Project Proposal: Sentiment analysis of events via Twitter EECS 4415

Edwin Gonzalez Dos Santos
York University
Toronto, Canada
edwin96@my.yorku.ca

Amir Younesi
York University
Toronto, Canada
kxn@my.yorku.ca

Abdullah Basulaib
York University
Toronto, Canada
basulaib@my.yorku.ca

Paul Kim
York University
Toronto, Canada
pyjaekim@my.yorku.ca

KEYWORDS

twitter, sports, sentiment analysis, text analysis, word analysis

ACM Reference Format:
Edwin Gonzalez Dos Santos, Amir Younesi, Abdullah Basulaib, and Paul Kim. 2019. Project Proposal: Sentiment analysis of events via Twitter EECS 4415. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DOMAIN DESCRIPTION AND MOTIVATION

The data domain for our project will consist of the textual content of a series of tweets containing one or more of the user provided hashtags along with their respective timestamps. We will also be using a dictionary containing words taken from imdb reviews along with the average score of the review to gauge sentiment for which words are used in tweets.

The goal of our project is to analyze the textual content of tweets that use a hashtag related to a sports event to map out the average sentiment over the course of the sports event. The tweets will be divided into upto 3 groups, those who are supporters of the home team, supporters of the away team and those using a neutral hashtag. This will allow us to judge the sentiment of different sets of supporters independently.

We consider this subject worthy of rigorous data analysis due to supporter sentiment often being an overlooked statistic in the sports world. This application attempts to better understand if supporter sentiment fluctuations are the only result of major events in the field of play or if they can be used as a predictive measure of events that are yet to come. The utility of this application can be used to better understand why supporters sentiment changes, and could also be used by the management of a sports team to address negative fluctuations in supporter sentiment and attempt to correct that before it gets too negative.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Being able to understand public sentiment via word analysis is important because it allows outside forces to better understand a group without having to rely on them filling out a survey or waiting for a groups sentiment become so negative that it begins to demonstrate itself in other ways beyond their word choice in tweets.

2 ARCHITECTURE OF PROPOSED SOLUTION

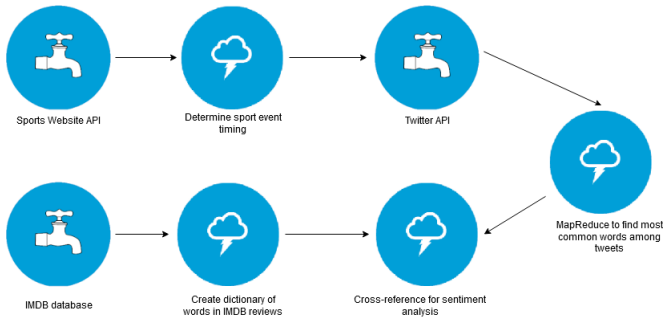

Fig. 1

Overall architecture and data flow

In order to get the tweets relating to the sports event we will be using Twitter's API to search for tweets containing one or more of the requested hashtags that were posted between 30 minutes before and after the event. The data returned will be processed to only keep the textual content of the tweets along with the timestamp of when they were posted. We will not be keeping any user data or any non textual content that is connected to their tweet including pictures, videos and gifs. Up to 3 sets of tweets will be returned from the Twitter API the first will include those that used the hashtag of the home team the second will include tweets using the hashtag of the away team, the third will include a non team specific hashtag relating to the sports event. Using the NHL hockey game played on November 2nd 2019 between the Toronto Maple Leafs and the Philadelphia Flyers as an example the home team hashtag would be FlyOrDie the away team hashtag would be LeafsForever and the neutral hashtag would be TORvsPHI.

Within each set of tweets they will be divided into a series of files stored on disk relating to the time when the tweets were posted.

The size of the time segments is dependant on the volume of tweets so it will be left up to the end user to specify the size they wish to use. Once the tweets have been segregated a MapReduce will be done on each of the segments to find the most commonly used words within that time span using the specific hashtag.

In order to gauge the positive or negative connotation of the most commonly used words we will cross reference those words with a set of words and their respective ratings. This will be done by the IMDB dictionary mentioned earlier. This will allow us to assign a value from 1-10 to the most commonly used words. That value will then be used to calculate the average sentiment of tweets within that timespan using a particular hashtag. Once this process has been applied to each time slot within each of the hashtag categories the result will be plotted onto a graph that will show fluctuations in sentiment ranging from 10 being the most positive to 1 being the most negative sentiment along the y axis and the time being represented along the x axis. We also intend to use a popular sports website API (e.g., SportsNet) to get the time and date when the event begins as well as which time important events take place during the game. Those values will be plotted on the graph as well in order for it to be easier to contextualize fluctuations.

Some of the challenges that we face is that we will not be able to understand the context for which some terms words might be used and that will affect how reliable our results are. Secondly, our analysis does not capture tweets relating to the sports event that are not using the relevant hashtags and this will shrink our dataset. Thirdly, since we have no way to interpret emojis, or reaction gifs it is possible we will be overlooking data points that could be used to gauge groups sentiment.

3 SYSTEM EVALUATION AND DATA ANALYSIS

Our application will be evaluated on efficiency and accuracy. The results we expect to get is a graph that lists the average sentiment of tweets under up to 3 specific hashtags over the course of a sports event. Sentiment will be rated on a scale from 1-10 where 1 represents a negative comment as indicated by its high usage in reviews with low ratings as opposed to a 10 rating which would indicate that the most used words are the same words which are used in imdb reviews with high ratings. This graph would also list important points in the game as markers on the graph represented by a vertical line at the time which the event took place. What is considered an important event will vary by sport but sticking with the example used earlier an important event in a hockey game would include the start and end of a period of play, a goal, and a penalty being called against one of the teams. These markers will serve to contextualize any fluctuations in the results.

Time permitting we will be looking to expand the types of datasets that can be handled to include other sports events and if possible other platforms such as Facebook comments and Instagram comments. While this application is designed to be used for sports, it could also be used to measure public sentiment for many other events. For example, a campaign or events such as the Oscars and the Grammys. If it could be assumed that supporters would be more likely to use a candidate's campaign hashtag, then we could measure supporter sentiment. If this tool could be modified to handle real time datasets, then that could be used to provide a rolling average of public sentiment during events instead of just being able to look back on events that already happened.

Given that this application relies on Hadoop to run its MapReduce operations, this application could easily handle datasets with a higher volume of tweets if more nodes are added to the cluster. Due to the simple nature of the architecture of the application, we should have no issue scaling the application to handle a larger volume of tweets. If this application was to be modified to handle live datasets that would make scaling more complicated as completing tasks within a specific time to avoid growing backlogs would be of greater importance than in our version which does not handle live datasets. This issue could still be resolved by only doing partial MapReduce operations, shrinking the amount of words that need to be looked up in the dictionary or by growing the time slots as that will result in fewer larger MapRecuce and dictionary lookups. A combination of these solutions would ensure that the application could still be scaled up with minimal impact to the quality of the data.