Predicting Wins and Spread in the Premier League Using a Sentiment Analysis of Twitter

Robert P. Schumaker[1], A. Tomasz Jarmoszko[2] and Chester S. Labedz Jr.[3]

[1]Computer Science Dept., University of Texas at Tyler, Tyler, Texas 75799, USA

[2]Management Information Systems Dept., Central Connecticut State University, New Britain, Connecticut 06050, USA

[3]Management & Organization Dept., Central Connecticut State University, New Britain, Connecticut 06050, USA

rob.schumaker@gmail.com, jarmoszko@ccsu.edu and clabedz@gmail.com

Word Count: 8,311

Abstract

Can the sentiment contained in tweets serve as a meaningful proxy to predict match outcomes and if so, can the magnitude of outcomes be predicted based on a degree of sentiment?

To answer these questions we constructed the CentralSport system to gather tweets related to the twenty clubs of the English Premier League and analyze their sentiment content, not only to predict match outcomes, but also to use as a wagering decision system. From our analysis, tweet sentiment outperformed wagering on odds-favorites, with higher payout returns (best $2,704.63 versus odds-only $1,887.88) but lower accuracy, a trade-off from non-favorite wagering. This result may suggest a performance degradation that arises from conservatism in the odds-setting process, especially when three match results are possible outcomes. We found that leveraging a positive tweet sentiment surge over club average could net a payout of $3,011.20. Lastly, we found that as the magnitude of positive sentiment between two clubs increased, so too did the point spread; 0.42 goal difference for clubs with a slight positive edge versus 0.90 goal difference for an overwhelming difference in positive sentiment. In both these

cases, the cultural expectancy of positive tweet dominance within the twitter-base may be realistic. These outcomes may suggest that professional odds-making excessively predicts non-positive match outcomes and tighter goal spreads. These results demonstrate the power of hidden information contained within tweet sentiment and has predictive implications on the design of automated wagering systems.

Keywords: Business Intelligence, Sentiment Analysis, Twitter, Sports Analytics, English Premier League, Crowdsourcing

1. Introduction

Predicting the outcomes of sporting events has a long and rich tradition. Since ancient times people have designed methods to divine natural and physical events. Today the urge to successfully predict still grips gamblers and academics alike. Prediction is no longer an art and probability is now considered a complex science. The most difficult aspect of prediction rests with identifying the relevant parameters and separating them from the noise of the event. Critical parameters are sometimes difficult to identify or measure, are constantly changing, or are not yet fully explored. The inability to correctly identify the most relevant parameters can sometimes lead to crippled systems relying on unimportant data or, worse, may create forecasts not based on sound science (e.g., basing predictions on the color of a uniform).

One way to simplify this problem of choosing and weighting parameters is to implement crowdsourcing as a forecasting tool. In James Surowiecki's seminal book, The Wisdom of Crowds [1], he made claim that large groups of individuals are better at making forecasts in conditions of uncertainty than are domain experts. This stems from collective intelligences, on the whole, being better able to properly sift through and analyze data than an individual. About the same time, another milestone book, Moneyball [2], popularized the use of statistics and sabermetric techniques (a quasi-scientific methodology of identifying relevant sports metrics,

applying and refining them) in sports. Academic focus was not too far behind as the field of sports analytics gained popularity [3].

Twitter has been a boon to academic research with rich crowd-based datasets that can be easily collected and analyzed. Its data have been used to make predictions on phenomena as diverse as crime [4], the stock market [5], political elections [6], public opinion polls [7], public health [8] and movie sales [9]. The lure of Twitter for academic research is two-fold. It provides a rich topical memory in the form of author-annotated hashtags, and provides a record of trends in public perception. Coupled together, academics can mine the twitterverse (i.e., universe of Twitter data) and identify valuable insights.

Our research aims to demonstrate a crowdsourced system that can extract sentiment information from Twitter to make match and point spread predictions in the English Premier League. Further, we analyze specific sentiment components such as tone and polarity, use them to calculate the degree changes in club-level sentiment and predict the magnitude of match goal differentials.

The rest of this paper is framed as follows. Section 2 provides an overview of literature concerning crowdsourcing, sentiment analysis and relevant studies. Section 3 presents our research questions. Section 4 introduces the CentralSport system and explains its various components. Section 5 sets up the Experimental Design. Section 6 details the Experimental Results and discussion. Finally Section 7 presents study conclusions and suggests further extensions of this stream of research.

2. Literature Review

Crowdsourcing is a tool through which the average of crowd forecasts is used to predict future events [1]. In sports, this forecasting behavior generally equates to wagering on favorites

and has been found to be a fairly accurate and reliable indicator of expectations. In a study of UFC fights, crowds were better able to predict wins (85.7%) than were bookies (67.6%) [10]. In a study of the wagering on matches in the Bundesliga (Germany's premier football league), crowds were found to be more accurate in their forecasts than bookies [11]. In a study of the FIFA World Cup 2006 tournament, crowds were also better able to predict winners than were pre-tournament rankings or random chance [12]. All three studies suggest that crowds were able to collectively make more accurate forecasts by weighting the data, not scientifically, but naïvely. Their decision-making contrasts with the weighting schemes designed by experts, which are driven by experience and previously seen patterns of data and profits generation. While crowdsourcing has demonstrated itself as an effective prediction tool, critics observe that some bettors may simply select the crowd favorite rather than evaluate the data independently [13]. This reinforcing behavior could lead to over-valuing the crowd favorite and can have an impact on accuracy. However, empirical evidence has shown that this typically encompasses a minority of wagering activity [13].

2.1 Odds-Makers and Wagering

Before two clubs take to the soccer playing field, or pitch, odds-makers will set a betting line in an attempt to draw an equal currency amount of wagers on each club. By balancing the wagers, in effect the losing side of the wager pays the winning side minus the sportsbook's[1] commission. Should the line become unbalanced, the sportsbooks are responsible for the difference and this imbalance may cause them a monetary loss. If one club is heavily favored, the sportsbook will increase odds on the less favored club to give bettors an incentive to wager longshots and rebalance the line.

[1] We use the terms odds-makers and sportsbooks interchangeably.

One type of popular wagering system is the Moneyline. In this system, clubs with negative values are favored and clubs with high positive values are longshots. Odds and payouts are based on a unit of £100. For example, Arsenal and Swansea may have a Moneyline of -220 and +550 respectively. For the bettor on Arsenal (the favorite) they would need to wager £220 to win £100. For the Swansea bettor, they would wager £100 in a bid to win £550. The odds-makers attempt to gauge betting interest on the match and adjust the Moneyline to balance the monetary amounts wagered on both clubs.

Once odds are initially set, odds will move in response to the currency amount of wagers to continually balance the odds-makers match balance sheet. Because there are a variety of odds-makers with which to place wagers, the amount of currency wagering between clubs may differ between sportsbooks. This will lead to differences in odds between books. Typically the sportsbook with the more favorable odds will attract more wagering and will thus force their odds to return to market equilibrium.

2.2 Social Media and Prediction

There has been much academic interest in using social media to make predictions. These predictions have crossed a diverse number of domains because of social media's rich crowd-based datasets that can be easily collected and analyzed. These areas have included crime, movie sales, politics, the stock market and sports. In a study of social media prediction and crime, Twitter content was topically clustered into distinct discussion areas, correlated with the geo-location of the tweet and fed into a crime prediction model to demonstrate better predictive performance in 19 of 25 crime types [4]. Even though the tweeters were not making predictions themselves of crimes, their topics of discussion were a decent predictor.

In a study that correlated social media attention to movie sales, Twitter content was found to be a good predictor [9]. In particular, positive twitter content was associated with higher movie sales whereas negative content was associated with lower movie sales. The authors also noted that tweets expressing an intention to watch a particular movie had the strongest predictive effect. In this case, tweeters were expressing their intention to watch or not watch a particular movie. This differs from the crime prediction study where the topics of the tweets themselves were used for prediction.

In US politics, Twitter tweet counts and sentiment have been used to predict voter outcomes. In a study of the German Federal elections, Tumasjan et. al. used a simple and easy to implement method of counting tweets that mention a candidate or political party [14]. Their reasoning was that tweets mentioning a candidate or party indicated their voting intention. This method was fairly accurate when applied against German federal elections with an error rate of 1.65%. When more complex methods were investigated such as using a sentiment analyzer to further determine voter intention, the results were not as precise [15]. Although the results are dependent upon the methods of how sentiment was captured and analyzed. It was further noted that sentiment polarity methods, at the time, were not sophisticated enough to recognize political language nuances, had poor performance and produced unacceptable errors [16].

Another political study investigated using a moving average of candidate, or elected official, tweet sentiment as a replacement to traditional polling services [7]. This work noted that natural language processing techniques achieved an 80% correlation.

In a study of the sentiment of financial news articles and the stock market, Schumaker et. al. used the article sentiment as a method for predicting the magnitude of stock price movements

immediately following article release [17]. Their work found that articles with a negative sentiment were easiest to predict, netting a 3.04% trading return using a simple trading engine.

Fans post tweets in order to express their personal feelings, most fundamentally (as we collected them) about the strengths, weaknesses and prospects of the team they follow and its next opponent. Admittedly, the tone and polarity of fans' tweets likely do not affect match results, except in cases in which extraordinary fan base sentiment might exceptionally motivate or demotivate a team. Fans' tweets by themselves are not likely to influence betting lines offered to bettors, which of course potentially includes those fans. Nonetheless, the tones and polarities of opponent fans' tweets may modestly affect initial betting line odds or later adjustments thereto, as we note elsewhere.

As tweets suggest the expected outcome on the field (i.e., the full-time score line), the wisdom of crowds premise becomes more credible. Many thousands of fans, well versed through years as footballers themselves before advancing age and injuries transformed them into amateur pundits, bring considerable collective intelligence to sentiment crowdsourcing. This fan base is sufficiently diverse, decentralized through the reach of the Internet, able to be summarized and rabidly independent. In expressing their sentiments about real events on the turf pitch from odds-distorted results on the shadow field of wagering, fans' tweeted views contain useful raw information about future score lines.

In many of these studies, the act of tweeting was treated as an intention to act even though specific predictions were not solicited. The tweets themselves were used as predictive proxies. Putting this in terms of sports prediction from tweets, fans can choose to express their emotions towards their team as positive, negative or neutral (an intention to act), or choose not to tweet at all. This sentiment content can then be treated as an extra dimension of information.

2.3 Sentiment Analysis

Investigating the role of sentiment as a predictor further, identifying fan or club-based sentiment could help in odds-setting or refinement. In sentiment analysis the focus is on analyzing direction-based text to determine tone (whether the author is positioning the text as objective/factual or subjective/opinion-based) and polarity (whether the author's word choice is positive or negative) [18].

Sentiment analysis techniques have been well-studied in stock prediction [19, 20] , online product sales [21] and corporate reputation [22]. One of the major findings was that negative sentiment was a better predictor of downward moves in firm value than were other sentiment-based techniques [23]. Further work identified positive and negative polarity in financial news to be consistent with human judgment [24] on firm performance [25-27].

To measure sentiment, one well-known and tested tool is OpinionFinder. This tool can identify sentence-level tone and polarity based on user-selected terms [28]. OpinionFinder was developed by Wiebe et. al. based on a series of publications, such as the subjective sentence classifier [29, 30], and the polarity classifier [31]. It performs well compared to the baseline MPQA Opinion Corpus, with an accuracy of 74%, subjective precision of 78.4%, subjective recall of 73.2% and a subjective F-measure of 75.7%, as compared to baseline MPQA's accuracy of 55.3%.

2.4 Sentiment Analysis in Sports

The use of Twitter for prediction has become more popular among researchers. Schoen et. al. posits that social media projects an impression "as a widely accepted and reliable source of data for predicting future outcomes [16]." Following up on this social media impression for prediction in sports, two studies have tackled using social media the prediction of outcomes of North American football games. Hong and Skiena [32] use Lydia – a text analytics system – to

analyze four sources of online text streams: LiveJournal blogs, RSS blogs captured by Spinn3r, Twitter and traditional news media. Using indicators of positive and negative sentiment within each message, the authors develop a measure of relative favorableness for teams which is then translated into a match prediction. They report the accuracy of their predictive method – when applied to 30 games between 2006 and 2009 -- as 60%. Sinha et. al. [33] undertake a study of the relationships between North American football games and tweets which mention the teams involved. The authors predict game outcomes for: 1) straight wins, 2) wins with/against the spread and 3) over-under point totals based on 10% of tweets exchanged during the 2010-2012 National Football League (NFL) seasons. Tweets were classified into weekly, pre-game and post-game categories and linked to specific teams via hashtags. Using logistic regression the authors determined prediction accuracy of 56%.

In European football (i.e., soccer), the most relevant studies are Godin et. al. [34] and Radosavljevic et. al. [35]. Although both studies aim to predict outcomes of English Premier League (EPL) games played during the 2013-2014 season, their methods differ. First, Godin et. al. establish baseline predictive indicators of naïve predictions (home team always wins), expert predictions (BBC pundit views) and bookmaker predictions (the averaged odds of some 50 bookmakers). The three baseline methods lead to predictive accuracy of 51%, 60% and 67% respectively. Godin et. al. then employ a variety of individual and combined methods including: two versions of statistical analysis, twitter volume, sentiment analysis, two versions of user prediction analysis, and combined methods of majority voting, early fusion and late fusion. Although they did not provide details of the number derivations, they claim predictive accuracy of 52% to 68%. It was also reported that a theoretical profit of 30% could have been realized in betting on EPL games during the second half of the 2013-2014 season.

Twitter is not the only possible source for input data for sentiment analysis. Radosavljevic et. al. developed a method based on Poisson regression which used 83.1 billion posts in Tumblr to predict outcomes of the 2014 World Cup [35]. This method estimates the likelihood of win/draw/loss outcomes from vector elements that are based on the number of mentions of teams and players. The researchers trained their method on two years of international game data leading up to World Cup games. Application of their method to the World Cup resulted in a success rate of nearly 50% which is quite good for a three class problem like football.

2.5 Odds Crowdsourcing versus Sentiment Crowdsourcing

The act of wagering on a match has the potential to influence odds movement. If enough actors in the domain engage in this activity, it can be considered a type of crowdsourcing where actors are collectively predicting match outcomes. Surowiecki describes how this collective intelligence works:

> There are four key qualities that make a crowd smart. It needs to be diverse, so that people are bringing different pieces of information to the table. It needs to be decentralized, so that no one at the top is dictating the crowds answer. It needs a way of summarizing peoples opinions into one collective verdict. And the people in the crowd need to be independent, so that they pay attention mostly to their own information, and not worrying about what everyone around them thinks [1].

For odds movement, crowdsourcing easily fulfills two of the requirements: diversity of individuals and summarization of verdict. The second and fourth requirements, decentralization and paying attention to their own information, could be argued. While sportsbooks are not consolidated entities, their combined odds are generally on par with market equilibrium. We argue that it is not so much a matter of decentralization as it is market feedback that can lead to an arbitrage opportunity.

Taking this idea of market feedback further, the fourth part of Surowiecki's definition of crowd intelligence is being independent in decision-making and paying attention to their own information. While some bettors may behave in this manner, the public display of odds is meant to balance currency wagering and is not a true representation of crowd expectation. The reasons for wagering from a bettor's standpoint include luck and entertainment, desperation (e.g., wagering longshots) or wagering favorites in cases of laziness. Thus we cannot expect an odds-market to behave in a completely independent crowd-sourced manner.

For sentiment of tweets, crowdsourcing fulfills all four of the requirements. Tweeters are a generally diversified group, linked by their interest in the EPL. It is decentralized from the standpoint that there is no centralized authority establishing sentiment or providing instant feedback. It can be used to summarize sentiment (i.e., the focus of this study) and it is mostly independent (although it could be argued that threaded discussion in Twitter can ensue).

Given these arguments, we believe that crowdsourced sentiment may be a better predictor of match outcome than wager crowdsourcing.

2.6 Research Gaps

Our review of the literature identified several opportunities not previously pursued, notably a lack of sentiment studies in football. Although Godin et. al. performed some research within this domain, we seek to extend the body of work to investigate both tone and polarity measures, and conduct a deeper investigation of wagering activity as an evaluative factor.

Another gap was that prior studies focused on a binary prediction of winners. We seek to leverage twitter sentiment and look for signals in the data such as the magnitude of polarity that may lead to predicting in-match goal differential. Our intent is to uncover sentiment information and apply it to match prediction in a novel and interesting way.

3. Research Questions

These gaps in the literature led to the following research questions:

1. *What signals exist in Twitter data that may provide match predictions?*

Following Surowiecki's crowdsourcing approach, we believe that tweeters (i.e., tweet authors) are better able to make match predictions than bettors and are able to convey those predictions through the sentiment of their word choice. Tweets that are positive may be indicative of a favorable match outcome whereas negative tweets may reflect pessimism and predict a potential loss. We feel that the normalized aggregate view of this sentiment information for each club and match may have predictive value.

2. *What role does sentiment magnitude have on successful match prediction?*

Similarly we believe that a normalized imbalance of positive vs negative sentiments may help predict goal differential. We plan to investigate the magnitude of match sentiment versus the respective club averages. We reason that a significant surge or drop in match-level sentiment versus their club average may be a predictive signal of potential goal difference. It is further believed that a club with a normalized surge in positive tweets may win by a larger margin than one in which the difference is less. Conversely, a club with a surge in negative tweets may be expected to lose by a larger margin than one where the difference is less than average.

3. *What is the impact of sentiment-based prediction on wagering returns?*

From prior studies, accuracy and wagering profit have been observed to have an inverse relationship (i.e., high accuracy/low payout or low accuracy/high payout). Does the same hold true in the EPL, and if so, how are crowds able to identify longshot wins over odds-makers?

4. System Design

To address our research questions, we built the CentralSport system as shown in Figure 1.

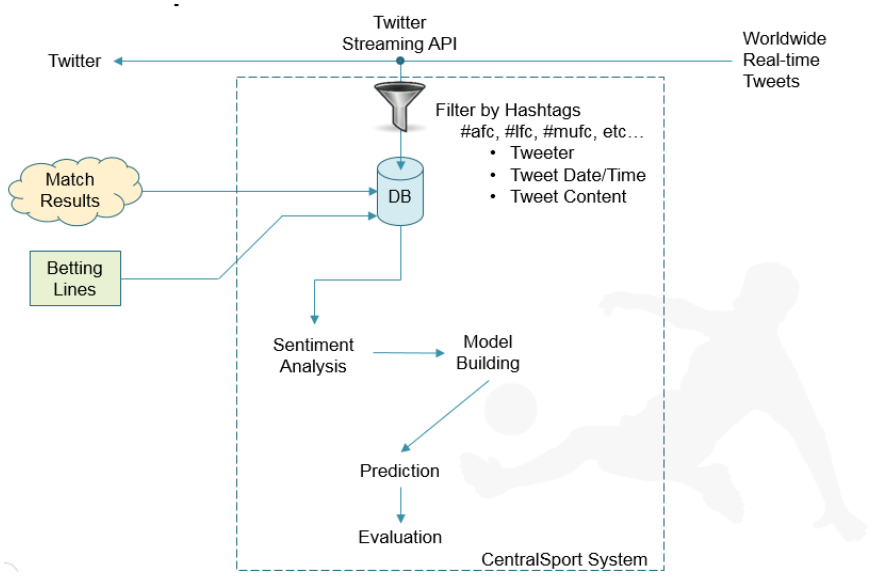

Figure 1. The CentralSport System

The CentralSport System is a twitter collector and sentiment analysis tool that interfaces directly with Twitter's streaming API and captures desired tweets in real-time based on the hashtag filter. Each tweet is composed of specific information, such as the Twitter handle (i.e., chosen name of the tweet author a.k.a. tweeter), the date/time it was tweeted and the tweet content. This feed is stored in a database along with match results and betting lines at match start, for each match, as shown in Table 1.

HomeClub	AwayClub	MatchStart	HomeScore	AwayScore	HomeOdds	DrawOdds	AwayOdds
#lfc	#nufc	2014-05-11 10:00	2	1	-556	706	1345
#mcfc	#whufc	2014-05-11 10:00	2	0	-588	760	1303
#ncfc	#afc	2014-05-11 10:00	0	2	396	300	-145
#saintsfc	#mufc	2014-05-11 10:00	1	1	175	273	141
#sufc	#swans	2014-05-11 10:00	1	3	13	244	225

Table 1. Sample Match Results and Betting Lines

Match results are obtained from ESPN.com and are manually entered into the system. Betting lines are acquired from OddsPortal.com which is an aggregation of 15 different sports books and follows the Moneyline wagering approach. Moneyline odds are unitless, meaning the

monetary unit (e.g., dollar, pound, euro, etc.) is irrelevant. We chose to use dollars for this research. Specific wagering details are further described in the Experimental Design section.

Once collected, each tweet is analyzed using OpinionFinder to identify sentiment information. It categorizes tweets in two axes, tone and polarity. Tone evaluates whether a tweet is subjective or objective. OpinionFinder classifies each sentence within the tweet, and the designation of tone for the tweet follows the majority of the individual sentences. In cases of a tie or ambiguity, the tweet is marked tone neutral. Here are examples of tweet tone:

Objective: *Andre Schurrle celebrates his first Premier League goal for Chelsea #cfc*

Subjective: *Has someone just took the batteries out of our players ???? Our players have just stopped functioning ???? #mufc*

Polarity evaluates the positive or negative bias of a tweet. Like tone, polarity classifies each sentence and uses majority rules. Tweets can be marked polarity neutral in cases of ambiguity or a tie. Examples of tweet polarity include:

Negative: *Supporting Newcastle is actually making me hate football. #nufc*

Positive: *Were delighted to confirm the signing of @R9Soldado from Valencia after successfully completing his medical. #thfc*

For tweets with more than one hashtag, we elected to label it with the first club hashtag. We reason that if a tweet mentions two or more clubs, the tweeter may have intended more emphasis towards the first club mentioned. While not perfect, we felt this to be an adequate automation compromise for the large volume of tweets collected.

5. Experimental Design

5.1 The Experiment

For this study we used data from the final three months of the 2013-2014 English Premier League season, February 16 through May 11, 2014. Tweets were gathered from the Twitter streaming API using one team-specific hashtag per club, identified by a domain expert. While we recognize that using only one hashtag per club may be a study limitation compared to gathering an entire universe of club-related tweets, the volume of tweets gathered offsets the limitation.

During this period, 122 matches were played. For each match we used tweets for the ninety-six hours up to match start, consistent with Hong and Skiena's work. From these data we constructed a baseline model that used aggregated odds-only data from OddsPortal.com to predict outcomes, and eight sentiment models that describe the data based on the axes of tone and polarity. Figure 2 depicts the sentiment models with an explanation to follow.

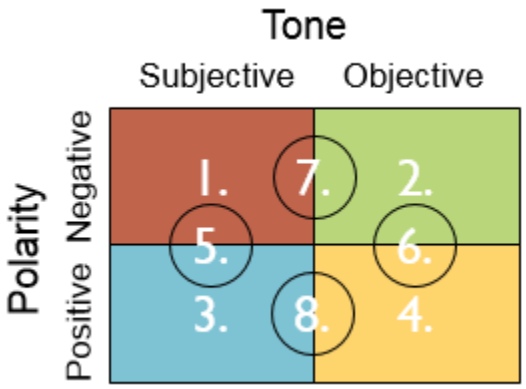

Figure 2. Sentiment Models

From this figure we develop eight models across the axes of tone and polarity; Model 1 – Subjective Negative tweets, Model 2 – Objective Negative, Model 3 – Subjective Positive, Model 4 – Objective Positive, Model 5 – All Subjective, Model 6 – All Objective, Model 7 – All Negative, and Model 8 – All Positive.

To address those tweets identified as either neutral tone or neutral polarity, we chose not to use them in the models. While it could be argued that neutral tone tweets should not have an impact on Models 7 and 8, All Negative and All Positive polarity respectively, we felt that using identical data across all models would provide a more robust and equal comparison between models.

For our first research question, we test sentiment between clubs of unequal tweets by normalizing the sentiment model data versus tweets for the particular club and match, and used it for comparison purposes as shown in Equation 1.

$$Max\left(\frac{\Sigma(Tweets \mid Model_n, Club_1, Match_m)}{\Sigma(Tweets \mid Club_1, Match_m)}, \frac{\Sigma(Tweets \mid Model_n, Club_2, Match_m)}{\Sigma(Tweets \mid Club_2, Match_m)}\right) \qquad \text{(Equation 1)}$$

For models using negative polarity sentiment (Models 1, 2 and 7), we expect that the club with the higher match normalized value would lose, following the logic that negative sentiment indicates anxiety and a potential for loss. Whereas for all other sentiment models we expected the club with the highest value to win. Table 2 demonstrates how match normalization works using Model 8 – All Positive tweets.

HomeClub	AwayClub	Date	#Htweets	#Atweets	#HMod8	#AMod8	HNrmlz	ANrmlz
Man United	Liverpool	3/16/2014	10,138	5,034	8,503	3,501	0.8387	0.6955

Table 2 – Example of match normalization for Model 8 (All Positive)

From this table, the home club variable *HNrmlz* is the number of positive tweets for Manchester United (8,503) in the ninety-six hours before match start, divided by the overall number of tweets for Man United (10,138) during the same period, excluding neutral categories. A similar calculation was performed for *ANrmlz*. Comparing the two values (0.8387 versus 0.6955 for Home and Away respectively), the Home team has greater subjective positive sentiment and is predicted to win the match.

For our second and third research questions, we use averaged club-based sentiment, where the number of sentiment-based tweets for the match is compared against an average value for the club. This measure indicates if a surge or drop in sentiment is occurring for a particular match, which may indicate predictive value. We then analyze each match, comparing values using the formula in Equation 2.

$$Max\left(\frac{\Sigma(Tweets \mid Model_n, Club_1, Match_m)}{\Sigma(Tweets \mid Model_n, Club_1) / \Sigma(m \mid Club_1)}, \frac{\Sigma(Tweets \mid Model_n, Club_2, Match_m)}{\Sigma(Tweets \mid Model_n, Club_2) / \Sigma(m \mid Club_2)}\right) \text{ (Equation 2)}$$

We expect models with higher values for Models 1, 2 and 7, to lose. For all other sentiment models we expected the club with the highest value to win. As an example using Manchester United from earlier, we use the number of positive tweets in the ninety-six hours before match start (8,503) and divide it by the average number of tweets for Model 8 – All Positive (8,317.2) for a normalized value of 1.0223 versus Liverpool at 8.0314. These values indicate that both clubs were more positive in their tweets than average, however, Liverpool was 8x their typical positive sentiment and the model would predict them to win the match (Liverpool did win and it was an odds upset).

5.2 The Collection

For the study period, 18,027,966 tweets were gathered. The average tweet length was 105.0 characters with a standard deviation of 34.9. Removing the neutral tone and polarity tweets and using tweets only within the ninety-six hours before match start left a dataset of 1,026,569 tweets as broken down by club and sentiment values in Table 3.

Club Name	Hashtag	# Tweets	# Objective	# Subjective	# Positive	# Negative
Arsenal	#afc	45,807	44,194	1,613	29,176	16,631
Aston Villa	#avfc	9,794	9,442	352	6,435	3,359
Cardiff City	#cardiffcity	727	705	22	397	330
Chelsea	#cfc	70,912	69,008	1,904	50,781	20,131
Crystal Palace	#cpfc	6,603	6,360	243	4,615	1,988
Everton	#efc	10,059	9,791	268	6,000	4,059
Fulham	#ffc	3,205	2,748	457	2,279	926
Hull City	#hcafc	2,214	2,126	88	1,385	829
Liverpool	#lfc	86,636	83,419	3,217	62,228	24,408
Manchester City	#mcfc	35,504	34,402	1,102	22,185	13,319
Manchester United	#mufc	171,499	166,459	5,040	121,310	50,189
Newcastle	#nufc	11,763	11,256	507	6,421	5,342
Norwich	#ncfc	4,513	4,319	194	2,830	1,683
Southampton	#saintsfc	4,682	4,596	86	3,407	1,275
Stoke City	#scfc	2,491	2,432	59	1,589	902
Sunderland	#sufc	4,620	4,294	326	3,668	952
Swansea	#swans	1,601	1,554	47	1,052	549
Tottenham	#thfc	12,450	12,013	437	7,219	5,231
West Bromwich	#wbafc	366	338	28	228	138
West Ham	#whufc	3,583	3,448	135	2,184	1,399

Table 3. Breakdown of club level sentiment

From this table, Manchester United, Liverpool, Chelsea, Arsenal and Manchester City had the most tweets, consistent with expectations based on club size and popularity. Clubs promoted to the 2013-2014 Premier League (i.e., Cardiff City, Crystal Palace and Hull City) had some of the fewest, but not the least, number of tweets.

This table indicates that the tone of tweets was mostly objective, 96.7% of tone. For polarity (e.g., whether the tweet is positive or negative), every club's tweeters were more positive (average 68.6%) than negative, expressing a mostly optimistic attitude. The most optimistic club tweeters followed Sunderland (79.4%), Southampton (72.8%) and Liverpool (71.8%). Looking at the optimistic club records (win-draw-loss), Sunderland was 4-2-7, Southampton 5-2-5 and Liverpool 10-1-1. It was interesting to note that it was not necessarily the clubs with the best records that had the highest fan optimism (as demonstrated by

Sunderland). We speculate that the optimism/pessimism differences may be due to regional cultural differences. Although interesting, this was not the focus of our research.

5.3 The Metrics

We evaluated the models on accuracy (i.e., how correct the models were versus actual results), payout (i.e., constructing a simple wagering algorithm to measure hypothetical payouts) and betting efficiency (i.e., the averaged payout per wager). Depending on the sentiment model to be tested, the relevant axes of tone (objective/subjective) and polarity (positive/negative) were used.

5.3.1 Accuracy

Accuracy is a measure of predicted match outcome versus actual outcome. Aggregating the average of the 122 matches for each model led to this measure. Based on the values of the normalized match data and the model, the system would select either the home or away club to win. If either the number of Home or Away tweets for the model was 0, then the match was not considered for that model. A Draw wager was considered, however, it was found that sentiment was unable to fully recognize a draw outcome and we chose to ignore this category. Even so, the performance returns from predicting just Home and Away match outcomes offset the need for Draw. Within our dataset, there were 6 draw occurrences.

5.3.2 Payout

Payout is a summed value of returns on hypothetical $100 wagers made on predicted match outcomes for the 122 matches. For models using negative polarity sentiment (Models 1, 2 and 7), the wagering engine would bet on the club with the highest normalized value to lose. All other models bet on the club with the highest normalized value to win. Further, a third decision of *No Bet* was used if either the number of Home or Away tweets for the model was 0.

5.3.3 Wagering Efficiency

Wagering efficiency is the aggregated wagering payout for all wagered matches divided by the number of matches wagered upon for each model. This value allows us to identify which models produce the best return with the least bankroll.

6. Experimental Findings and Discussion

6.1 What Signals Exist in Twitter data

To answer our first research question, *what signals exist in Twitter data that may provide match predictions*, we looked at the accuracy and payout of predictions. Table 4 presents the results.

	Correct	Incorrect	No Bet	Accuracy	Payout	Excess Return
Baseline	80	42	0	65.57%	$1,887.88	
Model 1	52	51	19	50.49%	$1,934.70	$46.82
Model 2	55	67	0	45.08%	$1,946.53	$58.65
Model 3	41	61	20	40.20%	$823.38	($1,064.50)
Model 4	58	64	0	47.54%	$2,270.55	$382.67
Model 5	44	72	6	37.93%	($195.54)	($2,083.42)
Model 6	61	61	0	50.00%	$2,704.63	$816.75
Model 7	57	65	0	46.72%	$2,614.53	$726.65
Model 8	55	67	0	45.08%	$1,708.52	($179.36)

Table 4. Accuracy and Payout results of models

The Baseline odds-only approach was correct on 80 of the 122 matches in our study for an accuracy of 65.57% holding consistent with Godin et. al.'s observation of 67%. None of the sentiment models outperformed Baseline on accuracy and only two models (Model 1 – Subjective Negative and Model 6 – All Objective) were at or exceeded 50.0%. A careful reader will observe that Model 5 – All Subjective does not appear to reflect a weighted average of Model 1 – Subjective Negative and Model 3 – Subjective Positive, because the tweet limit threshold per match is exceeded for some matches in Model 5. The same applies to Model 6 – All Objective.

While the accuracy results are not attractive, accuracy only presents one facet of the results whereas the more interesting metrics to the gambler or betting house is payout and betting efficiency. Payout measures the hypothetical return on a $100 wager for either the home or away club, with the exception of the *No Bet* category. *Payout* is the amount of return derived from wagering, minus the initial wager amount. The column *Excess Return* is the payout of the model minus Baseline. Positive *Excess Return* values indicate a return greater than Baseline. From the table, Baseline had a $1,887.88 payout. Seven of the eight sentiment models also exhibited a positive payout. Only Model 5 (All Subjective) with a $195.54 payout loss did not. This is the result of fan sentiment skewing slightly towards longshot wagers, decreasing accuracy with the trade-off of better payouts.

Further, models that incorporate some level of positive and subjective tweets (Model 3 – Subjective Positive, Model 5 – All Subjective and Model 8 – All Positive) had negative excess returns. The sentiment models with the best excess returns were Model 6 (All Objective) with an excess return of $816.75 above baseline, Model 7 (All Negative) with $726.65 and Model 4 (Objective Positive) with an excess return of $382.67 over Baseline.

While the models that incorporated sentiment did not exhibit better prediction accuracy than the odds-only approach, two of the models that used subjective-only data reported negative excess returns versus the models that incorporated some form of objective data. This observation was unexpected and counter-intuitive at the surface. In prior sentiment work on financial news articles and their impact on stock price, researchers had found that subjective articles were better predictors of price movement than objective articles [17]. It was believed that the tone of the articles was influencing traders and consequently price. However, for this study, tweets of a subjective nature have a negative effect, and objective tweets are more

meaningful predictors. We believe this to be the case for two reasons. First, tweets are more descriptive, a reflection of individual expression, and not prescriptive. In other words, tweets are generally not reporting previously unknown events that may impact play which may be the case for financial news sentiment. The second reason is noise in the medium. With financial news articles, there are fewer articles to apply to a longer event horizon (i.e., trading day). With sports-related tweets, there are many (sometimes thousands) more tweets to correspond to an event of a much more limited duration. As a consequence of the deluge of information (one could argue the quality of the information too) and shorter event duration, the noise from tweets would be greater and hence less influential (if at all) than would financial news articles.

6.2 What Role Does Polarity Magnitude Have on Match Prediction?

To answer our second research question, *what role does sentiment magnitude have on successful match prediction*, we looked at the polarity models, Models 7 and 8, All Negative and All Positive tweets respectively. For the analysis we normalized data by dividing the number of tweets for each match by the average number of tweets for the model over the period of study. The normalized values tell us if there is a surge or drop in positive/negative sentiment, and by how much. Next, we compared the values between clubs for each match (e.g., Manchester United (1.0223) versus Liverpool (8.0314) is a difference of 7.0090 on Liverpool's behalf) and used these differences as sentiment magnitudes. A sliding threshold from 0 to 10 in 0.1 increments was introduced, where only sentiment magnitude differences greater than or equal to the threshold value was used for accuracy and payout calculations. Figure 3 shows the accuracy and payout results of the models versus the sliding sentiment magnitude thresholds.

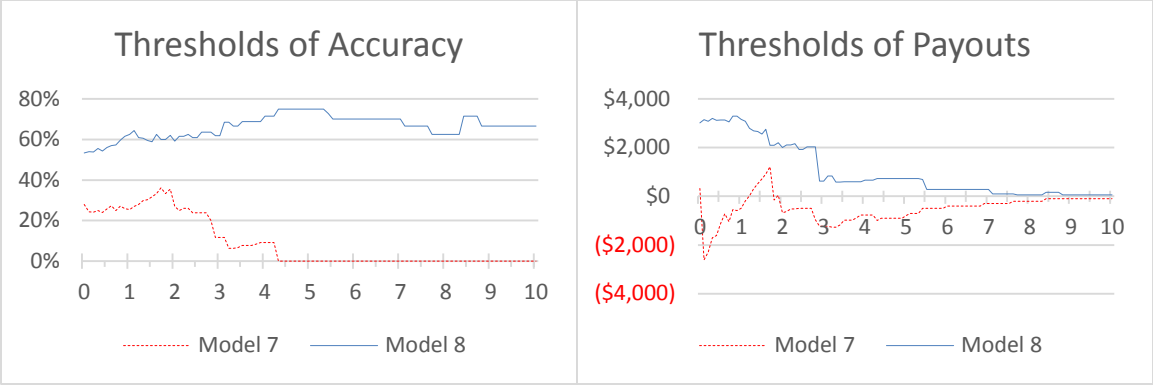

Figure 3. Accuracy and Payout of Models 7 and 8 versus Sentiment Thresholds

For a sliding threshold (sentiment magnitude) of zero, we wager on all matches (excluding *No Bet*). Whereas for sentiment magnitude of 1.0, we wager only on those matches where the difference in sentiment magnitude is at least 100% greater. Positive tweets at sentiment magnitude zero had an accuracy of 53.28% and payout of $3,011.20, versus negative polarity with 27.87% accuracy and $315.17 payout. Model 8 (All Positive) had good accuracy results that improved to 75.00% on thresholds between 4.3 and 5.3 inclusively, on 12 matches, before declining. This same model also had sizeable payouts, peaking at $3,295.24 at 0.9 threshold on 52 matches, before declining. Model 7 (All Negative) had comparatively worse accuracy, peaking at 36.1% at 1.7 threshold on 36 matches, before descending to 0% accuracy at thresholds 4.3 and greater. Model 7 also exhibited worse payouts by comparison, maxing out at $1,215.03 at 1.7 threshold on 36 matches. From these data it would appear that positive sentiment was a better predictor of match outcomes. We believe that coupled with the fan optimism finding from earlier, positive sentiment may be the cultural expectancy for the English Premier League twitter-base, with the majority of tweeters from the UK. Looking into this further, we also note an excessive number of All Positive tweets (276,628) versus All Negative (132,906).

Next we examined if the sentiment magnitude was a predictor of match goal differential. It was expected that as sentiment magnitude increased between clubs it would have a similar increase on expected goals. Table 5 depicts the average of goals broken down by sentiment magnitude.

Sentiment Magnitude	Model 7	Model 8
0.0 >= x < 1.0	+0.60	+0.42
1.0 >= x < 2.0	+0.69	+1.07
x >= 2.0	+1.15	+0.90

Table 5. Model 7 and 8 Average of Goals versus Sentiment Magnitude

For both models, as the magnitude of sentiment difference between clubs increased, so too did the goal differential. However, Model 7 (All Negative) showed a contrarian relationship. As tweeters became more negative towards the club, the number of goals scored for that club increased leading to that club winning. We speculate that this is attributable to opposing club tweeters disparaging a stronger opponent. Looking deeper into the data, 76.9% of the negative sentiment magnitude greater than or equal to 2.0 was directed towards the odds-favorite club. Seven times this phenomenon was observed against Liverpool, eight times against Manchester United, four times against Chelsea and one time by Chelsea against Liverpool, four times by Aston Villa, and two times by Cardiff City, Manchester City and Sunderland against stronger opponents. It would appear that it wasn't the clubs' fans disparaging their own team, but rather spirited conversation from the weaker opponent.

Model 8 (All Positive) similarly showed an increase in goals, however, for sentiment magnitudes over 2.0x different, the average goals decreased (0.90). While still positive and consistent with expectations, this deviation could indicate an overconfidence in the club.

6.3 What is the Impact of Prediction on Wagering Returns?

To answer our third research question, *what is the impact of sentiment-based prediction on wagering returns*, we looked at balancing accuracy and payout with betting efficiency, which provides the average return per wager minus the initial bet.

In looking at the models, both Models 7 and 8 were profitably engaged in betting against favorites and longshot wagering. We define betting against the favorites as seeking a return between 1 to 2 times the wager, based on the odds (e.g., Moneyline odds between +100 and +199 inclusive). We further define longshot wagers as seeking a return of 2 or more times the wager (e.g., Moneyline odds of +200 or greater). While we recognize that some matches will have positive Moneyline odds for both clubs (e.g., meaning a wager either way would be against the favorites using our definition), we feel that breaking apart the wagering activity into these distinct buckets will provide additional insight as shown in Table 6.

Model 7	Baseline	M7 Against		Baseline	M7 Longshots
Accuracy	50.00%	50.00%		65.45%	23.64%
Payout	$405.00	$412.00		$907.36	$1,627.00
Bet Efficiency	$14.46	$14.71		$16.50	$29.58
Model 8	Baseline	M8 Against		Baseline	M8 Longshots
Accuracy	50.00%	50.00%		68.52%	22.22%
Payout	$422.00	$429.00		$1,189.37	$1,003.00
Bet Efficiency	$15.07	$15.32		$22.03	$18.57

Table 6. Results of Wagering Activity Against Favorites and Longshots

From this table several interesting trends emerge. First, for wagering *Mx Against*, against the favorites across models (left-side of the table), the accuracy, payouts and betting efficiency values are fairly similar to Baseline on the same matches, with a slight edge to both Models 7 and 8. This would indicate a weak relationship between both all positive and all negative tweet sentiment and a better return on match prediction than following an odds-only approach ($14.71 versus $14.46 for Model 7 and $15.32 versus $15.07 for Model 8).

Second, in looking at *Mx Against* and *Mx Longshots* (left to right), we notice that accuracy decreases, and payouts increase, consistent with prior results. What is interesting is that Baseline did not follow the same pattern. For Baseline, both Accuracy and Payout increased. We theorize that this is a result of the three classes of outcomes: win, draw and loss. Because of the three outcomes, sometimes Moneyline wagering will show strong positive values for all classes (especially if clubs are evenly matched). Looking through the data this did appear to be the case and helps to explain the discrepancy.

Third, in wagering on just *Mx Longshots* (seeking returns 2 or more times the wager), Model 7 outperformed Baseline on payout and betting efficiency, but for Model 8, Baseline garnered the higher values. Returning to our earlier comment in the second research question, we speculate that this may indicate a sentiment overconfidence in the wagered clubs, especially given the strong positive sentiment that *M8 Longshots* portrays.

From our observations it would appear that twitter sentiment can be effectively used to uncover arbitrage wagering opportunities.

7. Conclusions and Future Directions

From our study we found that crowdsourced sentiment can be a better predictor of match outcomes than crowdsourced odds. In looking at accuracy and payout, the crowdsourced odds-only (Baseline) approach had the highest accuracy versus the eight sentiment models tested. However, in terms of payout, five of the eight sentiment models had higher returns (Subjective Negative $46.82, Objective Negative $58.65, Objective Positive $382.67, All Objective $816.75 and All Negative $726.65). The three models with returns less than Baseline were all clustered around Subjective Positive (Subjective Positive -$1,064.50, All Subjective -$2,083.42 and All Positive -$179.36). Of the three, only All Subjective lost money, -$195.54. We believe that

crowdsourced sentiment was better at identifying longshot wagers as evidenced by five of the sentiment models. For the other three we found the subjective positiveness harmed the results and believe this to be a reaction to events and overconfidence in club performance, rather than rational prescriptive observation transcribed to tweet sentiment.

In looking specifically at the models of All Positive and All Negative sentiment and evaluating the surge/drop of match sentiment versus their club average, we found that this technique led to higher accuracy and payouts for the All Positive model (53.28% accuracy and $3,011.20 payout). Conversely, All Negative showed a marked decline in performance (27.87% accuracy and $315.17 payout). This result indicates that positive surges (above average club levels) in tweet sentiment generally leads to better match predictability. Tweet authors recognize some factor in their clubs' performance and express it through tweet sentiment. While we expected a similar result with negative sentiment (e.g., recognize something wrong with the club and expect a loss), this was not the case. However, upon a deeper analysis it appears that tweeters from weaker clubs were purposefully injecting negative sentiment into the feeds of their stronger opponents.

Lastly, in examining All Positive and All Negative's wagering behavior against favorites and on longshots, both models exhibited a decrease in accuracy and increase in payouts when wagering on longshots. While the system sacrificed accuracy, it made up for it in payouts. All Negative increased payouts from $412.00 to $1,627.00 and All Positive increased from $429.00 to $1,003.00. When compared to the odds-only Baseline, All Negative outperformed Baseline in both payout and betting efficiency (betting efficiency of $14.71 versus $14.46 Baseline against the favorites and $29.58 versus $16.50 on longshots). All Positive showed a similar gain towards Baseline against the favorites ($15.32 versus $15.07) but not on longshots ($18.57

versus $22.03). This again was found to be the result of weaker opponents negatively tweeting against their stronger rivals.

There are many potential extensions to this research as the system we created could be ported to other sports domains. One such extension would be the inclusion of Draw categories. While we ignored this category in our study and still managed good results, future work should look into ways of algorithmically identifying Draws with good accuracy. Another extension would be to analyze tweets during a match or briefly thereafter. It might provide some additional insight into tweet author behavior based on goal differences such as more/less interest in matches with more/less goal differentials. A third extension would be to analyze tweets and performance in the first half of the season versus the second half. Sinha et. al. discovered season-half differences in the NFL and perhaps similar differences exist in the Premiership. Fourth, an analysis of tweets and retweets may prove interesting in answering the question of what conditions lead to the most retweets? Lastly, it would be interesting to merge sentiment and social network theory to identify the tweeters with the greatest say on setting the sentiment mood for a particular hashtag. It is quite clear that within this domain there are plenty of opportunities for further research.

References

1. Surowiecki, J., 2004. *The Wisdom of Crowds* New York: Doubleday.

2. Lewis, M., 2003. *Moneyball: The Art of Winning an Unfair Game*: WW Norton & Company.

3. Schumaker, R.P., O.K. Solieman, and H. Chen, 2010. Sports Knowledge Management and Data Mining. *Annual Review of Information Science and Technology*, 44(1):115-157.

4. Gerber, M., 2014. Predicting Crime using Twitter and Kernel Density Estimation. *Decision Support Systems*, 61:115-125.

5. Bollen, J., H. Mao, and X. Zeng, 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1-8.

6. Gayo-Avello, D., 2013. A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, 31(6):649-679.

7. O'Connor, B., et al., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *The Fourth Annual International AAAI Conference on Weblogs and Social Media*. Washington, DC.

8. Paul, M.J. and M. Dredze, 2011. You are what you Tweet: Analyzing Twitter for Public Health. *The Fifth Annual International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain.

9. Rui, H., Y. Liu, and A. Whinston, 2013. Whose and What Chatter Matters? The Effect of Tweets on Movie Sales. *Decision Support Systems*, 55:863-870.

10. Wise, S., 2009. Testing the Effectiveness of Semi-Predictive Markets: Are Fight Fans Smarter than Expert Bookies? *Collaborative Innovation Networks Conference*. Savannah, GA.

11. Spann, M. and B. Skiera, 2008. Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters. *Journal of Forecasting*, 28(1):55-72.

12. Luckner, S., J. Schroder, and C. Slamka, 2008. On the Forecast Accuracy of Sports Prediction Markets, in Negotiation, Auctions, and Market Engineering, H. Gimpel, et al., Editors, Springer: Berlin. p. 227-234.

13. Qiu, L., H. Rui, and A. Whinston, 2011. A Twitter-Based Prediction Market: Social Network Approach. *International Conference on Information Systems (ICIS)*. Shanghai, China.

14. Tumasjan, A., et al., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *The Fourth Annual International AAAI Conference on Weblogs and Social Media*, 10:178-185.

15. Gayo-Avello, D., 2011. Don't Turn Social Media into Another Literary Digest Poll. *Communications of the ACM*, 54(10):121-128.

16. Schoen, H., et al., 2013. The Power of Prediction with Social Media. *Internet Research*, 23(5):528-543.

17. Schumaker, R.P., et al., 2012. Evaluating Sentiment in Financial News Articles. *Decision Support Systems*, 53(3):458-464.

18. Wiebe, J., et al., 2004. Learning Subjective Language. *Computational Linguistics*, 30(3):277-308.

19. Hill, S. and N. Ready-Campbell, 2011. Expert Stock Picker: The Wisdom of (Experts in) Crowds. *International Journal of Electronic Commerce*, 15(3):73-101.

20. Schumaker, R.P. and H. Chen, 2009. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. *ACM Transactions on Information Systems*, 27(2).

21. Baek, H., J. Ahn, and Y. Choi, 2012. Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues. *International Journal of Electronic Commerce*, 17(2):99-126.

22. Li, T., G. Berens, and M. de Maertelaere, 2013. Corporate Twitter Channels: The Impact of Engagement and Informedness on Corporate Reputation. *International Journal of Electronic Commerce*, 18(2):97-125.

23. Tetlock, P., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139-1168.

24. Devitt, A. and K. Ahmad, 2007. Sentiment Polarity Identification in Financial News: A Cohesion-Based Approach. *Association of Computational Linguistics*. Prague, Czech Republic.

25. Das, S. and M. Chen, 2007. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9):1375-1388.

26. Davis, A., J. Piger, and L. Sedor, 2006. Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases, in *Technical Report*, Federal Reserve Bank of St. Louis.

27. Ghose, A., P. Ipeirotis, and S. Arun, 2007. Opinion Mining Using Econometrics: A Case Study on Reputation Systems. *Association of Computational Linguistics*. Prague, Czech Republic.

28. Wilson, T., et al., 2005. OpinionFinder: A System for Subjectivity Analysis. *Human Language Technology Conference*. Vancouver, Canada.

29. Riloff, E. and J. Wiebe, 2003. Learning Extraction Patterns for Subjective Expressions. *Conference on Empirical Methods in Natural Language Processing*. Sapporo, Japan.

30. Wiebe, J. and E. Riloff, 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Sixth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico.

31. Wilson, T., J. Wiebe, and P. Hoffman, 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, Canada.

32. Hong, Y. and S. Skiena, 2010. The Wisdom of Bookies? Sentiment Analysis Versus the NFL Point Spread. *The Fourth Annual International AAAI Conference on Weblogs and Social Media*. Washington, DC.

33. Sinha, S., et al., 2013. Predicting the NFL using Twitter. *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.

34. Godin, F., et al., 2014. Beating the Bookmakers: Leveraging Statistics and Twitter Microposts for Predicting Soccer Results. *KDD Workshop on Large-Scale Sports Analytics*. Sydney, Australia.

35. Radosavljevic, V., et al., 2014. Large-scale World Cup 2014 outcome prediction based on Tumblr posts, in *KDD Workshop on Large-Scale Sports Analytics*: Sydney, Australia.