Draft Conclusion

David Freed and Samuel Green

May 1, 2016

Our work expands on previous work applying Twitter for predictive sports analysis to demonstrates the predictive power of relevant Twitter data in forecasting both short-term and long-term outcomes in NCAA March Madness games. We begin by providing evidence that, over short time periods, Twitter is responsive to game events; we demonstrate in a prior section that the volume of Tweets (and the sentiment of those Tweets) about a particular team will increase in proportion to the squad's quality of play in prior periods.

Next, we show that Twitter is not only a reactive mechanism, but a useful forecasting tool. In Section 4.2, we show that the volume and sentiment of Tweets in a given period is a statistically significant predictor of game events in the next period. We moved beyond this to demonstrate that Twitter data has significant predictive power in forecasting the eventual winner of the game. Our model that incorporated Twitter sentiment outperformed the standard FiveThirtyEight model at every point in the game. 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.

Our results are significant, but carry some caveats and areas of potential development. 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. As mentioned earlier, many words that are conventionally negative, like “dirty” or “filthy,” are often positive in the context of basketball. Our classifier initially classified the Tweet “`Hell yes, Syracuse is going to the Final Four`” as a ‘negative’ Tweet, but it is evident by inspection that the author is elated at the prospect of Syracuse advanced to the semifinals of the tournament. We hypothesize that were we to re-train the model with a better training set, our model would do a better job of extracting sentiments from Tweets. This would give a better idea of the current opinion of the crowd and provide a good check on our current analysis. We consider this to be a promising avenue for future work.

The second primary caveat to our analysis is the effectiveness of our relevance classifier. The process of separating noisy and relevant Tweets remains a challenge; our initial pull from Twitter contained 27,000 Tweets about GOP presidential candidate and international celebrity Donald Trump. Many of the hashtags that are commonly associated with teams (e.g. `#Maryland` or `#Wisconsin`) have many other uses. We did our best to eliminate these tweets, as previously discussed, but finding a mechanism to collect a more targeted set for classification would likely also contribute to improvement.[1]

The third caveat to the analysis is the way in which we projected game events onto real time. There was no available source of data which provided a reliable timestamp for each event, and our modified uniform approximation of real time is not a perfect mapping. A possible extension or refinement of this process would be to institute many more check points in the data (by tapping Tweets that contained the time left in the game and using those as time signposts) to get a better approximation. Through spot-checking, our approximation appeared to work fairly well, but it remains an important caveat when evaluating our

[1] One related issue that we had is that many Tweets would show up multiple times in our dataset because they had been ‘reTweeted’, a Twitter functionality that allows users to send another account’s Tweet from their own account. We opined that any individual who chose to ‘reTweet’ a message shared the sentiment of the original author of the Tweet.

analysis.

There are two immediate extensions that fall out of the work that we have already done. It would be interesting to consider a more granular separation of the "crowd" into trusted and untrusted "sub-crowds." That is, could predictive improvements be made by considering crowd-sourced opinions from known experts, like commentators or individuals prominent in the network, separate from or more heavily weighted than standard or low influence individuals? Presumably, experts are ex ante more qualified to predict outcomes based on game progress but it is not immediately apparent that has to be the case.

The second obvious extension would have been to conduct a systematic study of Twitter's ability to incorporate specific types of unobservable information. Earlier in the paper, we reference the cases of a star player getting injured mid-game or coming out of the game with a fourth foul as situations where the margin may not accurately reflect a swing in the game. If we had an expanded dataset, we could take advantage of Twitter's ability to detect momentum to conduct a more thorough event study on this basis.

The final trivial extension would be to apply this to other leagues and sports. The National Basketball Association (henceforth, "NBA") would be a logical extension. One other interesting extension would be to gauge the predictive power for Twitter in smaller markets—how many people need to be following the game closely for their aggregate opinion to become predictive? Given the vast difference in attendance between games in the Big 12 and the Ivy League athletic conferences, this poses an interesting question for practitioners, who can often win substantial sums in less liquid markets (e.g. Ivy League athletics gambling).

In sum, the results presented above are a novel extension of previous work. We distinguish ourselves from the previous literature on this subject by measuring the sentiment of Twitter, not only the volume, at any point in time. Our results about the predictive power of models that incorporate Twitter data provide the basis for improved in-game win probability models and provoke interesting questions about how to measure unobservable information.

To our knowledge, the results presented here are novel. We provide the first models using Twitter data as input to a real-time model. We are also the first to show that a model of this variety has power when predicting future events and full game outcomes in the NCAA. Above, we note a series of logical extensions to the work, which can extend the analysis above and continue digging into the interesting empirical questions raised by the paper.