Understanding Twitter Activity During Live Sporting Events

Stephen Armstrong
sjarmstr@ucalgary.ca

ABSTRACT

In this paper, we will be investigating the relationships between Twitter, a microblogging and social networking service, and sporting events. Sporting events have always been a social occasion and now with Twitter people are able to socialize remotely. The problem proposed in this paper is does momentum, as reflected on social media, correspond with in-game events? The approach we will use is to collect tweets posted during a Canadian Football League (CFL) game and apply sentiment analysis, term frequency and several other analyses to the dataset. From this analysis we will attempt to determine who is in control of the game based on the twitter data. In terms of football, the team with more time in possession of the ball can be said to have more "momentum" and thus is in control of the game. For the purposes of the work presented here this will be how we will quantify momentum. We will compare our data to human-generated summaries of the game's progression. We hope to see that the information gathered by our tool reflects the same game progression as the human generated summaries. Unlike the other papers exploring similar topics, this paper will show that using multiple low level analyses can be used to draw conclusions about the sporting event.

KEYWORDS

Twitter, Sports, Momentum

1. INTRODUCTION

Twitter is a social networking and microblogging service which enables its users to post and read 140 character, text-based messages on a public forum. As can be seen in Figure 1, these posts, called tweets, contain a wealth of information about what the user is thinking, doing and feeling. During sporting events, Twitter users regularly post status updates about the game, as well as feelings and opinions about the game's progress [1, 2]. We will be collecting Twitter data relating to sporting events in order to find relationships between the data and actual events within the game. Then, using the tool we have developed we will explore these relationships. Specifically we will determine which team has the most momentum at any given time.

Momentum can be very difficult to define and varies between sports. However, commonalities can be identified to provide a simplistic definition of Momentum. For instance, nearly every sport involves some sort of object to be used as a scoring device. It could be argued that the team which has possession of this object, usually a ball, is in a more advantageous position than the other team at that particular time. Thus, we can say that the team with more time in possession of the ball on offence has more momentum than the other team. We intend to build a tool which will use sentiment analysis to attempt to identify which team has more momentum at any given time or over a period of time. As illustrated in Figure 2, we will require three parts to complete the tool. The first is a collector to gather the Twitter data. The second is an analyzer which will perform the sentiment analysis as well as other forms of analyses to interpret the data. The third and final part will be a visualizer to graph the data.

Figure 1: CFL News Feed on Twitter (https://twitter.com/CFL)

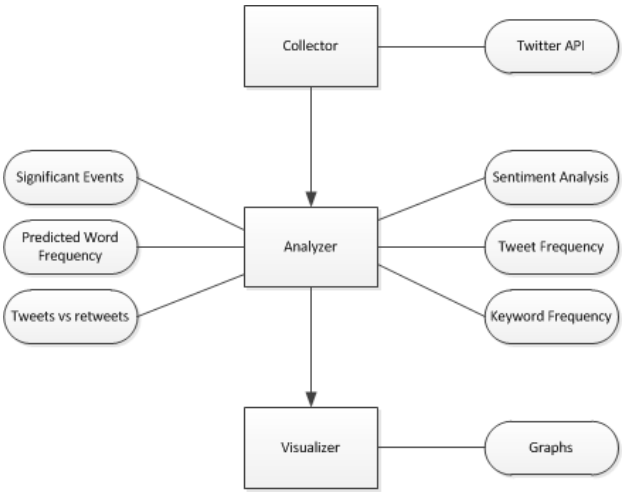

Figure 2: Program Diagram

2. PREVIOUS WORK

Similar work has been conducted in several papers. One example of such is a paper on "Summarizing Sporting Events Using Twitter" [1]. The problem presented was to create an algorithm which can create a summary of events using only Twitter status updates from the Twitter streaming Application Programming Interface (API). The authors wanted to see if it was possible to create a summary from the Twitter data which was as good as summaries published by sports writers. The authors collected tweets posted during soccer games and identified important events which occurred during the match. Summaries were then generated by copying the text from tweets which correlated to the events. These summaries were then compared to the ones written by sports writers for the same game. The conclusion reached is that the algorithm does produce a summary that could be interpreted by someone who was not watching the game [1]. Our paper will utilize similar means of filtering the data collected as was presented in "Summarizing Sporting Events Using Twitter".

Another paper, presenting a similar problem is "Personalized and automatic social summarization of events in video" [2]. The problem presented was to create a video highlight reel from time stamped Twitter posts. Like in the previous paper, Twitter posts were collected using the Twitter streaming API based on key terms and hash tags relating to the world cup. Summaries were generated by either using frequency based data or content based data. Frequency based method would select the top number of documents with the highest number of tweets and assign a time slice. These time slices were concatenated together to produce the highlight reel. For content based method user based terms were provided to limit the content of the tweets available. Then the highlight reel is built in the same way using the largest number of tweets. The results found that the summaries were satisfying, but had some evident flaws. One example of such a flaw is that if the user submitted poor queries for the content based approach the results would be poor [2]. Like both of the previously discussed papers this paper will be utilizing the Twitter streaming API to collect tweets for analysis.

Similar work can also be seen in the paper "TwitInfo: Aggregating and Visualizing Micoblogs for Event Exploration" [3]. Their system, TwitInfo, automatically identifies peaks in Twitter data and marks them as events. Users can then go and view these events as well as summaries of what has happened. TwitInfo also makes us of sentiment analysis, tweets are classified as positive, negative or neutral and then a sentiment analysis is shown along with the search results. Users however, found the sentiment to be misleading as when they searched for topics such as "earthquake" they would often see a positive sentiment, not the negative one expected. Upon further investigation users found this was mainly caused by tweets which offer well wishes to the people affected by the earthquake [3]. In our paper we will also be using sentiment analysis; however, we will be using the SentiWordNet [4] database to determine the polarity of the tweets.

SentiWordNet, as described in "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining" is an automatic annotation of "WORDNET" according to the notions of "positivity", "negativity", and "neutrality" [4]. The idea is to assign the definition of each word with three values: a positive score, a negative score and an objective score. Each score ranges from 0.0 to 1.0 and the sum of these scores must always equal 1.0. The user can then evaluate these scores and categorize the words based on sentiment [4]. This is crucially important to our work as it will be the basis of our sentiment analysis of the tweets. By using this database we hope to be able to identify which team has the favor of the crowd at any given time.

Others have used SentiWordNet in similar ways. For instance the paper "Sentiment Polarity Identification in Financial News: A Cohesion-based Approach" uses SentiWordNet to quantify positive and negative sentiments in financial news reports. The authors wanted to determine if the polarity of news reports could be determined which was consistent with human judgment. Sentiment analysis is applied by comparing each word in the news report to the SentiWordNet database then building several Metrics to compare with the judgments of humans [5]. The authors then evaluated how closely the sentiment analyses correlated to the human analyses.

3. PROPOSED WORK

As previously discussed the work proposed in this paper is to build a tool which is capable of determining which team carries the momentum of the game. Specifically we will be using CFL football games to build the data set. Three parts will be required in all: a collector to collect the tweets, an analyzer to interpret the data and a visualizer to show the data. In addition, game statistics will be manually collected for use as a baseline to compare the tweets to.

3.1 COLLECTOR

The collector will use the Twitter streaming API to collect tweets posted during CFL football games. If possible we will also collect tweets from other sports to see if the tool is applicable to other sports. Search terms such as "CFL", "BCLions", "Stamps" will be entered and all tweets posted with these search terms in the will be captured by the collector. Using search terms, as opposed to collecting all tweets posted during the game, will help to filter out unwanted and erroneous tweets. The data collected will be saved into a .csv file for analysis. Alongside the automated collector, we will be collecting in-game events with timestamps. This will be done manually based on the live TV feed.

3.2 ANALYZER

The analyzer will open the .csv file and perform analyses on the contents of the file. The data in the file is organized in a specific way. Each line denotes a separate tweet. The first item on the line is the user name, then the date and time at which the tweet was posted. Finally are the tweet id number and the tweet body. The tweet body may contain several important pieces information. Tweets posted in response to another tweet are characterized by "RT" at the start of the tweet. Also, if the tweet is directed at a specific user it will contain a "@" followed by the user's name. The last part of information in the tweet body is the message itself.

We can use the username of the tweets to see how many posts are made by each user, the time stamp will allow us to determine the tweet frequency over time as well as help determine the time of significant events. The body of the tweet can be analyzed with sentiment analysis using the SentiWordNet database [4]. The sentiment analysis will allow us to categorize tweets into positive and negative categories in relation to each team. Then we can determine which team has the most support and thus the most momentum. We can also use the body for determining

significant events. A significant event, in the case of football, would be something like a touchdown, or possibly a penalty. A significant event would show an increase in tweet frequency as well as an increase in the number of keywords contained in the body such as “touchdown”, “TD” or “penalty”. The data collected can then be passed to the analyzer.

3.3 VISUALIZER

The visualizer will display the data produced by the analyzer. Each of the analysis methods will have to be visualized. Momentum will likely be characterized by a variety of graphs; one such graph can be seen in Figure 3. We will also attempt to collate similar data between representations. For instance if a specific piece of data is highlighted it would be highlighted in other visual representations of the data.

Figure 3: Tweet Frequency over Time

4. TIMELINE

Sept 14: Collector started

Sept 21: Proposal Draft Due

Data collection started

Collector Complete

Sept 28: Proposal Due

Data collection continues

Word frequency and Tweet frequency complete

Oct 5: Data collection continues

Sentiment Analysis complete

Oct 12: Data collection complete

Keyword Frequency and Significant Events Complete

Oct 19: Analyzer Complete and Visualizer started

Oct 26: Word frequency visualization complete

Nov 2: Tweet frequency visualization complete

Nov 16: Sentiment Analysis visualization complete

Nov 23: Visualizer complete

Nov 30: Final Paper Due

Dec 2-7: Presentation

5. REFERENCES

[1] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. *Summarizing sporting events using twitter.* In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12). ACM, New York, NY, USA, 189-198. DOI=10.1145/2166966.2166999 http://doi.acm.org/10.1145/2166966.2166999

[2] John Hannon, Kevin McCarthy, James Lynch, and Barry Smyth. 2011. *Personalized and automatic social summarization of events in video.* In Proceedings of the 16th international conference on intelligent user interfaces (IUI '11). ACM, New York, NY, USA, 335-338. DOI=10.1145/1943403.1943459 http://doi.acm.org/10.1145/1943403.1943459

[3] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. *Twitinfo: aggregating and visualizing microblogs for event exploration.* In Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11). ACM, New York, NY, USA, 227-236. DOI=10.1145/1978942.1978975 http://doi.acm.org/10.1145/1978942.1978975

[4] Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204.

[5] Devitt A, Ahmad K: *Sentiment Polarity Identification in Financial News: A Cohesion-based Approach.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, June 2007, 984-991