Text-based Forecasting

Jaime Arguello
INLS 613: Text Data Mining
jarguell@email.unc.edu

November 12, 2014

Text-based Forecasting

- So far, we've used text analysis to predict:
  - properties of the text (e.g., science- vs. sports-related)
  - the author's opinion (e.g., positive vs. negative)
  - the author's emotional state (e.g., happy vs. sad)
  - the author's stance (e.g., pro-life vs. pro-choice?)
- Text analysis can also be used to detect on-going "real-world" events or to predict future events

Detecting on-going Events

- Detecting on-going "real-world" events
  - consumer confidence
  - candidate approval ratings
  - newsworthy events (e.g., natural disasters)
  - drug side-effects
  - demographic information
  - people's habits and moods
  - consumer engagement with a product (viewers)
  - identifying influential "players"
  - traffic
  - ....

Detecting on-going Events

- There exist alternative methods for detecting on-going events (e.g., polls, surveys, eye-witness reports, hospital records, financial reports, ...)
- However, they have limitations
  - expensive
  - delayed response
  - localized
  - intrusive/desruptive
  - ....

Predicting Future Events

- Predicting future events
  - stock price movements
  - election results
  - voter turnout
  - product sales or, more generally, product demand
  - consumer spending
  - socio-political unrest
  - ....

Sources of (Textual) Evidence

- Webpages
- News articles
- Blogs
- Tweets
- Search engine queries
- Facebook posts, comments, likes, connections, etc.
- Linked-in actions (e.g., cross-company connections)
- Event transcriptions (e.g., http://www.fednews.com/)
- ....
- Discussion: how are these different and what are they good for?

Examples

Researchers Use Twitter To Predict When New Yorkers Will Catch The Flu With 90% Accuracy

Alyson Shontell | Aug. 1, 2012, 2:34 PM | 1,474 |

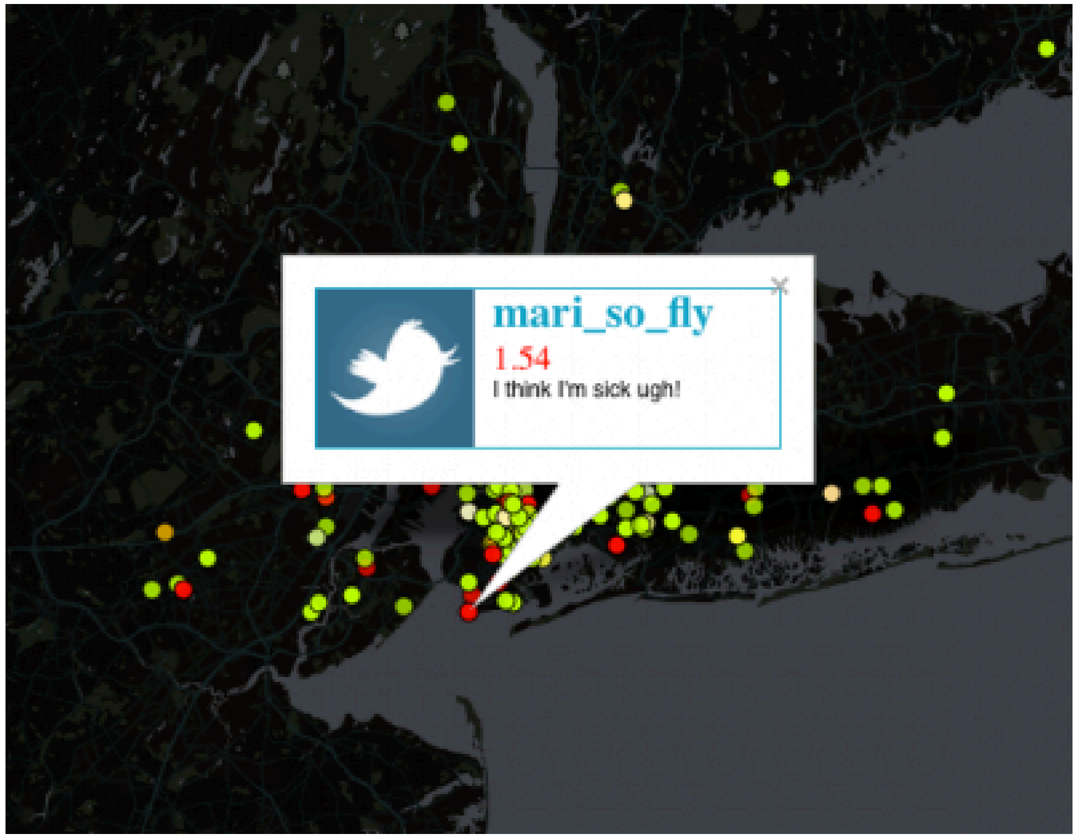
Recommend 34 | Share 11 | Tweet 139 | +1 2 | Email | More

The University of Rochester's Adam Sadilek and his colleagues conducted a Twitter experiment.

Like Google Flu, they used Twitter data to try and predict when New Yorkers would fall ill.

They were successful.

After examining 4.4 million tweets from more than 630,000 New York Twitter users in 2010, they could predict when someone would get sick up to eight days prior with 90% accuracy.



Twitter Health

If you're near Twitter user @mari_so_fly right now, you may fall sick very soon.

Examples

Twitter mood maps reveal emotional states of America

- 12:14 21 July 2010 by Celeste Biever
- For similar stories, visit the US national issues and The Human Brain Topic Guides

Video: Twitter mood map

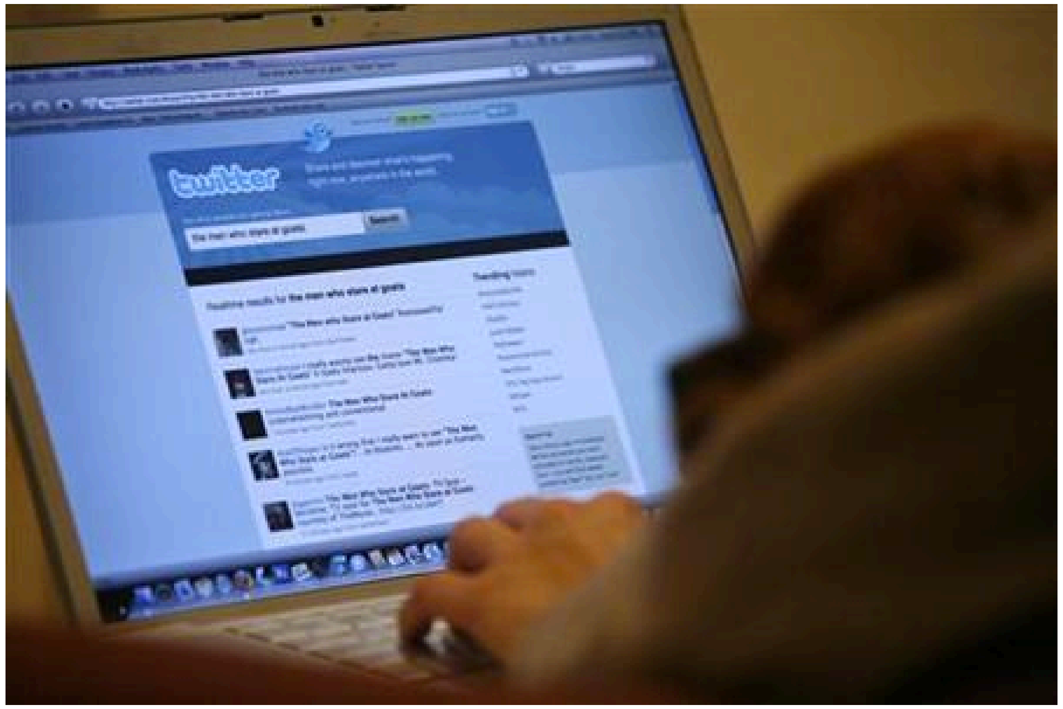
America, are you happy? The emotional words contained in hundreds of millions of messages posted to the Twitter website may hold the answer.

Computer scientist Alan Mislove at Northeastern University in Boston and colleagues have found that these "tweets" suggest that the west coast is happier than the east coast, and across the country happiness peaks each Sunday morning, with a trough on Thursday evenings. The team calls their work the "pulse of the nation".

Examples

Could Twitter predict the stock market?

Recommend 49 people recommend this. Be the first of your friends.



Tweet 160

Share

Share this

+1 5

Email

Print

Analysis & Opinion

Essential tax and accounting reading: Santorum's tax returns, progress on payroll taxes, Wegelin, and more

U.S. Catholic bishops plan aggressive expansion of birth-control battle

Related Topics

Money »

Investing Simplified »

By Chris Taylor

NEW YORK | Thu Feb 16, 2012 4:43pm EST

(Reuters) - When Richard Peterson first started meeting with hedge funds about eight years ago to pitch using social media to predict market movement, investment managers looked at him as if he had just arrived from outer space.

Back then, what he was pitching them seemed pretty insane. Peterson, managing director of Santa Monica-based MarketPsych, said that social media can be mined for data about what people are thinking and feeling. And that, in turn, could translate into powerful investment ideas.

Basic Ingredients

- Stream of textual data + target signal
- Temporal window (depends on the task, on-going or future outcome)
- Method for identifying the ‘relevant’ elements
  - Can be tricky (e.g., predicting Facebook stock price using tweets)
- Sentiment or topic analysis of individual datapoints
- Data point aggregation
- Classification or regression algorithm

General Assumptions

- The text contains enough signal to predict the outcome
- Correlation, not causation
- Errors at the micro-level do necessarily translate to errors at the macro-level as long as the errors are independent given the target outcome value
  - example: mood prediction

Reading the Markets

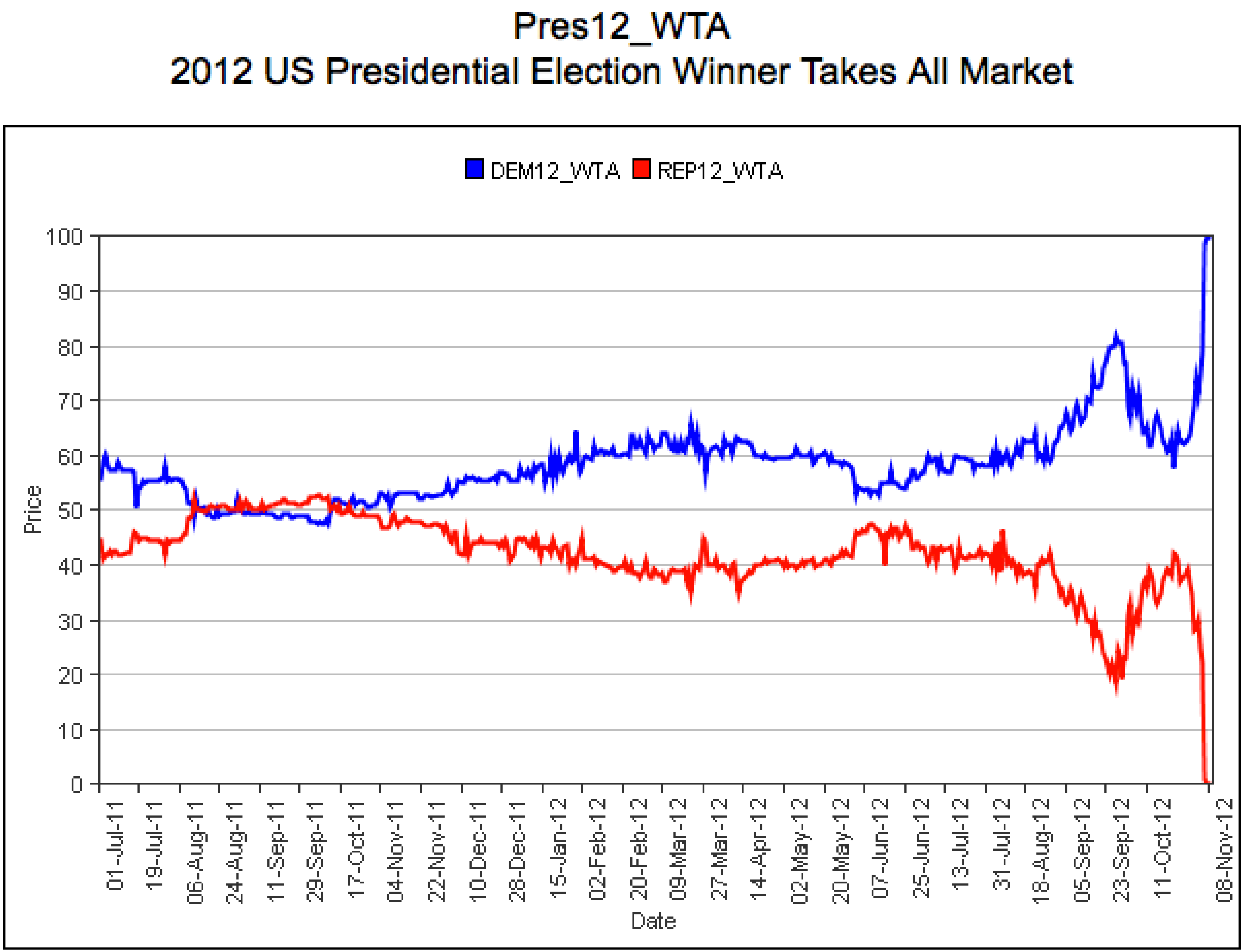
- K. Lerman, A. Gilder, Mark Dredze, and F. Pereira. Reading the Markets: Forecasting Public Opinion of Political Candidates by News Analysis. In *Coling '08*.

Reading the Markets

- Input: news articles
- Outcomes:
  - public opinion about presidential candidates in the 2004 election (e.g., Kerry, Bush)
  - public opinion surrogate: on-going “stock” price for a candidate ($1 awarded for every winning stock) in a prediction market
- Motivation: public opinion can be predicted based on the topics covered in the news (not just sentiment)

Prediction Markets

http://tippie.uiowa.edu/iem/markets/data_pres12.html



Reading the Markets

- Task: predict whether the average daily price of a candidate's stock will go up/down from today to tomorrow.
- Outcomes: news articles and market data up to today

Reading the Markets

(1) unigram features

- Motivation: public opinion may depend on the topics covered in the media
  - e.g., mentions of “iraq” are bad for Bush
- Method: term counts generated from all of the day’s news articles (big document)

Reading the Markets

(2) news focus features

- Motivation: while the news may cover an event for several days, public opinion may not shift. Thus, it seems important to model shifts in news focus (term frequencies)
- Method: compare each term's frequency today with the average frequency in the past three days
- Values > 0 indicate increase in focus; values < 0 indicate decrease in focus

$$\Delta f_i^t = \log\left(\frac{f_i^t}{\frac{1}{3}(f_i^{t-1} + f_i^{t-2} + f_i^{t-3})}\right)$$

Reading the Markets

(3) entity features

- Motivation: public opinion may depend on the topics associated with a particular candidate
  - e.g., the term "scandal" may be bad for Bush, but only if it is associated with Bush (and not Kerry)
- Method: identify sentences that mention the candidate (e.g., Bush) and construct features by combining the candidate with all content words in the sentence
- Example: "Bush is facing another scandal" would be associated with features **bush_facing** and **bush_scandal**

Reading the Markets
(4) dependency features

- Motivation: the previous feature representation cannot handle sentences that mention more than one entity
  - e.g., "Bush defeated Kerry in the debate"
- Method: generate features from a *dependency parse* of the sentence

Typed dependencies

```
nsubj(defeated-2, Bush-1)
root(ROOT-0, defeated-2)
dobj(defeated-2, Kerry-3)
prep(Kerry-3, in-4)
det(debate-6, the-5)
pobj(in-4, debate-6)
```

(output from stanford parser: http://nlp.stanford.edu:8080/parser/)

Reading the Markets

(5) market history feature

- Motivation: the market has a “natural” flow (independent of news).
  - e.g., a candidate who is doing well will continue doing well.
- Method: train a classifier to predict the current day’s market price based on the market price in the past few days and use this classifier’s prediction as a feature

Evaluation Methodology

- On-line Evaluation: Given data for day t, make a prediction for day $t + 1$. Move to $t + 1$ and increase training set.
- Metric: percentage of best possible profit. Takes into account direction and magnitude. In the range [0,1]

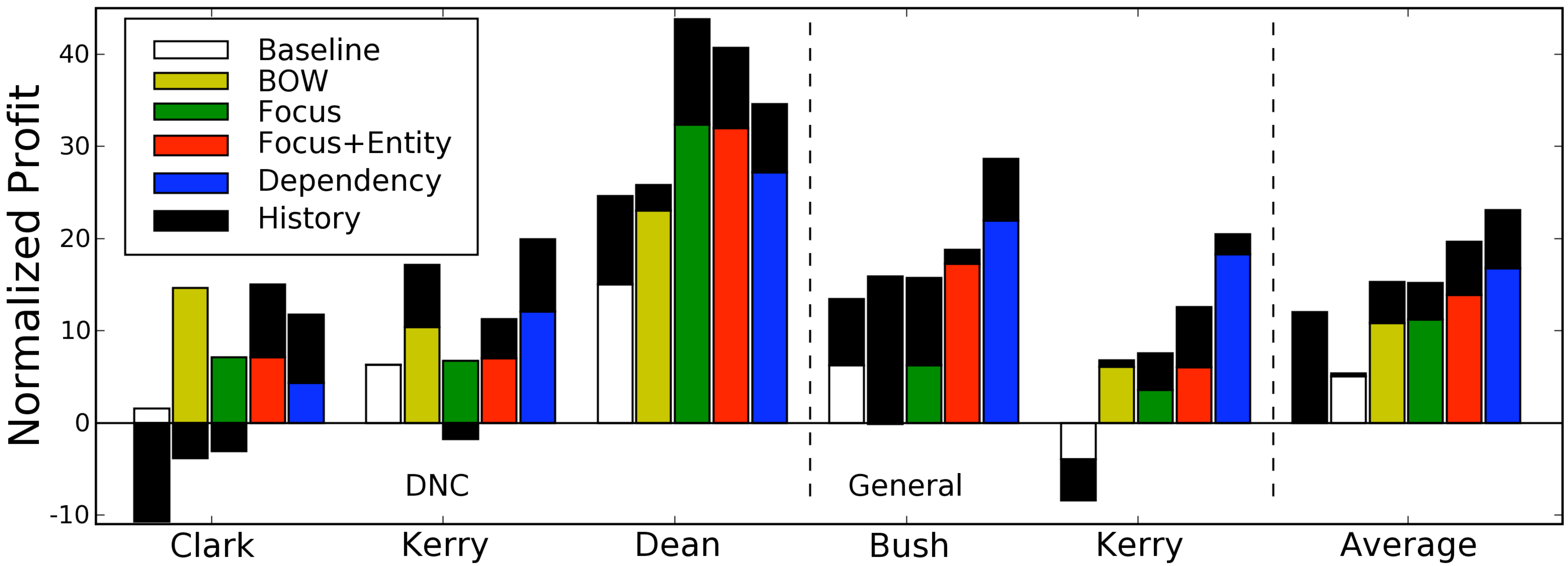
Reading the Markets
results

- History: predict that the market will do what it did today
- Baseline: # of mentions of each entity as features

Market		*History*	*Baseline*
DNC	Clark	20	13
	Clinton	38	-8
	Dean	23	24
	Gephardt	8	1
	Kerry	-6	6
	Lieberman	3	2
General	Kerry	2	15
	Bush	21	20
Average (% omniscience)		13.6	9.1

Reading the Markets

results



From Tweets to Polls

- B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *ICWSM '10*. AAAI.

From Tweets to Polls

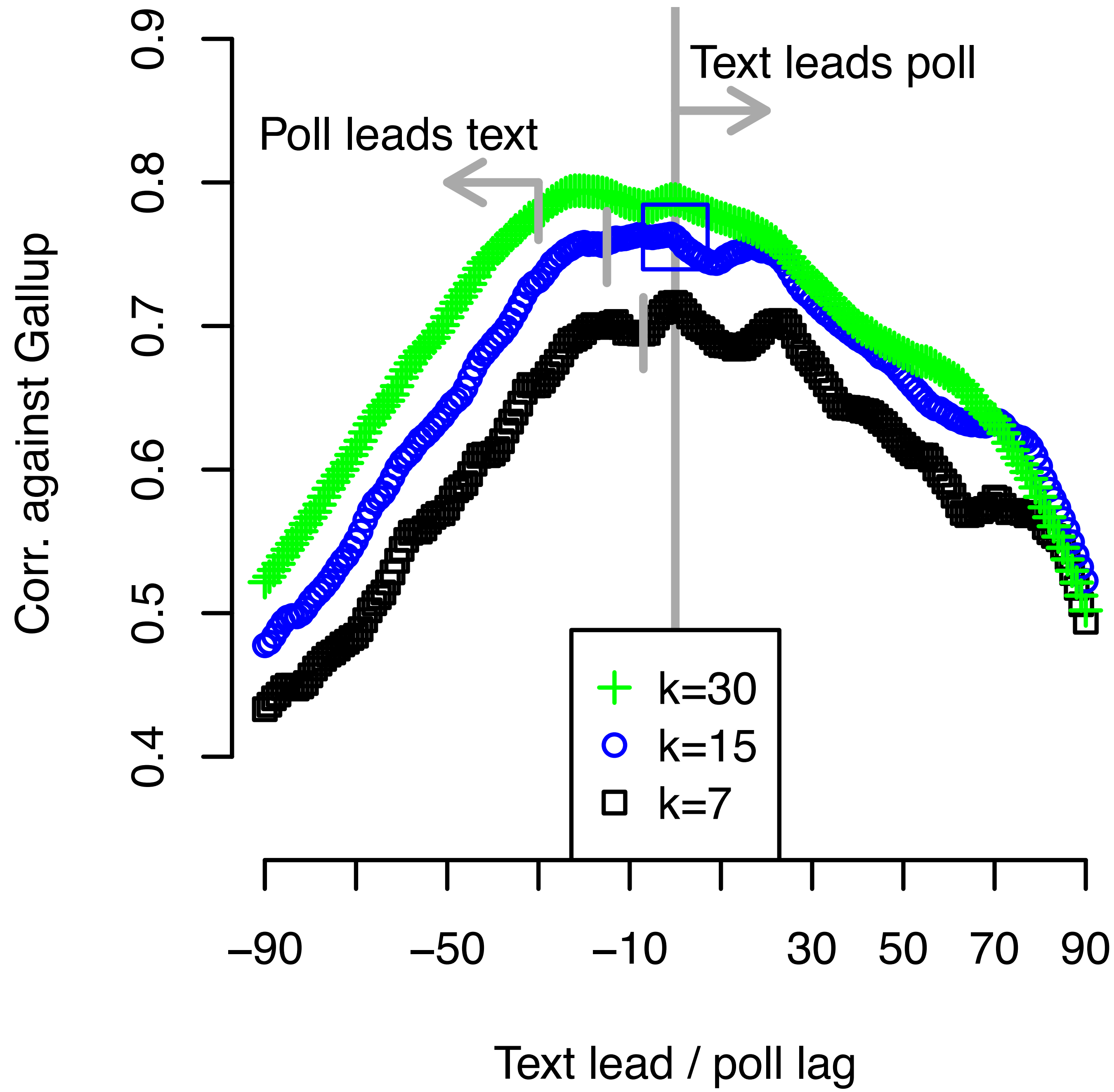
- Input: random sample of tweets from 2008 and 2009
- Outcomes:
  - index of consumer sentiment
  - approval ratings for candidate obama and mccain
  - approval ratings for president obama

From Tweets to Polls
method

- Relevant tweets identified using key words:
    - index of consumer sentiment: *economy, job(s)*
    - candidate obama and mccain: *obama, mccain*
    - president obama: *obama*
- Tweet sentiment predicted using a sentiment lexicon
    - A tweet is positive(negative) if it contains at least one positive(negative) sentiment term
- Daily positive-sentiment score predicted using the ratio of positive to negative tweets
- Daily sentiment smoothed by averaging the daily sentiment of the previous *k* days

From Tweets to Polls

results



$$y_{t+L} = b + a \sum_{j=0}^{k-1} x_{t-j} + \epsilon_t$$