UCLA ECE219 - Project 5

March 22nd, 2019

Authors

Ning An - 205034447
Zaurbek Tsorojev - 805029443
Endi Xu - 005030030
Minya Yao - 704161217

Introduction

The goal of this project was to analyze the tweet activity for given hashtags from Twitter in order to predict its future popularity. We train a regression model with the given data, and use it to make predictions for other hashtags. The dataset that we are using was obtained by querying popular Twitter hashtags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game.

Question 15

In this question, we use the textual content of the tweet to predict the location of the author of a tweet (Washington or Massachusetts), by training a binary classifier. We consider all the tweets including `#superbowl`.

Point 1

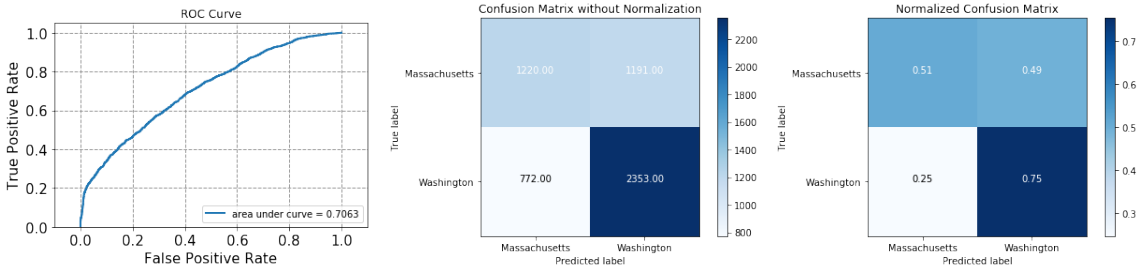
In order to only select tweets whose location is in Washington or Massachusetts, we got through all the tweets and check if their location is any of the cities of these states. If it is, we assign the tweet a label of "0" for Massachusetts and "1" for Washington. Processing this way, we collected 55,355 tweets.

Point 2

In this part, we will train 4 binary classifiers to predict the location of the author of a tweet.

a. Linear SVM Classifier

We start with a linear SVM classifier. The ROC curve and confusion matrix are reported below.

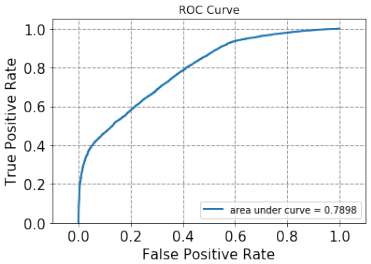


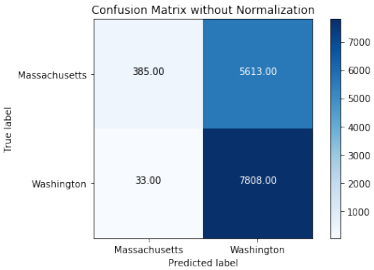
We get the following metric values :

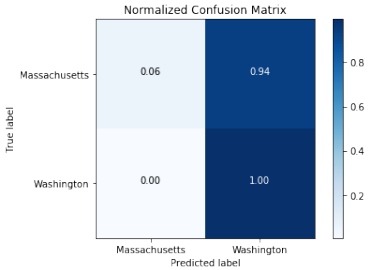
Accuracy	0.64541
Recall	0.75296
Precision	0.66394

b. MultinomialNB

Next, we train a Naïve Bayes classifier with the MultinomialNB model.



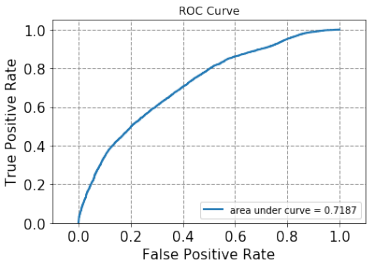


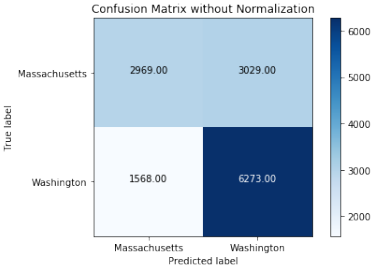


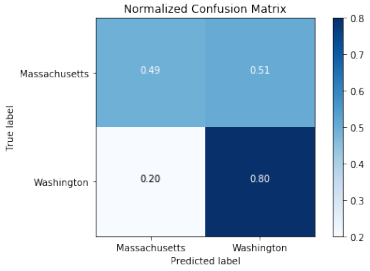
Accuracy	0.59202
Recall	0.99579
Precision	0.58177

c. GaussianNB

We try a Naïve Bayes classifier again, but this time with the GaussianNB model.



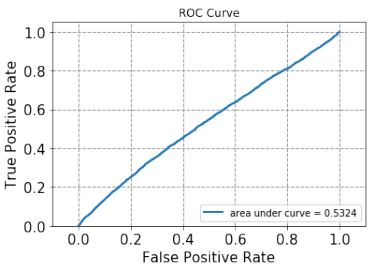


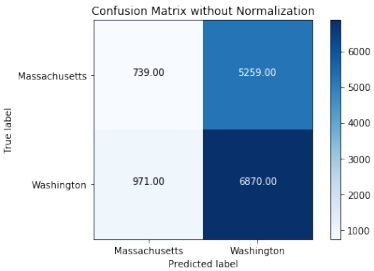


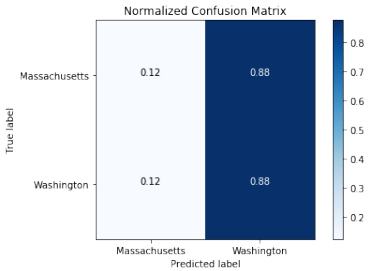
Accuracy	0.66782
Recall	0.80003
Precision	0.67437

d. Logistic Regression

Finally, we try a logistic classifier.







Accuracy	0.54982
Recall	0.87616
Precision	0.56641

Conclusion

From these 4 classifiers, the best results were achieved with the Linear SVM and GaussianNB classifiers. All of the classifiers were able to predict the location of Washington with good accuracy, however they all predicted Massachusetts with bad accuracy. A reason for this could be that we didn't perfectly assigned the right labels when selecting the cities in the two states.

Question 16

In this last part, we will analyze the tweet sentiments for several hashtags in order to understand how people are feeling (positve or negative).

We chose 4 different hashtags : `#nfl, #superbowl, #gopatriots, #gohawks`
For each hashtag, we will plot the number of positive and negative tweets over the two weeks of data that we have.

Hashtag #nfl

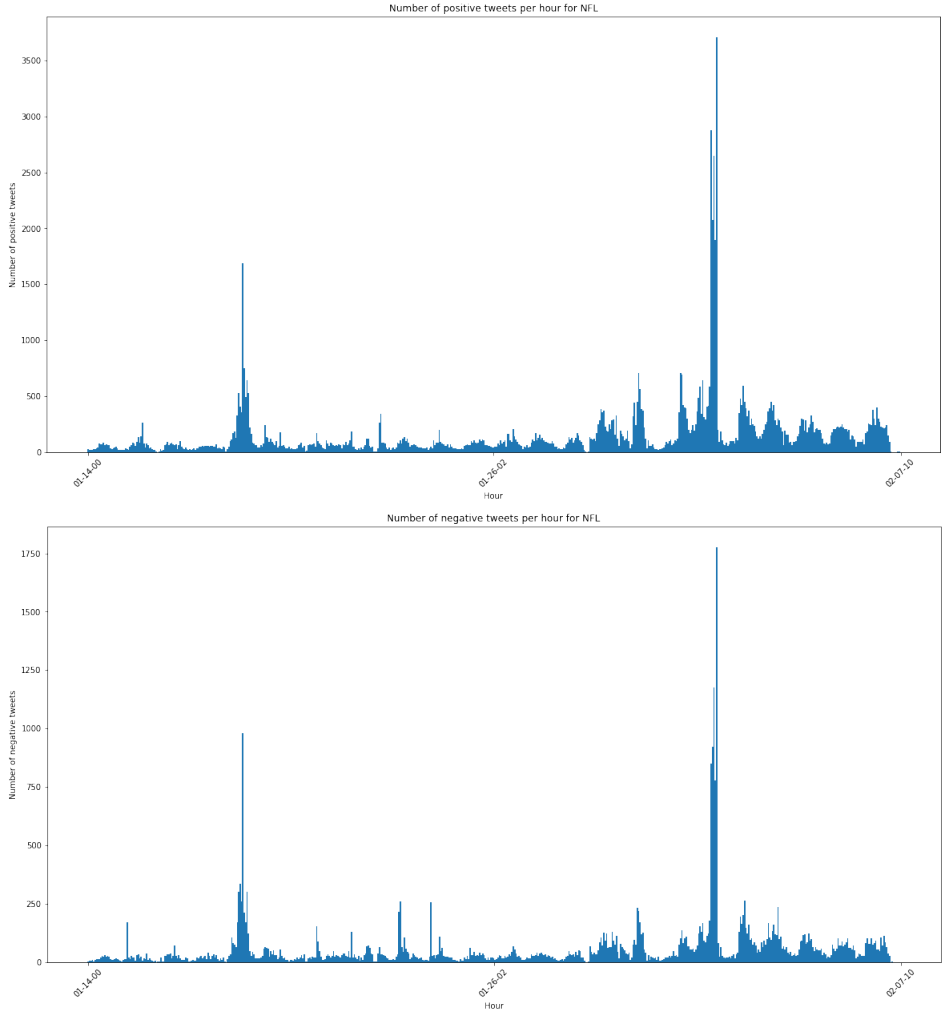
We see that we have two main spikes for both plots. This show a high number of tweets during or after an important match. There is double number of positive tweets than negative tweets, which shows that the majority of people had a positive feeling about the event.

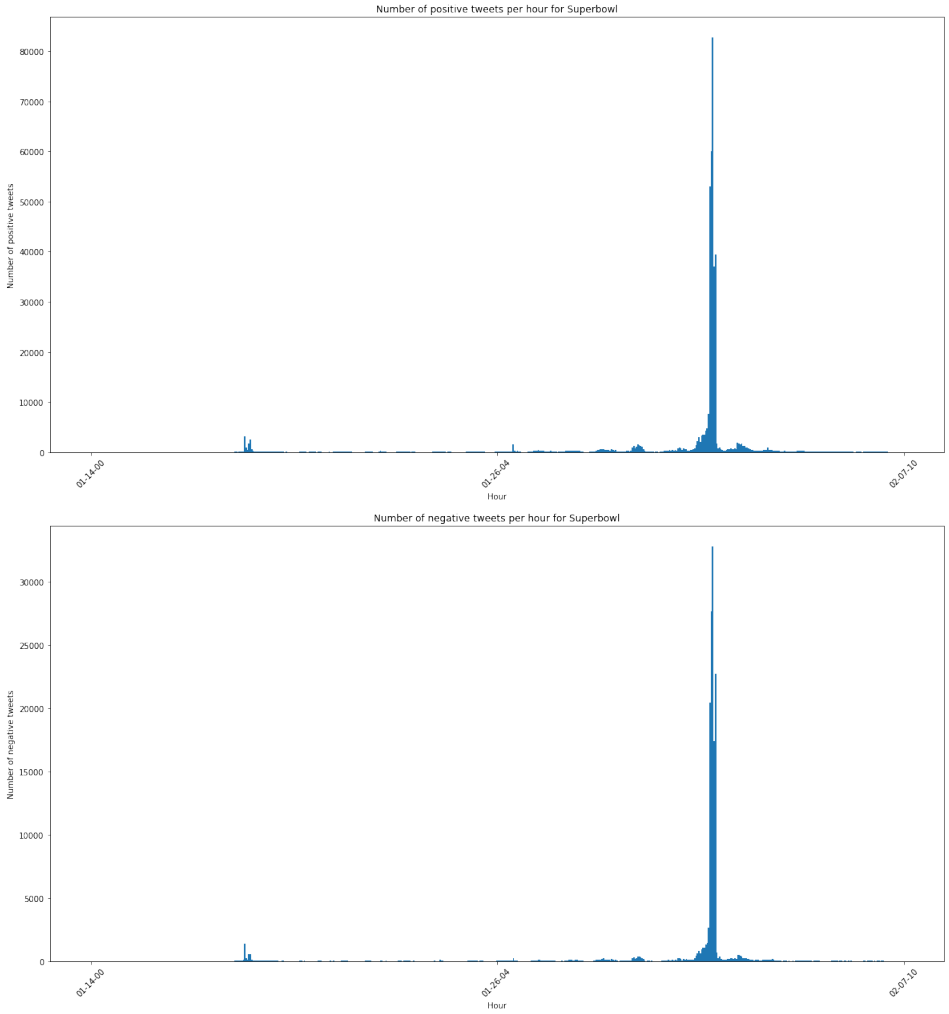
Hashtag #superbowl

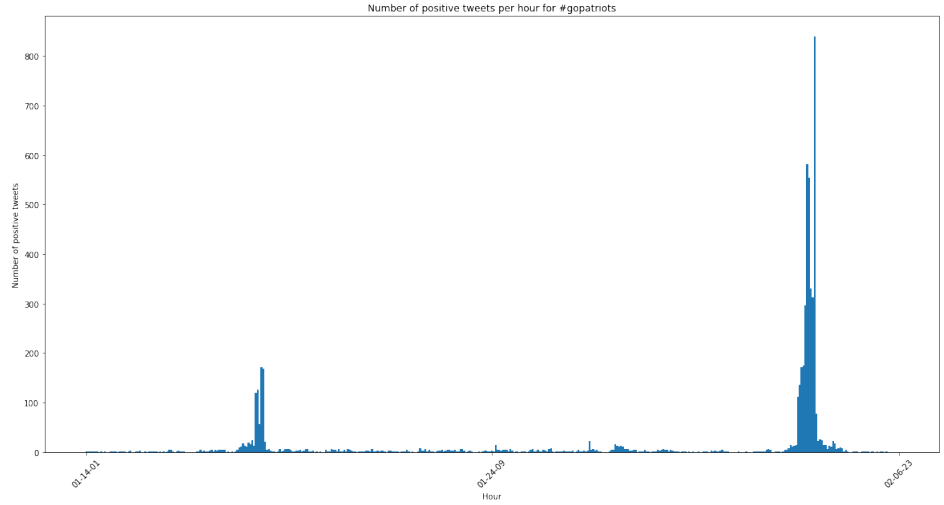
Here we have one main spike. At that moment in time, there are about 85,000 positive tweets and 35,000 negative tweets. The moment corresponds to the final of the Superbowl. The positive tweets are probably all the people happy that their team won, and the negative ones are those who are disappointed.

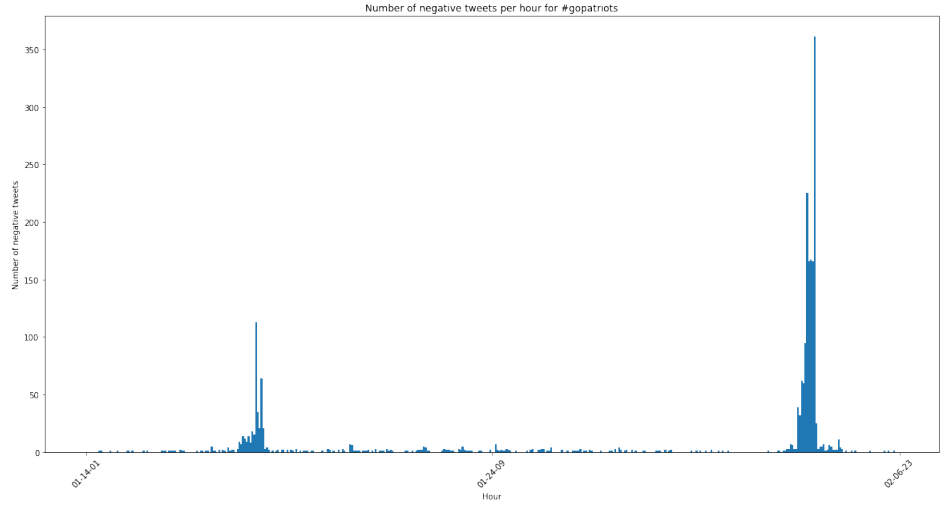
Hashtags #gopatriots and #gohawks

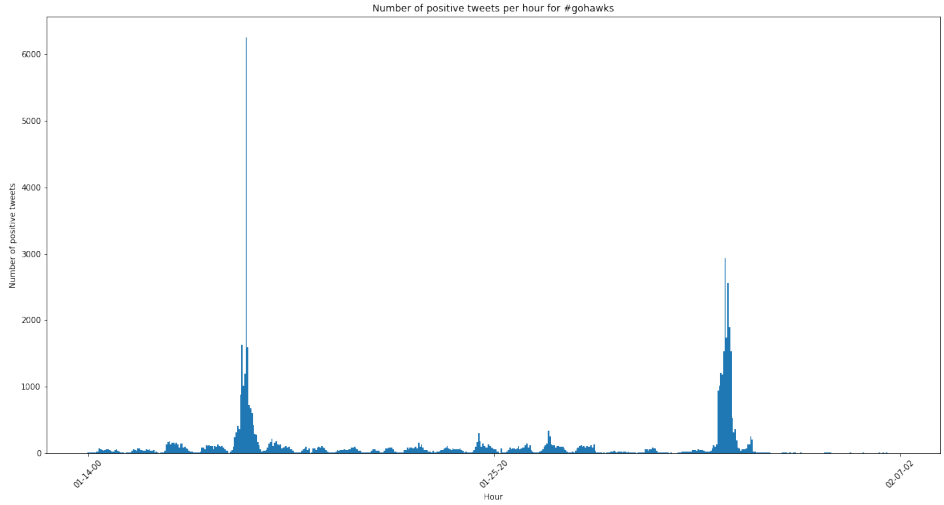
What is interesting is that we can also analyze the sentiment of tweets linked to a specific match to know who is winning. In this example, we see that Patriots have much more positive hashtags than Seahawks, and their ratio of positive to negative hashtags is lower. This indicates that the Patriots probably won the finals, and this is indeed what happened.

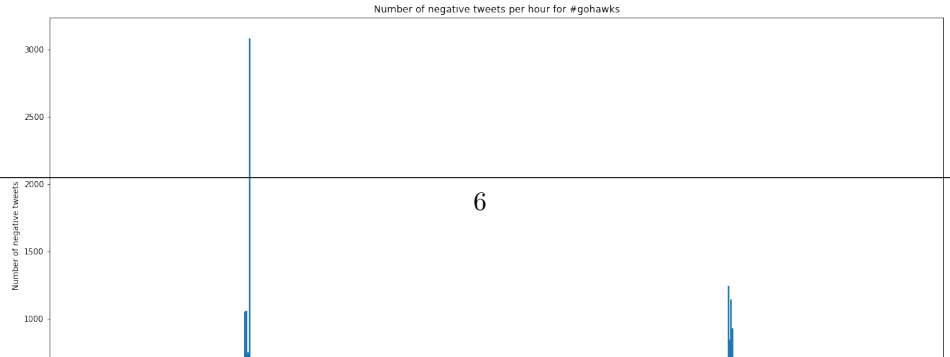

Number of positive tweets per hour for NFL
Number of positive tweets
0
500
1000
1500
2000
2500
3000
3500
01-14-00
01-26-02
02-07-10
Hour
Number of negative tweets per hour for NFL
Number of negative tweets
0
250
500
750
1000
1250
1500
1750
01-14-00
01-26-02
02-07-10
Hour


Number of positive tweets per hour for Superbowl
Number of positive tweets
80000
70000
60000
50000
40000
30000
20000
10000
0
01-14-00
01-26-04
02-07-10
Hour
Number of negative tweets per hour for Superbowl
Number of negative tweets
30000
25000
20000
15000
10000
5000
0
01-14-00
01-26-04
02-07-10
Hour


Number of positive tweets per hour for #gopatriots
Number of positive tweets
0
100
200
300
400
500
600
700
800
01-14-01
01-24-09
02-06-23
Hour


Number of negative tweets per hour for #gopatriots
Number of negative tweets
0
50
100
150
200
250
300
350
01-14-01
01-24-09
02-06-23
Hour


Number of positive tweets per hour for #gohawks
Number of positive tweets
0
1000
2000
3000
4000
5000
6000
01-14-00
01-25-20
02-07-02
Hour


Number of negative tweets per hour for #gohawks
Number of negative tweets
1000
1500
2000
2500
3000