Event Identification and Sentiment Detection with the 2018 World Cup Twitter Textual Data

Xurui Chen (xc1454@nyu.edu) | Yunhe Cui (yc3420@nyu.edu)

Word Count: 2583

Introduction

More than 4 million people in the world consider themselves football fans which makes football the world's most popular sport. According to the survey conducted by Nielsen in 2017 (Nielsen, 2018), over 40% interviewees said they were "interested" or "very interested" in football which made it well ahead of nearest rival sports (i.e. 37% for basketball). Nowadays, football has a strong impact on people's life and their culture rather than simply being a kind of sport. For example, in UK, football is frequently linked with politics and national pride.

Since there are thousands of leagues in the world (professional, semi-professional, amateur, youth and more), it is impossible for us to conduct a general study. In this case, we decided to conduct our sentiment detection and event identification on twitter data, and it is important for us to obtain a sufficient number of sample to dig out meaningful patterns. Thus, we need to choose a match that owns a large number of audience and is widely discussed. World Cup, one of the games gains the widest attention (over 3.5 billion viewers were engaged during the 2018 World Cup (Reed, 2018)), will be our project target.

Twitter is a popular micro-blogging service used for social interaction online which allows any registered user to publish. The posts are limited in 140 characters and are called tweets. During the 2018 World Cup, Twitter held over 115 billion relevant tweets and the most-Tweeted matches are shown as follows (Bavishi, 2018). In this project, we are interested in finding out the quantitative relationship between textual tweets and the matches using Latent Dirichlet Allocation (LDA) model. Additional to the event detection, we will also focus on sentiment analysis during the day of the final to see the emotional change in audiences.

Literature Review

Research works have been carried out for sentiment analysis over decades as well as event detection using social media post. Yang and Zhang (2018) used both LDA topic mode and sentiment analysis (using syuzhet package) to discern the most popular topics in the movie-related twitter data and extract as well as analyze hidden useful information from them. They compared their result with reviews aggregated from other websites such as IMDb and Rotten Tomatoes and found that they got a similar emotional tendency.

Jai-Andaloussi et al proposed a framework for automatic textual soccer summarization based on real-time sentiment analysis of twitter event using machine learning and KDD processes which enable them to extract the knowledge from data in large databases (Jai-Andaloussi, Mourabit, Madrane, Chaouni, and Sekkakim, 2015). Their research outcome allows them to verify the feasibility and efficiency of soccer events summarization through sentiment analysis.

Data and Events Brief

We retrieved our research dataset from Kaggle (Kaggle, 2018) and the glance of raw data is shown in Figure 1. Kaggle team used API to download the English tweets containing any references to FIFA or the World Cup in the date of June 29th - July 04th, and July 10/11/15th. They also did some pre-processing to facilitate further analysis while trying to keep the original tweet information. To be specific, they removed nuances such as website name, hashtags, and special characters with “BeautifulSoup” and “regex” libraries. Furthermore, the abbreviations (i.e. “I’ll”) were transformed into their proper English language expressions.

On the basis of their, we then removed stopwords from the dataset using Quanteda (R library) English stopwords dictionary. We created a word cloud plot (Figure 2) for Hashtags to get the quick image of what are the topics that drew the most attention. Next, based on the tweet word frequency, we generated another word cloud plot (Figure 3).

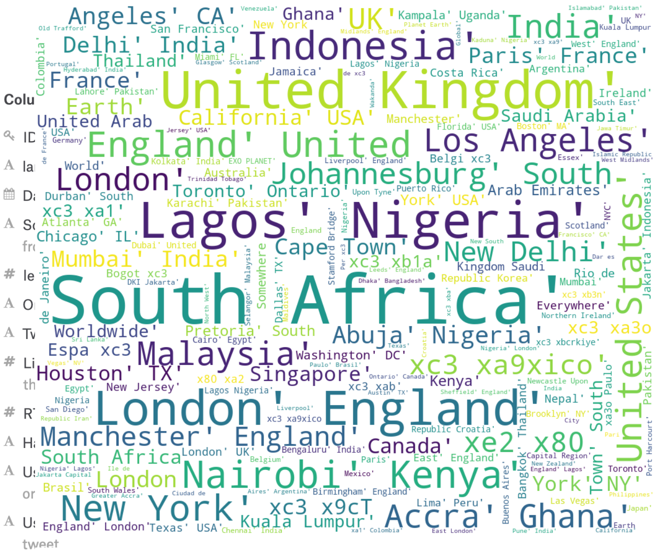


Figure 2 Hashtag Word Cloud

Figure 1. Kaggle Dataset Glance

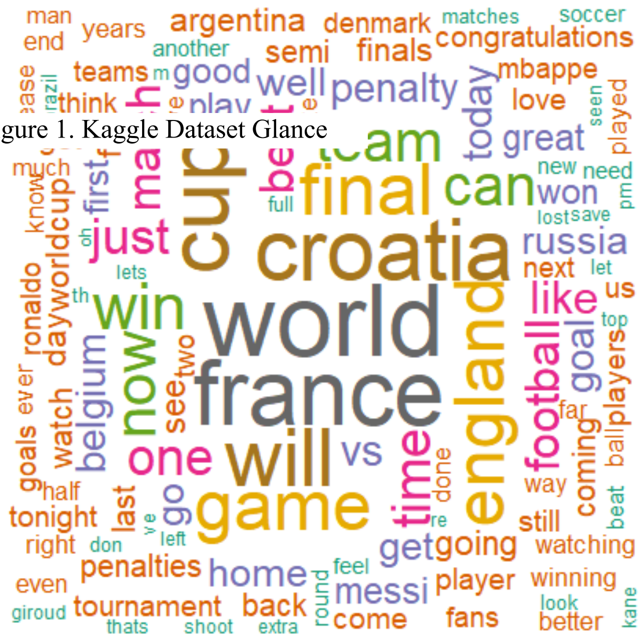
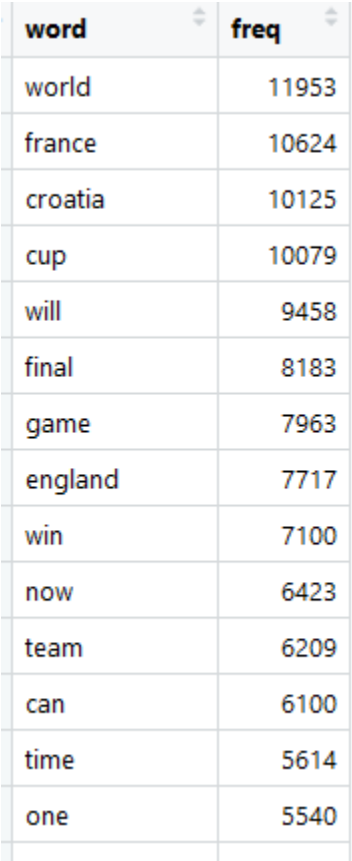


Figure 3 Tweet Word Frequency Word Cloud



word	freq
world	11953
france	10624
croatia	10125
cup	10079
will	9458
final	8183
game	7963
england	7717
win	7100
now	6423
team	6209
can	6100
time	5614
one	5540

Figure 4 Tweets Word Frequency

It is not surprising to see France and Croatia appeared frequently (both over 10,000 times in the data) since they are champion and second-place for the 2018 World Cup (Figure 4). Finally, for our next-step analysis, we removed punctuations, numbers, and emoji for our textual analysis and turned all words into lower case.

Since the research topic is event detection and sentiment analysis, the match-related information such as when is the match, what are the two teams and what is the score are necessary and that information served as our reference (Table 1).

Date	Match Day？(Y/N)	Match Type	Team(s) and Score(s)
June 29	N	N.A	N.A
June 30	Y	Eighth-final	France 4: 3 Argentina Uruguay 2:1 Portugal
July 01	Y	Eighth-final	Spain 1: 1 Russia Croatia 1:1 Denmark
July 02	Y	Eighth-final	Brazil 2:0 Mexico Belgium 3:2 Japan
July 03	Y	Eighth-final	Sweden 1:0 Switzerland Columbia 1:1 England
July 04	N	N.A	N.A
July 10	Y	Semi-final	France 1:0 Belgium

July 11	Y	Semi-final	Croatia 2:1 England
July 15	Y	Final	France 4:2 Croatia

Table 1. Brief Event Introduction

Hypothesis

Our research hypothesis is that using the LDA model, we could find out the topic model embedded in the social media post. In addition to that, we could detect the change in people's emotion and enable automatic event summary with the tested model. With the reference of other resources, we could then obtain the quick image of the event and know the reaction of people.

Theory and Model

We applied LDA for topic analysis and event detection. LDA is a text model method proposed by Blei in 2003 (Blei, Ng, and Jordan, 2003). The precondition of LDA is that the feed-in data should be a collection of words while the grammar and word order are ignored. In LDA, documents are seen as a distribution and the topics are distribution over words (Juan, Tian, Li, Zhang and Sheng, 2009). The LDA based topic model is used to cluster the collection of tweets to find important categories of major match event in an unsupervised way and then associates each tweet with a single topic.

In this project, we did not use LDA and TF-IDF together. As LDA is a probabilistic model which estimates probability distributions for topics in documents and words in topics, the TF-IDF looks unnecessary. Besides, according to Blei's paper in 2003 (Blei, Ng, and Jordan, 2003), LDA addresses the shortcomings of the TF-IDF model and leaves this approach behind.

We then randomly selected 5000 data samples from our dataset to reduce computation time as well as memory cost. Then we compared three evaluation indexes in terms of topic number: CaoJuan, Griffiths, and Deveaud. The data sample LDA model evaluation plot is shown below in

Figure 5. We decided to use 16 as our topic model number, since, at this point, Cao has a relatively low value while Griffiths has a relatively high one. The top 10 words for 16 different topic models and the corresponding topic label are shown in Figure 6 and Table 2 separately. We also noticed that the hot topic changes dramatically in the duration of June 29 - July 15 because the new hot topics came out (or burst) near or during each game and people's interests change accordingly. Figure 7 shows top 3 models for the research dates while Figure 8 shows the change of tweet count in topic model over time.

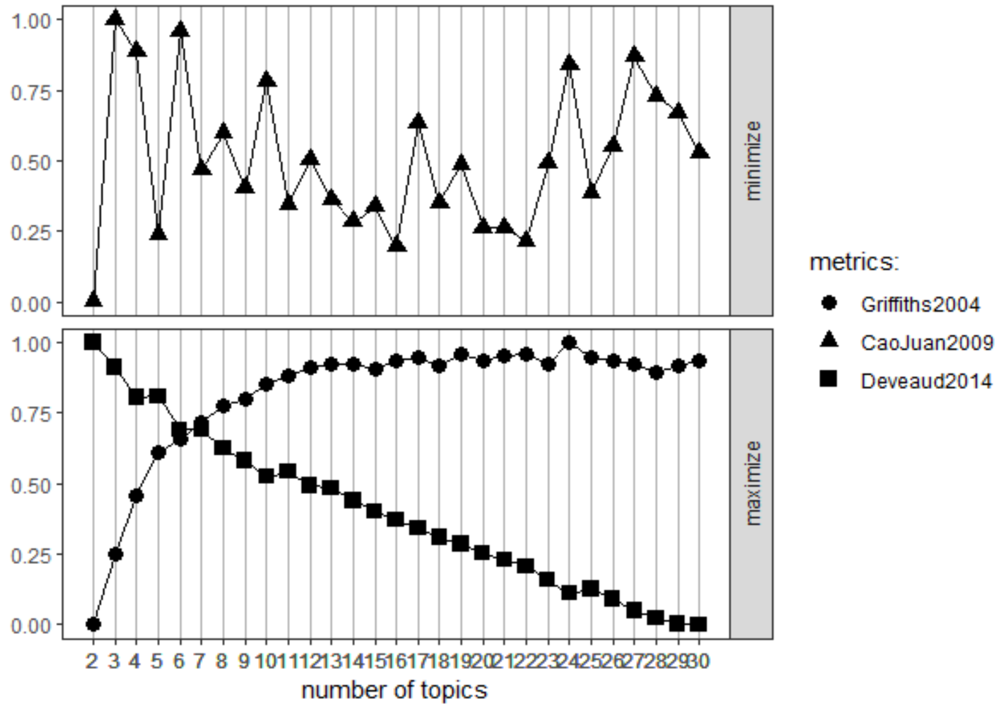


Figure 5 Topic Number Evaluation

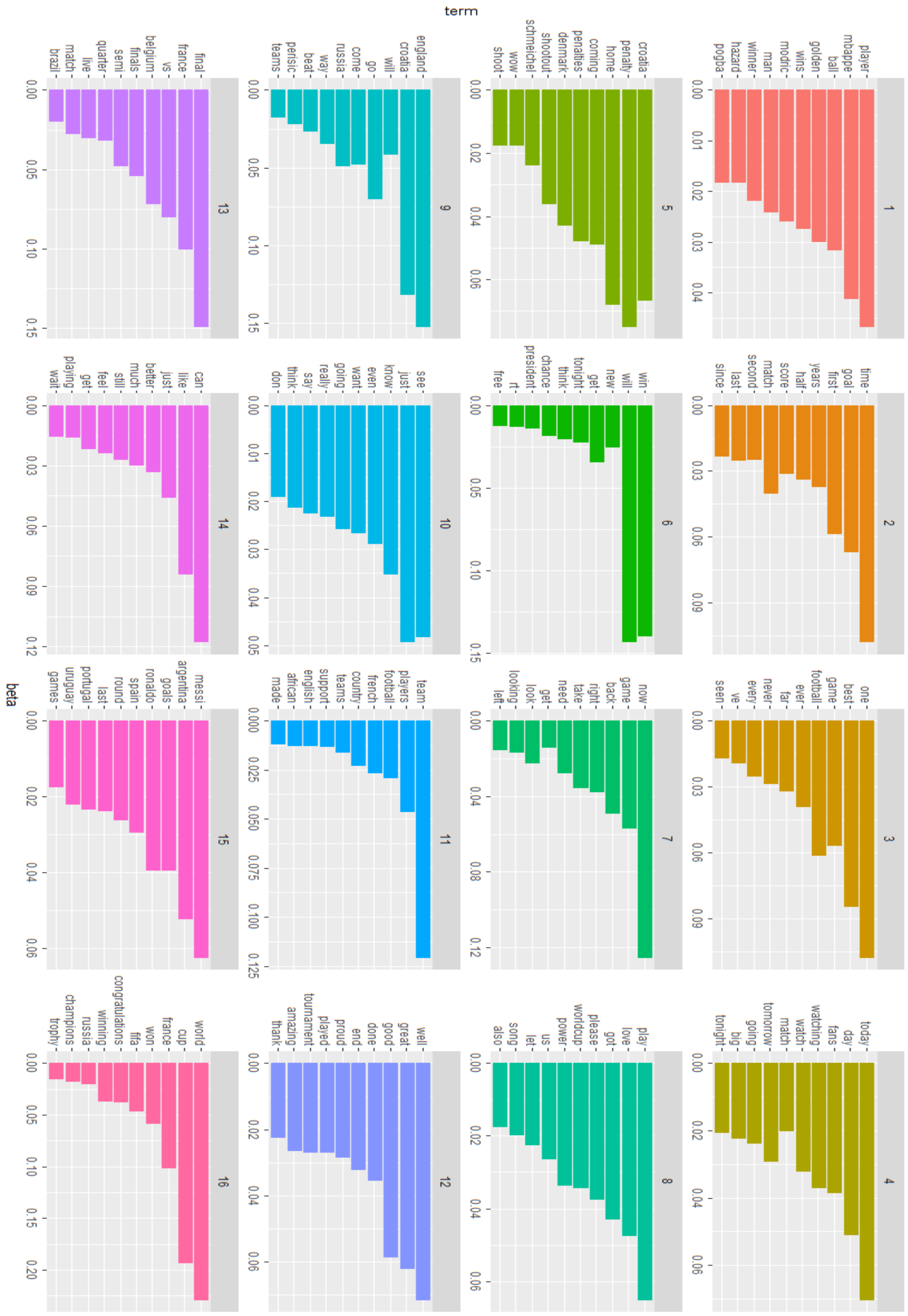
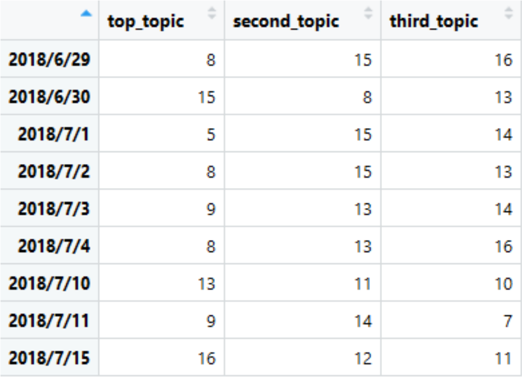


Figure 6 Top 10 Words for 16 Topics

Topic Number	Label	Topic Number	Label
1	Final Game	9	Croatia and England Match
2	Match Time	10	General comment on the match
3	Best Game	11	French
4	Expectations	12	Praise of Games
5	Croatia and Denmark Match	13	Match Schedule
6	Celebrate	14	Disappointment
7	Tactics	15	Messi and Ronaldo
8	**World Cup Lover**	16	French Champion



	top_topic	second_topic	third_topic
2018/6/29	8	15	16
2018/6/30	15	8	13
2018/7/1	5	15	14
2018/7/2	8	15	13
2018/7/3	9	13	14
2018/7/4	8	13	16
2018/7/10	13	11	10
2018/7/11	9	14	7
2018/7/15	16	12	11

Figure 7 Top 3 topic models

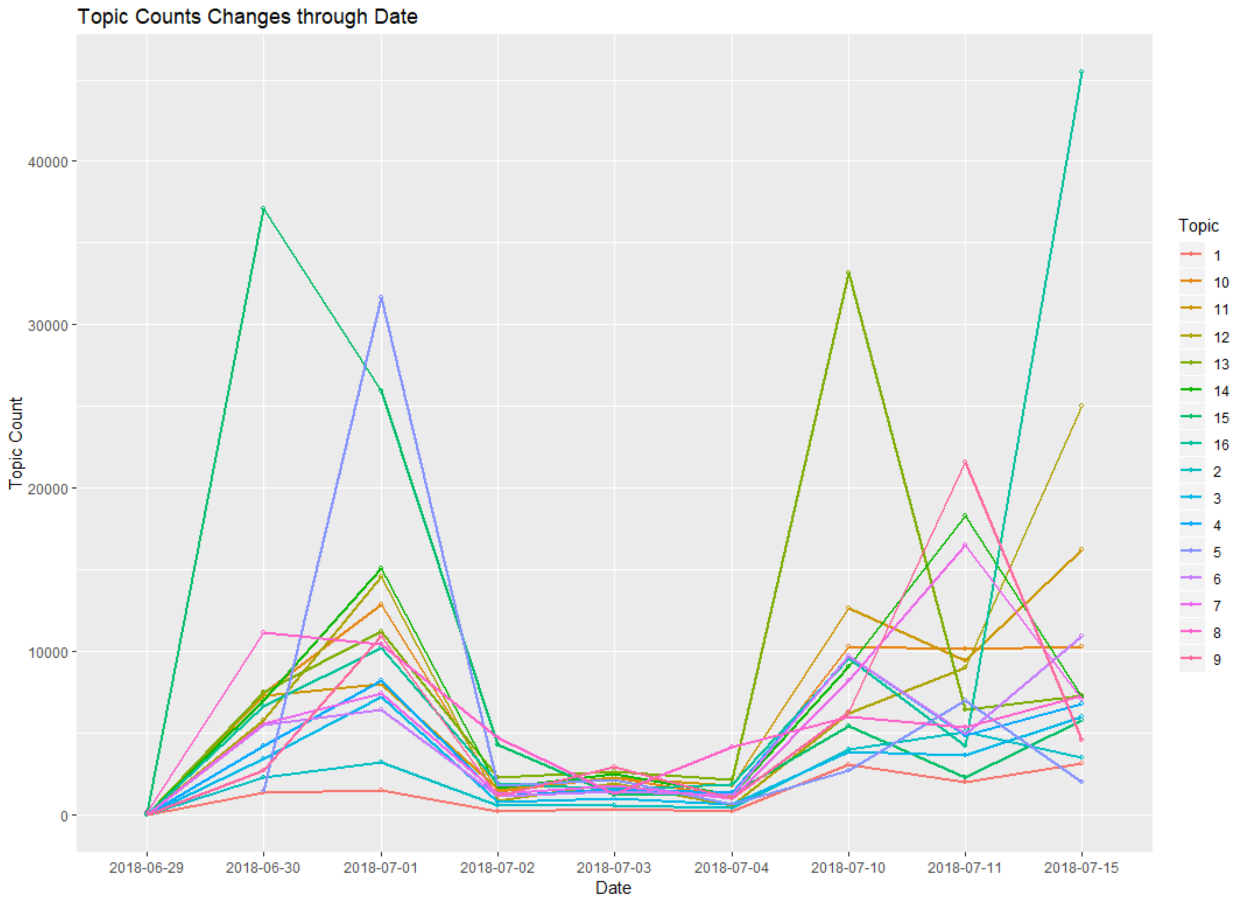


Figure 8 Topic Model Count Change, June 29 to July 15

For the sentiment analysis, we used the positive/negative words dictionaries that discussed in the Hu and Liu's (2014) customer review research. The sentiment score is calculated using the formula in Figure 9. We extracted the subset of tweets that posted on July 15 (the final match day) both because of the higher tweet density (highest tweets day count in our research date range) (Figure 10) and potentially heightened emotion changes. The final between France and Croatia delivered many of twists, turns, riveting moments, individual superlatives and extraordinary overall performance for both teams and it was thought to be the one that would enter the archives (probably is one of the best in the World Cup history).

$$\text{tone of document } i = \sum_{m=1}^{M} \frac{s_m w_{im}}{N_i}$$

Figure 9 Sentiment Formula

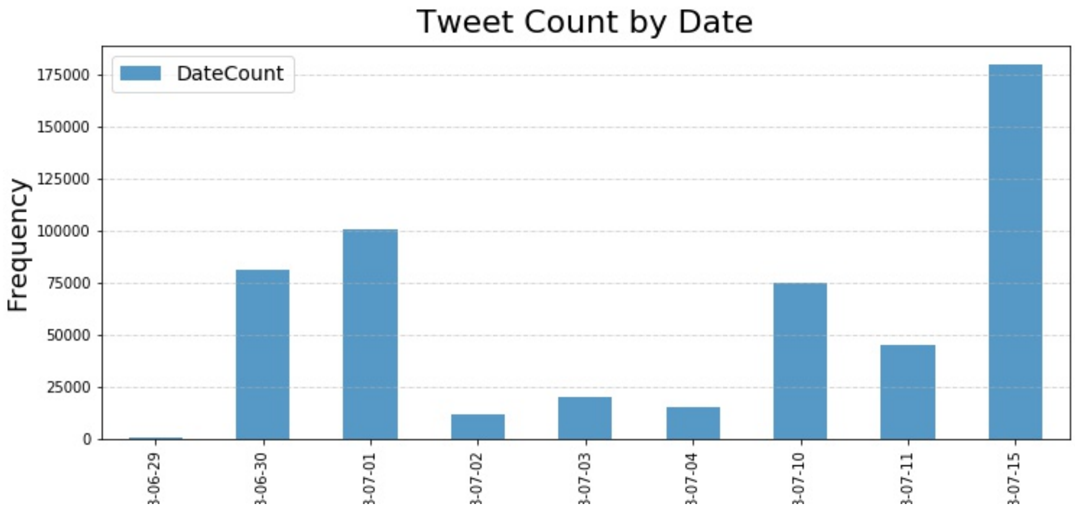


Figure 10 Tweet Day Count, June 29 to July 15

From the plots that show sentiment analysis result (Figure 11 and Figure 12), we could see that people were generally calm and the positive and negative emotion expressions were at a relatively low level. People were getting increasingly excited when the time approached the match start. After the match, we could notice that the positive emotion soared in a short time period while negative ones remained stable which reflected that people were most satisfied with the match result.

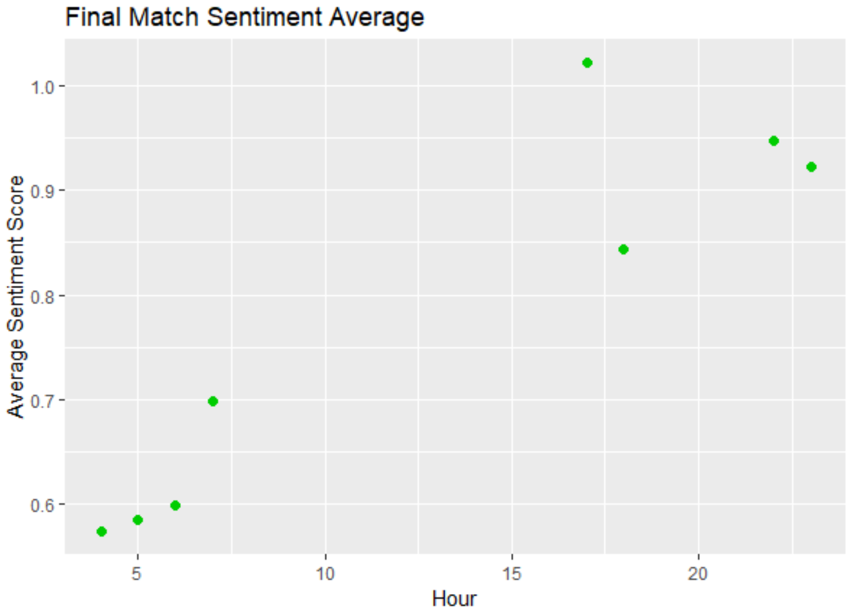


Figure 11. Overall Sentiment Count

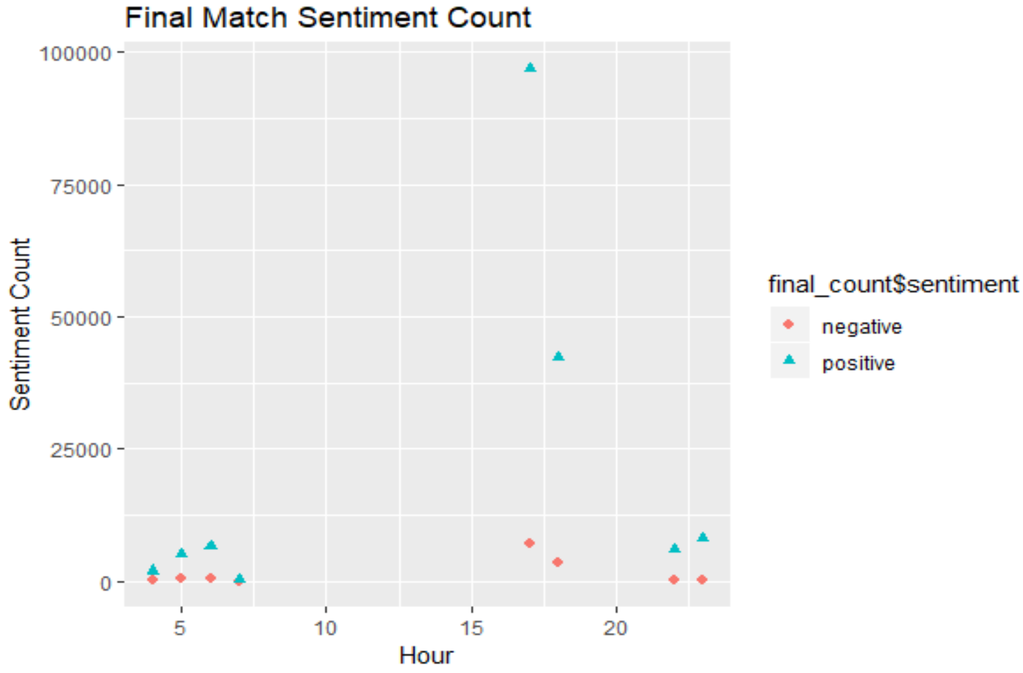


Figure 12 Sentiment Count for positive and negative words

Result and Discussion

Based on the top topics for each day, we could extract the potential football event (or match result) based on known facts. Here we took several examples to show how we related topic and event. As a “control group”, we could see that there is no significant high topic caught in no match days (June 29 and June 04). On June 30, the count of topic 15 is much higher than others. After checking the event calendar, we saw that two games of France vs Argentina and Uruguay vs Portugal are scheduled at that time. Those two games gained relative high attention is because three favorites of Championship and some world’s best players (i.e. Lionel Andres Messi and Cristiano Ronaldo) are involved. Compared to the matches happened between traditional top teams, Croatia 2:1 England gives an unexpected result. Croatia was thought to be the dark horse of the World Cup, and after the match, people intended to talk more about their tactics (Topic 7) and generally think that England could perform better in the match (Topic 14). In other words, in the future, if we detect a peak value in tactics as well as “better performance”, combining with other techniques, we could give reasonable inference (i.e. an unexpected result of the match) and act accordingly.

The result of sentiment analysis allowed us to verify the feasibility and efficiency of soccer event summarization. In today’s world, people’s attention and interest are one of the determining factors for whether a merchant could make a profit through online media. Thus, event detection and sentiment analysis techniques could be utilized to help the merchant track hotspot and plan ads and push notifications to attract people’s attention and turn the traffic to cash. For the media platforms, sentiment analysis could also help them to better understand their user’s taste so that they could adjust their push to attract existing users constantly.

The combination of event detection and sentiment analysis in football-related topics enable us to quantify the social presence and the level of public attention toward a certain football star and/or a team. Social presence, in today’s world, means brand value and profit.

Limitation and Future Work

In this research project, we only used English language tweets as our research objects. However, as the World Cup audiences are from all over the world, it is unrealistic to expect everyone tweets in English. In the plot below (Figure 13), we could see that there is only slightly over 50% of World Cup tweets are posted in English. Thus, this research on event detection and sentiment analysis is limited. In the future, we should expand our range of study to extract a less biased conclusion.

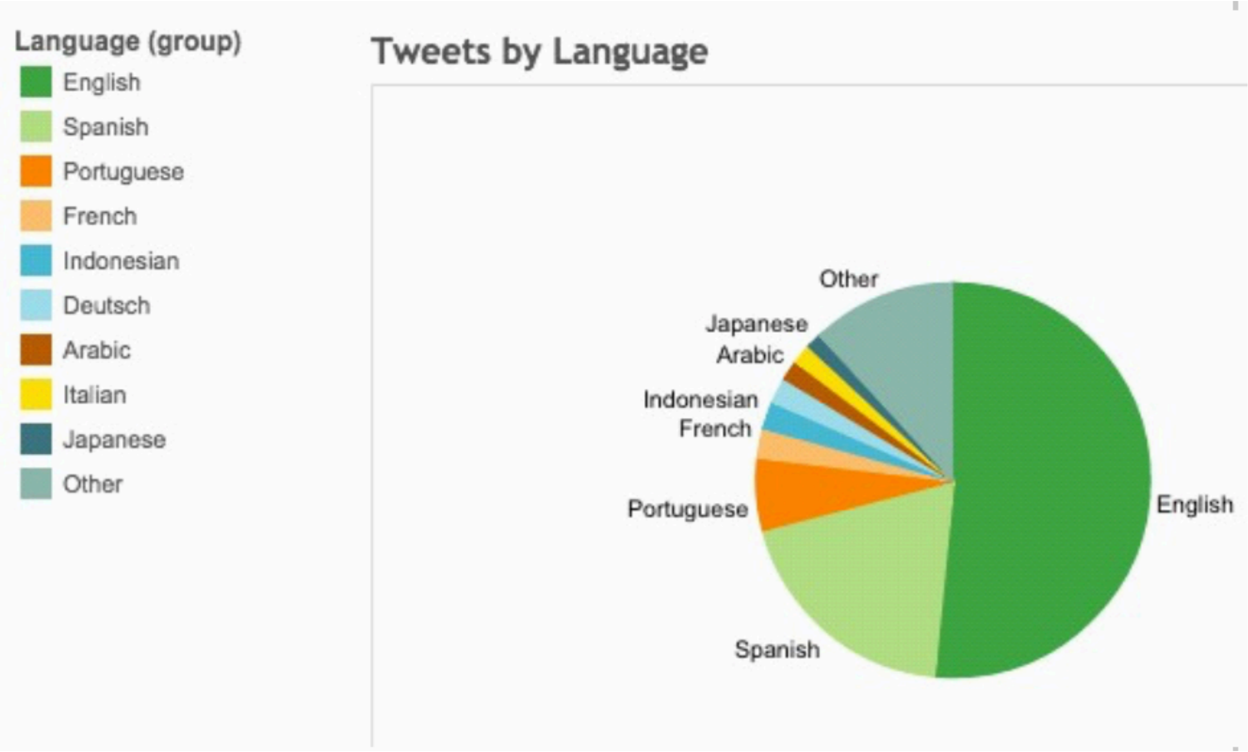
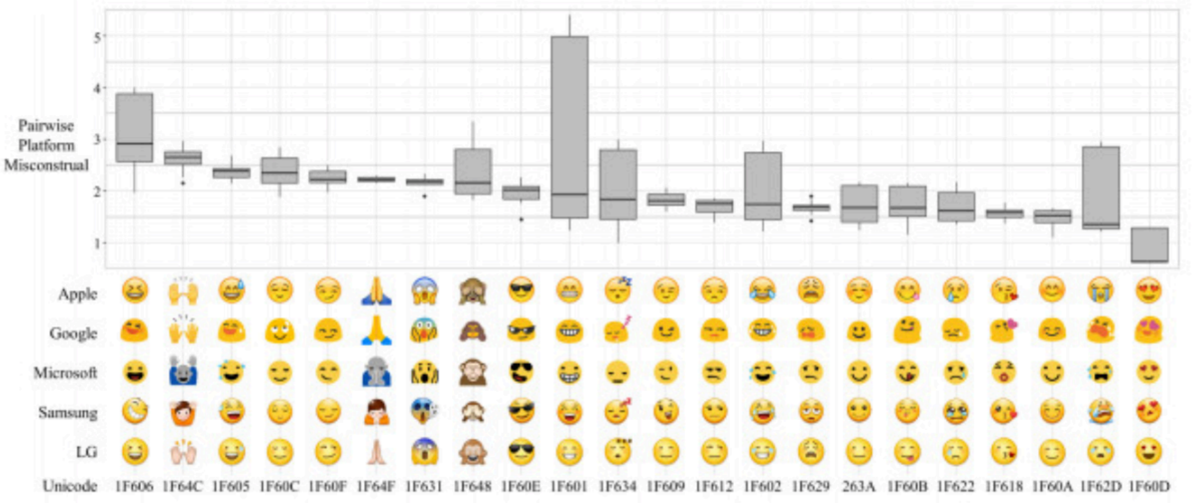


Figure 13 Tweets by language (Ghaffari, 2015)

In this project, we only considered the textual tweet data and neglected the embedded emojis. Emojis are initially introduced as expressive, non-verbal add-ons to the textual contents, mirroring the role played by facial expressions (Pozzim, 2018). In the new era of online visual communication, especially with micro-blogging such as Twitter, emoji becomes a fast and clear "language" that allows users to communicate with others globally, regardless of language and culture barriers. Emotion detects and analysis could be more precise with the emoji analysis involved. However, the emoji meanings changed as time goes by and the emojified text might be drastically misinterpreted (Figure 14). Even worse, in the case that the emoji rendering selected by the sender is exactly the same as the recipient sees, there is still a high probability of misconstruction. Thus, there will be a challenge for the researcher to dig out the true emoji meaning and apply the emotional analysis with limited misinterpretation.



Emoji interpretation differences, ranked by decreasing misconstrual scores from left to right. | GroupLens

Figure 14 Emoji interpretation differences, (Henwood, 2016)

Even for the pure textual information analysis, we should admit that this model is oversimplified the situation. For example, we considered all the text as same "weight" in terms of emotional analysis. However, the use of lengthening words and upper-class words in twitter and other micro-blogs is now pervasive and frequent. It is clear that "This match is so great!" and "This match is SOOOOOOOOOO GREAT" tweets are emotionally different even they convey the same attitude toward a certain event. Although the different emotion is straightforward to people, the researchers should make more efforts in optimizing relevant algorithms (Brody and Diakopoulos, 2011).

Reference List

Bavishi, J. and Filadelfo, E. (2018, July 17), Insights into the 2018 #WorldCup conversation on Twitter. Retrieved from https://blog.twitter.com/en_us/topics/events/2018 /2018-World-Cup-Insights.html

Blei D.M., Ng A. Y., and Jordan M. (2003, March). I. Latent Dirichlet Allocation. Journal of Machine Learning Research[J]. 993 – 1022

Brody, S. and Diakopoulous, N. (2011, July). N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. Retrieved from https://aclweb.org/anthology/pa pers/D/D11/D11-1052/

Cao, J., Tian, Xia., Li, J., Zhang, Y., and Sheng, T. (2009, March). A density-based method for adaptive LDA model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008* 72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

Ghaffari, P. (2015, January). Data Mining and Text Analytics of World Cup 2014. Retrieved from https://www.kdnuggets.com/2015/01/data-mining-text-analytics-world-cup-2014.html

Henwood, B. (2016, April 13). Same face, different meaning: a new study reveals how people interpret emoji [Web log post]. Retrieved from https://www.vox.com/2016/4/13/11422886/emoji-interpretation-different. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008*72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

Hu, M. & Liu, B. (2004). Mining and SummarizingCustomer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22–25, 2004, Seattle, Washington, USA*

Jai-Andaloussi, S., Mourabit I. E., Madrane, N., Chaouni S. B., and Sekkakim, A. (2015). Soccer Events Summarization by Using Sentiment Analysis, Retrieved from https://american-cse.org/csci2015/data/9795a398.pdf

Kaggle, (2018), FIFA World Cup 2018 Tweets, Retrieved from https://www.kaggle.com/rgupta09/world-cup-2018-tweets

Nielsen. (2018, November 06). World Football Report [Web log post]. Retrieved from https://www.nielsen.com/uk/en/insights/reports/2018/world-football-report.html

Reed, A. (2018, December 21). Half the world's population tuned into this year's soccer World Cup. Retrieved fromhttps://www.cnbc.com/2018/12/21/world-cup-2018-half-t he-worlds-population-tuned-in-to-this-years-soccer-tournament.html

StackOverflow, (n.a). Necessary to apply TF-IDF to new documents in gensim LDA model? [Web log post]. Retrieved from https://stackoverflow.com/questions/44781047/necessary-to-apply-tf-idf-to-new-documents-in-gensim-lda-model/44789327#44789327

Pozzi, F. A. (2018, July 16). The role of emojis in sentiment analysis. Retrieved from https://blogs.sas.com/content/hiddeninsights/2018/07/16/role-emojis-sentiment-analysis/

Yang, S. and Zhang, H., (2018). Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*. V. 12, No.7, Retrieved from https://waset.org/publications/10009246/text-mining-of-twitter-data-using-a-latent-dirichlet-allocation-topic-model-and-sentiment-analysis