The Voice of the Crowd:

Modeling Real-Time Win Probabilities With Twitter

David Freed and Samuel Green*

Applied Mathematics 221

Harvard University

John A. Paulson School of Engineering and Applied Sciences

May 1, 2016

*We thank Yaron Singer, Thibaut Horel, Lior Seeman, and Rajko Radovanovic for their sage advice and constant support throughout the process of taking the class and writing this paper.

Abstract

The practice of aggregating independent opinions into a collective prediction is a standard one: the "wisdom of the crowds" is touted across multiple disciplines. Our paper applies this concept to sports, providing a novel method of measuring in-game crowd sentiment. We aggregate relevant Tweets sent during 43 NCAA March Madness games and use a sentiment classifier to judge Twitter sentiment towards each team at any point in time. By matching this data to game data, we find that Twitter is both reactive to prior game events and predictive of future ones. We conclude by showing that including Twitter sentiment will improve the predictive power of in-game win probability models, implying that Twitter actors capture information that cannot be gleaned from the box score.

Contents

1 Introduction

The wisdom of the crowds is not a new phenomenon. Economists treat the collective opinion of independent rational agents as sacrosanct; efficient market hypotheses are, at their core, expressing a fundamental belief in the wisdom of the crowds. Statisticians can explain the idea as the reduction of systemic risk—the variance of the sum of n independent and identically distributed random variables is smaller than the variance of any individual r.v. by a factor of $\frac{1}{n}$. In politics, prediction markets have quickly become a better predictor of election results than polls or the opinions of the expert.

The last example evidences the importance of being able to properly assess the wisdom of the crowds. While betting on political elections is illegal in the United States, the number of bets on the current presidential race is up fourfold since 2012 in Ireland. Betfair, the largest online betting exchange for U.S. presidential elections, has nearly one million users and records nearly seven million transactions a day, which it claims is "more than all European stock exchanges combined". In this cutthroat industry, any edge a gambler can get matters.

The betting market on politics pales next to the worldwide sports betting market—a colossal enterprise whose size experts ballpark at around $3 trillion dollars a year. In a market with roughly the GDP of the United Kingdom, understanding the underlying statistics (in this case, the relative quality of the teams) is fundamental. Any bettor who can consistently outperform the crowds can turn a predictable profit; one has to win only 52.4 percent of his bets to break even in Vegas.

In our paper, we tackle the common questions about the wisdom of the crowds from a different angle: instead of asking whether crowds or experts, alone or in aggregate, can predict games before they happen, we look at their ability to understand and predict games that are ongoing. We consider Twitter data from 43 games in the National Collegiate Athletic Association (henceforth, "NCAA") Men's Basketball Tournament (henceforth, "March Madness") to identify metrics that gauge the level of interest and the sentiment of the crowd at any point in time.

The justification for using Twitter as a predictive mechanism is simple: it can update far quicker than standard predictors like the margin of the game. If a star player goes down with injury or picks up a fourth or fifth foul—forcing him out of the game—the immediate effect will not be seen in the margin, but Twitter users will be able to accurately process the effect it will have on the game. Likewise, while Twitter can distinguish margins that are unsustainable (e.g. leads built on fluky plays or low-percentage shots) and those that are not, standard statistical models have a difficult time doing so.

Prior research has taken a cursory look at these questions; previous studies of National Football League (henceforth, "NFL") games identified that the volume of Tweets before a game is predictive of the final outcome and that Twitter is reactive to big plays in the game. Our work builds upon these analyses, not only by substantiating prior results for a different sport[2] but also by distinguishing between Tweet volume and Tweet sentiment. Prior papers considered only the volume of Tweet in each contest to measure the wisdom of the crowd; we use a sentiment classifier to determine how Twitter feels about both teams and how that evolves over the course of the game. In our Empirical Results section, we demonstrate that Twitter sentiment has predictive power that Twitter volume does not.

Our paper has three main results. First, to establish a baseline for our future analysis, we demonstrate empirically that Twitter is responsive to important game events—Twitter volume and sentiment will increase in response to salient changes in the score. This provides evidence for our initial claim that Twitter is responsive to the events currently going on in the game.

After showing that Twitter can capture what has happened in the past, we argue that Twitter is a useful predictor of what happens in the future. Our second result demonstrates that Twitter sentiment and Twitter volume is a statistically significant predictor of future

[2]Given the differences between football and basketball, this is a non-trivial contribution. Since basketball has far fewer breaks than football, one might imagine that Twitter would be significantly slower to reach to big events (fewer timeouts and breaks of play with which to process what happened). We show this not to be true in the Empirical Results section.

changes in margin; the aggregate Twitter sentiment in any period[3] is predictive of the change in margin in the next period. Here we find evidence for our prior claim about unobservable data—while margins tend to show mean reversion (i.e. teams that are up by a lot in period t tend to see their margin shrink in period $t+1$) as a whole, Twitter sentiment helps to distinguish which leads will continue to increase.

Our final result tests whether Twitter data can be used as an effective predictor of the final result at any point during the game. We use logistic regression models to incorporate the sentiment in a given period. We find that the Twitter sentiment is only statistically significant for about the final quarter of the game, but models that include Twitter sentiment outperform standard models[4] at every single minute of the game. We conclude by showing that incorporating Twitter sentiment is more valuable than simply incorporating raw Twitter volume, as other papers have done.

The remainder of the paper proceeds as follows. In Section 2, we review the brief literature on the subject, demonstrating both the advances and the gaps in prior research. In Section 3, we detail our data collection process, explaining how we matched gametime data (e.g. "7:30 remaining in first half") to real-time data (e.g. "9:30 PM") and discussing the construction of both our relevance and sentiment classifiers. In Section 4, we provide an overview of our empirical results and a thorough discussion of the aforementioned three main results and their significance. In Section 5, we conclude and discuss the important caveats and extensions to our work.

2 Overview of Related Literature

Previous literature in this space has established that Twitter data can contain useful information for making predictions about the future. This section proceeds as follows. First, we review the literature on using sentiment classification for Twitter. Second, we reference

[3]This is roughly measured as the support for one team minus the support for another team.

[4]For the purposes of our analysis, we cite the logistic prediction models rolled out by FiveThirtyEight during this year's March Madness as the standard model for predicting the end of the game

previous work using Twitter data to conduct sports analysis.

Go et al. (2009) provide a useful overview of the empirical work on classifying Twitter data. Prior work established that it is possible to accurately classify Twitter data using a variety of machine learning approaches to sort data according to the express sentiment. Go et al. demonstrate that you can achieve reasonably high classification rate through either a Naive Bayes, Maximum Entropy, or Support Vector Machine approach. In each case, their work proved to be relatively significant. Their work inspires our classification strategy, which is elaborated on later in the paper.[5]

Ibrahim and Yousef (2015) build on the work expressed by Go et al., demonstrating that reasonably accurate Naive Bayes classifiers can in fact be trained by using trivially small datasets of labeled Tweets. This analysis is crucial to the work that we do later in the paper, as it indicates that our corpus is of sufficient size to train an efficient classifier. The work of Ibrahim and Yousef draws on that of Yu and Salton (1976) and Roberson and Spark (1976), some of the original pioneers of the Naive Bayes model.

This sentiment analysis, to our knowledge, has not yet been applied to sports in terms of an in-game probability model. The majority of win probability models, such as those used by Chase Stuart of Football Outsiders and Nate Silver of FiveThirtyEight, are based on the margin of the game and some ex ante measure of how good each team is. We discuss these models later in our paper.

This is not to say that there has not been prior work on using Twitter to model the outcomes of sports contests. Sinha et al. (2013) assembled a large dataset of Tweets related to National Football League (henceforth, "NFL") games and used the Twitter volume for each team to predict the outcome of each game. The authors found that their model performed better than the standard win probability models, with a success rate of %55 percent—above the aforementioned threshold for breaking even in Vegas. We draw heavily upon the methodology used by Sinha et al. in assembling out dataset, modeling our relevance classifier off

[5]While the authors were able to demonstrate that a reasonably accurate training set could be derived without necessarily hand-labeling information in the Tweets, we do not reproduce that analysis in our paper.

much of what they express in their paper.

However, what is absent in the Sinha et al. paper is at the crux of our analysis: measuring Twitter sentiment in real time. To do this, we draw on results from Zhao et al. (2012). In this paper, the authors use Twitter to identify significant plays in real time (with intervals on the order of 1-2 minutes), a substantial improvement over previous approaches that required access to full dataset. Their identification procedure did not rely on sentiment classification, but rather on Tweet rates from specific users and grouping Tweets based on whether they were most likely to have been sent by humans or machines. This piece of previous work provides evidence that Tweets respond in real time to progress in sporting events, informing our hypothesis that Tweets can be predictive and responsive to real time events.

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, "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,[6] is the NCAA's primary source of revenue—in 2015, it comprised 90 percent of the organization's total revenue.

[6]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

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 2015), raising interest and attention in the games.

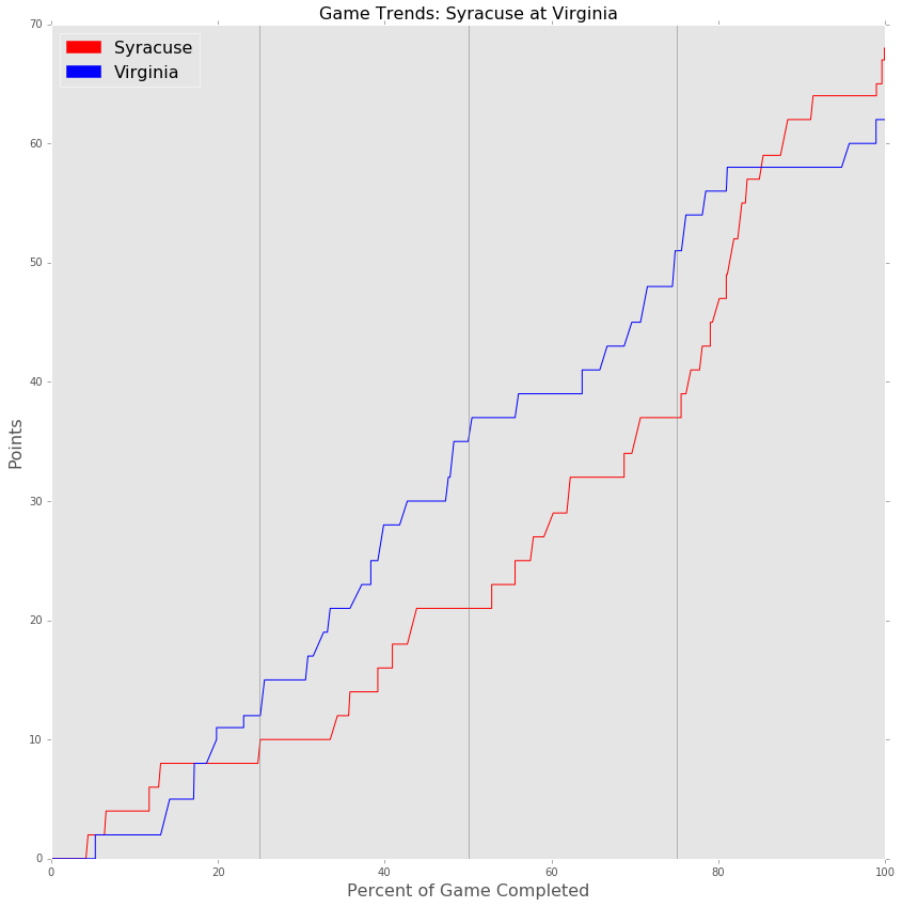

Figure 1: Score over time for Syracuse-Virginia Elite Eight contest

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.[7] The other hashtags were taken directly from the official Twitter accounts of the two

[7]With the caveat that the final Tweet, "#SyracusevsVirginia", was altered in each case to refer only to

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. Forty-eight 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.

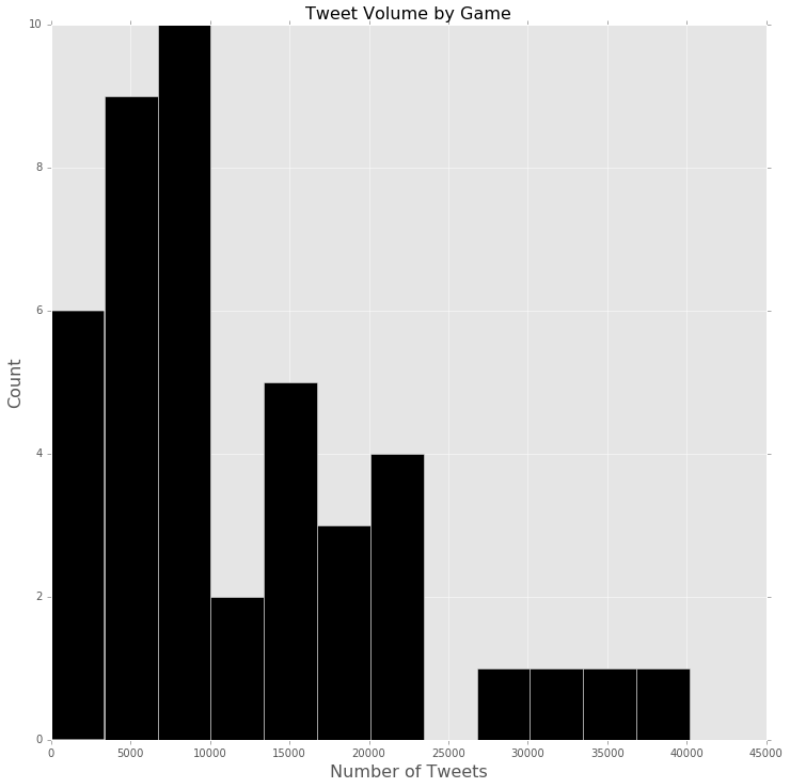

Figure 3: Tweet volume by contest

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

the teams playing in the game.

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.

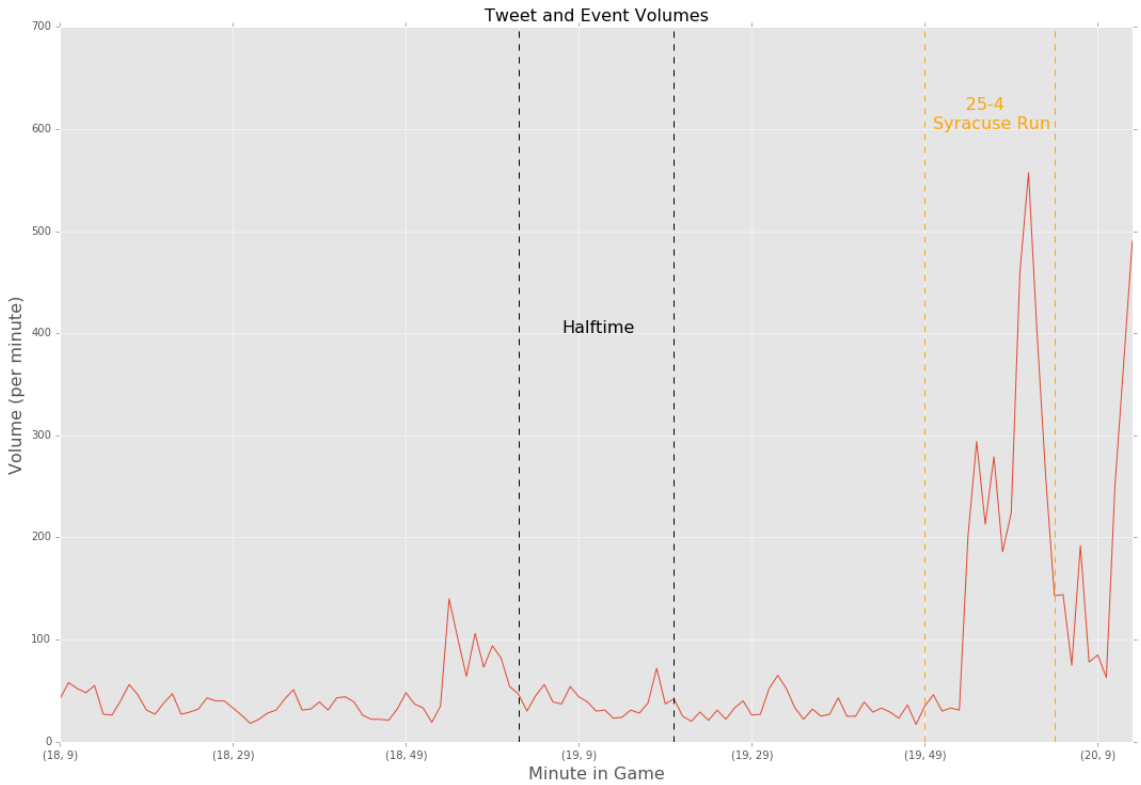

Figure 4: Tweet counts over time for UVA-Syracuse

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[8] 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).

[8]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.

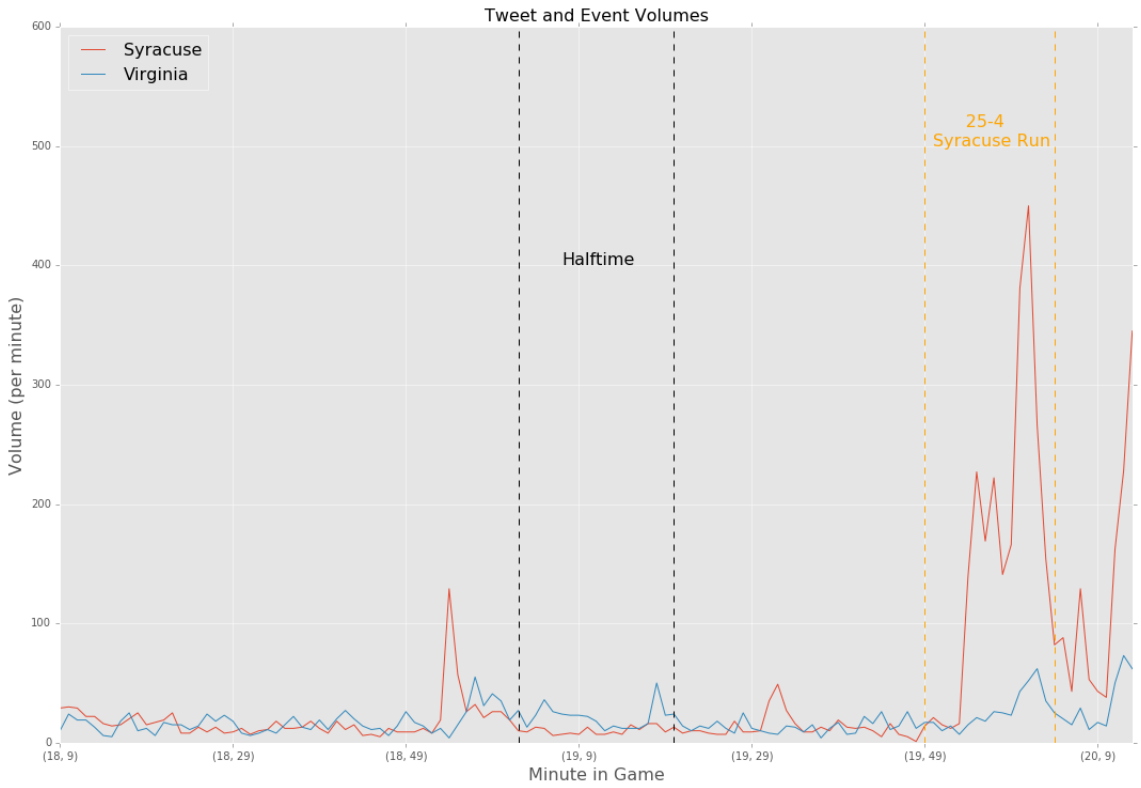

Figure 5: Tweets counts by subject over time for UVA-Syracuse

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 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[9].

[9]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

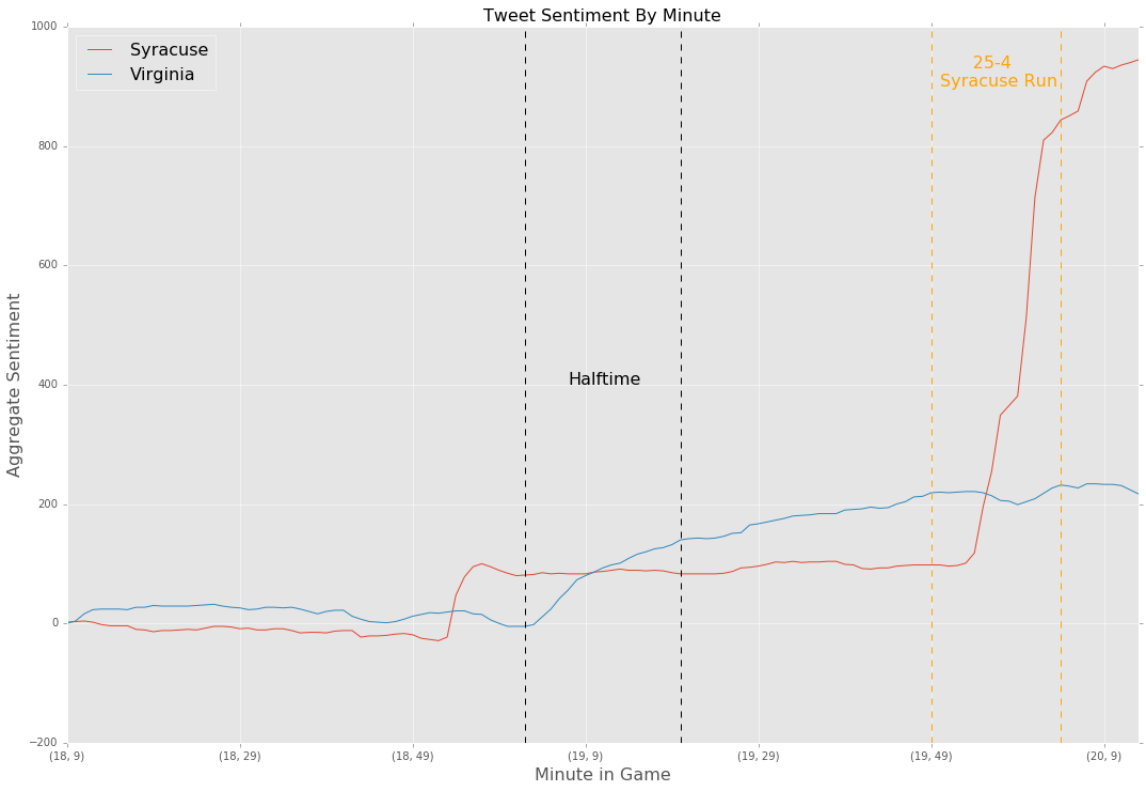

Figure 6: Tweet sentiment over time for UVA-Syracuse

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.[10].

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.

[10]Since our classifier was more sensitive to positive speech than negative speech, we gave a higher coefficient to the amount of negative Tweets, assuming that they were an under-representation of the general sentiment

4 Empirical Results

This section lays out support for the theoretical hypotheses presented earlier. Section 4.1 demonstrates that Twitter is responsive to large swings in the game and can effectively incorporate information about game events. Section 4.2 demonstrates that both Twitter volume and aggregate Twitter sentiment can predict future events in the game. After showing the predictive power of our variables, we conclude in Section 4.3 by demonstrating that they constitute an improvement on prior models. We compare standard logistic regression models with those incorporating Twitter information, showing the latter is more predictive of the final outcome at each point in the game.

4.1 Twitter's Ability to Process Prior Events

The first idea that we want to assess is whether Twitter is able to process game events as they happen. As discussed in the previous section, we use a modified uniform approximation to map game times to real times, allowing us to capture subsets of Tweets that occurred in the time surrounding an event. To empirically test this result, we divided up our dataset for each game into a series of one-minute intervals. For each interval, we not only measured the change in margin (i.e. how Team 1's lead/deficit changed over the course of the minute) but also the change in Twitter sentiment and change in Twitter volume for both teams.

If Twitter is responsive to events in the game, then Twitter sentiment and volume for a given team should spike as their play improves. We have already seen evidence of this in previous figures, which showed that Tweet sentiment and volume about Syracuse increased drastically as they made their comeback against Virginia. Over the same time, sentiment and volume plateaued for the Cavaliers.

In Table 1, we look at whether the difference in margin[11] in period $t-1$ can predict the sentiment difference in period t. The variable `Margin_Period_Lag_1` represents the change

[11] We define margin in period t as the difference between the number of points scored by Team 1 in period t and the number of points scored by Team 2 in period t.

in the lead/deficit in the prior period, while `SentDiff_Period` shows how Twitter sentiment changed in the current period.[12] In Columns (1) and (2), we find that the margin in period t is a significant positive predictor of sentiment different in period $t+1$ even after controlling for time fixed effects and measures of team quality and popularity.[13] In Column (3), we find that even if we lag the margin by two time periods, it remains significant even with our standard controls.The coefficients on all the lagged margins are positive, indicating that the better Team 1 does in the prior period, the more Twitter responds in the next period. This is exactly what we would have expected to see.

In Table 2, we see very similar results for volume of tweets, which we establish by regressing `VolDiff_Period`, a measurement of Twitter volume, on `Margin_Period_Lag_1`.[14] We find the exact same results as we did in Table 1, demonstrating that Twitter both gets increasingly positive about a team as it does better and begins to Tweet more about that team. We measure volume in thousands of tweets, so we can interpret the coefficient at the top of Column (2) as saying that for every additional point that Team 1 pulls ahead, there will be, on average, nearly six thousand more Tweets sent in the next period about Team 1.

One interesting question is how this responsiveness changes at the end of close games, when game events have a significantly larger impact on the final outcome of the game. In Table 3, we limit our dataset to one-minute intervals in the final 10 minutes of the game

[12]Specifically, this measures the difference in sentiment scores between the two teams in the current period. If the public assigns Syracuse a sentiment score of 130 and Virginia a sentiment score of 110, then `SentDiff_Period` will equal 20. The higher the metric, the more the public favors one team over the other. We find this metric to be correlated with the actual margin.

[13]Our time fixed effects variable, `Min_End`, refers to the end of the current interval. We have two measures of team quality. The first is `Vegas_Line`, which refers to the pre-game betting line in Las Vegas (as sourced from ESPN). The second metric is `Quality_Diff`, which refers to the difference in Ken Pomeroy scores for the two teams. Ken Pomeroy scores are a widely accepted advanced statistical measurement of team quality, so here they are used as a proxy for any quality differences that the betting lines do not account for. Last, we use `TwitterDiff`, or the difference in the number of Twitter followers for each team's official Twitter account, as a measure of the spread of popularity between the two teams. Since both of our Twitter metrics (volume and sentiment) are aggregate numbers, this helps control for the size of the fan base.

[14]We evaluate changes in volume in the same way we do changes in sentiment: on a relative basis. Instead of looking at how volume for Team 1 increases over time, we de-trend for increases in volume for Team 2 to isolate shifts in opinion on Twitter. Without de-trending the data, we would find that at the end of close games (where Twitter volume naturally increases as viewers tune in to watch the final minutes) volume could increase substantially for a team that is falling father and farther behind.

and find some interesting results. Columns (1) and (2) indicate that as the game winds down, Twitter is responsive not to past events, but to current events. The coefficient on `Margin_Period` in Column (1) is more than three times as large as any of the coefficients in Table 1, indicating that not only does Twitter process events at the end of the game quicker, but its reaction is far stronger. Columns (3) and (4) provide reinforcing evidence for this claim, demonstrating that Twitter responds quicker and more vigorously to margin changes at the end of the game.

4.2 Twitter's Ability To Predict Near-Term Events

After showing that Twitter can incorporate information about what is happening in the game, we next want to demonstrate that it is incorporating predictive information. If Twitter is just incorporating information about the margin in period $t-1$–which is very uncorrelated with the margin in period t—then it will be a poor predictor of future performance. However, if we think that Twitter is incorporating some unobservable information, then it should be predictive of results that we see in the future.

In Table 4, we test this by regressing the sentiment difference in period $t-1$ on the margin in period t. We find in Columns (1) and (2) that the sentiment difference is a statistically significant predictor and positive in predicting the margin in the next period. We can interpret this result as saying that the more bullish Twitter is on a team, the better that team will perform in future periods on average. In Column (3), we see that Twitter sentiment in the most recent period is the only sentiment that matters, but that it remains significant even after controlling for prior sentiment differences.

Table 5, which tests the same concept with a lagged difference in Twitter volume instead, reinforces the same idea: Twitter's reaction in period $t-1$, as measured by volume of tweets related to each team, tends to be predictive of changes in the game in period t.[15] Given that the margin in period $t-1$ is not predictive, this supports the argument that Twitter

[15]Column (3) would appear to indicate that Twitter volume is a better predictor than Twitter sentiment, which could be explained by the problems with our sentiment classifier (see Discussion).

contains some predictive ability beyond what is observed in the game.

There is an obvious counterpoint: Twitter is picking up not only the margin in period $t-1$, but the total margin for the game. If a team is up by 16 and then is outscored by 4 points in a one-minute interval, Twitter will "understand" that is it still up by 12, while simply regressing on the margin in period $t-1$ will not capture that information. To test whether Twitter is just picking up the effect of the total margin, in Table 6 we investigate how the regression changes when we control for the total margin in period $t-1$.

The results still appear to indicate that Twitter does have additional predictive ability. In Column (1), we find an interesting tendency towards mean reversion—since `Margin_TOT_Lag1`, our measure of the total margin in the previous period, has a negative coefficient, we can infer that teams that are ahead overall are projected to give back part of the lead in the next period. Even more than that, we can say that the larger the lead, the more the lead will shrink in the next period—an interesting result that holds up in Column (3) even if we control for the quality of the team.[16]

Our most interesting results come in Columns (2) and (4) however, which appear to indicate that even after controlling for the margin, Twitter sentiment has a statistically significant positive coefficient. This implies that Twitter is picking up information that is not already baked into the total margin, giving it some additional predictive power into the future. One way to express this result is that Twitter can differentiate between sustainable and unsustainable leads—while the total margin does not demonstrate whether a team will keep its lead into the future, the aggregate Twitter sentiment does. Said another way, if a team is ahead and Twitter sentiment is positive, it can expect to lengthen its lead in future periods. However, if a team is ahead but Twitter sentiment is negative, then its lead is likely to regress.

For obvious reasons, this result is very interesting. It implies that the crowd is able to sort between early leads that are built on fluky shots or unsustainably high levels of play

[16]One might have hypothesized that this reflects a tendency for weaker teams to give back leads that they get early, since the talent of a better team will 'win out' over a longer time period.

and leads that are generated by teams that are simply better than their opponent. While our dataset is not large enough for us to evaluate whether this is due to the presence of unobservables (an avenue for future research mentioned in Section 5), it does indicate that Twitter can read information from the game that standard box score metrics cannot.

4.3 Twitter's Ability To Predict Game-End Winners

Given that we have established Twitter's ability to predict events in the immediate future, we now want to attack the question that practitioners (read: gamblers) care about: is Twitter useful in predicting the outcome of games?[17] To find the answer to this question, we first look at if and when Twitter sentiment becomes significant in predicting the final outcome of the game, and then assess whether this relationship is an improvement over more standard models.

[17]From a practical standpoint, while the majority of betting markets involve ex ante bets (i.e. those cast before the game begins), a growing subset of international markets incorporate during-game betting and Betfair recently rolled out an 'in-play' betting service, indicating that there is an opportunity for practitioners to potentially put these results into action.

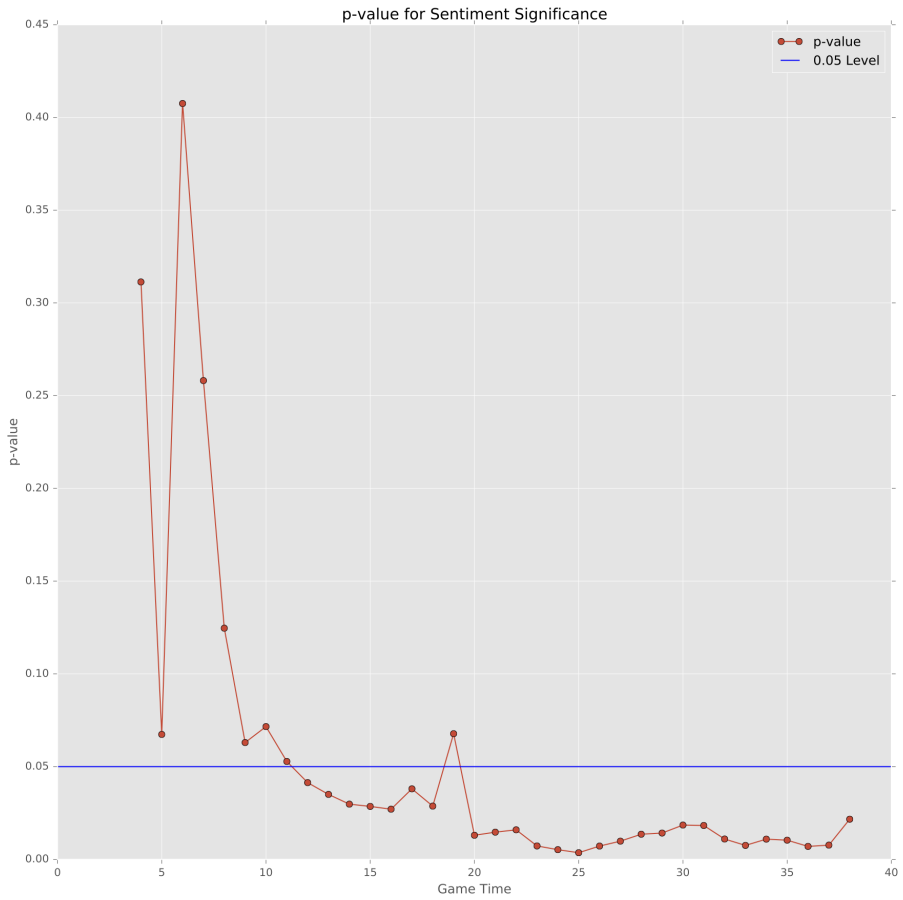

Figure 7: Predictive power of sentiment difference over time

To create Figure 7 above, we ran logistic regressions for the winner of the game with varying amounts of time remaining. We found that the total sentiment difference between the two teams only started being consistently significant at the 0.05 percent level in the second half, where it was an important predictor of the final outcome. We can see from the graph that in the early stages of the game, Twitter's opinion about a team is not nearly as effective at predicting the final outcome.

From this, we can infer that Twitter does have some predictive power to evaluate the winner of each game from an intermediate stage. To properly assess this, we compare the

standard in-game prediction models to those that utilize Twitter sentiment. The canonical mid-game models consist of logistic regressions that use just two variables: the current margin and an ex ante predictor of team quality (typically the betting line provided by Vegas before the game). This was the basis of the models that FiveThirtyEight popularized during the 2016 tournament while creating their in-game prediction models.

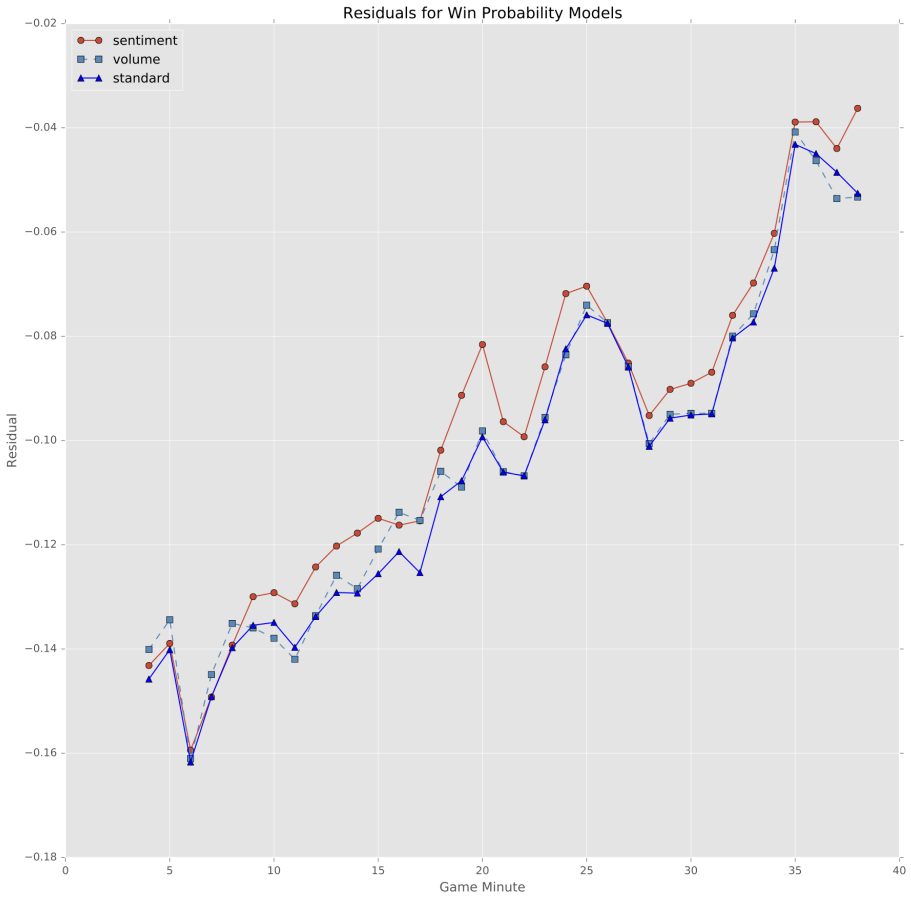

Figure 8: Relative quality of standard and updated models at each minute of the game

In Figure 8 above, we look at how the standard model measures against the logistic regression model that includes Twitter sentiment and Twitter volume as explanatory factors. As can be inferred from the above graph, a logistic regression model that adds in Twitter

sentiment outperforms the standard model at every point in the game—the residual for the updated model is consistently smaller than the residual for the standard model. We note that these results do appear to reinforce a result we first saw in Table 5, which indicates that Twitter sentiment carries more predictive power than Twitter volume.[18]

These results provide more substantial evidence for what we had inferred before: Twitter captures important predictive information that previous models did not. Table 9 shows the results of global regressions that merely confirm the relevance of both Twitter sentiment and volume. In Columns (2), (3), and (4) the two metrics are statistically significant at the 0.01 percent significance level and carry positive coefficients, implying that the higher Twitter is on a team, the more likely they are to eventually come out victorious.

5 Discussion

Our work expands on previous work applying Twitter for predictive sports analysis to demonstrates the predictive power of relevant Twitter data in forecasting both short-term and long-term outcomes in NCAA March Madness games. We begin by providing evidence that, over short time periods, Twitter is responsive to game events; we demonstrate in a prior section that the volume of Tweets (and the sentiment of those Tweets) about a particular team will increase in proportion to the squad's quality of play in prior periods.

Next, we show that Twitter is not only a reactive mechanism, but a useful forecasting tool. In Section 4.2, we show that the volume and sentiment of Tweets in a given period is a statistically significant predictor of game events in the next period. We moved beyond this to demonstrate that Twitter data has significant predictive power in forecasting the eventual winner of the game. Our model that incorporated Twitter sentiment outperformed the standard FiveThirtyEight model at every point in the game. These results support previous work demonstrating "wisdom of the crowds" effects on Twitter and show that Twitter analytics can be usefully applied to the NCAA.

[18]Recall that in section 4.1 we had shown that Twitter volume is more responsive to past events, however.

Our results are significant, but carry some caveats and areas of potential development. The foremost point for improvement sits at the pivot point of this study: sentiment classification. We trained our classifier using a corpus of labelled tweets related to Apple and Google product launches, given the impracticality of hand-constructing or commissioning a labelled training set for the sports-specific domain.

While hand tests showed that our classifier was able to discern conventionally positive language from conventionally negative language, this is inherently inadequate for the sports-specific domain. As mentioned earlier, many words that are conventionally negative, like "dirty" or "filthy," are often positive in the context of basketball. Our classifier initially classified the Tweet "`Hell yes, Syracuse is going to the Final Four`" as a 'negative' Tweet, but it is evident by inspection that the author is elated at the prospect of Syracuse advanced to the semifinals of the tournament. We hypothesize that were we to re-train the model with a better training set, our model would do a better job of extracting sentiments from Tweets. This would give a better idea of the current opinion of the crowd and provide a good check on our current analysis. We consider this to be a promising avenue for future work.

The second primary caveat to our analysis is the effectiveness of our relevance classifier. The process of separating noisy and relevant Tweets remains a challenge; our initial pull from Twitter contained 27,000 Tweets about GOP presidential candidate and international celebrity Donald Trump. Many of the hashtags that are commonly associated with teams (e.g. `#Maryland` or `#Wisconsin`) have many other uses. We did our best to eliminate these tweets, as previously discussed, but finding a mechanism to collect a more targeted set for classification would likely also contribute to improvement.[19]

The third caveat to the analysis is the way in which we projected game events onto real time. There was no available source of data which provided a reliable timestamp for each

[19]One related issue that we had is that many Tweets would show up multiple times in our dataset because they had been 'reTweeted', a Twitter functionality that allows users to send another account's Tweet from their own account. We opined that any individual who chose to 'reTweet' a message shared the sentiment of the original author of the Tweet.

event, and our modified uniform approximation of real time is not a perfect mapping. A possible extension or refinement of this process would be to institute many more check points in the data (by tapping Tweets that contained the time left in the game and using those as time signposts) to get a better approximation. Through spot-checking, our approximation appeared to work fairly well, but it remains an important caveat when evaluating our analysis.

There are two immediate extensions that fall out of the work that we have already done. It would be interesting to consider a more granular separation of the "crowd" into trusted and untrusted "sub-crowds." That is, could predictive improvements be made by considering crowd-sourced opinions from known experts, like commentators or individuals prominent in the network, separate from or more heavily weighted than standard or low influence individuals? Presumably, experts are ex ante more qualified to predict outcomes based on game progress but it is not immediately apparent that has to be the case.

The second obvious extension would have been to conduct a systematic study of Twitter's ability to incorporate specific types of unobservable information. Earlier in the paper, we reference the cases of a star player getting injured mid-game or coming out of the game with a fourth foul as situations where the margin may not accurately reflect a swing in the game. If we had an expanded dataset, we could take advantage of Twitter's ability to detect momentum to conduct a more thorough event study on this basis.

The final trivial extension would be to apply this to other leagues and sports. The National Basketball Association (henceforth, "NBA") would be a logical extension. One other interesting extension would be to gauge the predictive power for Twitter in smaller markets—how many people need to be following the game closely for their aggregate opinion to become predictive? Given the vast difference in attendance between games in the Big 12 and the Ivy League athletic conferences, this poses an interesting question for practitioners, who can often win substantial sums in less liquid markets (e.g. Ivy League athletics gambling).

In sum, the results presented above are a novel extension of previous work. We distinguish ourselves from the previous literature on this subject by measuring the sentiment of Twitter, not only the volume, at any point in time. Our results about the predictive power of models that incorporate Twitter data provide the basis for improved in-game win probability models and provoke interesting questions about how to measure unobservable information.

To our knowledge, the results presented here are novel. We provide the first models using Twitter data as input to a real-time model. We are also the first to show that a model of this variety has power when predicting future events and full game outcomes in the NCAA. Above, we note a series of logical extensions to the work, which can extend the analysis above and continue digging into the interesting empirical questions raised by the paper.

6 Tables and Figures

Table 1: Twitter Responsiveness to Game Events (pt. 1)

	Dependent variable:		
	SentDiff_Period		
	(1)	(2)	(3)
Margin_Period_Lag1	0.891**	0.881**	0.910**
	(0.367)	(0.366)	(0.370)
Margin_Period_Lag2			0.864**
			(0.377)
Vegas_Line		−1.008***	−0.973***
		(0.231)	(0.237)
QualityDiff		35.390***	32.811***
		(10.041)	(10.304)
TwitterDiff		−0.004	−0.006
		(0.011)	(0.011)
Min_End		0.068	0.061
		(0.075)	(0.079)
Observations	1,306	1,306	1,266
R^2	0.005	0.021	0.023
Adjusted R^2	0.004	0.017	0.018
Residual Std. Error	30.443	30.238	30.562
F Statistic	5.897**	5.548***	4.972***

Table 2: Twitter Responsiveness to Game Events (pt. 2)

	Dependent variable:		
	VolDiff_Period		
	(1)	(2)	(3)
Margin_Period_Lag1	6.253***	5.877***	6.028***
	(1.259)	(1.237)	(1.246)
Margin_Period_Lag2			4.779***
			(1.268)
Vegas_Line		−1.454*	−1.416*
		(0.781)	(0.797)
QualityDiff		129.964***	122.143***
		(33.914)	(34.662)
TwitterDiff		0.165***	0.168***
		(0.036)	(0.037)
Min_End		−0.630**	−0.679**
		(0.253)	(0.267)
Observations	1,306	1,306	1,266
R^2	0.019	0.063	0.073
Adjusted R^2	0.018	0.060	0.069
Residual Std. Error	104.377	102.132	102.812
F Statistic	24.685***	17.550***	16.599***

Table 3: Twitter Responsiveness to Game Events (Closing Minutes)

	Dependent variable:			
	SentDiff_Period		VolDiff_Period	
	(1)	(2)	(3)	(4)
Margin_Period	3.080***		9.598***	
	(0.898)		(2.980)	
Margin_Period_Lag1		1.325		7.570**
		(0.934)		(3.071)
Vegas_Line	−1.691***	−1.802***	1.061	0.671
	(0.585)	(0.594)	(1.941)	(1.953)
QualityDiff	54.306**	61.013**	44.282	60.681
	(25.706)	(26.013)	(85.282)	(85.559)
TwitterDiff	−0.012	−0.012	0.451***	0.447***
	(0.028)	(0.028)	(0.092)	(0.092)
Min_End	−0.526	−0.719	−2.755	−3.466
	(0.740)	(0.751)	(2.455)	(2.470)
Observations	316	316	316	316
R^2	0.068	0.039	0.133	0.121
Adjusted R^2	0.053	0.024	0.119	0.107
Residual Std. Error	37.840	38.427	125.537	126.387
F Statistic	4.535***	2.520**	9.490***	8.531***

Table 4: The Accuracy of the Crowds (pt. 1)

	Dependent variable:		
	Margin_Period		
	(1)	(2)	(3)
SentDiff_Period_Lag1	0.006***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)
SentDiff_Period_Lag2			0.001
			(0.003)
SentDiff_Period_Lag3			−0.003
			(0.002)
Vegas_Line		0.009	0.003
		(0.018)	(0.019)
QualityDiff		0.983	1.239
		(0.780)	(0.809)
TwitterDiff		0.0003	0.0002
		(0.001)	(0.001)
Min_End		0.004	0.007
		(0.006)	(0.006)
Observations	1,306	1,306	1,226
R^2	0.006	0.012	0.015
Adjusted R^2	0.005	0.008	0.009
Residual Std. Error	2.342	2.338	2.344
F Statistic	7.250***	3.106***	2.638**

Table 5: The Accuracy of the Crowds (pt. 2)

	Dependent variable:			
	Margin_Period			
	(1)	(2)	(3)	(4)
VolDiff_Period_Lag1	0.002***	0.002***	0.002**	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
SentDiff_Period_Lag1			0.004	
			(0.002)	
VolDiff_Period_Lag2				0.0004
				(0.001)
VolDiff_Period_Lag3				−0.002**
				(0.001)
Vegas_Line		0.006	0.009	0.0004
		(0.018)	(0.018)	(0.018)
QualityDiff		0.902	0.837	1.224
		(0.780)	(0.781)	(0.806)
TwitterDiff		−0.00003	0.0001	0.00005
		(0.001)	(0.001)	(0.001)
Min_End		0.005	0.005	0.008
		(0.006)	(0.006)	(0.006)
Observations	1,306	1,306	1,306	1,226
R^2	0.009	0.014	0.016	0.021
Adjusted R^2	0.008	0.010	0.011	0.015
Residual Std. Error	2.338	2.335	2.334	2.337
F Statistic	11.852***	3.686***	3.435***	3.675***

Table 6: Reversion to the Mean

	Dependent variable:			
	Margin_Period			
	(1)	(2)	(3)	(4)
Margin_TOT_Lag1	-0.018^{**}	-0.023^{***}	-0.019^{**}	-0.024^{***}
	(0.008)	(0.008)	(0.008)	(0.008)
SentDiff_Total_Lag1		0.001^{**}		0.001^{**}
		(0.0004)		(0.0004)
VolDiff_Total_Lag1			0.00004	0.00005
			(0.0001)	(0.0001)
Vegas_Line	0.005	0.016	0.005	0.017
	(0.018)	(0.019)	(0.018)	(0.019)
QualityDiff	1.679^{**}	1.364^{*}	1.631^{**}	1.303
	(0.802)	(0.815)	(0.805)	(0.819)
TwitterDiff	0.0004	0.0005	0.0003	0.0003
	(0.001)	(0.001)	(0.001)	(0.001)
Min_End	0.002	0.001	0.002	0.002
	(0.006)	(0.006)	(0.006)	(0.006)
Observations	1,306	1,306	1,306	1,306
R^2	0.011	0.014	0.011	0.015
Adjusted R^2	0.007	0.010	0.007	0.009
Residual Std. Error	2.339	2.336	2.340	2.337
F Statistic	2.865^{**}	3.111^{***}	2.454^{**}	2.744^{***}

Table 7: Logistic Regressions for Winner Prediction

	Dependent variable:			
	Winner			
	(1)	(2)	(3)	(4)
Vegas_Line	-0.041^{***}	-0.034^{**}	-0.044^{***}	-0.038^{***}
	(0.014)	(0.014)	(0.014)	(0.014)
Margin_TOT	0.173^{***}	0.166^{***}	0.168^{***}	0.160^{***}
	(0.012)	(0.012)	(0.012)	(0.012)
SentDiff_Total		0.001^{***}		0.001^{***}
		(0.0005)		(0.001)
VolDiff_Total			0.0002^{**}	0.0002^{**}
			(0.0001)	(0.0001)
Observations	1,346	1,346	1,346	1,346
Log Likelihood	−594.303	−590.365	−591.049	−587.436
Akaike Inf. Crit.	1,194.606	1,188.730	1,190.097	1,184.873

7 Bibliography

American Free Press. "Global Sports Gambling worth 'up to $3 Trillion'" Mail Online. Associated Newspapers, 15 Apr. 2015. Web. 01 May 2016.

"Betfair Facts." Betfair. N.p., n.d. Web. 01 May 2016.

Davis, Owen. "March Madness 2015: Getting To The NCAA Finals Costs A Lot, But The Rewards For Most Are Slim." International Business Times. N.p., 18 Mar. 2015. Web. 01 May 2016.

Go, Alec, Richa Bhayani, and Lei Huang. "Feature Extraction for Sentiment Classification on Twitter Data." IJSR International Journal of Science and Research (IJSR) 5.2 (2009): 2183-189. Stanford University. Web. 1 May 2016.

Ibrahim, Mohd Naim Mohd, and Mohd Zaliman Mohd Yusoff. "Twitter Sentiment Classification Using Naive Bayes Based on Trainer Perception." 2015 IEEE Conference on E-Learning, E-Management and E-Services (IC3e) (2015): n. pag. Web.

Kreutz, Liz, Video Ali Dukakis, and Tom Thornton. "In Las Vegas, No Bets on Hillary Clinton (Literally)." ABC News. ABC News Network, 9 May 2015. Web. 01 May 2016.

Ota, Kevin."ESPN Tournament Challenge: 13 Million Brackets Set New All-Time Record - ESPN MediaZone." ESPN MediaZone. N.p., 17 Mar. 2016. Web. 01 May 2016.

"Report for Selected Countries and Subjects." International Monetary Fund. N.p., 1 Oct. 2015. Web. 01 May 2016.

Silver, Nate, and Jay Bruce. "How FiveThirtyEight Is Forecasting The 2016 NCAA Tournament." FiveThirtyEight. ESPN, 13 Mar. 2016. Web. 1 May 2016.

Sinya, Shiladitya, Chris Dyer, Kevin Gimpel, and Noah Smith. "Predicting the NFL Using Twitter." Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics. (2013): n. pag. Print.

S. J. K. and R. S. E., "Relevance weighting of search," Journal of the American Society for Information Science, vol. 27, no. 3, pp. 129-146, 1976.

"Sports Betting Math." The Sports Geek. N.p., n.d. Web. 01 May 2016.

Stossel, John. "Prediction Markets More Accurate than Polls." WND. N.p., 05 Jan. 2016. Web. 01 May 2016.

Zhao, Siqi, Lin Zhong, Jehan Wickramasuriya, Venu Vasudevan, Robert LiKamWa, and Ahmad Ahmad Rahmati."SportSense: Real-Time Detection of NFL Game Events from Twitter." Technical Report (2012): n. pag. Print.

Yu, C. and Salton, G., "Precision Weighting?An Effective Automatic Indexing Method," Journal of the ACM (JACM), vol. 23, no. 1, pp. 76-88, 1976.

8 Appendices

Table 8: Official Twitter Accounts For Relevant Teams

Team	Twitter Account	Team	Twitter Account
Arkansas-Little Rock	@LittleRockMBB	Notre Dame	@NDmbb
Butler	@ButlerMBB	Oklahoma	@OU_MBBall
Cal State Bakersfield	@CSUB_MBB	Oregon	@OregonMBB
California	@CalMensBBall	Oregon State	@OregonStateMBB
Cincinnati	@GoBearcatsMBB	Pittsburgh	@HailtoPittHoops
Connecticut	@UConnMBB	Providence	@PCFriarsmbb
Dayton	@DaytonMBB	Saint Joseph's	@SJUHawks_MBB
Duke	@Duke_MBB	South Dakota State	@GoJacksMBB
Gonzaga	@ZagMBB	Stephen F. Austin	@SFA_MBB
Green Bay	@gbphoenixmbb	Syracuse	@Cuse_MBB
Hawaii	@HawaiiMBB	Temple	@TUMBBHoops
Holy Cross	@HCrossMBB	Texas	@TexasMBB
Indiana	@IndianaMBB	Texas A&M	@AggieMensHoops
Iowa	@IowaHoops	UNC Asheville	@UNCAbasketball
Iowa State	@CycloneMBB	Utah	@Runnin_Utes
Kansas	@KUHoops	VCU	@VCU_Hoops
Kentucky	@KentuckyMBB	Villanova	@NovaMBB
Maryland	@TerrapinHoops	Virginia	@UVAMensHoops
Miami	@CanesHoops	Weber State	@WeberStateMBB
Michigan	@umichbball	West Virginia	@WVUhoops
Michigan State	@MSU_Basketball	Wichita State	@GoShockers
Middle Tennessee	@MT_MBB	Wisconsin	@BadgerMBB
North Carolina	@UNC_Basketball	Xavier	@XavierMBB
Northern Iowa	@UNImbb	Yale	@Yale_Basketball