Analyzing Real-time Tweets to Generate Match Reports

Deepak Gautam, Sandra Gibson

ABSTRACT

Due to the nature of Twitter, tweets are able to provide real-time data for various world occurrences. We decided to focus on sports, since we believed that the tweets posted by the users watching sporting events can provide valuable insight to key events. Due to the wide range of sports that exist in modern world, we decided to limit our research to the most commonly followed game, soccer. In order to accomplish this, tweets were collected from Twitter's streaming API and placed into a MongoDB in which tweets were filtered by official game hashtags and team names. By performing a post-hoc analysis of spikes in the volume of tweets, we were able to detect match events such as goal scored, the scorer's name, the scorer's team, and penalties throughout the game. Likewise, we tested the use of sentiment analysis to further correlate the relationship of the tweets and overall match progress.

KEYWORDS
Social Media, tweet analysis, time series analysis, sentiment analysis

INTRODUCTION

Social media refers to the tools such as Twitter and Facebook that enable people to share, or exchange information in virtual internet communities. Everyday, the plethora of status updates are posted every day in such social networks. There are about 500 million tweets posted on twitter everyday which means about 5700 tweets per second [1]. Many of the tweets relates to the events people getting involved via various media sources such as television. Because of real-time nature of twitter, people instantly share their emotions, knowledge or information as soon as they have the information. There are numerous tweets that directly relate to the events such as election debates, sporting events like soccer game, natural disasters. The facts and figures posted in the form of Twitter status updates during such events have significant information and detail about the event itself. Such information can be useful in analyzing, predicting similar events or their details in future. The main talking points and details of such events can be summarized by analyzing the tweet posted during the time that event, the earthquake location detection is an example of this [2]. Sometimes, people even conduct such events through twitter itself where a twitter user can participate directly to the event source. The presidential debates via live tweets is an example of this [3]. If the details of the event can be presented by automated system, the information can be passed through to public more quickly without waiting for news articles or something similar.

In this research, we tried to build an automated system that detects main moments from a sporting event based on the tweets posted in relation to the event. Due to the wide range of sporting events that exist in the world, we had to limit our research on more specific, free flowing game, soccer. Soccer, being the most popular sporting event across europe and rest of the world [4], has significant sets of fans who share game moments via Twitter status updates. We tried to detect the moments of the soccer games that have been played mostly in England and other parts of europe. We used the following two approaches:

- With the team names and player squad that is playing a game, we analyzed the volume of tweets collected during the game that was being played to obtain information about the game.
- We analyzed the sentiment values for the tweets by dividing tweets into two halves for each team, and calculated sentiment values for each team to get overall result of the game.

RELATED WORK

A considerable amount of work on extraction of moments from the events exists in the field of natural disasters such as earthquake [2]. Another work was carried out in Twitter itself during the soccer world cup tournament that they simultaneously watched the soccer game of USA vs Belgium and the real-time tweet per second (TPS) graph plot of the related tweets and able to see the relation of tweet volume with the game moments [5].

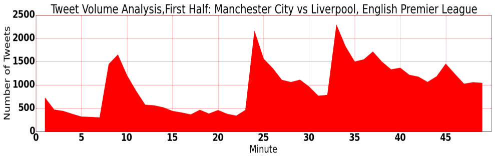
DATASET

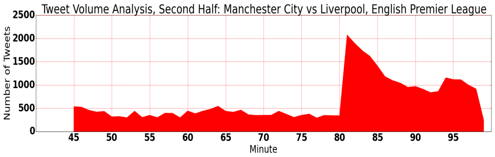
We used Twitter streaming API for collecting tweets. With its filtering feature based on keywords, we were able to collect tweets related to several English Premier Soccer League games and other European soccer games for analysis. Keywords were official game hashtags from the league's, team names, and the team nicknames (e.g. saints for Southampton football club of England) if any. The team starting lineup is also recorded from the involved teams' official twitter accounts for analyzing the player involvement in the game. The game kickoff time is also recorded for translating the timestamps into the minutes of the game for analysis. Initially, we collected most of the datasets without language tag that twitter assigns in every tweets. Later, we realized that we need language tag for both sentiment and volume analysis, we had to re-collect the datasets in later stages. This is why, we were only able to test our algorithm in six different games. Fortunately, these games had most of the moments that we wanted to detect.

We collected datasets into MongoDB from the streaming API. Then we converted each tweet entry in MongoDB into a row in csv. The volume analysis is performed on the csv files but for the sentiment analysis we had to restructure the tweets into smaller json files to get sentiments. The text commentary from FotMob application (www.fotmob.com) has been used to obtain actual game highlights which were needed for comparison to detected moments.

DETECTION OF IMPORTANT MOMENTS

The soccer game consists of several moments, and every moment is represented by actions from players and the team involved. The algorithm makes the use of Twitter users and their tweets to identify the moments within a soccer game. The algorithm captures the sudden increases, or "spikes," in the volume of the tweets, because such spikes indicate something important has happened in the game that the fans turned into Twitter to share their emotions and thoughts regarding the particular moment. Also there can be lots of minor rises in the volume of tweets, we have defined a threshold based on tweet per minute (TPM) of each game. Our post-hoc implementation of the algorithm runs after game finishes and all of the tweets are collected. So the threshold value could be calculated based on total tweets collected within the game period. We have marked "spikes," with slope greater than 15% of TPM as useful spike. This value is based on analysis and results we got from several games. This algorithm was also thought of developed to detect moments in real-time in that case the threshold should be improved as time passes based on the number of tweets collected in real-time.





Thus obtained spikes contain significant amount of tweets that explain the moments in the game. We created a vector of TF-IDFs [] of the words encountered in the tweets within each of the moments. We have used TF-IDF vectorizer because it accounts the importance of each word in a tweet, or in overall tweets related to the moment. We have used english stop words provided by scikit-learn library and added the keywords we used to filter the game tweets as our own game specific stop words. We also ignored punctuations. For a given moment, we calculated top ten important words based on the tf-idf values of the words. Thus obtained words if match to the predefined list of possible moments such as goal, penalty etc., then we declare the moment as it is matched, otherwise we throw it away as garbage moment. We also have predefined a list of players that are playing that specific game. If the detected top words fall in the players list, we mark that player as player of the moment. Usually, top words obtained match one of the predefined moments and one or more players. For multiple matches in case of players, we mark top most player as player of the moment and is responsible for the moment. For example, if moment matched is ‘goal,’, and there are multiple players match due to various reason such as remarkable pass that assisted the goal or remarkable mistake by opposition that made the team score. In such cases, rarely the algorithm fails on detecting the player of the moment. We get team name after we match the player detected to our predefined list, and increment the score if the ‘goal’ moment has occurred.

SENTIMENT ANALYSIS

We wanted to use sentiment analysis to see if the tweets about a specific team would coincide with that teams highlights for the game. We originally tried to use the sentiment analysis tool, Algorithmia. This tool was determined to not be a good fit for our needs based on the volume of tweets we would be analyzing. We were concerned that data would be lost because of Algorithmia’s request limits imposed.

We ultimately chose to use Sentiment140 for our sentiment analysis. Sentiment140 determines whether a tweet has a positive, negative, or neutral connotation. This API only supports English and Spanish language. With so many of the games in Europe, it was difficult to find games that had English and Spanish language. Our early data captures did not have a language parameter set, therefore this data needed to be discarded. After the new data was collected and the sentiment analysis was tested, it was determined that the analysis was not giving the results anticipated. We had very few tweets that showed positive or negative, and most tweets registered as neutral. The positive tweets ended up being nothing related to the soccer game. Sentiment140 did not yield the results we were hoping for.

EVALUATION

Our initial thought was to detect every moment of soccer games. But from the analysis of the tweets we collected, we were able to find that only top highlights can be detected through this approach. Top highlights included overall match result and score correctness, goal and scorer’s name, red cards, and penalty. To evaluate correctness of the algorithm, we needed list of actual game moments for each game. We took six games taken from European leagues and international friendlies. We collected them from the text match commentary of the games commented on FotMob (fotmob.com) application. We calculated precision and recall for each type of moments to evaluate the accuracy of our algorithm. Precision is the percentage of correctness among detected moments, whereas, Recall is the percentage of detected moments among all moments that occurred during the game. For example, precision of 0.95 in Goal detection means, 95% of the goals we detected are true goals and recall of 0.86 means we were able to detect 86% of the total goals.

Game	Tweet Count	Match Result	Goal Detected (FP)	Missed Goals
Real Madrid vs Barcelona	114289	0 - 4	3(1)	1
Chelsea vs Arsenal	119029	2 - 0	2(0)	0
England vs France	79381	2 - 0	2(0)	0
Chelsea vs Southampton	91972	1 - 3	3(0)	1
Manchester City vs Liverpool	83946	1 - 4	5(0)	0
Swansea vs Bournemouth	5368	2 - 2	3(0)	1

The above table shows the detected goals with false positives. Some of the goals were missed because there have been two or more goals in quick succession that our algorithm evaluated it as a one goal. Other missed goals are due to their less importance because of the game status the time the goal scored. Other wrongly identified goals were because of huge goal scoring chance or disallowed goals.

	Goal	Penalty	Red Cards
Precision	0.95	1	1
Recall	0.86	0.33	1

The overall precision and recall values for other type of moments are shown above in the table. We have less recall value for penalty because our algorithm detected most of the penalty appeals as penalty. The red card detection is perfect as it has both recall and precision value as 1.

	Result	Correct Score	Player Involvement
Accuracy	0.83	0.33	0.89

Overall results shown above are based on the goal detected and correctness of player involvement. It has been hard to detect own goals in terms of results because we detected goal scorer and team name to which scorer plays. It adds score to wrong team because it’s an own goal. Also the multiple goals within single spike has also affected the accuracy figures of overall results.

NOISE ELIMINATION

It is the fact that the Twitter data is the most noisy data among the social network because of its real-time nature. In Twitter, people want to share as soon as they can when they encountered anything special while following the event via other media sources. We encountered following types of noises in tweets:

- The tweets contain repeating letter within a word, this maybe due to utter excitement in users. We tried the algorithm by removing repeating letters, but the results didn’t vary much because of the presence of top words with correct spelling is significant.
- We tried volume analysis with the tweets with their language marked as English by twitter. We also analyzed by ignoring the language tag(which means - analysis on all the tweets). The results look better if only English tweets are considered.
- It was impossible to remove the garbage tweets that are collected because advertisers use game hashtags to advertise their products.
- The language tag assigned by twitter is not so reliable that it sometimes assigns English as the language of a tweet rather than actual language or null. This makes lots of garbage tweets appearing in analysis even though tweets are filtered by language.

CONCLUSION AND FUTURE WORK

Our volume analysis seems to capture most of the top moments in soccer games. We believe that if proper noise removal techniques applied, we could also detect minor events such as player substitutions, and yellow cards.

The sentiment analysis in tweets related to sporting event is difficult in a sense that the tweets are short in length and have lots of noise in them. The sentiment analyzer API, we chose, might be the problem for analyzing sentiments in tweets. Whether we can relate sentiments in tweets to the actual game moments is still a topic of research in future. The proper research on sentiment analyzer for such tweets is also the part of future research work.

REFERENCES

[1] Blog.twitter.com, 'New Tweets per second record, and how! | Twitter Blogs', 2013. [Online]. Available: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how. [Accessed: 04- Dec- 2015].

[2] T. Sakaki, M. Okazaki and Y. Matsuo, 'Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development', *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919-931, 2013.

[3] L. Moraes, L. Moraes, L. Moraes, L. Moraes and L. Moraes, 'Donald Trump Joins CNN's Democratic Debate Via Live Tweets â€“ Update', *Deadline*, 2015. [Online]. Available: http://deadline.com/2015/10/donald-trump-democratic-debate-cnn-hillary-clinton-twitter-1201579052/. [Accessed: 05- Dec-2015].

[4] Topendsports.com, 'World's Most Popular Sports by Fans', 2015. [Online]. Available:http://www.topendsports.com/world/lists/popular-sport/fans.htm. [Accessed: 06- Dec- 2015].

[5] Forbes.com, 'Forbes Welcome', 2015. [Online]. Available: http://www.forbes.com/sites/jeffbercovici/2014/07/02/watching-team-usas-swan-song-from-twitters-world-cup-nerve-center/. [Accessed: 06- Dec- 2015].