Draft Conclusion

David Freed and Samuel Green

May 1, 2016

Expanding on previous work applying Twitter for predictive sports analysis, this study demonstrates the beneficial link between relevant Twitter data and both short-term and long-term outcomes in NCAA March Madness games. We provided evidence first that Twitter is responsive in a short timeperiod to events in games, both in the volume of tweets sent and in the sentiment contained in those tweets. We then showed that Twitter data can be used to make statistically significant predictions about future game events and showed that, particularly as games progress, Twitter data can be used to improve standard models typically used to predict game winners. 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.

Given the significance of our results, we discuss some potential flaws in the analysis and methodology and avenues for future work.

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. For example, many words that are conventionally negative, like "dirty" or "filthy," are often positive in the

context of basketball. As an additional example, our classifier initally called the tweet `Hell yes, Syracuse is going to the Final Four` negative, though it is clearly positive in the context of the NCAA tournament. We hypothesis that retraining the model with a more appropriate training set would improve model performance and highlight this as an obvious opportunity for future work.

Relevance classification, namely identifying noisey tweets from tweets that contained information relevant to the game, remains a challenge. Though we collected tweets only with specific hashtags, tweets that weren't related to the NCAA tournament were still collected (which we observed by inspection of the dataset). We did our best to eliminate these tweets, as previously discussed, but finding a mechanism to collect a more targetted set for classification would likely also contribute to improvement.

A further training concern is that our models rely on a noisey projection of game events onto real time, since we did not discover a source of game data that included real timestamps. Our quasi-uniform projection of game time onto real time results in a reasonable approximation, but he models should be reconstructed using more reliable game data if this becomes possible in the future.