239AS - Special Topics in Signals and Systems
Project 4 - Popularity Prediction on Twitter

Mansee Jadhav - 204567818
Mauli Shah - 004567942
Ronak Sumbaly - 604591897

March 19, 2016

Introduction

Twitter, with its public discussion model, is a good platform to predict future popularity of a topic or event. Knowing current and previous tweet activity for a **hash-tag** (#) , we can predict if it became more prominent and trendy in the future and if yes by how much.

Twitter data is collected by querying popular hash-tags related to the **2015 Super Bowl** spanning a period starting from 2 weeks before the game to a week after the game. We use this data to train a regression model and then use the model to making predictions for other hash-tags. The test data consists of tweets containing a hash-tag in a specified time window, and we have then used our model to predict number of tweets containing the hash-tag posted within one hour immediately following the given time window.

Question 1 - Tweet Data Statistics

The training tweet data was loaded and statistics for each hash-tag was calculated in this question. In order to keep track of the hour count we have employed a hour-window approach. Since the tweets are all in sorted order of their posting time (**firstpost_date**). The first tweet is considered and the 1^{st} hour-window is created using the formula

$$end\_time = start\_time + 3600 \tag{1}$$

We loop through each tweet in the file and compare the post-time of the tweet with the end time of the present window. If it lies within the window we increase the hour-count if it doesn't we create a new window by using eq^n.(1) and adding 3600 (1 hour in UNIX time) again to the end-time. At the same time a count is kept for the number of followers of users (**author/followers**) and the number of re-tweets (**metrics/citations/total**) for each tweet. The statistics calculated using the above procedure are listed below.

Hashtag	**Total Tweets**	**Avg. # Tweets/hr**	**Avg. # of Followers of Users**	**Avg. # of Retweets**
#gohawks	188135	193.5438	1596.443	2.0146
#gopatriots	26231	38.3832	1292.2031	1.4001
#nfl	259019	279.5503	4394.2539	1.5385
#patriots	489710	499.4200	1607.4407	1.7828
#sb49	826905	1419.886	2229.6948	2.5111
#superbowl	1348766	1401.2445	3675.3394	2.3882

Table 1: Statistics for Each Hashtag

Analysis of the Statistics

1. **Most Tweeted Hashtags per hour**: **#sb49 and #superbowl**
2. **Most Followers of Users for Hashtag** : **#nfl and #superbowl**
3. All of the tweet data collected comprise of tweets that are not re-tweeted or are re-tweeted by very few users hence making the average re-tweet count ≈ 2.

In order to visualize the number of tweets in an hour a histogram was plotted for #SuperBowl and #NFL. A **steep-rise** can be seen for both the graphs at the same time which indicates the **hour of the event**.

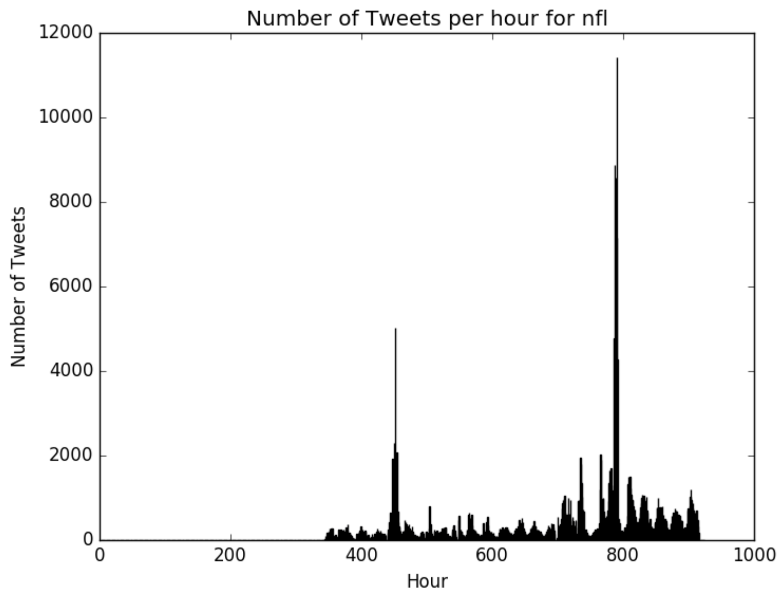

Figure 1: Number of tweets in hour : #NFL

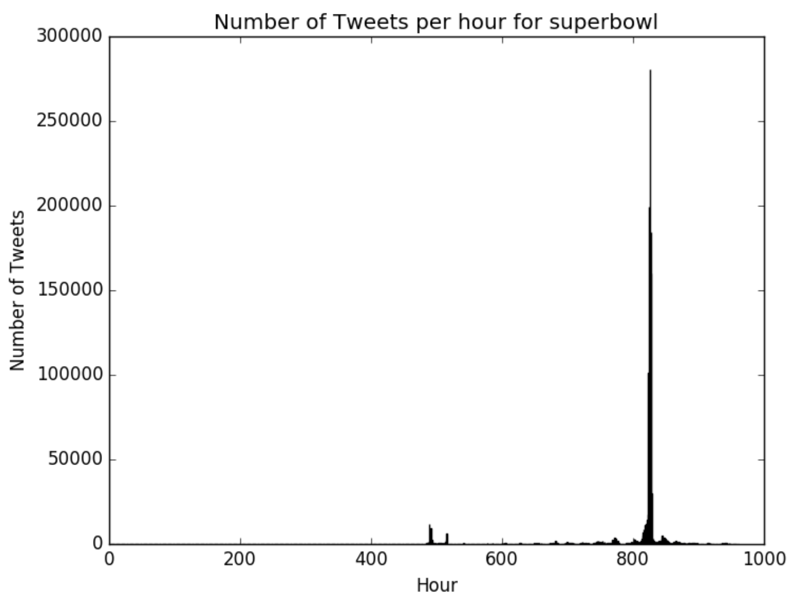

Figure 2: Number of tweets in hour : #SuperBowl

Question 2 - Linear Regression

A **linear regression model** was created using 5 features to predict number of tweets in the next hour, with features extracted from the tweet data in the previous hour. The **features** used to create the model were,

1. **Numbers of Tweets (Class Variable)**
2. **Total Number of Re-tweets -** (metrics/citations/total)
3. **Sum of the number of followers of the users -** (authors/followers)
4. **Maximum number of followers of the users posting the hashtag**
5. **Time of the data -** Obtained using the post-time of the tweet

The same hour-window approach was employed to calculate all the features. The output variable for each hour-window was the tweet count for the next hour-window. The model was trained using the OLS statsmodel library. The results obtained for each of the hashtag are as follows,

HashTag	**Accuracy**
#gohawks	41.78
#gopatriots	43.15
#nfl	54.69
#patriots	43.72
#sb49	58.54
#superbowl	66.13

Table 2: Model Accuracy for each Hashtag

The (p-value [1], t-value [2]) for each attribute was recorded as well, the results are as follows,

Hashtag	**# of Retweets**	**$\sum$ of # of followers of users**	**Max. # of followers**	**Time of the data**
#gohawks	$(2.115*10^{-5}, 4.273)$	$(1.066*10^{-7}, 5.355)$	$(1.290*10^{-6}, -4.871)$	$(4.230*10^{-3}, 2.867)$
#gopatriots	$(8.732*10^{-27}, 11.247)$	$(3.517*10^{-16}, -8.386)$	$(1.048*10^{-12}, 7.278)$	$(8.643*10^{-1}, -0.170)$
#nfl	$(3.525*10^{-16}, 8.266)$	$(7.424*10^{-2}, 1.787)$	$(4.185*10^{-1}, -0.809)$	$(2.314*10^{-5}, 4.253)$
#patriots	$(4.432*10^{-63}, 18.053)$	$(6.807*10^{-14}, -7.602)$	$(2.209*10^{-5}, 4.263)$	$(4.776*10^{-1}, 0.710)$
#sb49	$(4.681*10^{-56}, 17.647)$	$(2.329*10^{-31}, -12.374)$	$(1.320*10^{-15}, 8.222)$	$(6.402*10^{-2}, -1.855)$
#superbowl	$(4.897*10^{-149}, 31.256)$	$(1.612*10^{-116}, -26.504)$	$(4.853*10^{-52}, 16.145)$	$(7.136*10^{-2}, -1.805)$

Table 3: p-value & t-value for Model Parameters

Analysis of Results

- According to the definition of *p-value and t-value* it can be seen that the **most contributing feature** towards the regression model in all hash-tag files is the **Number of Re-tweets** posting a hash-tag.
- A fairly **low accuracy** is obtained for most of the hash-tag. This can be attributed to the window-size of one-hour as in the initial hours the average number of tweets are pretty low and creating a model for these sparse features is more difficult.

[1] A predictor that has a low p-value is likely to be a meaningful addition to your model.
[2] The larger the absolute value of t, the less likely that the actual value of the parameter could be zero.

Question 3 - Regression Model with Extra Features

A **new regression model** was created using custom extra features (including original features considered in Question 2.) that were considered based on various papers and observation of the data. The new features considered were as follows,

1. **Numbers of Tweets (Class Variable)**
2. **Total Number of Re-tweets** - (metrics/citations/total)
3. **Sum of the number of followers of the users** - (authors/followers)
4. **Maximum number of followers of the users posting the hashtag**
5. **Time of the data** - Obtained using the post-time of the tweet
6. **Ranking Score** - (metrics/ranking_score)
7. **Impression Count** - (metrics/impression) - Measures the number of times a user is served a Promoted Tweet either in time-line or on search
8. **Favorite Count** - (tweet/favorite_count) - Number of tweets favourite's by users
9. **Number of Users per hour** - (tweet/user/id) - Counted number of users posting per hour
10. **Number of Long Tweets per hour** - (title) - Counted the number of tweets with length > 100 characters.

A total of **9 features** were used to create the new regression model and after employing the same methodology of Question-2, features were collected using one-hour window. Since the last hour window cannot predict a tweet-count value it has been removed while creating the model. The model was tested and the results obtained were as follows,

HashTag	Accuracy
#gohawks	78.384
#gopatriots	53.118
#nfl	64.840
#patriots	58.793
#sb49	70.623
#superbowl	77.089

Table 4: Model Accuracy for each Hash-tag

As seen from the above results we have a **significant increase in the accuracy of the model** for each of the hash-tag, this can be attributed to features that are not sparse and have a well defined distribution through-out the period of the SuperBowl. Metrics employed in the tweet-data have been used to model the importance of the tweet for a given window frame thereby increasing the accuracy. In order to better visualize the contribution of the features in the model a scatter plot was created of the **Top 3 features** for each hash-tag. Since the initial hours have less number of tweets, all of the graphs exhibit clustering of values near low of tweets/hour.

#gohawks

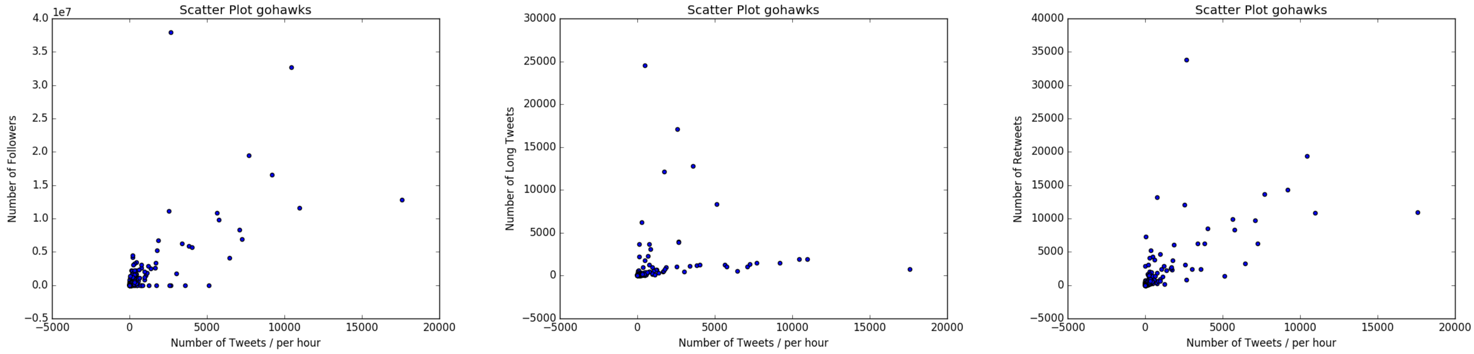

Figure 3: Top 3 feature for #gohawks (# of Followers, # of Re-tweets, # of Long Tweets)

#gopatriots

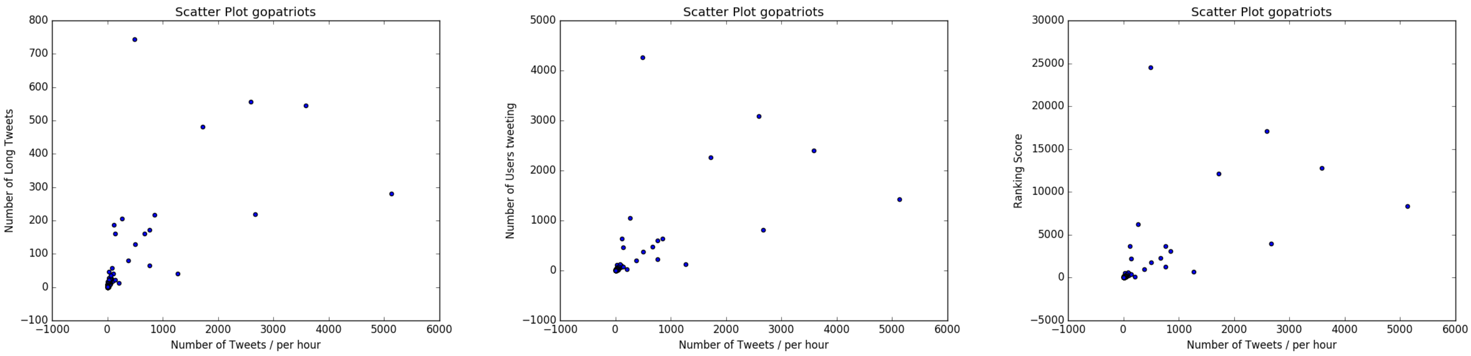

Figure 4: Top 3 feature for #gopatriots (# of Long Tweets, # of Users Tweeting, Ranking Score)

#nfl

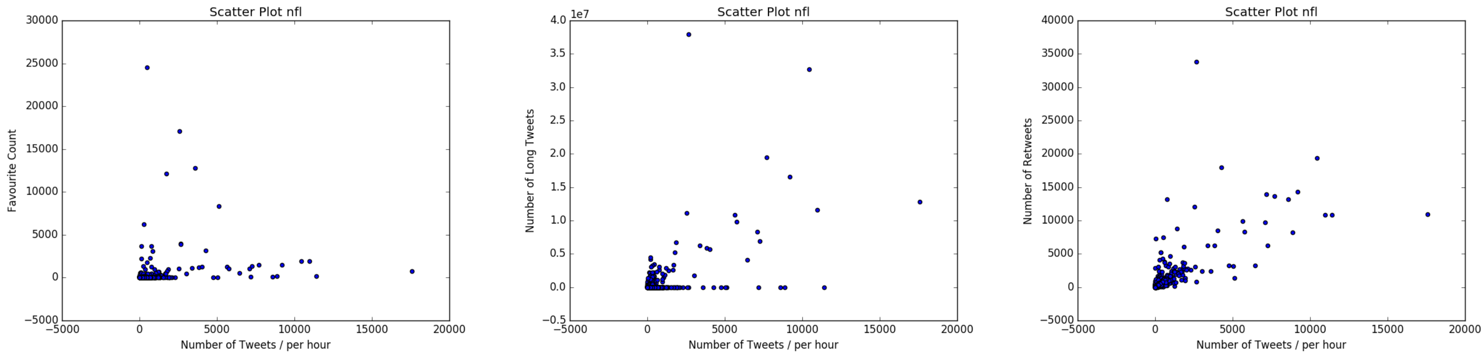

Figure 5: Top 3 feature for #gohawks (Favorite Count, # of Re-tweets, # of Long Tweets)

#patriots

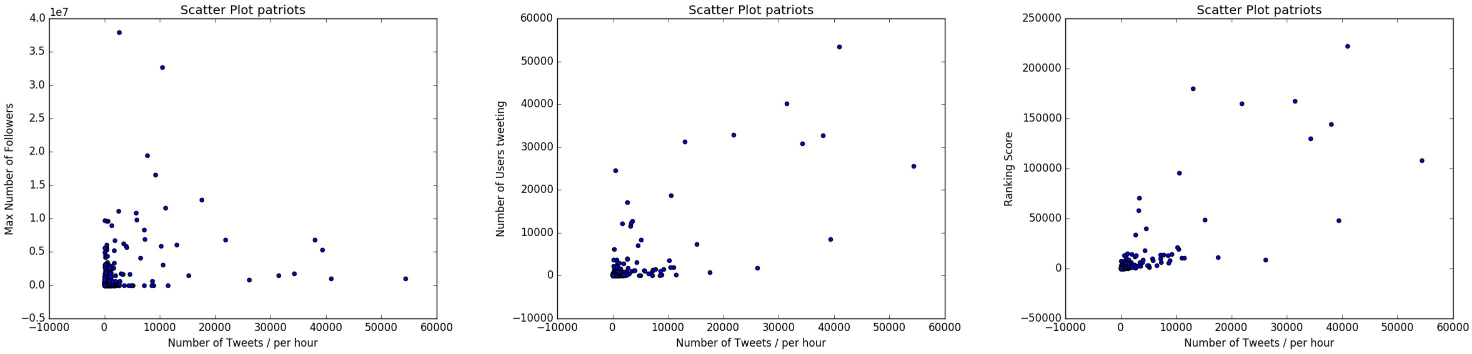

Figure 6: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

#sb49

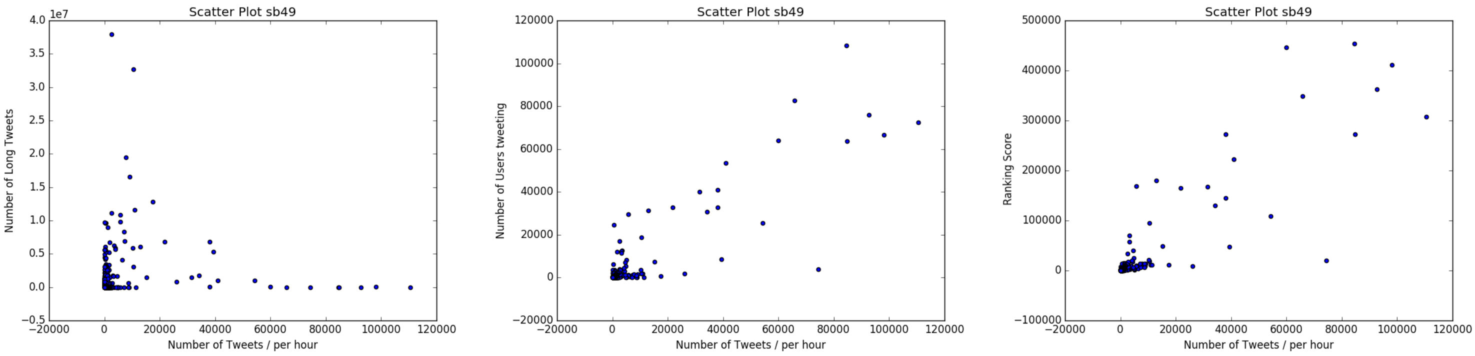

Figure 7: Top 3 feature for #gohawks (# of Long Tweets, # of Users Tweeting, Ranking Score)

#superbowl

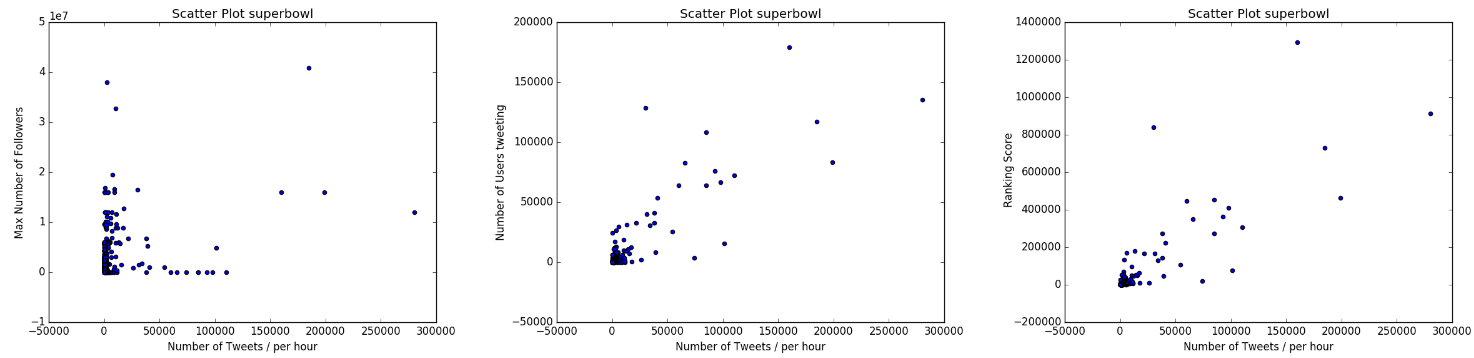

Figure 8: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

Analysis of Scatter Plots

HashTag	Analysis
#gohawks	A linear proportionality can be seen in the scatter plots signifying a good relationship between all the 3 features
#gopatriots	Almost identical scatter plots with clustering towards the region of the origin
#nfl	A constant relationship can be seen for features (Favorite Count, # of Long Tweets) while a linear relationship is visible for # of Retweets
#patriots	Constant relationship for Max # of Followers feature while other two show linear proportionality
#sb49	Similar analysis to #patriots
#superbowl	Clustered regions with a very small linear deviation. Large number of instances fits a better regression model hence the higher accuracy

Table 5: Analysis of Top 3 Features Scatter Plot

Question 4 - Cross Validation

The first part of Question-4 requires the usage of same features used in Question-3 and to perform **10-fold Cross Validation** across data. The accuracy results obtained across various hash-tags and over every fold given below,

Fold Number	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
(1)	7.782	20.127	23.921	180.855	31.417	229.980
(2)	8.438	46.514	1.376	84.489	61.089	255.881
(3)	10.145	4.814	3.181	31.927	99.079	337.870
(4)	204.985	2.245	28.109	52.189	64.583	397.136
(5)	15.497	117.978	185.833	265.855	124.529	361.339
(6)	41.759	629.267	133.980	997.125	301.904	2506.928
(7)	19.302	147.079	93.183	687.341	881.058	1168.849
(8)	18.391	171.120	194.827	466.046	2854.875	2756.248
(9)	30.380	850.131	524.838	2046.537	1032.974	19664.687
(10)	247.476	5.099	137.612	176.498	321.142	1661.469
Average Error	**60.415**	**199.437**	**132.686**	**498.886**	**577.265**	**2934.039**

Table 6: Average Error of 10 Fold Cross Validation

Analysis of Results

- We can see that there is a relationship between the number of tweets for a hash-tag and the average error of cross validation. **Greater the number of tweets leads to a higher absolute average error for the hash-tag**.
- In particular it is seen that for each hash-tag the error of one of the cross-validation fold is too high due to the the **uneven distribution of the data-set**. A fold might consider a split wherein the test-data has all high values for the class (tweets during the time of the SuperBowl) and training-data has all low values for the class (tweets before and after the SuperBowl), hence producing a high error value for that fold (e.g Fold 9 for #gopatriots).

Question 4 - Cross Validation with Time Periods

The second part of Question-4 deals with analysis of regression models created for different time-periods during the SuperBowl. Three different time-periods were considered to create the regression models,

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.

Each tweet was segregated based on the time it was posted and split into windows of one-hour. The models were tested using 10-fold Cross Validation and the average errors for all folds obtained were as follows,

HashTag	**Before**	**Between**	**After**
#gohawks	167.189	7022.163	2607.692
#gopatriots	16.217	238.102	1760.682
#nfl	75.919	753.944	533.593
#patriots	190.869	93528.077	9745.065
#sb49	39.833	51166.878	12012.449
#superbowl	203.754	12861.877	11834.395

Table 7: Average Error of 10 Fold Cross Validation for each Time-Period

Analysis of Results

- It can be clearly seen that due to the **Between time-period** having only **12 one-hour window** the number of instances in this time-period to create a model is very low. Hence the model created is giving **very high average error values**.
- Since the **Before time-period** has a greater number of instances a better model is created hence giving **low average error values**.

Question 5 - Testing Data

The testing data was downloaded and for each file in the testing data features were collected using methods employed in the previous questions. Since the entire data are of 6-hour window instances each testing data-set have **less than 6 instances**. Each period was compared with the corresponding model that was created in Question 4 for each hashtag.

Since the test data comprises of all the hashtags mixed we need to apply only those models that fit appropriately. An alternative approach is to apply all the models and check the error of predicted values of the first 6 hours to estimate the performance of the 7th hour. The **predicted values for the 7th hour** is provided below and the value with the least error with respect to the 6 hour data is highlighted indicating the estimated predicted value.

HashTag	S1_P1	S2_P2	S3_P3	S4_P1	S5_P1
#gopatriots	290.547	2383960.826	34608.248	1602.209	393.140
#gohawks	365.351	-858140.983	-1814.662	93.041	409.403
#nfl	174.141	**2178909.082**	-1567.131	**284.860**	**280.906**
#patriots	242.045	173637.896	6571.395	219.402	231.576
#sb49	**111.779**	-1486543.408	**1488.885**	143.832	172.105
#superbowl	15.384	-1283467.82	1240.606	50.764	37.700

Table 8: Predicted Value for 7th Hour using Regression Model

HashTag	S6_P2	S7_P3	S8_P1	S9_P2	S10_P3
#gopatriots	-21993.469	-87.643	-109.984	57966.763	58.706
#gohawks	-3672.594	-32.583	295.120	-20145.225	-31.695
#nfl	**886782.04694**	102.665	105.748	**50099.683**	-17.310
#patriots	-91159.906	-50.581	151.477	-26168.779	932.511
#sb49	-87872.106	**197.338**	**40.019**	-28214.752	1939.845
#superbowl	-374372.274	207.131	81.62	-25818.844	**2789.611**

Table 9: Predicted Value for 7th Hour using Regression Model

Analysis of Results

- The highlight values in the table correspond to the predicted value for the 7th hour given the 6 hour data. The least error model value has been highlighted.
- As previously stated the **Between-period** or **P2** have a training dataset of 12 instances hence the values predicted for all the P2 test-data have a very high variation.

Question 6 - Twitter Ad-Celeb Week (Event Sequencing)

Evalutaing the flow of the events with Twitter

Problem Statement

The SuperBowl is a widely watched event supported by thousands of tweets online. The event acts as a publicity platform for various **high profile advertisments and celebrities**, a result of the game's extremely high viewership and wide demographics. The problem that we propose is that of **event-sequencing** and analytics. Given all the tweets can we recreate the flow of events that happened at the SuperBowl. Also since advertising and celebrity sightings are a part and parcel of the SuperBowl we want to analyze how the popularity of the two changed during the course of the event. Our end result is to provide a **brand comparison** which shows which brands are gathering the most attention and during which time phase of the super bowl along with **celebrity comparison** to get an insight into how involving a celebrity can impact the overall event.

Procedure

1. **Data Splitting -** All tweet text from a particular hash-tag are collected using the one-hour window concept.
2. **Data Preprocessing -** The tweet text is pre-processed by removing special characters and stop words
3. **Key Word Tokenization -** After preprocessing we tokenize the keywords into two categories:
   - HashTags
   - Non-HashTag Data
4. **Forming Bigrams -** The commonly occuring pairs are put into a counter which collaborates key word pairs for every hour
5. **Advertisement Classification -** The data is then classified into different advertisement categories. For this project we considered prominent brands like T-mobile, Budweiser, Snickers, McDonald's etc. The ads which are made of two words like Coca Cola and Dove's Mencare are searched using the bigrams counter created in the previous step. We also look for taglines in the bigrams counter created. The result of this classification is the hourly count of occurance of every advertisement.
6. **Celebrity Classification -** The same data is fed for celebrity classification wherein celebrities such as the popularity of celebs like Katy Perry, Missy Elliot, Idina Menzel etc. are analyzed on an hourly basis.
7. **Creating Topics -** In order to show the flow of events of the SuperBowl we have divided the flow into 4 topics of Teams Chatter, Goals/Touchdowns during the game Chatter,Advertisements and Celebrities. Each topic is analyzed on a hour to hour basis. (e.g. Any tweet about Missy Elliot goes to celebrities, while any tweet about T-Mobile goes into Ads. with classification based on one-hour window)
8. **Developing Event Flow : JSON -** Vincent library of python is used which generates a JSON file which has the indexes and data for every topic.
9. **Visualization -** D3 is used for final visualization. This is created using the JSON file created in the step above. The visualization can be seen below for all the three categorization: Event Sequencing, Brand comparison and Celebrity Popularity.

Event Flow Time Series

Figure 9 shows the flow of events between **19th January, 2015 to 7th February, 2015**. The visualisations are based on tweets collaborated over an hour. From the advertisements line graph, we see that, the **advertisements** have been a constant part of the whole event. It peaks to an extent during the SuperBowl finals in February. This can be attributed to the advertisements that are **telecasted during the half-time of the finals**. This is visible in the graph where there are sudden small peaks.

Similar to advertisement the **celebrities** has the maximum peak achieved on February 1st. This makes sense as it was the finals. Also, there were **multiple performances on that day** which can be inferred from the celebrities time series given in the following section. Thus the peak on 1st February defines the performances like those of Katy Perry, Missy Elliot and Idina Menzel.

The **orange colour from goal chatter** shows the peaks which represents every time there has been **goals or touchdowns**. Thus, the sudden peaks on graphs are during the games. If we were to change the date range to a particular game, it can give the distribution of the number of goals per hour.

Finally, the **team chatter** are the tweets which talks about the teams **Seattle SeaHawks and New England Patriots**. During the SuperBowl, the tweets are only filled with these two teams which is visible on the final distribution. On January, 19 we see a peak for the team chatter. This is because Seattle SeaHawks was playing a game which it won which is well explained by the peak.

The overall peaks during the finals shows how popular SuperBowl is. The peak in Advertisements show the extent to which advertising agencies are willing to be part of the SuperBowl and how much impact these Ads can have.

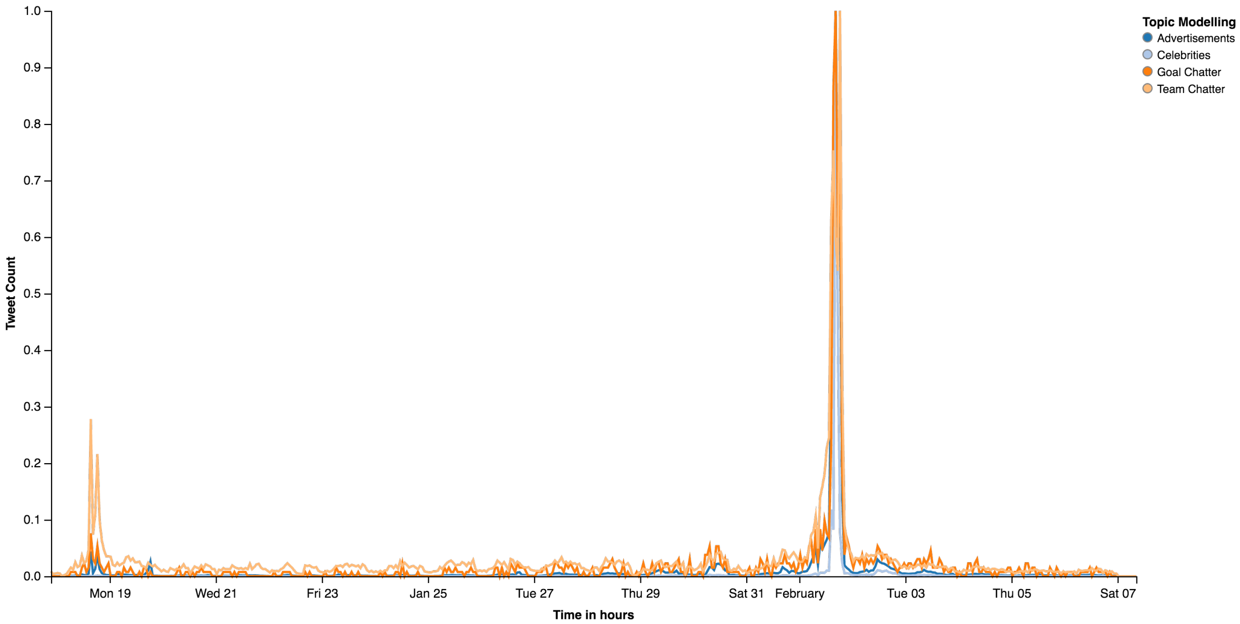

Figure 9: Time Series for Event Flow

Brand Comparison

Figure 10 shows the **popularity of different brands** during different periods of time of the SuperBowl. The brands seen in this graph are **Coca Cola, Budweiser, Doritos, McDonald's, Snickers, T-Mobile and Toyota**. From the graph we can clearly see that the dominant brands are T-mobile, Toyota and Snickers. While there has been some peaks before and after the SuperBowl for Doritos and T-mobile, the main advertisement impact is seen on during February 1-3. The twitter data basically visualizes the active marketing done during SuperBowl.

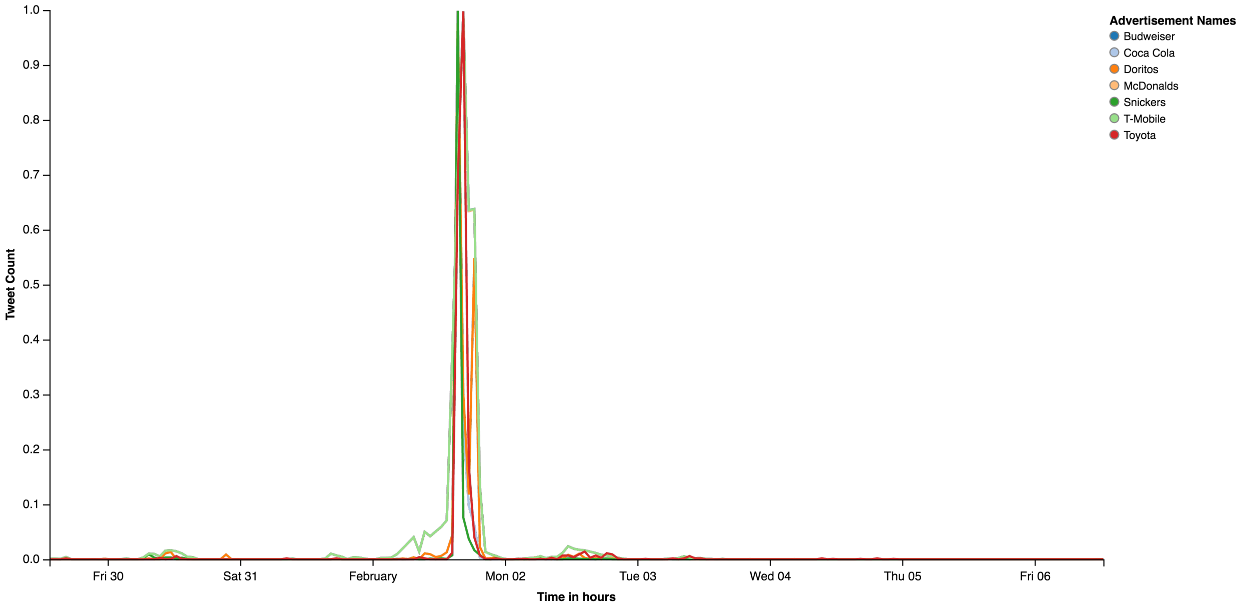

Figure 10: Time Series for Brand Comparison

Celebrity Popularity

Figure 11 shows the **popularity of celebrities** with respect to the SuperBowl. During the main event, we see maximum impact by **Katy Perry** from the visualizations. This is validated by the fact that she had a performance during half time. Also, there was a surprise performance by Missy Elliot, which can be seen from the light green line in the graph. Idina Menzel and John Legend performed before the game. Thus their peaks are shown slightly before Katy Perry and Missy Elliot. We can clearly see from the visualizations that the audience was majorly captivated by Katy Perry's performance.

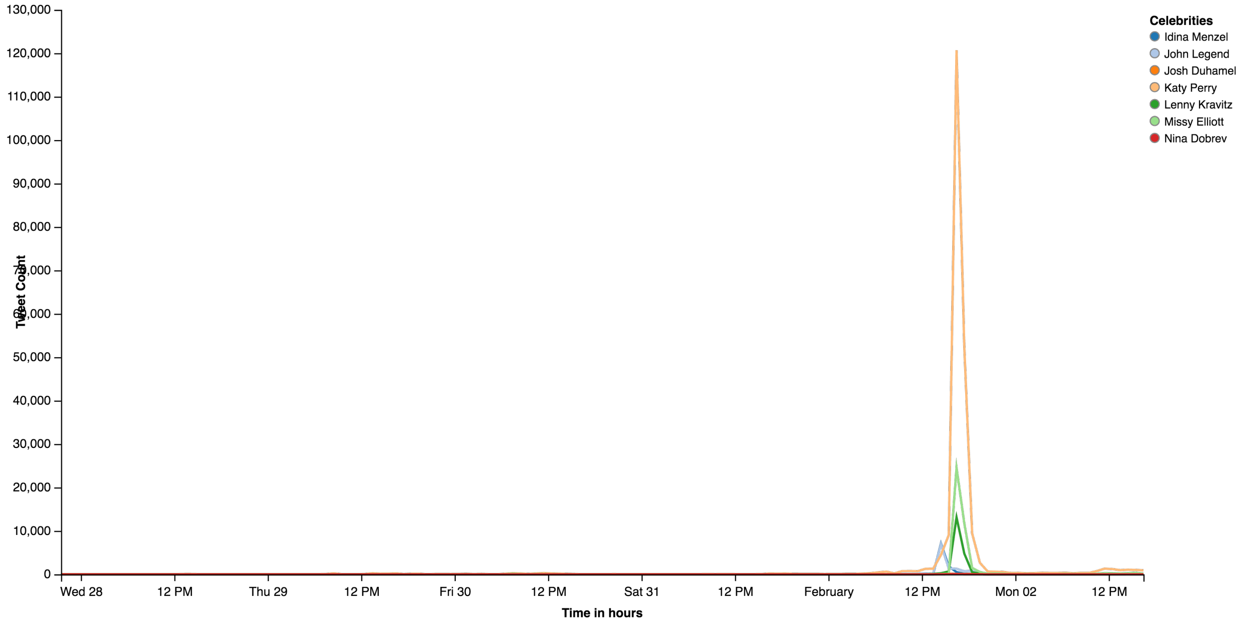

Figure 11: Time Series for Celebrity Popularity

Conclusion

As proposed a **Event-Sequencing and Ad-Celeb Popularity/Comparison Checker** was implemented and the results were presented above. The scope of the problem can be further be spread into areas of analytics for advertising agencies and for the celebrity PR teams. Sentiment analysis of the tweets collected can further represent the feelings of an advertisement or a celebritiesṕerformance during the SuperBowl.