X7,X2 and X8, which are all new features, are always chosen as top 3 important features for different datasets.

That means our new features play a significant role in improving the performance.

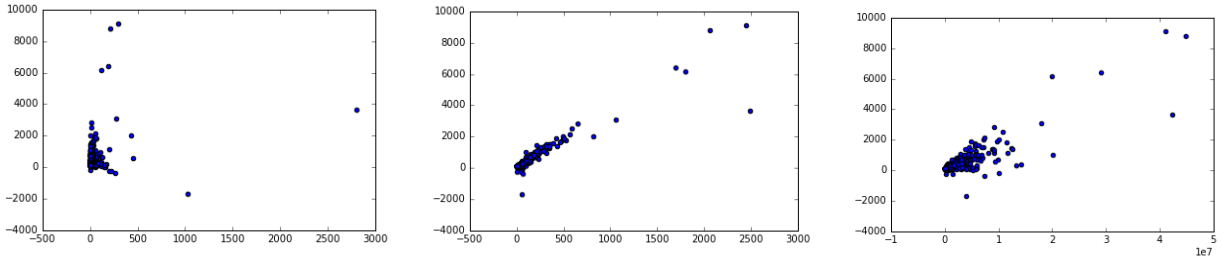
We plot the scatters of top three features of all the datasets as follows.



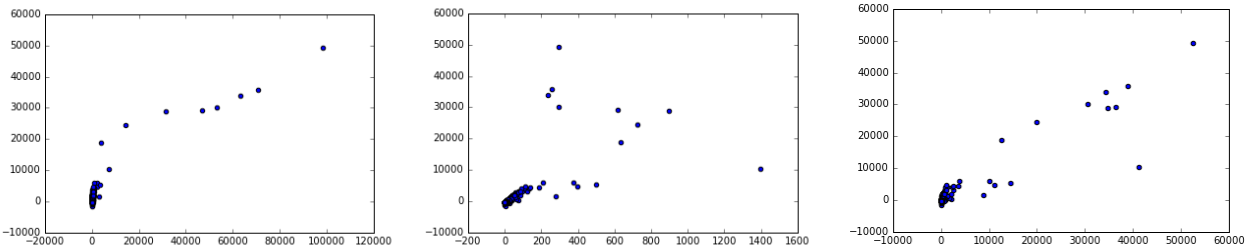
#gohawks, Feature: X6, X1, X7



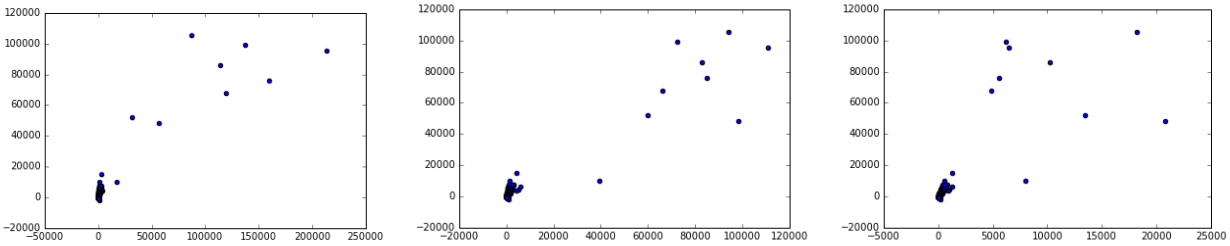
gopatriot Feature:X6 X1 X7



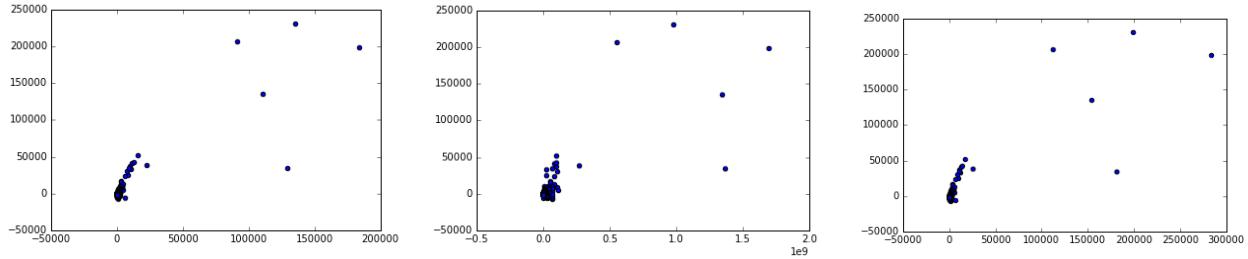
#NFL Feature:X2 X8 X3



#Patriot Feature: X8, X6 X1



SB49 Feature: X8,X1 X6



SuperBowl Feature X7 X3 X1

Fig 3-1. scatter plot of predictant (number of tweets for next hour) versus feature value

In Fig 3-1, predictant is positively correlated to the value of feature, especially the feature X8 in #NFL.. We also plot such figure of unimportant feature. The predictant isn't such strong positive relation with the feature value, such as the figure below. Since we use the linear regression model. We can roughly conclude that the important features have larger coefficients, which means changing of important features will significantly influence the output.

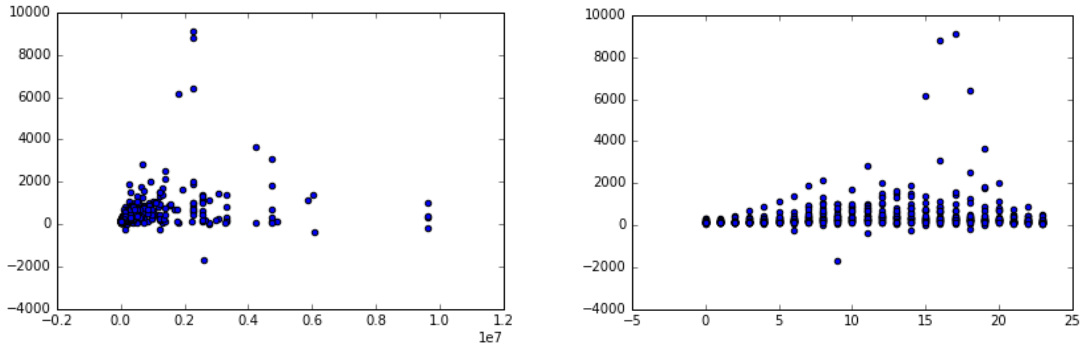


Fig 3-2 #gopatriot unimportant Feature:X4,X5

Besides the features above, we can also come up with other features which may be useful. First

is user passivity which is defined by

$$Psv(u_i) = \frac{N_d(u_i)}{1.0 + N_t(u_i)}$$

Where $Nd(ui)$ is the total number of tweets posted by the user.$Nt(ui)$ is the days since the user is registered. This feature can reflect the user's active level, which may influence the number of tweets in next hour. Because the limitation of time, we didn't test the performance after adding this feature.

Problem 4:

Create different regression models for the below time periods:

(1). Before Feb. 1, 8:00 a.m.

(2). Between Feb. 1, 8:00 a.m. and 8:00 p.m.

(3). After Feb. 1, 8:00 p.m.

Solution:

Note: For simplicity, the meanings of the abbreviations in the table are listed below:

lr=linear regression; rfr= random forest regression; poly=polynomial regression; lasso=lasso regression; ridge=ridge regression; svr=svm regression.

In Table 4-1, Table 4-2 and Table 4-3, row labels indicate different hashtag name, column labels indicate regression model names and the values in the table show the average prediction errors of the 10-fold cross validation. We will show results and analysis for each of the three time periods below.

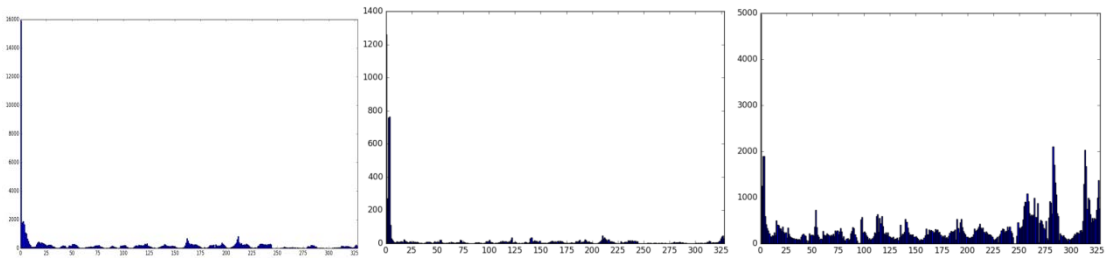
(1). Before Feb. 1, 8:00 a.m.

Table 4-1 Average prediction errors of different models in different hashtags before Feb. 1, 8:00 a.m.

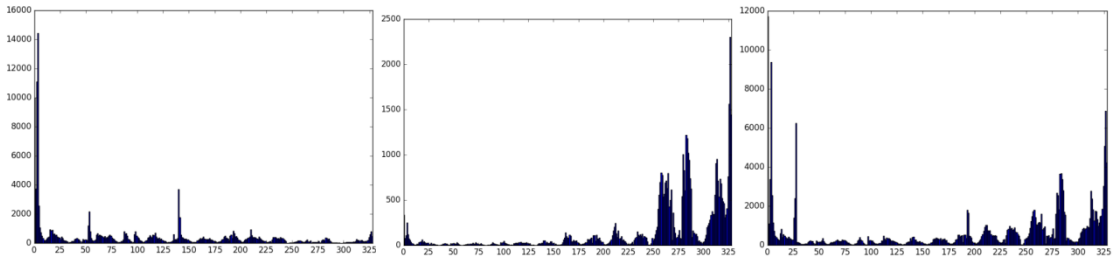
File\Avg_error\Model	lr	rfr	poly(deg=2)	lasso	ridge	svr
#gohawks	92	45	1700	92	90	116
#gopatriots	24	10	127	22	18	11
#nfl	128	111	257	127	121	182
#patriots	311	222	590	302	265	267
#sb49	59	54	146	58	57	125
#superbowl	321	272	1014	319	303	456

The models with the lowest average prediction errors for each specific hashtag are highlighted in the table above. As we can see, random forest regression achieves much better results in every hashtag, with the lowest average error of 10 in #gopatriots hashtag and highest average error of 272 in #superbowl hashtag.

The number of tweets versus hours histograms is depicted in Fig 4-1. With 329 hours covering the time period, we have a relative larger amount of data to train, which is responsible for the relative lower average prediction errors than the other two time periods.



#gohawks #gopatriots #nfl



#patriots #sb49 #superbowl

Fig 4-1 Number of different hashtag tweets per hour before Feb. 1, 8:00 a.m.

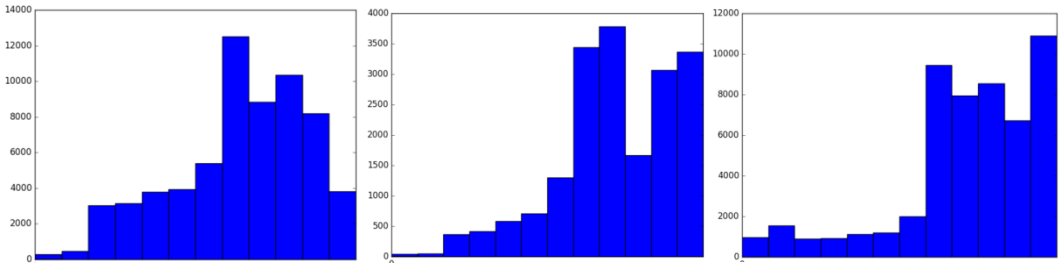
(2) Between Feb. 1, 8:00 a.m. and 8:00 p.m.

Table 4-2 Average prediction errors of different models in different hashtags between Feb. 1, 8:00 a.m. and 8:00 p.m.

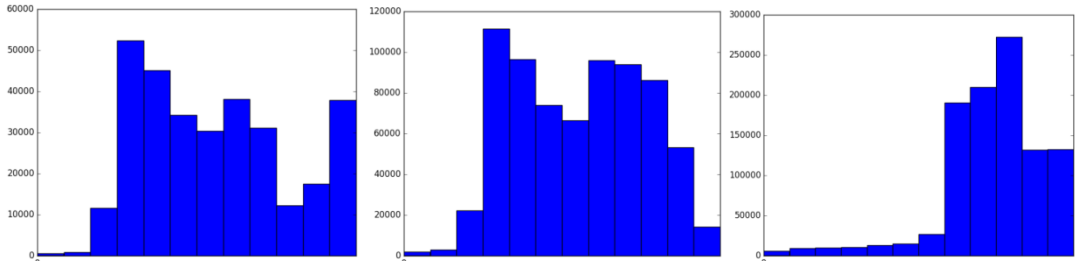
File\Model	lr	rfr	poly(deg=2)	lasso	ridge	svr
#gohawks	6808	2278	4294	1879	3635	3512
#gopatriots	6964	748	1679	2211	1492	1496
#nfl	36319	1927	6694	5881	4166	5203
#patriots	43889	12409	15562	11066	7545	13109
#sb49	601980	30175	171107	120157	88756	34030
#superbowl	1039579	50604	799483	282486	120242	121524

The models with the lowest average prediction errors for each specific hashtag are highlighted in the table above. As we can see, random forest regression achieves better results in 4/6 hashtags, lasso regression achieves better in 1/6 hashtag and ridge regression achieves better in 1/6 hashtag. The average prediction errors range from 748 in #gopatriots hashtag to 50604 in #superbowl hashtag.

The number of tweets versus hours histograms is depicted in Fig 4-2. With 12 hours covering the time period, we have a relative smaller amount of data to train, which is responsible for the obviously larger average prediction errors than the other two time periods.



#gohawks #gopatriots #nfl



#patriots #sb49 #superbowl

Fig 4-2 Number of different hashtag tweets per hour between Feb. 1, 8:00 a.m. and 8:00 p.m.

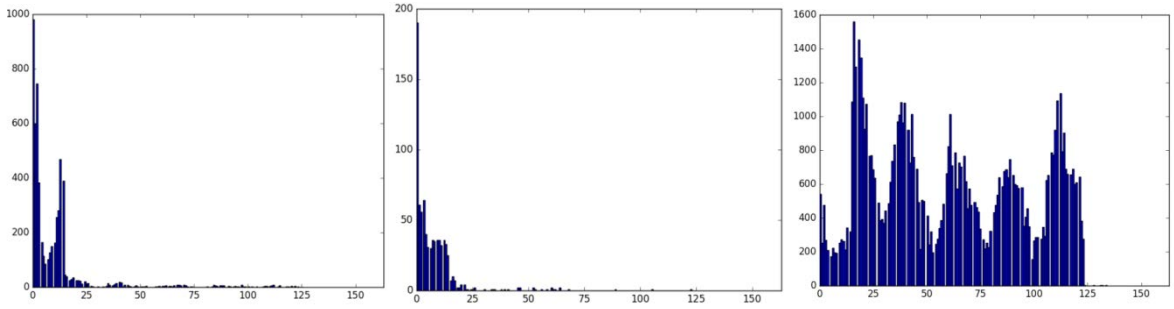
(3) After Feb. 1, 8:00 p.m.

Table 4-3 Average prediction errors of different models in different hashtags after Feb. 1, 8:00 p.m.

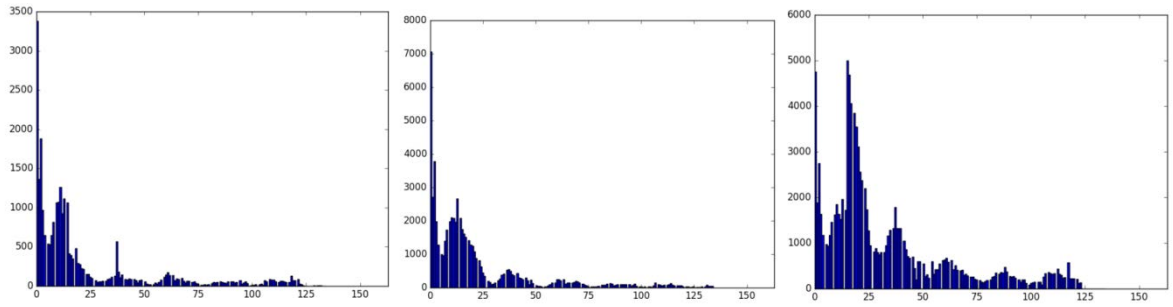
File\Model	lr	rfr	poly	lasso	ridge	svr
#gohawks	189	14	6379474	156	87	28
#gopatriots	2.731	1.225	931.584	2.289	2.439	3.779
#nfl	97.084	93.957	186.545	96.875	96.492	300.216
#patriots	82	45	7969	75	66	128
#sb49	140	76	8162	144	165	304
#superbowl	156	153	7950	154	149	523

The models with the lowest average prediction errors for each specific hashtag are highlighted in the table above. As we can see, random forest regression achieves better results in all 6 hashtags with average prediction errors ranging from 1.225 in #gopatriots hashtag to 153 in #superbowl hashtag.

The number of tweets versus hours histograms are depicted in Fig 4-2. With 164 hours covering the time period, we have a relative smaller amount of data to train than time period 1, but get better prediction results. This is possibly due to more obvious declining patterns than fluctuations patterns in period 1.



#gohawks #gopatriots #nfl



#patriots #sb49 #superbowl

Fig 4-3 Number of different hashtag tweets per hour after Feb. 1, 8:00 p.m.

Analysis of linear regression model in problem 4: Why linear regression model is not suitable for the period 1?

Take #gopatriots for example. For period 1, if we input the 5 features into the OLS model, then we can observe the low coefficients of feature 3 and 4, as well as low t value and high p value, which means these features are not so important in our current model. In addition, using the 10 folds cross validation in training linear regression model, we can both get very low score, 0.08, and the coefficients for the linear regression model are all zero. Furthermore, the random forest regression model also perform badly in this situation with score -0.10, showing that we have not picked the decisive features.

So we try to explain why feature 3 and 4 won't make a big difference on the prediction in the linear regression model , and what other features may have better performance for prediction in the period 1.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.520
Model:	OLS	Adj. R-squared:	0.513
Method:	Least Squares	F-statistic:	70.10
Date:	Thu, 17 Mar 2016	Prob (F-statistic):	1.71e-49
Time:	12:26:02	Log-Likelihood:	-1701.2
No. Observations:	328	AIC:	3412.
Df Residuals:	323	BIC:	3431.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	1.4113	0.316	4.466	0.000	0.790	2.033
x2	-0.5935	0.110	-5.390	0.000	-0.810	-0.377
x3	-5.039e-05	0.000	-0.203	0.839	-0.001	0.000
x4	-0.0001	0.000	-0.555	0.579	-0.001	0.000
x5	0.6536	0.200	3.271	0.001	0.260	1.047

Omnibus:	526.366	Durbin-Watson:	1.865
Prob(Omnibus):	0.000	Jarque-Bera (JB):	176046.654
Skew:	8.413	Prob(JB):	0.00
Kurtosis:	115.242	Cond. No.	2.38e+04

Fig 4-4. OLS regression summary of tweeter data

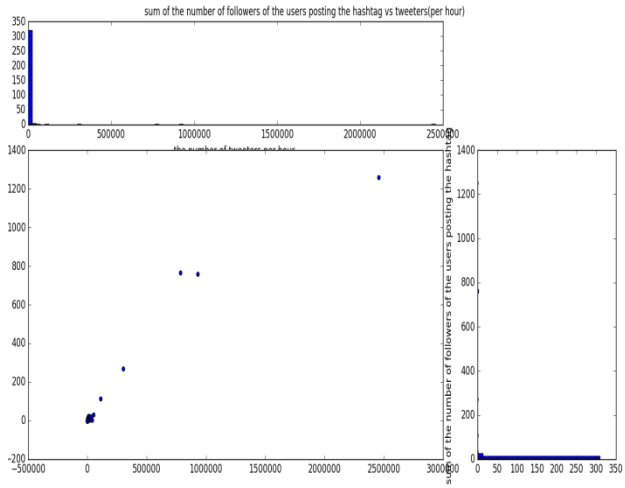


Fig 4-5. Sum of the number of followers versus tweeters/hour

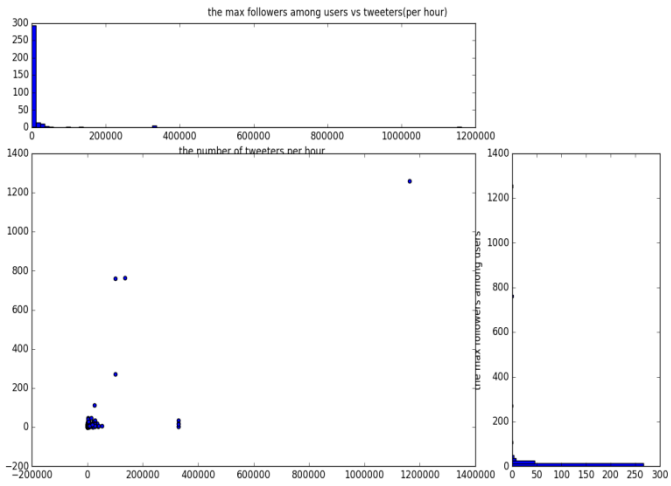


Fig 4-6. Maximum number of followers versus tweeters/hour

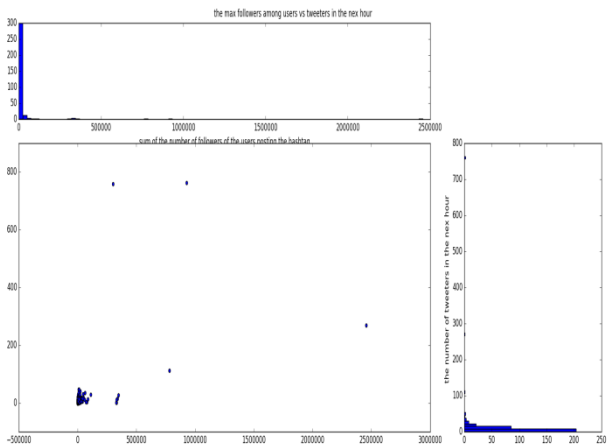


Fig 4-7. Maximum number of followers versus tweeters/hour

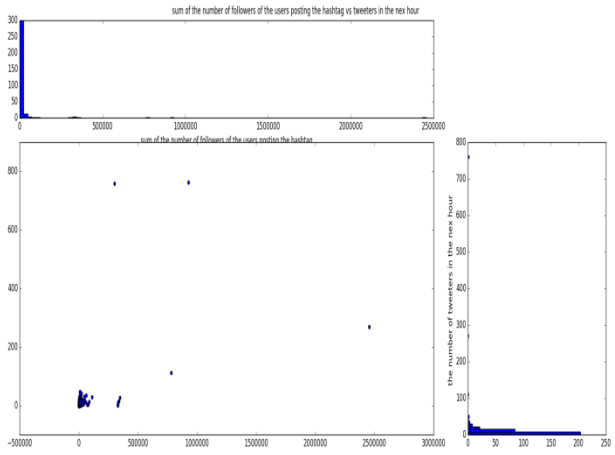


Fig 4-8. Sum of the number of followers versus tweeters next hour

As shown in the picture, there are not obvious linear correlations between the feature and the prediction. They clusters in the small range and scatter in the large range, which is not suitable for linear regression. If we keep track of the small range, we will have large deviation for the higher values, but if we try to model the trends, we will have large relative error for every small value. So there won't be a satisfactory solution.

Then come to the next question, can we make change of these variables to fit the linear model or can we explore other features in the raw data ,or even design new features that can be representative of the hidden information?

In this stage, we think these data are not representative enough to explore deeper, so we won't still fit the linear model. From the above figures, we can also observe that these two features are more correlated with the number of tweeters in the current hour than in the next hour, which means that they may not have a predictive ability for over one hour. This may be due to that they are the results of the popular topic instead of the main reasons for the popularity. They can show the results instead of causing the results. There is also another explanation that they are more effective among the current hour, lacking lasting influence on popularity.

Problem 5:

Download the test data3 and run your model to make predictions for the next hour in each case. Each file in the test data contains a hashtag's tweets for a 6-hour window. The file name shows sample number followed by the period number the

data is from. E.g. a file named sample5_period2.txt contains tweets for a 6-hour window that lies in the 2nd time period described in part 4.

Report your predicted number of tweets for the next hour of each sample window.

Solution:

	hashtag	prediction1	prediction2
SAMPLE1	sb49	364	1018
SAMPLE2	gopatriots+gohawks	68742	78632
SAMPLE3	superbowl	840	976
SAMPLE4	nfl	248	692
SAMPLE5	nfl	237	692
SAMPLE6	gopatriots+gohawks	33540	43829
SAMPLE7	nfl	202	107
SAMPLE8	nfl	213	395
SAMPLE9	gopatriots+gohawks	2013	3212
SAMPLE10	nfl	98	50

In this task, first we try to classify the samples into hashtags so that we can make the most use of the model we trained in the problem 4. If we can get the predictions based on each model, then we can calculate the average popularity by summing all the predictions up.But they are so many mixed hashtags that it's difficult to decide which one hashtag to put it into. In addition, not all hashtags are included in one sample, such as in the sample1, 4,there are no 'gopatriots'. There are uneven distributed hashtags. So then we try to find the dominate hashtag to represent the main trends.
According to problem 4, we observe that the Random Forest has the most steady performance, and during the game, we can not predict precisely, so we try to use the hashtag"gopatriots" and "gohawks" to predict directly.

Problem 6:

The dataset in hands is very rich as there is a lot of metadata to a tweet. Be creative and propose a new problem (something interesting that can be inferred from this dataset) other than popularity prediction. You can look into the literature of Twitter data analysis to get some ideas

Solution:

❖ Topic 1. Can we find the emotion behind the twitter?

We first want to do sentiment analysis about the tweets contents. We want to find how many users hold the positive attitudes and negative attitudes, respectively, towards the patriots and hawks by analysis the sentiment of tweets in #gohawks and #gopatriot.

- Model:

We use the sentiwordnet package in nltk in python to calculate the positive score, negative score and objective score of each word in each tweet. The tweets in language other than English are deleted for simplifying. According to our experiment, we find that most the tweets are in English. So deleting tweets in other language will not influence our results. Our model consists following two parts: preprocess text and calculate scores.

- Preprocessing

In preprocess part, for each tweet, we use ['tweet']['text'] to get the content of tweet. We redefine the tokenize function. We keep the emoticon (such as :)), mention (@Tom) and url (http://…) as one individual and also replace the #hashtag as hashtag . After tokenize, we get a list of string contain words, url, emoticon, mention, hashtag and punctuation. And then, we delete the url, the mention and punctuations since this item cannot give us any information of sentiment of this tweet. We also erase the duplicates of words. To simplify, we also ignore the emoji. The emoticon can always reflect the author's motion when writing a tweet. So we consider the emoticons as important feature to determine whether the tweet is positive. Since the sentiwordnet cannot handle the emoticons, we map it into words. For emotion with happy mouths like),],D,etc, we replace it with word 'happy' and for emoticons with mouths like (, we replace it with word 'sad'. The last step is to remove the stop words with stopwords function in nltk.

For example, the tweet :"Xinxin @xx: that's just an example! :D http://example.com #NLP" will become ['xinxin', "that's", 'example', 'happy', 'nlp'] after the preprocessing step. We use this list as the feature to calculate the score of this tweet.

- Classification

In score calculating part (classification part), we first calculate the positive score, negative score. Since most of the word have many different meanings. We choose the largest positive score of different meanings as the final positive score of this word. Similarly, we can get the negative score for this word. We consider the positive score and negative score separately. That is if the first meaning has highest positive score and the second meaning has highest negative score, we will use the positive score of first meaning as the final positive score and the negative score of second meaning as final negative score. Since the word with both high positive and negative score has different meanings, both of which denotes different strong motion. Such word may cause confusion when determining the sentiment of tweet. In our algorithm, the influence of such words are weakened. For each feature, we can get same number of positive scores and negative scores. We also use the maximum of positive scores and negative scores as the final score of the feature. The feature with larger positive score is labeled as positive tweet. The feature with larger negative score is labeled as negative tweet. Feature with same two scores is labeled as objective tweet.

- Results and analysis

We analyze the tweets sentiment on different time periods. They are before the game, during the game and after the game. For #gopatriot, we find the number of positive tweets and negative tweets are as follows. Total tweet number include the number of objective tweets

Table 6-1 For gopatriot, number of different labels of different time periods

Time period	positive	negative	neg/pos	total tweet
before	2290	1169	0.51048035	4617
during	6090	3501	0.57487685	14250
after	368	204	0.55434783	706

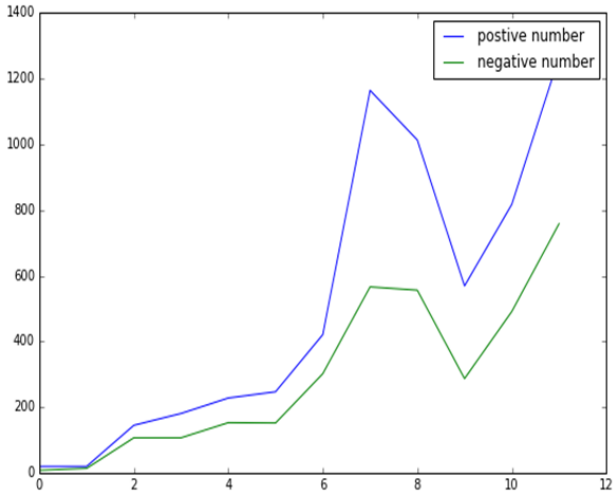


Fig 6-1. During game, the number of positive tweets each hour and negative tweets changing with time (hour)

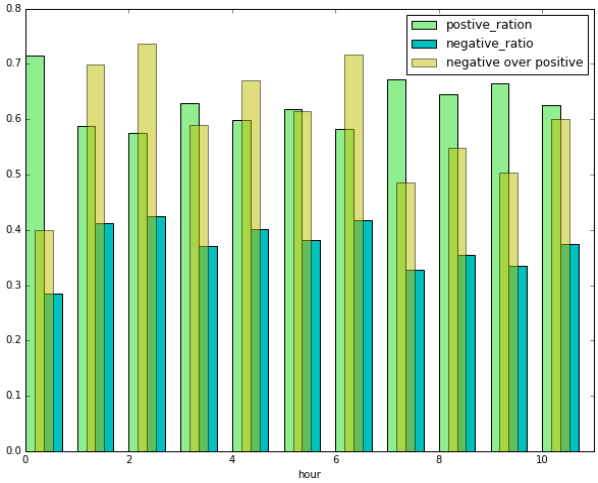


Fig 6-2. The ratio of positive and negative tweets of

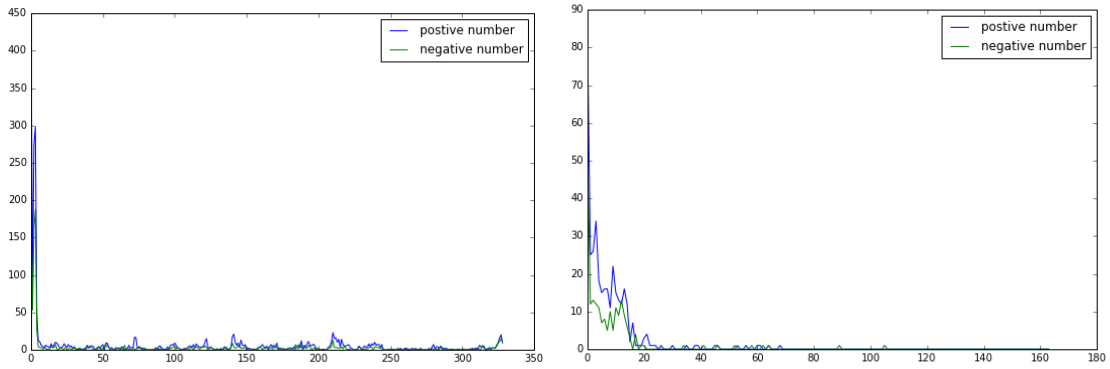


Fig 6-3. Number of two labels before(left figure) and after (right figure) the game

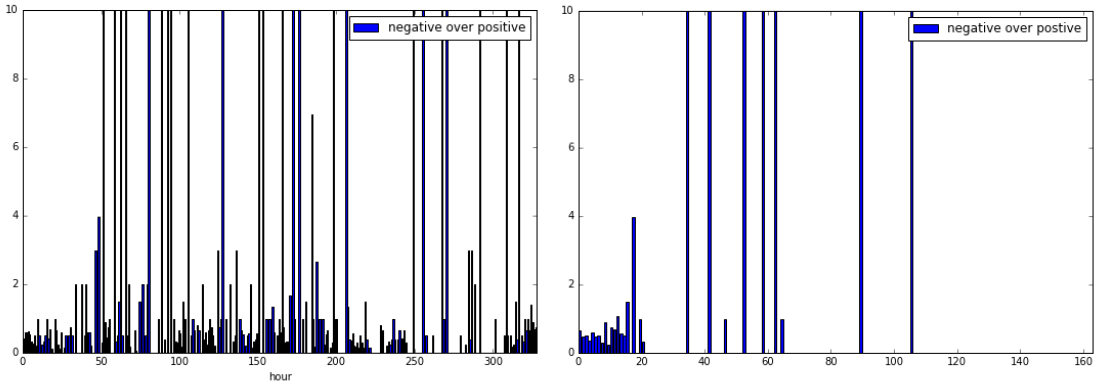


Fig 6-4. Ratio of negative over positive per hour before (left figure) and after (right figure) the game

For #gohawks, we also implemented such analysis.

Table 6-2 For gohawks, number of different labels of different time periods

Time period	positive	negative	neg/pos	total tweet
before	34867	16261	0.46637221	62265

during	27671	14243	0.51472661	54012
after	2972	1662	0.55921938	5447

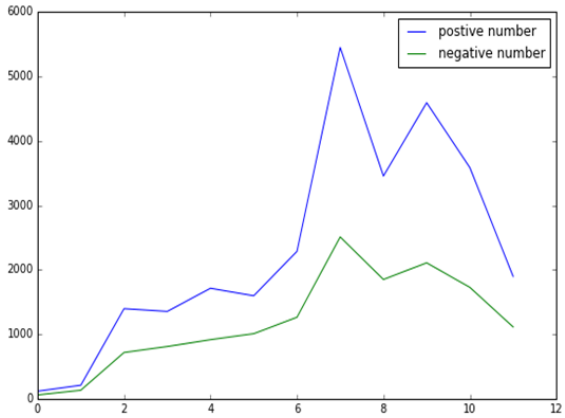


Fig 6-5. During game, the number of positive tweets of each hour tweets and negative tweets changing with time (hour)

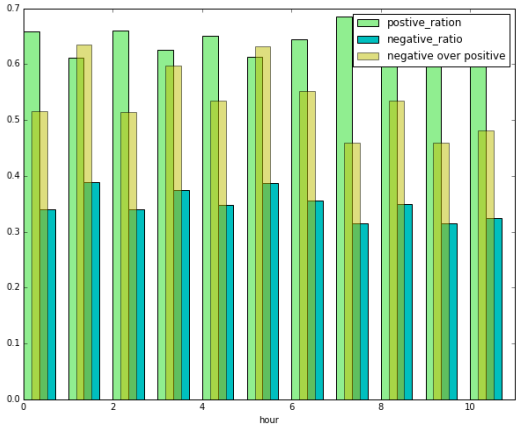


Fig 6-6. The ratio of positive and negative

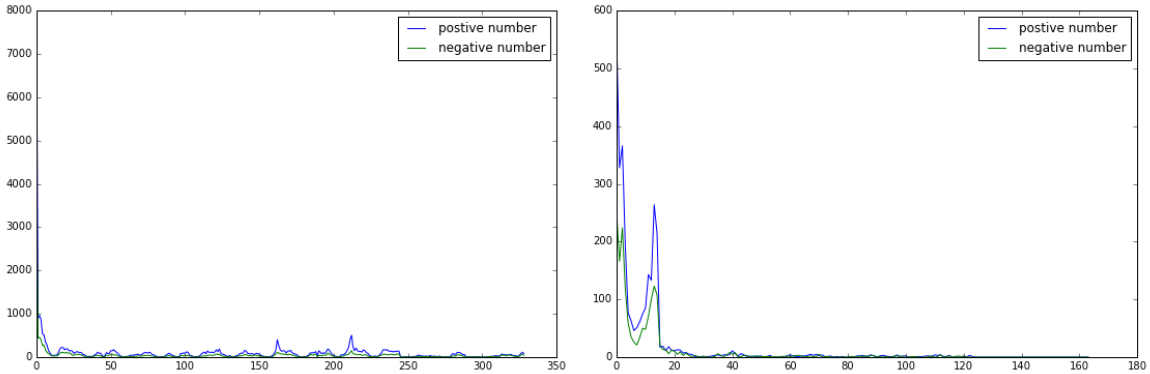


Fig 6-7. Number of two labels before(left figure) and after (right figure) the game

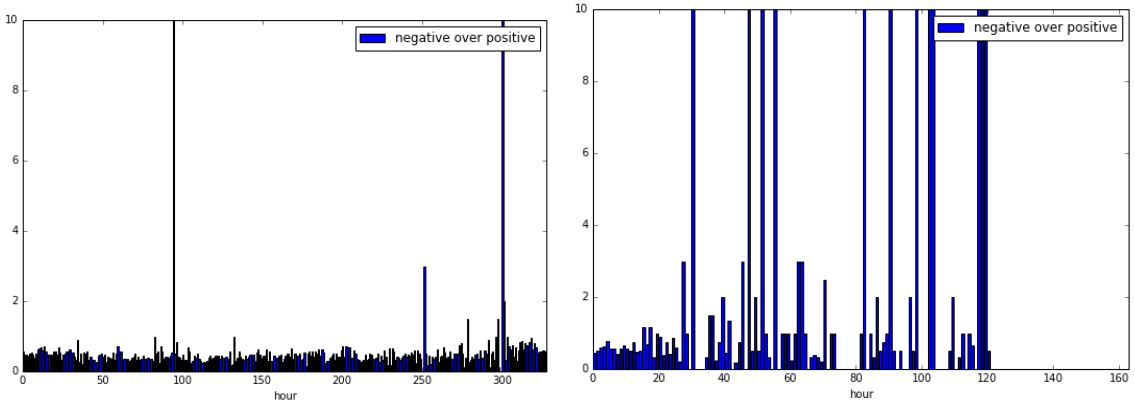


Fig 6-8. Ratio of negative over positive per hour before (left figure) and after (right figure) the game

In general, for both data sets, the positive tweets are more than the negative tweets. This two hashtag have the smallest amounts of tweets. Most of time, only fans will post tweets with unpopular hashtag #go… to cheer up their teams. So we can infer that most authors in these datasets are the fans of these two teams. This why the number of positive tweets is more than the number of negative tweets.

We call the ratio of negative tweets over positive tweets as n_p for short. As we can see, n_p of gopatriot get largest during the game and decrease after the game. But for hawks, n_p keep increasing from 0.46 before game to 0.56 after game. As we all know the patriot won the game, we can infer that, for hawks, losing the game is the reason why the negative tweets getting more. From figure 6-4 and 6-8, we can also find such pattern. In Fig 6-4, the tall histograms become less after game. However, in fig 6-8, the tall histograms become more after game.

Our model can roughly determine the sentiment of a tweet. However, there still many ways to improve the model performance. In the future, we can keep emojis as a feature, some of which also have strong sentiments. We didn't handle abbreviations, such as BTW, and character repetitions, such as Yesssssssss. We may loss some information in this way. So we can include algorithm to handle such problem in preprocess step in the future.

- Topic2: When talking about superbowl, what are they talking? Before, during and after the event, what changes?

If using tf-idf to analysis 'gopatriots' text, we can find the top 20 keywords as follows.

['bowl','brady','el','espntemsuperbowl49','game','gopatriots','gopats','http','la','nfl','patriots','pats','que','sb49','seahawks','super','superbowl','superbowlxlix','time','touchdown']

Besides topics above, we also come up with other interesting topics, which are worth to try in the future.

Advertisement: If we want to post an advertisement during superbowl this big event, when should we start best? End? If posted daily, when is the best time? 1. So we want to consider the embedded advertisement which needs considering retweets. 2. We need to find the time pattern as well as the users' habits of using twitter.

Time pattern: Find the most related word to the game's status. If we are given a tweeter, can we predict the game?

References:

[1] Kong S, Mei Q, Feng L, et al. On the Real-time Prediction Problems of Bursting Hashtags in Twitter[J]. arXiv preprint arXiv:1401.2018, 2014.

[2] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//LREC. 2010, 10: 2200-2204.

[3] Kouloumpis E, Wilson T, Moore J D. Twitter sentiment analysis: The good the bad and the omg![J]. Icwsm, 2011, 11: 538-541.