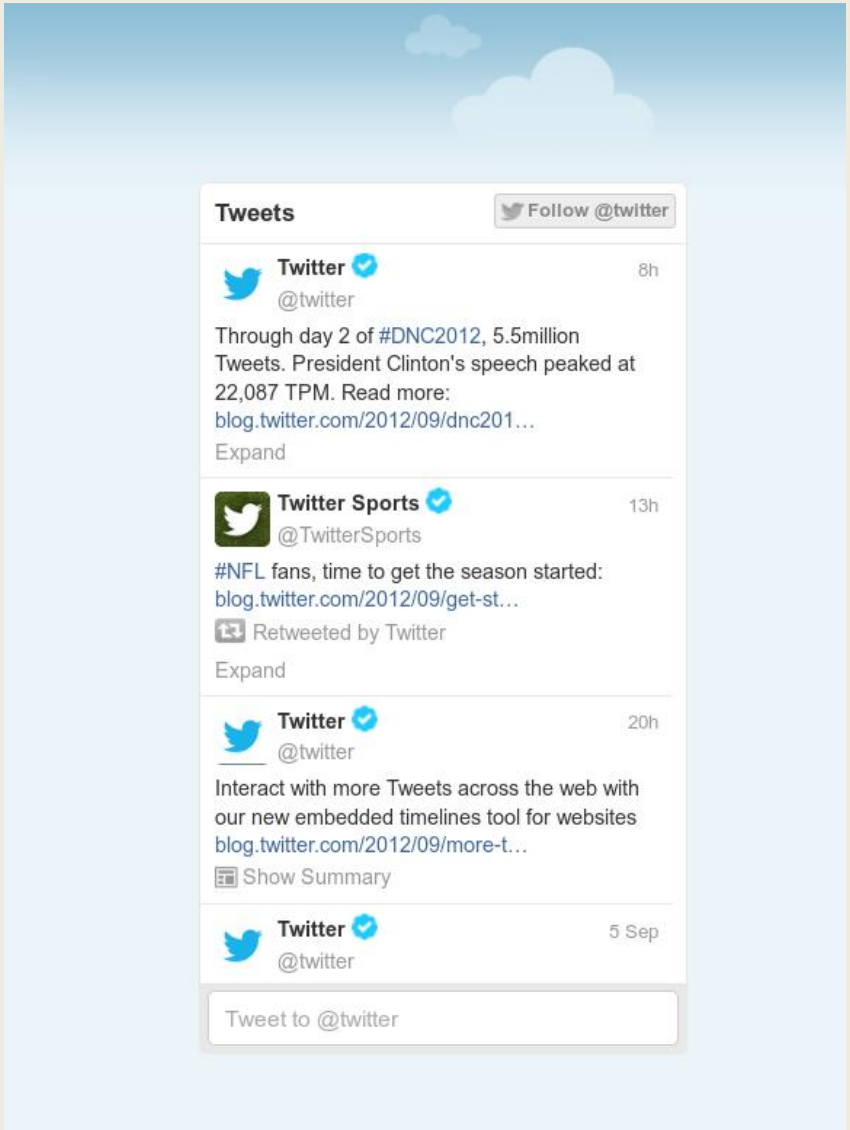
A Model for Determining Tweet Popularity via Prediction Methods

Presentation by Farhan Hormasji and Bonnie Reiff

The Question

Can we identify which features contribute most to a Tweet's popularity?

How well can we use these features to predict a Tweet's popularity?



Data Collection and Preprocessing

- Creation of 2 PHP Programs for interaction with the Public Stream and Search APIs
- Streaming program:
  - Outputs a CSV file with the chosen attributes for each Tweet
  - Preprocessing:
    - Creation of an array of the most popular hashtags for additional (manual) feature analysis
    - Generation of Binary or Integer Count attributes from String features
    - Data formatting
- Searches at 10 minutes, 30 minutes, 1 hour, and 20 hours after every data collection to collect Retweet and Favorite count information

Implementation

- Classification of Tweet as popular based on retweet/favorite count after 20 hours
- Features/Predictors:
  - User information (#followers, #statuses, etc.)
  - Top 10 popular hashtags
  - Retweet/favorite count *n* minutes after posting
- SVM binary classifier
- Sequential feature selection
- K-fold cross validation
  - Total number of Tweets used in largest dataset: 24980

Results

Raw dataset:

CDR = 95%

1258	87
961	20762

Filtered dataset:

CDR = 90%

920	191
154	2379

Conclusions

- Adding feature “retweet/favorite count after 10 minutes” increases correct detection rate from 49% to 86%
- Using only binary features resulted in ~70% correct detection rate
- Including top 10 hashtags didn’t improve error rate
- Future work: investigate popularity given a certain hashtag