

EE 239AS
Special Topics in Signals and Systems
Project 4
Popularity Prediction on Twitter
Winter 2016

Liqiang YU, Kaiming WANG and Jun FENG
904592975, 504592374, 304588434

03-18-2016

Contents

1	Introduction	3
2	Data Analysis	3
3	Description of the emotion change : Sentiment Analysis	3
3.1	Problem Statement	3
3.2	Data Preprocessing	5
3.3	Pointwise Mutual Information	5
3.4	Results and Discussions	6

1 Introduction

The Twitter website, as the most famous social network, is a good source to predict future popularity of a subject or event. In this project, we analyzed the data from twitter crawled during the period of 2015 superbowl, from two weeks before the game to one week after the game. We trained different regression models for tweets with different hashtags to predict the number of tweets in the next hour. We included features both from the tutorial and our own design. And we used the test set to evaluate our model's prediction results. Finally, we came up with a new idea based on the rich data from the twitter, that is try to describe the emotion change of fans from both teams during and after the game, based on the contents from their tweets.

The report is organized as follows : in section 2, we did a quick scan through the file and got some statistics information about the data....

2 Data Analysis

We have six hashtags for training : #gohawks, #gopatriots, #nfl, #patriots, #sb49, #superbowl. The data information is shown in table 1. The distribution for #superbowl and #nfl are shown in figure 1 and 2. We only consider the time period of interest, which is two weeks before Feb 1st, 2015 and one week after it. From the histograms we can see that both hashtags concentrated during the game time, especially for #SuperBowl. For #NFL, there was another peak during the last weekend before the superbowl final game, which showed that people would like to talk about the NFL game during weekends.

Table 1: The data information for each hashtag

	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Average number of tweets per hour	380.84	53.43	515.98	975.52	1647.31	2692.14
Average number of follower per users	1544.97	1298.82	4289.75	1650.32	2235.16	3591.60
Average number of retweets per tweet	2.01	1.40	1.54	1.78	2.51	2.39

3 Description of the emotion change : Sentiment Analysis

3.1 Problem Statement

When we get access to millions of twitter contents, what we are often interested in is what did the users say and what were their emotions at that time. This

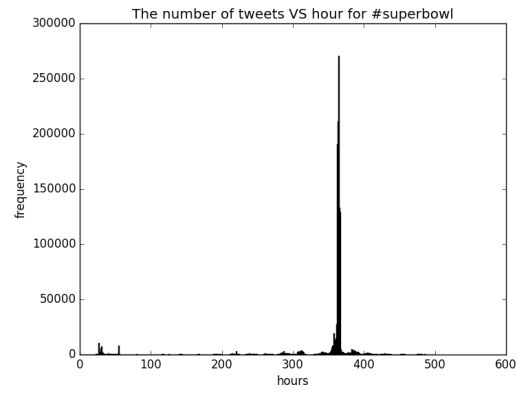


Figure 1: The histogram for #SuperBowl

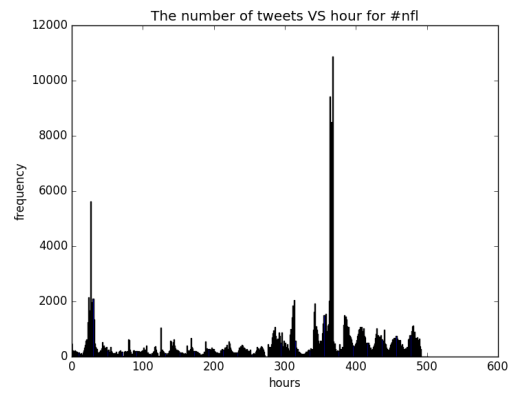


Figure 2: The histogram for #NFL

leads to the sentiment analysis and can be beneficial for lots of areas. For instance, during the president election campaign, the candidates want to know what the people’s reactions are towards their recent lecture or debate, which gives them instructions on how to improve it in the future.

Therefore, we tried to detect the emotion changes from fans of both teams during and after the final game with the tweet data we have. That is, we want to know whether we can detect more happiness from tweets of patriots fans and more sadness from fans of hawks.

3.2 Data Preprocessing

We used the ‘tweet’ metadata in json files, which represents the contents of that tweet. However some preprocessing steps need to be done before making the sentiment analysis. The steps include:

1. Tokenize the tweet text, which includes making all letters lowercase, getting rid of stop words and punctuations, and tokenizing the text with regular expressions which match common words, hashtags, urls, user mentions and simple emotion symbols like “:)”.
2. Calculate the term frequency for each term appeared in the tweet contents.
3. Calculate the term co-occurences for each term appeared in the tweet contents.

3.3 Pointwise Mutual Information

In this project, we define the Semantic Orientation(SO) of a word as the difference between its associations with positive and negative words. In practice, we want to calculate how close a word is with terms like *good* and *bad*. The chosen measure of closeness is Pointwise Mutual Information(PMI).

$$PMI(t_1, t_2) = \log \left(\frac{P(t_1 \cap t_2)}{P(t_1) * P(t_2)} \right)$$

Then the SO of a word is calculated against positive and negative terms. Let’s define V^+ a set of positive words and V^- a set of negative words, the SO of a term is defined as

$$SO(t) = \sum_{t' \in V^+} PMI(t, t') - \sum_{t' \in V^-} PMI(t, t')$$

We define the Document Frequency(DF) of a term as the number of documents where the term occurs, so the probabilities are defined as:

$$P(t) = \frac{DF(t)}{|D|}$$

$$P(t_1 \cap t_2) = \frac{DF(t_1 \cap t_2)}{|D|}$$

3.4 Results and Discussions

We choose to use #gohawks and #patriots as the support hashtag for two teams because the amount of data in #gopatriots is too small. We split the data into one hour period and only consider the 24 hours around the game. We plot the semantic orientation every hour for the twitter contents from two teams fans and make the comparison.

The results are shown in figure 3. From the figure we can see that at first seahawks fans are more confident than patriots fans, however there was a turning point at around 6:30pm, which is 3 hour after the game start. When we enlarge that area and show it in figure 4, we can see that after that turning point patriots fans are much happier than hawks fans. After referring it to the game facts, we think it makes sense because it describes a story of bouncing back from behind and win the championship.

The table 2 shows the game facts. From the game facts, we can see that at 190 minutes, which is nearly 3 hours after the game start, the hawks lead the patriots by 24:14, however after that the hawks got no point anymore and the patriots turned back and won the superbowl. Therefore it explained why hawks fans started to feel unhappy at the time 3 hours after the game start and why patriots felt happy.

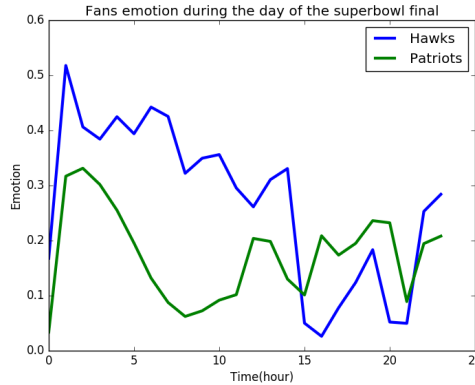


Figure 3: The emotion change during 24 hours

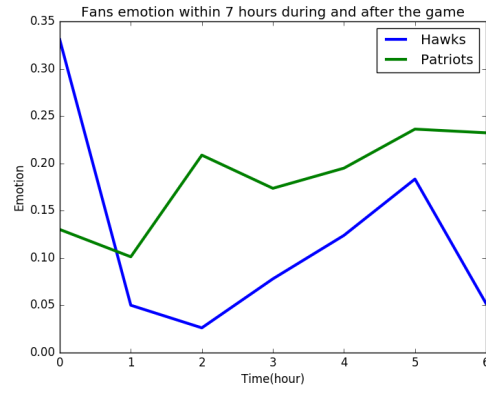


Figure 4: The emotion change from 2 hours to 8 hours after the game start

Table 2: Game facts for 2015 superbowl

Minutes after the game start	Scores(Patriots : Hawks)
88	7:0
105	7:7
120	14:7
130	14:14
175	14:17
190	14:24
227	21:24
243	28:24
END	28:24