Predicting the NFL Using Twitter

Shiladitya Sinha[1], Chris Dyer[1], Kevin Gimpel[2], and Noah A. Smith[1]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] Toyota Technological Institute at Chicago, Chicago IL 60637, USA

Abstract. We study the relationship between social media output and National Football League (NFL) games, using a dataset containing messages from Twitter and NFL game statistics. Specifically, we consider tweets pertaining to specific teams and games in the NFL season and use them alongside statistical game data to build predictive models for future game outcomes (which team will win?) and sports betting outcomes (which team will win with the point spread? will the total points be over/under the line?). We experiment with several feature sets and find that simple features using large volumes of tweets can match or exceed the performance of more traditional features that use game statistics.

1 Introduction

Twitter data has been used to predict and explain a variety of real-world phenomena, including opinion polls [18], elections [23], the spread of contagious diseases [20], and the stock market [2]. This is evidence that Twitter messages in aggregate contain useful information that can be exploited with statistical methods. In this way, Twitter may offer a way to harness the "wisdom of crowds" [22] for making better predictions about real-world events.

In this paper, we consider the relationship between National Football League (NFL) games and the Twitter messages mentioning the teams involved, in order to make predictions about games. We focus on the NFL because games are broadcast widely on television throughout the US and teams play at most once per week, enabling many to comment on games via social media. NFL football also has active betting markets. The most well-known is the point spread line, which is a handicap for the stronger team chosen by bookmakers to yield equal bets on both sides. Factoring in the bookmaker's commission, a betting strategy that predicts the winner "with the spread" in more than 53% of games will be profitable. In this paper, we build models to predict game and betting outcomes, considering a variety of feature sets that use Twitter and game statistical data. We find that simple features of Twitter data can match or exceed the performance of the game statistical features more traditionally used for these tasks.

Our dataset is provided for academic research at `www.ark.cs.cmu.edu/football`. It is hoped that our approach and dataset may be useful for those who want to use social media to study markets, in sports betting and beyond.

2 Problem Domain and Related Work

Each NFL regular season spans 17 weeks from September to January, with roughly one game played per week by each team. In each game, the **home team** plays at their own stadium and hosts the **away team**. The most popular wager in NFL football is to choose the team that will win given a particular handicap called the **point spread**. The point spread is a number set by bookmakers that encodes the handicap for the home team. It is added to the home team's score, and then the team with the most points is called the winner **with the spread (WTS)**. For example, if the NY Giants are hosting the NY Jets and the point spread is -4, then the Giants will have to win by at least 4 in order to win WTS. If the Giants win by fewer than 4, the Jets win WTS.[3] Also popular is to wager on whether the total number of points scored in the game will be above or below the **over/under line**.

Point spreads and over/under lines are set by sports betting agencies to reflect all publicly available information about upcoming games, including team performance and the perceived outlook of fans. Assuming market efficiency, one should not be able to devise a betting strategy that wins often enough to be profitable. In prior work, most have found the NFL point spread market to be efficient overall [16, 17, 4], or perhaps only slightly inefficient [6, 5]. Others pronounced more conclusively in favor of inefficiency [25, 9], but were generally unable to show large biases in practice [10].[4] Regardless of efficiency, several researchers have designed models to predict game outcomes [11, 21, 8, 15, 7, 1].

Recently, Hong and Skiena [12] used sentiment analysis from news and social media to design a successful NFL betting strategy. However, their main evaluation was on in-sample data, rather than forecasting. Also, they only had Twitter data from one season (2009) and therefore did not use it in their primary experiments. We use large quantities of tweets from the 2010–2012 seasons and do so in a genuine forecasting setting for both winner WTS and over/under prediction.

3 Data Gathering

We used Twitter (`www.twitter.com`) as our source of social media messages ("tweets"), using the "garden hose" (10%) stream to collect tweets during the 2010–2012 seasons. For the 2012 season, this produced an average of 42M messages per day. We tokenized the tweets using `twokenize`, a freely available Twitter tokenizer developed by O'Connor et al. [19].[5] We obtained NFL game statistics for the 2010–2012 seasons from NFLdata.com (`www.nfldata.com`). The data include a comprehensive set of game statistics as well as the point spread and total points line for each game obtained from bookmakers.

---

[3] If the Giants win by *exactly* 4, the result is a **push** and neither team wins WTS.

[4] Inefficiencies have been attributed to bettors overvaluing recent success and undervaluing recent failures [24], cases in which home teams are underdogs [5], large-audience games, including Super Bowls [6], and extreme gameday temperatures [3].

[5] `www.ark.cs.cmu.edu/TweetNLP`

Table 1. Hashtags used to assign tweets to New York Giants (top) and New York Jets (bottom). If a tweet contained any number of hashtags corresponding to exactly one NFL team, we assigned the tweet to that team and used it for our analysis.

```
#giants #newyorkgiants #nygiants #nyg #newyorkfootballgiants #nygmen #gmen
#gogiants #gonygiants #gogiantsgo #letsgogiants #giantsnation #giantnation
#jets #newyorkjets #nyjets #jetsjetsjets #jetlife #gojets #gojetsgo
#letsgojets #jetnation #jetsnation
```

Table 2. Yearly pregame, postgame, and weekly tweet counts.

season	pregame	postgame	weekly
2010	40,385	53,294	185,709
2011	130,977	147,834	524,453
2012	266,382	290,879	1,014,473

3.1 Finding Relevant Tweets

Our analysis relies on finding relevant tweets and assigning them to particular games during the 2010–2012 NFL seasons. We can use timestamps to assign the tweets to particular weeks of the seasons, but linking them to teams is more difficult. We chose a simple, high-precision approach based on the presence of hashtags in tweets. We manually created a list of hashtags associated with each team, based on familiarity with the NFL and validated using search queries on Twitter. There was variation across teams; two examples are shown in Table 1.[6] We discarded tweets that contained hashtags from more than one team. We did this to focus our analysis on tweets that were comments on particular games from the perspective of one of the two teams, rather than tweets that were merely commenting on particular games without associating with a team. When making predictions for a game, our features only use tweets that have been assigned to the teams in those games.

For the tasks in this paper, we created several subsets of these tweets. We labeled a tweet as a **weekly tweet** if it occurred at least 12 hours after the start of the previous game and 1 hour before the start of the upcoming game for its assigned team. **Pregame tweets** occurred between 24 hours and 1 hour before the start of the upcoming game, and **postgame tweets** occurred between 4 and 28 hours after the start of the previous game.[7] Table 3.1 shows the sizes of these sets of tweets across the three NFL seasons.

---

[6] Although our hashtag list was carefully constructed, some team names are used in many sports. After noticing that many tweets with `#giants` co-occurred with `#kyojin`, we found that we had retrieved many tweets referring to a Japanese professional baseball team also called the Giants. So we removed tweets with characters from the Katakana, Hiragana, or Han unicode character classes.

[7] Our dataset does not have game end times, though NFL games are nearly always shorter than 4 hours. Other time thresholds led to similar results in our analysis.

Table 3. Highly weighted features for postgame tweet classification. *home*/*away* indicates that the unigram is in the tweet for the home or away team, respectively.

predicting home team won			predicting away team won		
home: win	*home*: victory	*away*: loss	*away*: win	*away*: congrats	*home*: lost
home: won	*home*: WIN	*away*: lost	*away*: won	*away*: Go	*home*: loss
home: Great	*away*: lose	*away*: refs	*away*: Great	*away*: proud	*home*: bad

To encourage future work, we have released our data for academic research at `www.ark.cs.cmu.edu/football`. It includes game data for regular season games during the 2010–2012 seasons, including the point spread and total points line. We also include tweet IDs for the tweets that have been assigned to each team/game.

4 Data Analysis

Our dataset enables study of many questions involving professional sports and social media. We briefly present one study in this section: we measure our ability to classify a postgame tweet as whether it follows a win or a loss by its assigned team. By using a classifier with words as features and inspecting highly-weighted features, we can build domain-specific sentiment lexicons.

To classify postgame tweets in a particular week k in 2012, we train a logistic regression classifier on all postgame tweets starting from 2010 up to but not including week k in 2012. We use simple bag-of-words features, conjoining unigrams with an indicator representing whether the tweet is for a home or away team. In order to avoid noise from rare unigrams, we only used a unigram feature for a tweet if the unigram appeared in at least 10 tweets during the week that the tweet was written. We achieved an average accuracy of 67% over the tested weeks. Notable features that were among the top or bottom 30 weighted features are listed in Tab. 3. Most are intuitive ("win", "Great", etc.). Additionally, we surmise that fans are more likely to comment on the referees ("*away*: refs") after their team loses than after a win.

5 Forecasting

We consider the idea that fan behavior in aggregate can capture meaningful information about upcoming games, and test this claim empirically by using tweets to predict outcomes of NFL games on a weekly basis. We establish baselines using features derived from statistical game data, building upon prior work [7], and compare accuracies to those of our predictions made using Twitter data.

5.1 Modeling and Training

We use a logistic regression classifier to predict game and betting outcomes. In order to measure the performance of our feature sets, and tune hyperparameters

Table 4. List of preliminary feature sets using only game statistics, numbered for reference as F_i. *Denotes that the features appear for both the home and away teams.

point spread line (F_1)	over/under line (F_2)
avg. points beaten minus missed spread by in current season* (F_3)	avg. points beaten minus missed over/under by in current season* (F_4)
avg. points scored in current season* (F_5)	avg. points given up in current season* (F_6)
avg. total points scored in current season* (F_7)	avg. (point spread + points scored) in current season* (F_8)
home team win WTS percentage in home games in current season away team win WTS percentage in away games in current season (F_9)	avg. interceptions thrown in current season* avg. fumbles lost in current season* avg. times sacked in current season* (F_{10})

for our model as the season progresses, we use the following scheme: to make predictions for games taking place on week $k \in [4, 16]$ in 2012, we use all games from weeks $[1, 16]$ of seasons 2010 and 2011, as well as games from weeks $[1, k - 3]$ in 2012 as training data.[8] We then determine the L_1 or L_2 regularization coefficient from the set $\{0, 1, 5, 10, 25, 50, 100, 250, 500, 1000\}$ that maximizes accuracy on the development set, which consists of weeks $[k-2, k-1]$ of 2012. We follow this procedure to find the best regularization coefficients separately for each feature set and each test week k. We use the resulting values for final testing on week k. We repeat for all test weeks $k \in [4, 16]$ in 2012. To evaluate, we compute the accuracy of our predictions across all games in all test weeks. We note that these predictions occur in a strictly online setting, and do not consider any information from the future.

5.2 Features

Statistical Game Features We start with the 10 feature sets shown in Tab. 4 which only use game statistical data. We began with features from Gimpel [7] and settled upon the feature sets in the table by testing on seasons 2010–2011 using a scheme similar to the one described above. These 10 feature sets and the collection of their pairwise unions, a total of 55 feature sets, serve as a baseline to compare to our feature sets that use Twitter data.

Twitter Unigram Features When using tweets to produce feature sets, we first consider an approach similar to the one used in Sec. 4. In this case, for a given game, we assign the feature (*home/away*, unigram) the value log(1+unigram frequency over all weekly tweets assigned to the *home/away* team). As a means of

---

[8] We never test on weeks 1–3, and we do not train or test on week 17; it is difficult to predict the earliest games of the season due to lack of historical data and week 17 sees many atypical games among teams that have been confirmed or eliminated from play-off contention.

noise reduction, we only consider (*home*/*away*, unigram) pairs occurring in at least 0.1% of the weekly tweets corresponding to the given game; this can be determined before the game takes place. This forms an extremely high-dimensional feature space in contrast to the game statistics features, so we now turn to dimensionality reduction.

Dimensionality Reduction To combine the above two feature sets, we use **canonical correlation analysis** (CCA) [13]. We use CCA to simultaneously perform dimensionality reduction on the unigram features and the game statistical features to yield a low-dimensional representation of the total feature space.

For a paired sample of vectors $\mathbf{x}_1^i \in \mathbb{R}^{m_1}$ and $\mathbf{x}_2^i \in \mathbb{R}^{m_2}$, CCA finds a pair of linear transformations of the vectors onto $\mathbb{R}^k$ so as to maximize the correlation of the projected components and so that the correlation matrix between the variables in the canonical projection space is diagonal. While developed to compute the degree of correlation between two sets of variables, it is a good fit for **multi-view learning problems** in which the predictors can be partitioned into disjoint sets ('views') and each is assumed sufficient for making predictions. Previous work has focused on the semi-supervised setting in which linear transformations are learned from collections of predictors and then regression is carried out on the low dimensional projection, leading to lower sample complexity [14]. Here, we retain the fully supervised setting, but use CCA for dimensionality reduction of our extremely high-dimensional Twitter features. We experiment with several values for the number of components of the reduced matrices resulting from CCA.

Twitter Rate Features As another way to get a lower-dimensional set of Twitter features, we consider a feature that holds a signed representation of the level of increase/decrease in a team's weekly tweet volume compared to the previous week. In computing these **rate** features, we begin by taking the difference of a team's weekly tweet volume for the week to be predicted v_{curr}, and the team's weekly tweet volume for the previous week in which they had a game v_{prev} or the team's average weekly tweet volume after its previous game v_{prevavg}. We will use v_{old} to refer to the subtracted quantity in the difference, either v_{prev} or v_{prevavg}. This difference is mapped to a categorical variable based on the value of a parameter Δ which determines how significant we consider an increase in volume from v_{old} to be. Formally, we define a function $\text{rate}_S :$ $\mathbb{Z} \times \mathbb{Z} \times \mathbb{N} \to \{-2, -1, 0, 1, 2\}$, $(v_{\text{old}}, v_{\text{curr}}, \Delta) \mapsto \text{sign}(v_{\text{curr}} - v_{\text{old}}) \lfloor \frac{|v_{\text{curr}} - v_{\text{old}}|}{\Delta} \rfloor$ that is decreasing in its first argument, increasing in its second argument, and whose absolute value is decreasing in its third argument.

This idea of measuring the rate of change in tweet volumes is further generalized by categorizing the difference in volume $(v_{\text{curr}} - v_{\text{old}})$ by computing its percentage of v_{old}, or formally as a function $\text{rate}_P : \mathbb{Z} \times \mathbb{Z} \times (0, 1] \to \{-2, -1, 0, 1, 2\}$, $(v_{\text{old}}, v_{\text{curr}}, \theta) \mapsto \text{sign}(v_{\text{curr}} - v_{\text{old}}) \lfloor \frac{|v_{\text{curr}} - v_{\text{old}}|}{\theta \cdot v_{\text{old}}} \rfloor$ which has the same functional

Table 5. Example of how the rate_S feature is defined with $\Delta = 500$ (left) and how the rate_P feature is defined with $\theta = .2$.

v_{old}	v_{curr}	$\text{rate}_S(v_{\text{old}}, v_{\text{curr}}, 500)$	v_{old}	v_{curr}	$\text{rate}_P(v_{\text{old}}, v_{\text{curr}}, .2)$
2000	$(3000, \infty)$	2	2000	$(2800, \infty)$	2
2000	$(2500, 3000]$	1	2000	$(2400, 2800]$	1
2000	$[1500, 2500]$	0	2000	$[1600, 2400]$	0
2000	$[1000, 1500)$	-1	2000	$[1200, 1600)$	-1
2000	$[0, 1000)$	-2	2000	$[0, 1200)$	-2

properties as the rate_S function. Examples of how the rate_S and rate_P functions are defined are provided in Table 5. Thus, we may take $v_{\text{old}} = v_{\text{prev}}$ or $v_{\text{old}} = v_{\text{prevavg}}$, and categorize the difference using a static constant Δ or a percentage θ of v_{old}, giving us four different versions of the rate feature.

In preliminary testing on the 2010 and 2011 seasons, we found that the rate_S feature worked best with $v_{\text{old}} = v_{\text{prev}}$ and $\Delta = 500$, so we also use these values in our primary experiments below with rate_S. For rate_P, we experiment with $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $v_{\text{old}} \in \{v_{\text{prev}}, v_{\text{prevavg}}\}$.

5.3 Experiments

We consider three prediction tasks: winner, winner WTS, and over/under. Our primary results are shown in Tab. 7. We show results for all three tasks for several individual feature sets. We also experimented with many conjunctions of feature sets; the best results for each task over all feature set conjunctions tested are shown in the final three rows of the table.

The Twitter unigram features alone do poorly on the WTS task (47.6%), but they improve to above 50% when combined with the statistical features via CCA. Surprisingly, however, the Twitter unigram features alone perform better than most other feature sets on over/under prediction, reaching 54.3%. This may be worthy of follow-up research. On winner WTS prediction, the Twitter rate_S feature (with v_{prev} and $\Delta = 500$) obtains an accuracy above 55%, which is above the accuracy needed to be profitable after factoring in the bookmaker's commission. We found these hyperparameter settings (v_{prev} and Δ) based on preliminary testing on the 2011 season, in which they consistently performed better than other values; the success translates to the 2012 season as well. Interestingly, the Twitter rate features perform better on winner WTS than on straight winner prediction, while most statistical feature sets perform better on winner prediction. We see a similar trend in Tab. 6, which shows results with Twitter rate_P features with various values for θ and v_{old}.

We observed in preliminary experiments on the 2011 season that feature sets with high predictive accuracy early on in the season will not always be effective later, necessitating the use of different feature sets throughout the season. For each week $k \in [5, 16]$, we use the training and testing scheme described in Sec. 5.1 to compute the feature set that achieved the highest accuracy on average over

the previous two weeks, starting with week 3. This method of feature selection is similar to our method of tuning regularization coefficients. Over 12 weeks and 177 games in the 2012 season, this strategy correctly predicted the winner **63.8%** of the time, the winner WTS **52.0%** of the time, and the over under **44.1%** of the time. This is a simple way of selecting features and future work might experiment with more sophisticated online feature selection techniques. We expect there to be room for improvement due to the low accuracy on the over/under task (44.1%) despite there being several feature sets with much higher accuracies, as can be seen in Tab. 7.

Another simple method of selecting a feature set for week $k \in [4, 16]$ is choosing the feature set achieving the highest accuracy on average over *all* previous weeks, starting with week 3, using the same scheme described in Sec. 5.1. This feature set can be thought of as the best feature set at the point in the season at which it is chosen. In Fig. 1 we observe that the best feature set changes very frequently, going through 8 different feature sets in a 13-week period.

Table 6. rate$_P$ winner and winner WTS accuracies for different values of θ and v_{old}.

| θ	v_{prev}		v_{prevavg}	
	winner	WTS	winner	WTS
0.1	51.0	51.4	51.0	50.0
0.2	53.8	51.0	52.4	45.7
0.3	51.4	52.4	52.4	54.3
0.4	54.8	49.5	51.4	49.5
0.5	52.9	45.2	53.4	49.5

6 Conclusion

We introduced a new dataset that includes a large volume of tweets aligned to NFL games from the 2010–2012 seasons. We explored a range of feature sets for predicting game outcomes, finding that simple feature sets that use Twitter data could match or exceed the performance of game statistics features. Our dataset is made available for academic research at `www.ark.cs.cmu.edu/football`.

Table 7. Accuracies across prediction tasks and feature sets. Lower pane shows oracle feature sets for each task, with the highest accuracies starred.

	prediction tasks		
features	winner	WTS	over/under
point spread line (F_1)	60.6	47.6	48.6
over/under line (F_2)	52.3	49.0	48.6
F_3	56.3	50.0	50.0
F_4	52.3	54.8	50.5
F_5	65.9	51.0	44.7
F_{10}	56.7	51.4	46.6
$\bigcup_i F_i$	63.0	47.6	51.0
Twitter unigrams	52.3	47.6	54.3
CCA: $\bigcup_i F_i$ and Twitter unigrams, 1 component	47.6	50.4	43.8
2 components	47.6	51.0	43.8
4 components	50.5	51.9	44.2
8 components	47.6	48.1	42.3
Twitter rate$_S(v_{\mathrm{prev}}, \Delta = 500)$	51.0	55.3	52.4
$F_5 \cup F_9 \cup$ Twitter rate$_P(v_{\mathrm{prev}}, \theta = .2)$	65.9*	51.4	48.1
$F_3 \cup F_{10} \cup$ Twitter rate$_P(v_{\mathrm{prev}}, \theta = .1)$	56.3	57.2*	48.1
$F_3 \cup F_4 \cup$ Twitter rate$_S(v_{\mathrm{prev}}, \Delta = 200)$	54.8	49.0	58.2*
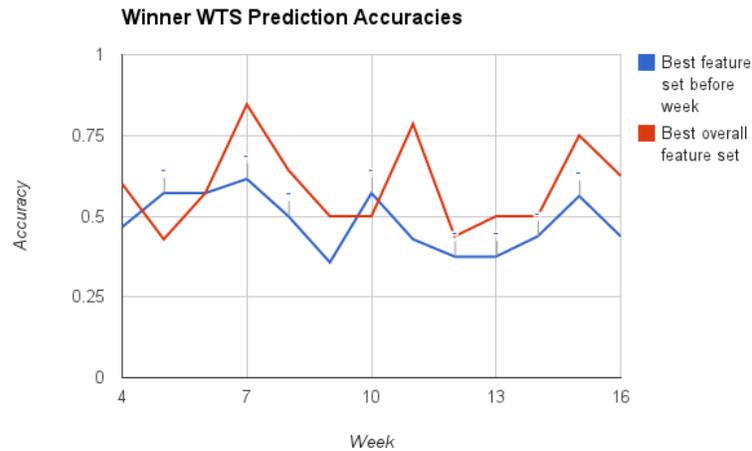
Fig. 1. Weekly accuracies for the best overall feature set in hindsight, and the best feature set leading up to the given week for winner WTS prediction. Marks above the 'Best feature set before week' line indicate weeks where the best feature set changed.

References

1. Baker, R.D., McHale, I.G.: Forecasting exact scores in national football league games. International Journal of Forecasting 29(1), 122–130 (2013)
2. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. Journal of Computational Science 2(1) (2011)
3. Borghesi, R.: The home team weather advantage and biases in the nfl betting market. Journal of Economics and Business 59(4), 340–354 (2007)
4. Boulier, B.L., Stekler, H.O., Amundson, S.: Testing the efficiency of the National Football League betting market. Applied Economics 38(3), 279–284 (February 2006)
5. Dare, W.H., Holland, A.S.: Efficiency in the NFL betting market: modifying and consolidating research methods. Applied Economics 36(1), 9–15 (2004)
6. Dare, W.H., MacDonald, S.S.: A generalized model for testing the home and favorite team advantage in point spread markets. Journal of Financial Economics 40(2), 295–318 (1996)
7. Gimpel, K.: Beating the NFL football point spread (2006), unpublished manuscript
8. Glickman, M.E., Stern, H.S.: A state-space model for National Football League scores. JASA 93(441), 25–35 (1998)
9. Golec, J., Tamarkin, M.: The degree of inefficiency in the football betting market : Statistical tests. Journal of Financial Economics 30(2), 311–323 (December 1991)
10. Gray, P.K., Gray, S.F.: Testing market efficiency: Evidence from the NFL sports betting market. The Journal of Finance 52(4), 1725–1737 (1997)
11. Harville, D.A.: Predictions for National Football League games via linear-model methodology. JASA 75(371), 516–524 (1980)
12. Hong, Y., Skiena, S.: The wisdom of bookies? sentiment analysis versus the NFL point spread. In: Proc. of ICWSM (2010)
13. Hotelling, H.: Relations between two sets of variates. Biometrika 28(3–4), 321–377 (1936)
14. Kakade, S.M., Foster, D.P.: Multi-view regression via canonical correlation analysis. In: Proc. of COLT (2007)
15. Knorr-Held, L.: Dynamic rating of sports teams. The Statistician 49(2), 261–276 (2000)
16. Lacey, N.J.: An estimation of market efficiency in the nfl point spread betting market. Applied Economics 22(1), 117–129 (1990)
17. Levitt, S.D.: How do markets function? an empirical analysis of gambling on the National Football League. Economic Journal 114(495), 2043–2066 (2004)
18. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From Tweets to polls: Linking text sentiment to public opinion time series. In: Proc. ICWSM (2010)
19. O'Connor, B., Krieger, M., Ahn, D.: Tweetmotif: Exploratory search and topic summarization for twitter. Proc. of ICWSM pp. 2–3 (2010)
20. Paul, M.J., Dredze, M.: You are what you Tweet: Analyzing Twitter for public health. In: Proc. of ICWSM (2011)
21. Stern, H.: On the probability of winning a football game. The American Statistician 45(3), 179–183 (1991)
22. Surowiecki, J.: The Wisdom of Crowds. Anchor (2005)
23. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: Proc. of ICWSM (2010)

24. Vergin, R.C.: Overreaction in the NFL point spread market. Applied Financial Economics 11(5), 497–509 (2001)
25. Zuber, R.A., Gandar, J.M., Bowers, B.D.: Beating the spread: Testing the efficiency of the gambling market for National Football League games. Journal of Political Economy 93(4), 800–806 (1985)