Draft Literature Review

David Freed and Samuel Green

May 1, 2016

1 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 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.[1]

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

[1]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.

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 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.