# Draft Literature Review

David Freed and Samuel Green

May 1, 2016

## 1 Overview of Related Literature

Previous literature has established that useful modeling information can be derived from Twitter data. We begin by reviewing some literature on sentiment classification for Twitter and then reference previous work on using Twitter data for sports analysis.

Previous work has established that accurate classification is possible for Twitter data. Multiple different methods have been shown to be reliable in practice, with multiple different approaches used to train models. Go et al. provide a useful overview and empirical work on classification of Twitter data. Their work demonstrated that reasonably high classification rates can be achieved using Naive Bayes, Maximum Entropy, and Support Vector Machin approaches. Their work also showed that a reasonablely accurate training set could be derived without hand-labeling using information embedded in Tweets, though we do not reapply that result here (Go et al., 2009)[1]. Ibrahim and Yusoff also provided evidence recently that reasonably accurate Naive Bayes classifiers could be trained using unexpectedly small datasets of labeled tweets, lending weight to the hypothesis that a large training set should provide even more useful information, as in the setting of this project (Ibrahim and Yusoff, 2015).[2].

[1] https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[2] http://ieeexplore.ieee.org.ezp-prod1.hul.harvard.edu/stamp/stamp.jsp?tp=&arnumber=7403510

Twitter sentiment has also been applied for sports analytics, though as far as we can tell, not in a real-time predictive model. Sinha et. al. assembled a large dataset of tweets related to NFL games in advance of these games taking place. They used the datasets to inform ex ante predictions about the outcomes of the games, after diong some classification work. They found that Twitter could be used effectively to build a potentially profitable winner with the spread prediction model. Their model produced a success rate of higher than 55%, the benchmark needed to turn a profit after factoring in bookmaker commissions. Sinha et. al. also built their dataset using targeted collections of hashtags, an approach that we re-apply to build our dataset (Sinha et al, 2013)[3].

Finally, the usefulness of Twitter in real-time identification of events during sports games has been established by previous work A real-time system was built in 2012 by Zhao et. al., which was able to use Twitter to identify significant plays in real time (on the order of 1-2 minutes), an 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 (Zhao et al., 2012).[4]. This piece of previous work provides evidence that tweets respond in real time to progress in sporting events. We base our hypothesis that tweets can be predictive and responsive in the NCAA in real time using this previous result.

[3] http://arxiv.org/abs/1310.6998

[4] http://arxiv.org/pdf/1205.3212v1.pdf