Wisdom of the Crowds:

Modeling Real-Time Win Probabilities With Twitter

David Freed and Samuel Green

Applied Mathematics 221

April 30, 2016

Abstract

Past research has considered the ... We ...

Contents

1 Introduction

The world-wide sports betting market, by some estimates, transacts between 700 billion and 1 trillion dollars per year.[1] Understanding this market, and doing so more effectively than others, makes a very enticing value proposition: beat the bookies, and one can turn a predictable profit. But is this possible? Sports fans, ever the arm-chair experts, might like to think so. However, individual fan opinions seem biased, or often even ridiculous.

In this paper, we explore whether a signal, or some "wisdom from the crowd," can be distilled from the noise of individual observers. In this study, we consider Twitter data from 42 NCAA March Madness games to expand on previous work that established a link between crowd-wisdom and prediction of game events. Ultimately, we show that volumes of relevant tweets and sentiment analysis of those tweets can be used to improve over standard models used to predict March Madness games.

After reviewing some previous work that used Twitter for sports analytics, we begin by describing the dataset and the data collection and cleaning process, which was non-trivial.

To establish a baseline for further analysis, we then show empirically that tweets relevant to a game in progress respond both in volume and in sentiment to events in that game. We use this result to justify the hypothesis that tweets contain information about game progress. From this result, we show that the relevant Twitter data can be used to establish statistically significant predictions about future events in the game. We demonstrate this relationship both between tweet sentiment and game events and between tweet volume and game events. We also show that predictive power accumulates in the course of a game: in other words, that the aggregated total volume and sentiment of tweets up to and including time $t-1$ can be used to make predictions about events in the game at time t.

After establishing the period-to-period significance of the Twitter data, we consider the usefulness of tweets in making period-to-period predictions of the winners of each game. We show that tweets, both sentiments and volumes, and be used to improve over the standard

[1] http://www.statista.com/topics/1740/sports-betting/

Vegas Line model used to form online probability curves during games. With this result, we conclude that that a properly filtered dataset of tweets contains practically useful information for real-time prediction during NCAA tournament games, presenting a promising avenue for further applications to the NBA and other sports.

2 Overview of Related Literature

Previous literature has established taht useful modeling information can be derived from Twitter data, some including work using sentiment analysis, with the body of work mostly focused on the NFL.

A real-time system was built in 2012 using volumes of tweets related to National Football League (NFL) games to isolate significant game events=. That project characterized a difference between human- and machine-generated tweets based on posting rates, and, by discriminating between different varieties of users, the system could identify events in near real-time. The event detection system leveraged pre-selected sets of hashtags to isolate relevant tweets as input to the system. (Zhao et al, 2012)[2].

Sentiment analysis has also been previously applied to Twitter data to a predictive model for NFL games. In 2013, Sinha et al. found that Twitter data collected in advance of weekly NFL games could be used effectively to predict the outcomes of games more successfully than methods using other traditional statistical models. Their work used a dataset of tweets collected over periods in advance of games and also collected tweets by building sets of hashtags related to participant teams (Sinha et al., 2013).[3]

3 Data Overview

Our dataset consists of approximately 1 million Tweets made during the 2016 National Collegiate Athletic Association (henceforth, "NCAA") Men's Basketball Tournament (henceforth,

[2]http://arxiv.org/pdf/1205.3212v1.pdf
[3]https://www.cs.cmu.edu/ nasmith/papers/sinha+dyer+gimpel+smith.mlsa13.pdf

"March Madness"). Before describing how we acquired and classified Tweets, we briefly describe March Madness and the basketball-related data we collected.

3.1 Game Data

March Madness, the largest single-elimination tournament in major American sports, is one of the most important events on the U.S. sporting calendar. The tournament takes place from mid-March to early April, with the 68-team field shrinking to sixteen after the first weekend. On the following weekend, the so-called "Sweet Sixteen" compete for a spot in the "Final Four"—the given name for the national semifinals. The 67-game tournament, which takes teams from across the country,[4] is the NCAA's primary source of revenue—in 2015, it comprised 90 percent of the organization's total revenue.

We chose to look at March Madness game data because of the level of excitement surrounding the event. Unlike other sports that share the same athletic calendar—most prominently, the National Basketball Association (henceforth, "NBA")—there are rarely more than two March Madness games occurring at once and so the entire focus is on the current game.

The level of interest surrounding an average March Madness game is much higher than that for a regular season mid-week NBA contest; the tournament encourages fans, regardless of their level of expertise, to fill out a bracket predicting the outcome. This cultural phenomenon has exploded in recent years, with ESPN receiving 13 million brackets this year. From the brackets arises an intense gambling market (more than $9 billion in 2016), raising interest and attention in the games.

[4]In fact, nearly every major school is eligible to compete in March Madness. The tournament reserves 33 spots for the winners of each major Division I athletic conference, leaving an automatic berth available for 351 colleges and universities across the nation

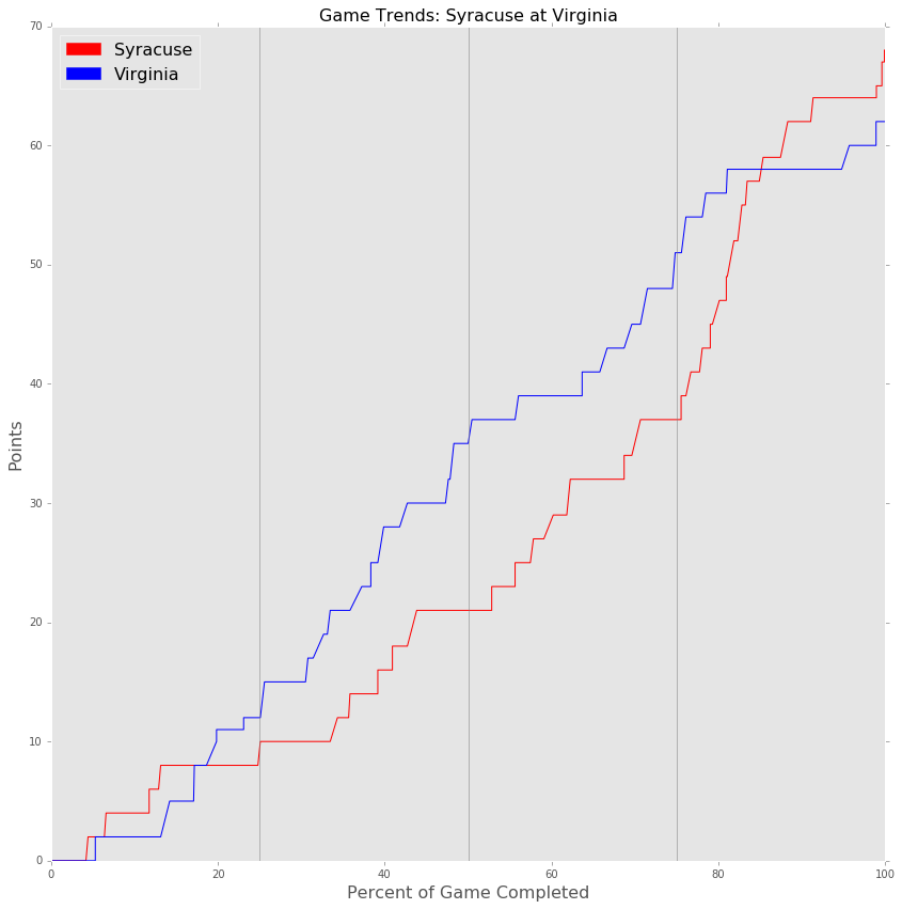

Figure 1: Score over time for Syracuse-Virginia Elite Eight contest

For each 2016 March Madness contest, we scraped `ESPN.com` to secure play-by-play data for each contest. From the website, we were able to get the scores for each team at every point in time, as well as a description of every event (i.e. "Paige, Marcus hits a three-pointer"). Figure 1 above shows the evolution of a Syracuse-Virginia Elite Eight contest, with Syracuse coming back from a 14-point deficit to take a late lead and storm into the Final Four.

3.2 Twitter Data

After obtaining our game data, we sought to collect a series of Tweets for each event to gauge public sentiment while the game was happening. In order to do this, we set up a Tweet listener while the game was actually going on, recording the Tweets as they were sent

for an hour before and an hour after each game.

In order to detect which Tweets were relevant, we only pulled Tweets that had a certain set of hashtags. Following the blueprint of Sinha et al. (2013), we constructed a set of Tweets associated with each game manually, scrolling through the official Twitter accounts of each individual team and adding the most commonly used hashtags to our list.

#NCAATournament #MarchMadness #LetsDance

#NCAATOURNAMENT #CBB #NCAAB

#SyracusevsVirginia #Syracuse #Cuse

#OrangeCrush #CuseMode #Virginia

#UVA #GoHoos #Cavaliers

Figure 2: Set of NBA hashtags for UVA-Syracuse Elite Eight game.

Figure 2 above demonstrates the set of hashtags used for the UVA-Syracuse game. Hashtags in black are hashtags that were not related to either team and common to all sets of tags.[5] The other hashtags were taken directly from the official Twitter accounts of the two schools and are colored in relation to which school they refer to.

To get a sufficient cross-section of data, we took Tweets corresponding to 43 separate games. Fifty-four of the 68 teams participated in at least one game in our dataset. All in all, we collected over 1 million Tweets, with an average of roughly 21,000 Tweets per game.

[5]With the caveat that the final Tweet, "#SyracusevsVirginia", was altered in each case to refer only to the teams playing in the game.

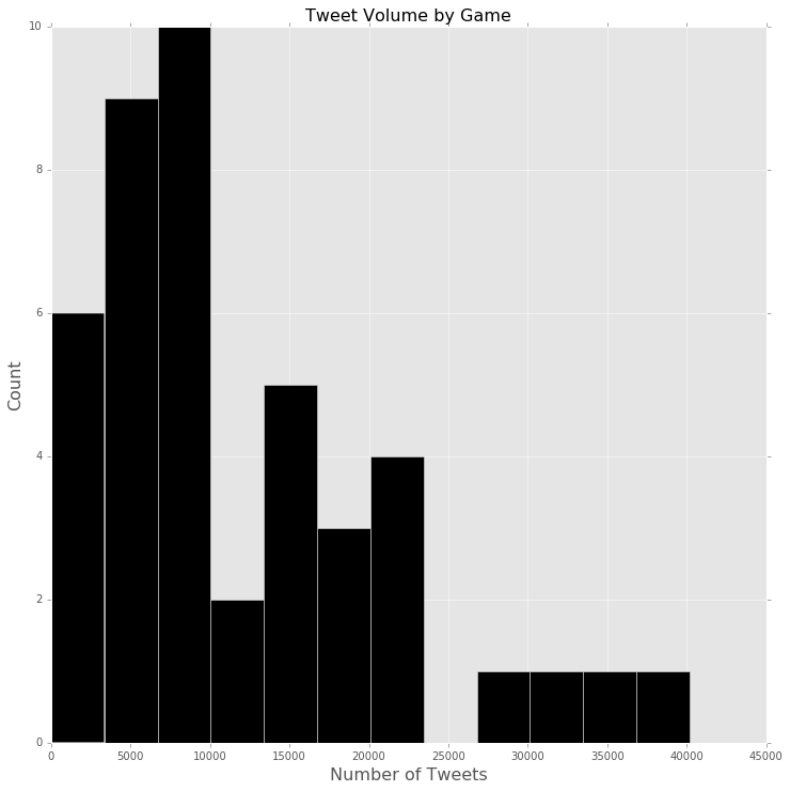

Figure 3: Tweet volume by contest

Figure 3 above shows the distribution of Tweets per game. The number of Tweets increased as the tournament went on; while first-round games had an average of just about 12.7 thousand Tweets a contest, there were an average of 15.8 thousand Tweets about each Sweet Sixteen contest in our dataset. Figure 3 excludes the national title game between Villanova and the University of North Carolina, which garnered just under 160,000 total tweets—by far the most in the dataset.

Once we had the two datasets, we set out to match the two to one another. Our source for game data did not log the exact moment at which each event occurred in real time, just in game time. To map game time (i.e. "11:30, first half") to real time ("9:30 PM"), we used a rough approximation algorithm. For each game, we manually took the beginning and end times of each game from `ESPN.com` and `@marchmadness`, the official Twitter handle of the NCAA Tournament and used that information to estimate the length of each half. From

that data, we estimated the time each event happened as a function of the length of each half and the time remaining in each half, using a modified uniform approximation to match game times to real times.

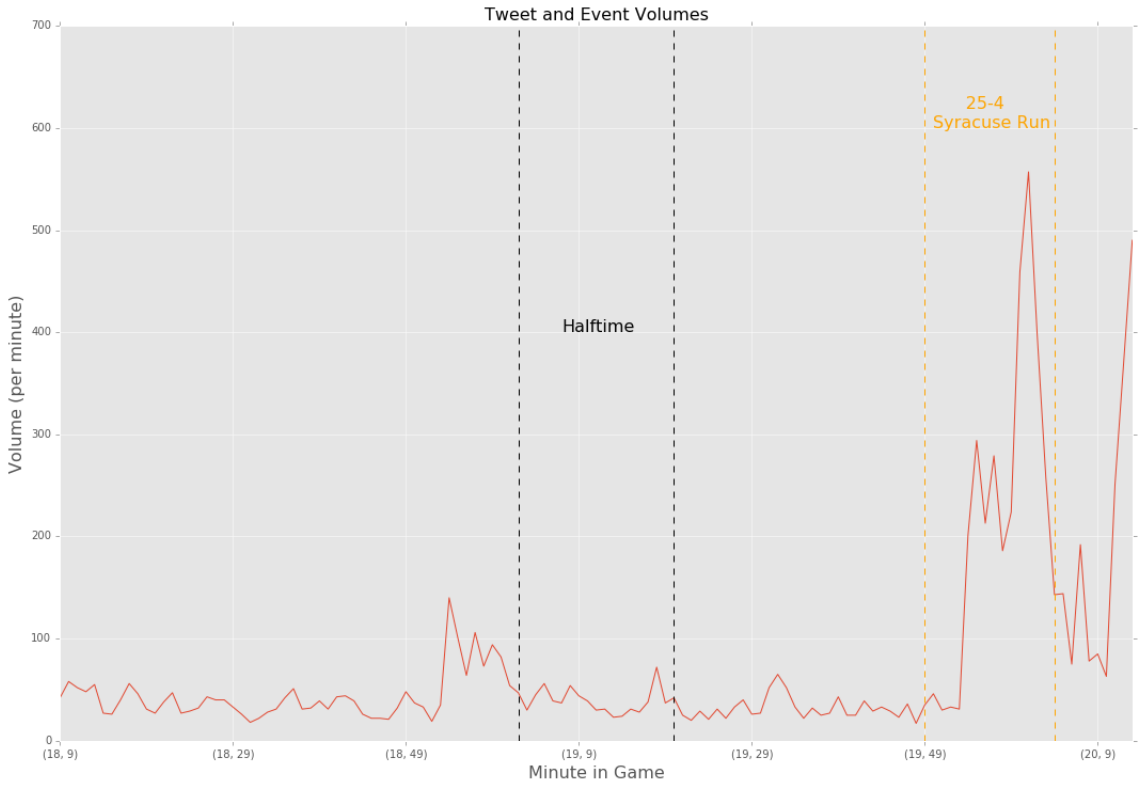

Figure 4: Tweet counts over time for UVA-Syracuse

Figure 4 above demonstrates the results of the mapping, which allowed us to identify exactly when halftime and key game events occurred in real time. As seen in the above graph, there was a significant spike in Twitter traffic during the pivotal moments of the game—a 25-4 run by Syracuse that brought the team from 14 points down into the lead. The associated large spike in Twitter traffic seen in the above figure reflects a common trend across the data: when the game got more exciting, Tweet volume spiked.

The next classification we made to the data was to classify each Tweet according to which team it related to. Since the eventual goal of the project was to be able to classify public sentiment towards any given team at any point in time, our intermediate step was to

associate each Tweet with a team based on the content of its message.

To classify the subject of each Tweet, we created a list of relevant tags[6] for each time and identified how often they showed up in the Tweet. From this, we computed a weighted relevance score for both teams, dividing the Tweets according to their relevance score for each team (i.e. those with a higher relevance score for Syracuse were tagged as 'Syracuse-related' Tweets).

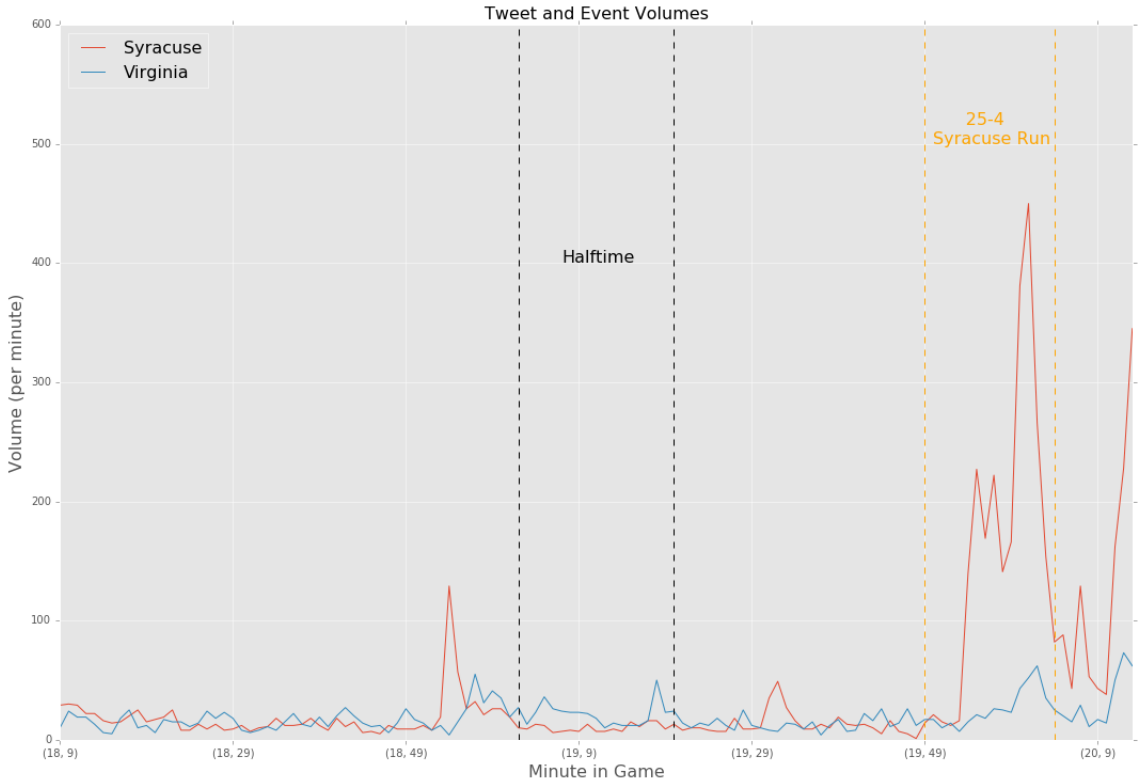

Figure 5: Tweets counts by subject over time for UVA-Syracuse

Figure 5 above breaks down the Tweet volume data by team, demonstrating that the spike in traffic during Syracuse' run comes almost entirely from people Tweeting about their comeback. We can see that when trailing early in the game, very few people were Tweeting about Syracuse; likewise, when Virginia opened up at a 14-point lead at halftime, they saw

[6]In addition to using the tags shown in Figure 2, we scraped the last names of the seven best players for each team and the coach of the team. In many cases, we found that Tweets included both team names (i.e. " ... #Virginia #Syracuse") but were actually about one team or the other. Including the last names of the players increased the accuracy of our classifier, since it better differentiated amongst these Tweets.

a brief bump in traffic during the 20-minute intermission.

The final step in our data collection was to classify the sentiment of each Tweet. In order to tell whether public opinion was positive or negative for each team over time, we constructed a sentiment classifier for individual Tweets. We chose a linear-kernel Support Vector Machine as our primary classifier, using a standard bag-of-words methodology and training the model on a prior labeled corpus of over 4000 words[7].

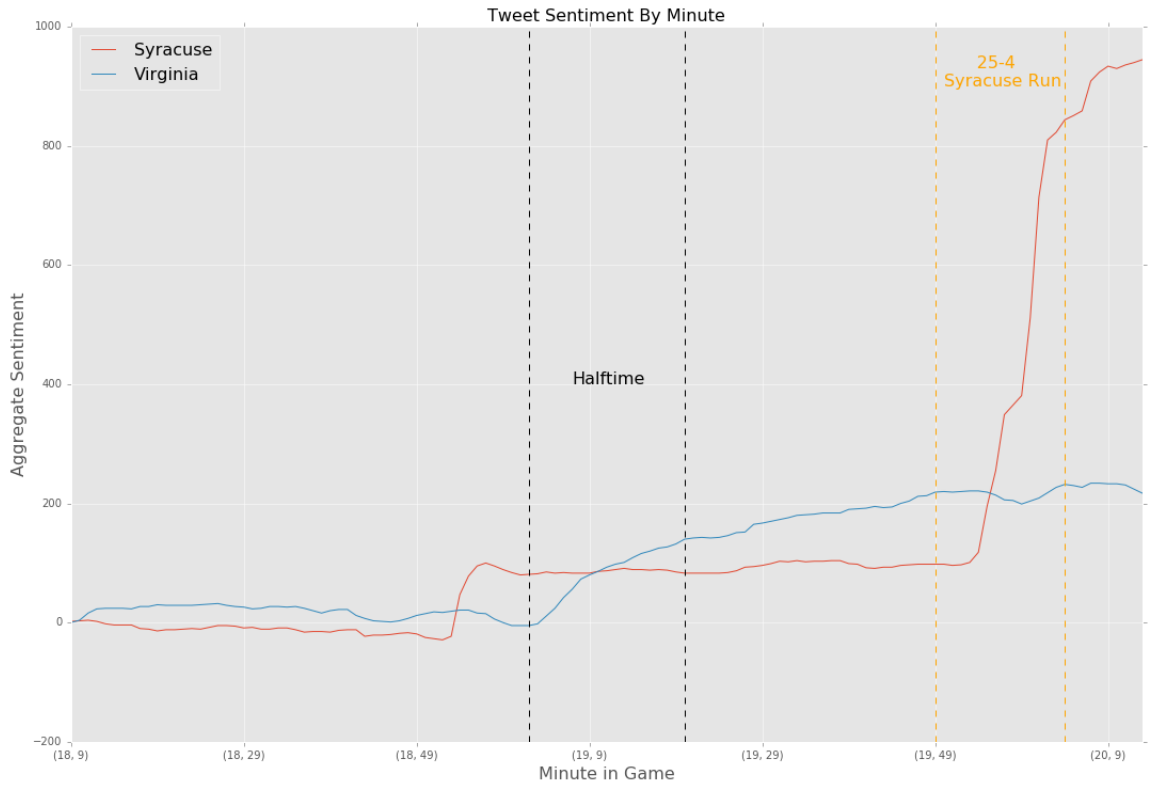

Figure 6: Tweet sentiment over time for UVA-Syracuse

Figure 6 above demonstrates the results of our classifier. We grouped Tweets into three categories: positive, negative, and neutral. In order to come up with an aggregate sentiment at any point in time, we took a simple linear combination of the three numbers—with our weights determined by the relative sensitivity of our classifier.[8].

[7]We had to make minor mechanical adjustments to the model due to the oddities of the language surrounding basketball; while words like 'dirty', 'filthy', and 'disgusting' would be classified as negative sentiments in almost any social context, they are the highest of compliments that can be paid on a basketball court

[8]Since our classifier was more sensitive to positive speech than negative speech, we gave a higher coefficient

The results, demonstrated above, were fascinating. During the Syracuse-UVA game, we can clearly see that as Syracuse falls behind, public sentiment drops into the negatives. An 8-0 run during the latter stages of the first half generates a lot of positive public sentiment, but during the half, the public begins to support Virginia (who has a 14-point lead). As Syracuse makes its comeback in the second half, however, Virginia plateaus and sentiment on Twitter shifts very strongly towards the Orange. As demonstrated in the forthcoming sections, this is evidence of the reactionary public sentiment to changes in the box score.

4 Empirical Results

5 Discussion

6 Tables and Figures

7 Bibliography

8 Appendices

to the amount of negative Tweets, assuming that they were an under-representation of the general sentiment