# Tweets and NCAA March Madness

David Freed and Samuel Green

{davidfreed,samuelgreen}@college.harvard.edu

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

## Real-Time Win Probability Modeling with Twitter

This project investigates the predictive power of Twitter data in modeling the outcomes and events in NCAA March Madness basketball games. As a platform designed for instant reactions, rather than publishing long-term opinions, one would expect useful information about the qualitative progress of a sporting event to be derivable from a dataset of tweets. We extend previous work that used only data collected in advance of games to make predictions about games, instead collecting live-action time series of tweets in-game. In preliminary investigation, we find correlations between volumes of tweets and number of events recorded in a game (rebounds, points scored, out of bounds) as well as between volumes of tweets and changes in game margin (one team moves farther ahead).

## Data Collection

Our dataset compiles approximately 1 million tweets from 42 games during the NCAA March Madness. For each game, we compiled hashtags relevant to the game, which we then used to collect tweets via the Streaming API. An set of example hashtags is:

#MarylandvsHawaii #WeWill #Maryland
#Terrapins #RainbowWarriors #Hawaii
#HawaiiMBB #RoadWarriors #GoBows

Summary statistics for the entire dataset are:

| N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|
| 42 | 20,754.810 | 27,273.890 | 1,930 | 171,600 |

## Relevance Analysis

Sentiment in this setting is much less meaningful without relating the sentiment contained in a tweet to one of the two teams. We therefore perform some heuristic classification of the relevance of a tweet to each team.

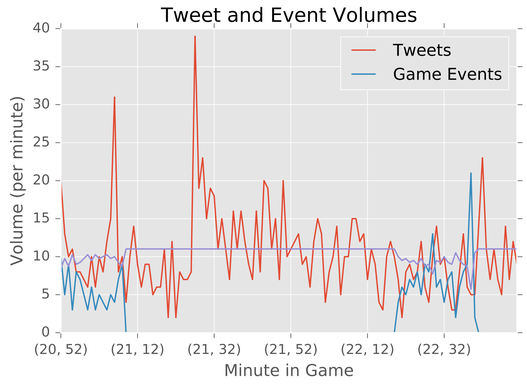
We base classification on the hashtags that were originally used to collect the tweets and off of rosters. To treat rosters uniformly, we included that listed last names for each of the 7 top players by playing time for each team, as well as the last name of the head coach, with all data from ESPN.

Tweets that were classified for both teams required tie-breaking, which we implement by choosing the first term to appear.

## References

[1] S. Sinha, C. Dyer, K. Gimpel, N. Smith. Predicting the NFL Using Twitter, Presented at ECML/PKDD 2013.

## Volumes Track Game Events



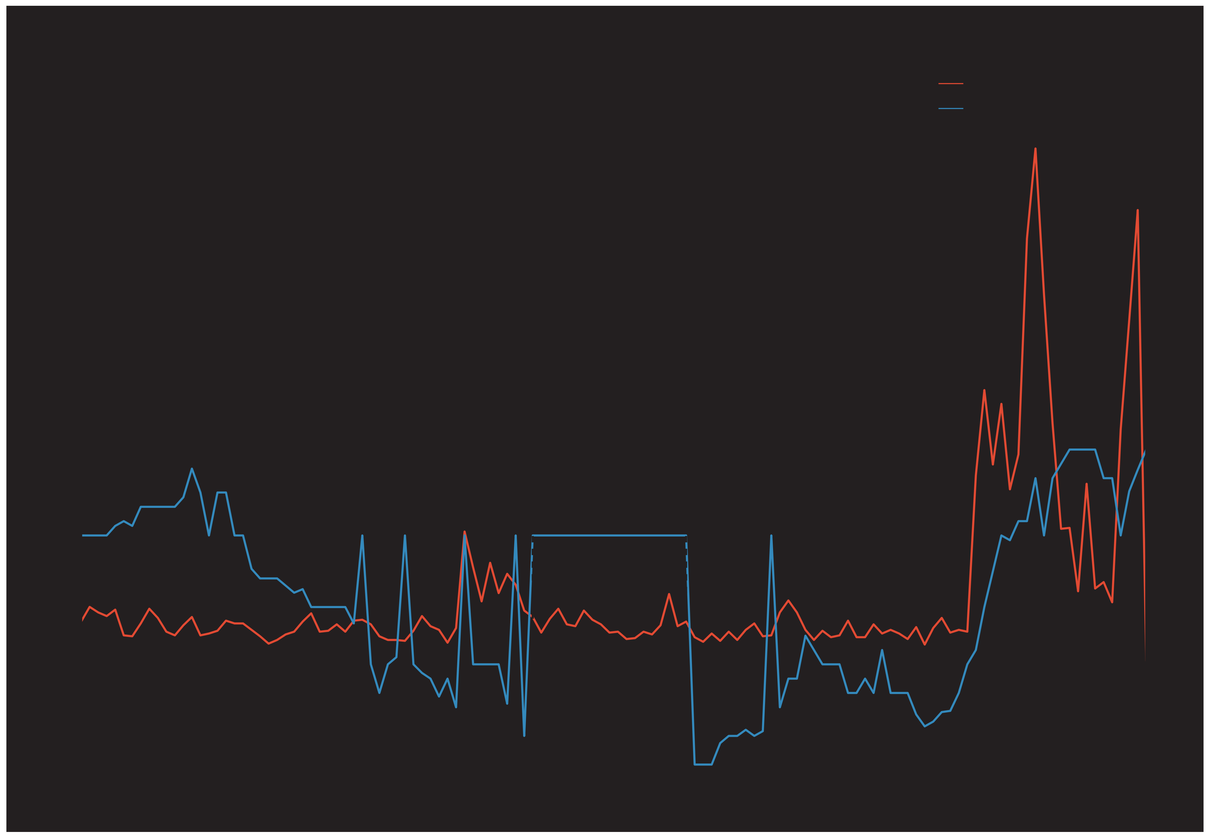

Tweets per Minute and Game Events per Minute, Virginia vs Syracuse

**Event volumes track tweet volumes.** A preliminary linear regression demonstrated a statistically significant relationship between volumes of tweets in a given minute and the number of in-game events that took place during that minute in the game. This result provides heuristic confirmation that tweets should could reflect qualitative changes in, for example, momentum.

## Sentiment Analysis

Sentiment classification for tweets forms a central pivot point for this project. To perform classification, we train a bag-of-words mdoel with a labeled training corpus of 4000 tweets and a linear-kernel Support Vector Machine. The classifier successfully identifies conventionally positive tweets but does not always capture tweets that are idiomatically positive towards sports teams.

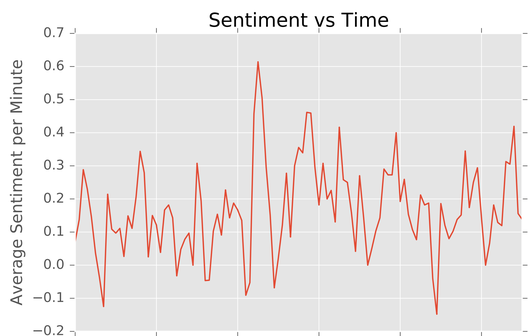
## Game Scores and Tweet Volumes



**Predicting Game Events with Tweet Volumes.** The chart above plots time series of z-scored tweet volumes and z-scored changes in game margins to explore whether we can use tweets, even naively, to predict minutes in the game that will bring large changes in the score. The plot above provides evidence that, while the fit is clearly imperfect, we can derive information about game proceedings using no prior information except the volumes of tweets recorded in a game.

This initial finding leaves room for significant improvement after incorporating sentiment scores and a more intelligent predictive model.

## Sentiment Game Plot




## Further Work

The major piece of remaining work for the scope of this project is to investigate the relationship between sentiment classifications and changes in-game. We plan also to benchmark the sentiment probability model against the industry-standard real-time probability model, which is based on historical data. We hope to apply a Hidden Markov Model to model the sequence of lead changes that take place during a game, as well as to model the likeliest overall winner at each time period during the game.