Political language

Chris Potts

Linguist 287 / CS 424P: Extracting Social Meaning and Sentiment, Fall 2010

Nov 16

Overview

Plan

- Feature selection
- On bias and bias features
- Classification (with social, relational structure)
- Twitter prognostication

Overview

Plan

- Feature selection
- On bias and bias features
- Classification (with social, relational structure)
- Twitter prognostication

Goals

- Confront the challenges of bias detection.
- Convey a sense for the social relevance of this work.
- Highlight some techniques that are of general utility (perhaps for your final projects)!

Research questions and challenges

- What are the markers of bias/partisanship?
- Do we see those features even when speakers are concealing their true biases?
- What is the role of the observer's perspective?
- How does the topic of discussion affect the expression of bias?

Some available data

- Thomas, Pang, and Lee: Congressional speech data:
  `http://www.cs.cornell.edu/home/llee/data/convote.html`
- 100K tweets from 40 political Tweeters, with party labels:
  `https://stanford.edu/class/cs424p/restricted/data/political-twitter.zip`
- CNN transcripts:
  `http://edition.cnn.com/TRANSCRIPTS/`
- Presidential and vice-presidential debate transcripts:
  `http://www.debates.org/index.php?page=debate-transcripts`
- Large political weblogs on AFS, in `mnt9/PottsCorpora`.
- Noah Smith's Political Blog Corpus (Yano et al. 2009):
  `http://www.ark.cs.cmu.edu/blog-data/`
- Policy Agendas Project
  `http://www.policyagendas.org/page/datasets-codebooks`

Feature selection

This section reviews the main points of Monroe et al.'s (2009) discussion of feature selection:

- Maximum Likelihood Estimates favor highly frequent words
- Log-odds ratios favor highly infrequent words
- Bayesian models with rich priors come closer to identifying characterizing words.

I also review log-likelihood ratio tests of the sort used by Yano et al. (2010) to find "sticky" bigrams.

Fightin' words

- Data: U. S. Senate speeches (1997-2004), categorized by party and topic and stemmed with the Porter stemmer.
- 118,000 speeches; 70,000,000 words
- The topics derive from an unsupervised topic model with similarities to Latent Dirichlet Allocation (LDA) but constrained to assign each document a single topic (Quinn et al. 2006).
- Intuitive labels for the inferred topics:
  - Abortion
  - Defense
  - Judicial nominations
  - Taxes
- General goal: define a technique for finding features that are truly indicative of specific sub-parts of the corpora (most prominently, of party–topic sub-corpora).

Maximum likelihood estimates (MLE)

- $y^i_{kw} \stackrel{def}{=}$ the number of tokens of word w in texts in topic k by individual/group i (with smoothing).

- Sub-corpus size:

$$n^i_k \stackrel{def}{=} \sum_{w \in Vocab} y^i_{kw}$$

- MLE:

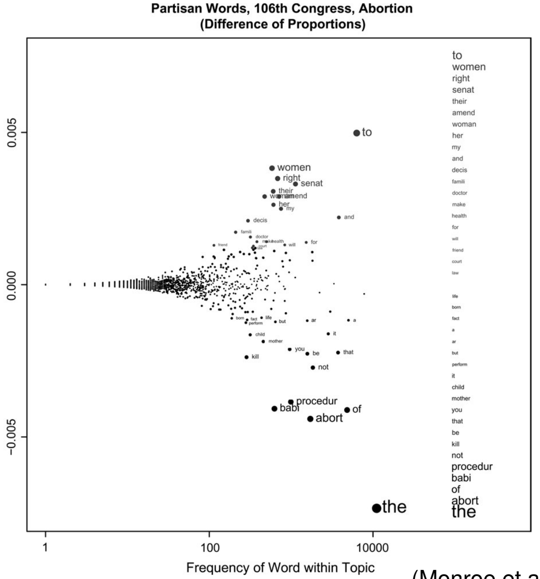
$$f^i_{kw} \stackrel{def}{=} \frac{y^i_{kw}}{n^i_k}$$

- Evaluation measure:

$$f^D_{kw} - f^R_{kw}$$

Maximum likelihood estimates (MLE)

$$f^D_{kw} - f^R_{kw}$$



(Monroe et al. 2009:fig. 1)

Maximum likelihood estimates (MLE)

Monroe et al. (2009:377–378)

> *The top Democratic word list is dominated by to and includes my, and, and for; the top Republican word list is dominated by the and includes of, not, be, that, you, it, and a. [. . .] the sampling variation in difference of proportions is greatest in high-frequency words. These are not partisan words; they are just common ones.*
>
> *A common response to this problem in many natural language processing applications is to eliminate "function" or "stop" words that are deemed unlikely to contain meaning. We note, however, the practice of stop word elimination has been found generally to create more problems than it solves [. . .]*

(We've seen that function words often carry sentiment information; see the 'Classification' handout especially.)

Odds ratios

- Odds:

$$O_{kw}^{i} \stackrel{def}{=} \frac{f_{kw}^{i}}{1-f_{kw}^{i}}$$

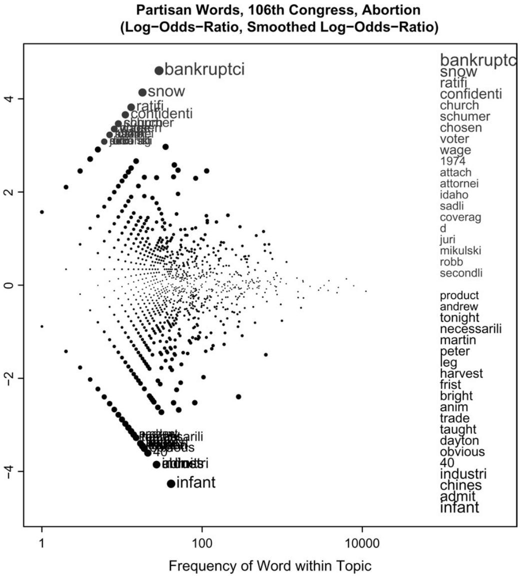
$$\stackrel{def}{=} \frac{y_{kw}^{i}}{n_{k}^{i}-y_{kw}^{i}}$$

- Evaluation measure:

$$\log(O_{kw}^{D} - O_{kw}^{R})$$

Odds ratios

$$\log(O^D_{kw} - O^R_{kw})$$



(Monroe et al. 2009:fig. 2)

Odds ratios

Monroe et al. (2009:379)

> *the most extreme words are obscure ones. These word lists are strange in a way opposite from that produced by the difference of proportions shown [for MLE]*

Odds ratios

Monroe et al. (2009:379)

> *A common response is to set some frequency "threshold" for features to "qualify" for consideration. Generally, this simply removes the most problematic features without resolving the issue.*
>
> *For an example of the latter, consider the list of Levitt and Dubner (2005; 194-8), in their freakonomic discussion of baby names, of "The Twenty White Girl [Boy] Names That Best Signify High-Education Parents." They identify these, from California records, with a list ranked by average mother's education. The top five girl names are Lucienne, Marie-Claire, Glynnis, Adair, and Meira; the top five boy names are Dov, Akiva, Sander, Yannick, and Sacha. A footnote clarifies that a name must appear at least 10 times to make the list (presumably because a list that allowed names used only once or twice might have looked ridiculous). It seems likely that each of these words was used exactly, or not many more than, 10 times in the sample.*

Log-likelihood ratios and the G-test

Definition (G-test)

$$G(w) \stackrel{def}{=} 2 \sum_{i,j} \text{Observed}_{ij} \log \left(\frac{\text{Observed}_{ij}}{\text{Expected}_{ij}} \right)$$

Significance testing as with χ^2. (Not esp. restrictive with large corpora.)

Definition (Expected counts)

$$\text{Expected}_{ij} \stackrel{def}{=} \sum_{j'} \text{Observed}_{ij'} * \left(\sum_{i'} \text{Observed}_{i'j} \Big/ \sum_{ij} \text{Observed}_{ij} \right)$$

Example

	Liberal	Conservative	Row totals
choice	265	97	362
not choice	1,001,136	690,987	1,692,123
Column totals	1,001,401	691,084	1,692,485

Bayesian priors with variance correction

- Observations from current sample of n words:

$$\mathbf{y} = \begin{bmatrix} \textit{choice} & \mapsto & 0.4 \\ \text{(all other words)} & \mapsto & 0.6 \end{bmatrix}$$

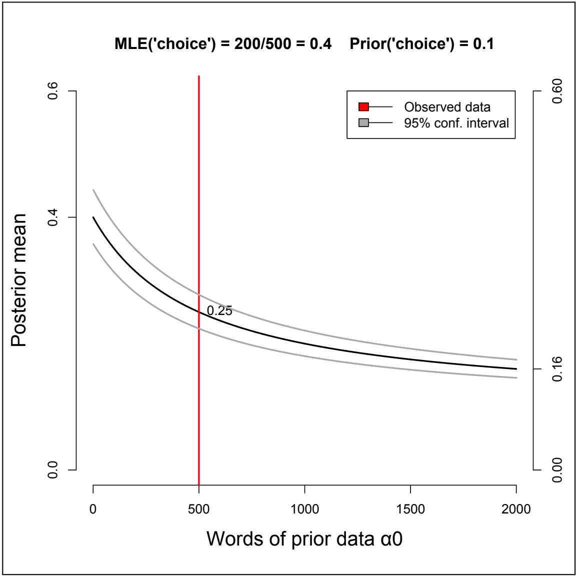
- Prior observations based on α_0 words:

$$\alpha = \begin{bmatrix} \textit{choice} & \mapsto & 0.1 \\ \text{(all other words)} & \mapsto & 0.9 \end{bmatrix}$$

- Posterior mean for *choice*:

$$\left(1 - \frac{\alpha_0}{\alpha_0 + n}\right) 0.4 + \left(\frac{\alpha_0}{\alpha_0 + n}\right) 0.1$$

Bayesian priors with variance correction



$$\left(1 - \frac{\alpha_0}{\alpha_0 + 500}\right) 0.4 + \left(\frac{\alpha_0}{\alpha_0 + 500}\right) 0.1$$

Less data means higher variance. Thus, low frequency words can be penalized by their high variance.

Bayesian priors with variance correction

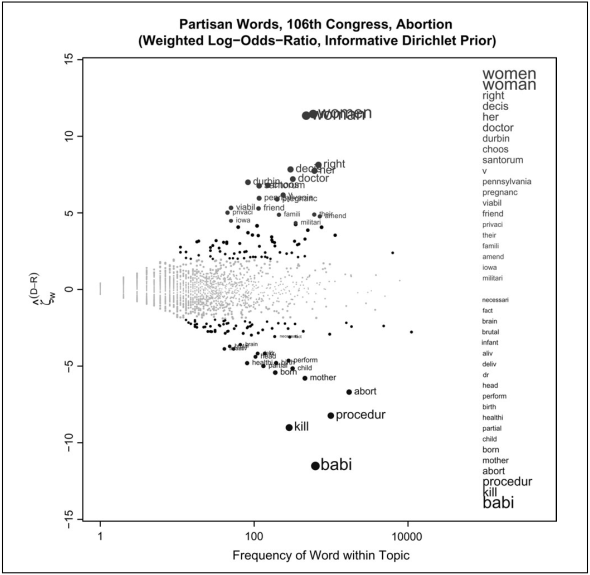
- Bayesian log-odds:

$$\delta_{kw}^{(D-R)} \stackrel{def}{=} \log\left(\frac{y_{kw}^{D} + \alpha_{kw}^{D}}{(n_k^D + \alpha_{k0}^{D}) - (y_{kw}^{D} + \alpha_{kw}^{D})}\right) - \log\left(\frac{y_{kw}^{R} + \alpha_{kw}^{R}}{(n_k^R + \alpha_{k0}^{R}) - (y_{kw}^{R} + \alpha_{kw}^{R})}\right)$$

- Evaluation measure:

$$\frac{\delta_{kw}^{(D-R)}}{\sqrt{\sigma^2\left(\delta_{kw}^{(D-R)}\right)}}$$

Bayesian priors with variance correction



(Monroe et al. 2009:fig. 7)

Some weblog results

DailyKos (liberal) and RedState (conservative), from the Yano et al. 2009 corpus.

Top conservative

	MLE	Odds	*Obs/Exp* LogLik $p \leqslant 0.05$	Bayesian 10K prior, IMDB	Bayesian 1M prior, IMDB
1	he	redstate	redstate	he	_meta_end_dot_
2	that	fns	fns	said	obama
3	is	tw	tw	mccain	mccain
4	to	steph	steph	obama	said
5	i	le	le	that	he
6	obama	securities	securities	romney	romney
7	mccain	lane	lane	read	hillary
8	not	mclaughlin	mclaughlin	tax	tax
9	_meta_end_dot_	stephanopoulos	stephanopoulos	is	senator
10	said	mtp	mtp	i	that
11	his	blackhedd	blackhedd	hillary	read
12	on	schieffer	schieffer	not	huckabee
13	romney	cianfrocca	cianfrocca	senator	barack
14	hillary	stearns	stearns	thompson	thompson
15	will	paulson	paulson	host	government
16	read	ftn	ftn	barack	host
17	be	erick	erick	market	president
18	would	daschle	daschle	huckabee	conservatives
19	tax	ayers	ayers	conservatives	conservative
20	senator	pejman	pejman	asked	barry

Some weblog results

DailyKos (liberal) and RedState (conservative), from the Yano et al. 2009 corpus.

	Top conservative				
	MLE	Odds	*Obs/Exp* LogLik $p \leqslant 0.05$	Bayesian 10K prior, IMDB	Bayesian 1M prior, IMDB
1	_meta_number_ref_	rangers	rangers	_meta_number_ref_	_meta_number_ref_
2	s	tracker	tracker	bush	!
3	in	moreville	moreville	!	bush
4	?	swoosh_meta_beep_ref_	swoosh_meta_beep_ref_	district	rescue
5	bush	gong_meta_beep_ref_	gong_meta_beep_ref_	fisa	fisa
6	t	rightnow	rightnow	update	diaries
7	!	susang	susang	s	district
8	clinton	telco	telco	diaries	movie
9	republican	srkp_meta_number_ref_	srkp_meta_number_ref_	blue	kos
10	so	pera	pera	daily	update
11	the	watercarrier_meta_number_ref_diogenes	watercarrier_meta_number_ref_diogenes	republican	film
12	republicans	avila	avila	republicans	daily
13	_meta_percent_ref_	taylormattd	taylormattd	kos	torture
14	for	grog	grog	?	rangers
15	update	vcmvo_meta_number_ref_	vcmvo_meta_number_ref_	torture	blue
16	ve	iob	iob	clinton	amnesty
17	race	shit	shit	rescue	caucus
18	vote	jennyjem	jennyjem	top	delegate
19	democratic	darcy	darcy	administration	maine
20	from	truthofangels	truthofangels	amnesty	cheney

Bias

- Techniques for finding textual indicators of bias.
- Close look at specific words and constructions that seem to carry subtle bias information.

On bias

Bias is not subjectivity

Yano et al. (2010): “A subjective sentence can be unbiased (*I think that movie was terrible*), and a biased sentence can purport to communicate factually (*Nationalizing our health care system is a point of no return for government interference in the lives of its citizens*).”

On bias

Bias is not always observed

Monroe et al. (2009:375) say that "Partisanship is perfectly observed" and thus conclude that there is no reason to build classifiers that predict it.

But this claim is just wrong; many people in public life (especially journalists and commentators) have a vested interest in trying to hide their partisanship, or even in being deceptive about it.

On bias

Bias is in the eye of the beholder

Yano et al. (2010): "In general, moderates perceive less bias than partisans [. . .], but conservatives show a much stronger tendency to label sentences as biased, in both directions. [. . .] Liberals in this sample are less balanced, perceiving conservative bias at double the rate of liberal bias."

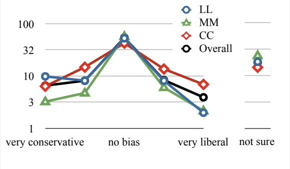


Figure 2: Distribution of bias labels (by judgment) for social and economic liberals (LL), social and economic moderates (MM), and social and economic conservatives (CC), and overall. Note that this plot uses a logarithmic scale, to tease apart the differences among groups.

Yano et al.'s annotations

Overall		Liberal		Conservative		Not Sure Which	
bad	0.60	Administration	0.28	illegal	0.40	pass	0.32
personally	0.56	Americans	0.24	Obama's	0.38	bad	0.32
illegal	0.53	woman	0.24	corruption	0.32	sure	0.28
woman	0.52	single	0.24	rich	0.28	blame	0.28
single	0.52	personally	0.24	stop	0.26	they're	0.24
rich	0.52	lobbyists	0.23	tax	0.25	happen	0.24
corruption	0.52	Republican	0.22	claimed	0.25	doubt	0.24
Administration	0.52	union	0.20	human	0.24	doing	0.24
Americans	0.51	torture	0.20	doesn't	0.24	death	0.24
conservative	0.50	rich	0.20	difficult	0.24	actually	0.24
doubt	0.48	interests	0.20	Democrats	0.24	exactly	0.22
torture	0.47	doing	0.20	less	0.23	wrong	0.22

Table 6: Most strongly biased words, ranked by relative frequency of receiving a bias mark, normalized by total frequency. Only words appearing five times or more in our annotation set are ranked.

Lakovian framing

Central tenets of framing (Lakoff 2003, 2004)

1. Every word has a frame.
2. Negating a frame evokes that frame.
3. Evoking a frame reenforces that frame.

Lakovian framing

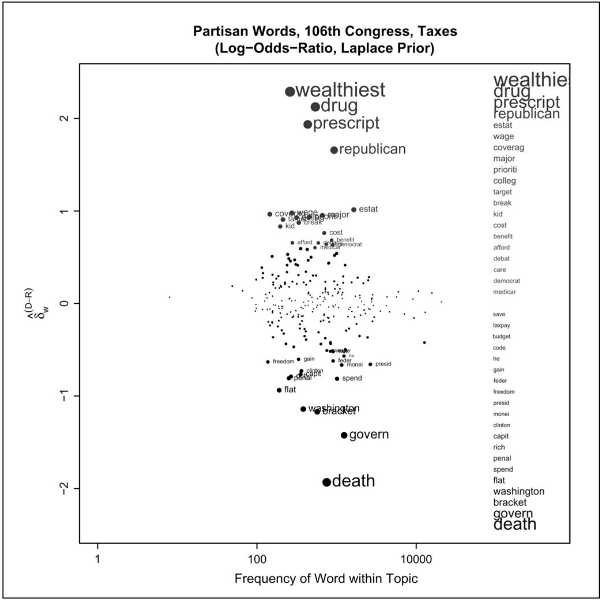
Central tenets of framing (Lakoff 2003, 2004)

1. Every word has a frame.
2. Negating a frame evokes that frame.
3. Evoking a frame reenforces that frame.

Examples (Relief)

1. Ed was relieved from his pain.
2. The pool hustler relieved Sally of her money.
3. hunger relief
4. We relieved Ed from his chores.
5. We relieved Ed from his vacation.
6. tax relief
7. $$\underset{\text{reliever-of-pain}}{\underset{\uparrow}{x}} \quad \text{relieves} \quad \underset{\text{blameless afflicted}}{\underset{\uparrow}{y}} \quad \text{from} \quad \underset{\text{cause}}{\underset{\uparrow}{z}}$$

Tax frames: Bayesian estimates from 'Fightin' words'



- prescription drug coverage
- college
- wealthiest
- breaks
- marriage penalty
- death tax

Meme identification and tracking

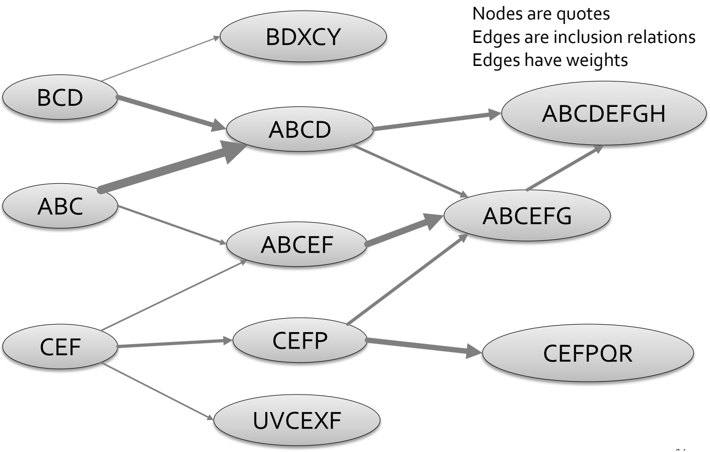
Leskovec et al. (2009:497–498)

> *short distinctive phrases that travel relatively intact through on-line text as it evolves over time.*
> *[. . .]*
> *From an algorithmic point of view, we consider these distinctive phrases to act as the analogue of "genetic signatures" for different memes. And like genetic signatures, we find that while they remain recognizable as they appear in text over time, they also undergo significant mutation. As a result, a central computational challenge in this approach is to find robust ways of extracting and identifying all the mutational variants of each of these distinctive phrases, and to group them together.*

Meme identification and tracking

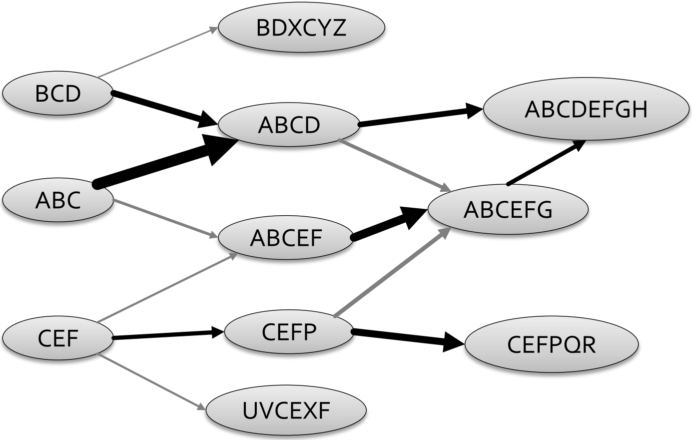
- Directed acyclic phrase graph G
  - Edge $(p, q) \in G$ iff p's words are a subset of q's and the word-level edit distance from p to q is less than a threshold (1 in their experiments).
  - Edge weights: decreases based on edit distance, increases based on the frequency of q.
  - Goal: Partition G into a series of subgraphs in which each path terminates at a single common node without any outgoing edges.
  - Problem is NP-complete; approximate approach: start with the leaves and move through to their roots, keeping, for each node n, only the edge (n, n'), where n' has the highest frequency of all edges (n, n'').

Meme identification and tracking



(graphics from Jure's slides)

Meme identification and tracking



(graphics from Jure's slides)

Meme identification and tracking

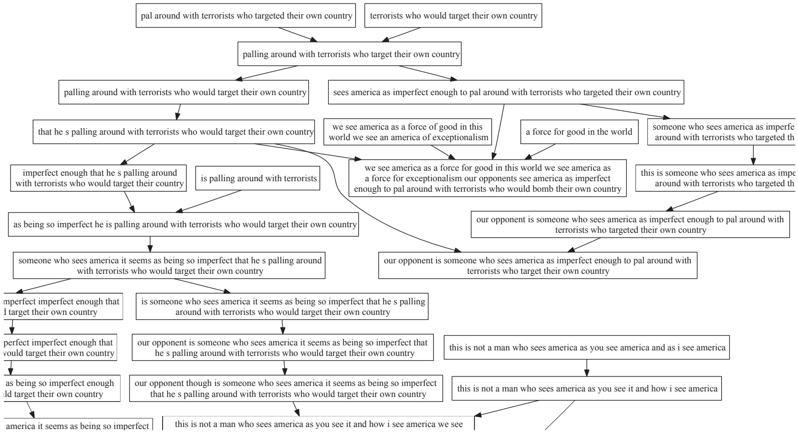


Figure 1: A small portion of the full set of variants of Sarah Palin's quote, "Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country." The arrows indicate the (approximate) inclusion of one variant in another, as part of the methodology developed in Section 2.

Meme identification and tracking

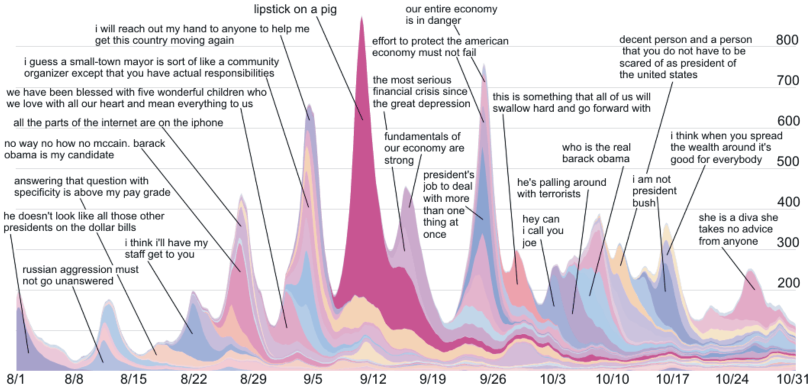


Figure 4: Top 50 threads in the news cycle with highest volume for the period Aug. 1 – Oct. 31, 2008. Each thread consists of all news articles and blog posts containing a textual variant of a particular quoted phrases. (Phrase variants for the two largest threads in each week are shown as labels pointing to the corresponding thread.) The data is drawn as a stacked plot in which the thickness of the strand corresponding to each thread indicates its volume over time. Interactive visualization is available at http://memetracker.org.

Partisan bigrams (Yano et al. 2010)

Log-likelihood ratios (plus stopword removal and some manual clean-up) to identify the top liberal and conservative bigrams.

Liberal	Conservative
thinkprogress org	exit question
video thinkprogress	hat tip
et rally	ed lasky
org 2008	hot air
gi bill	tony rezko
wonk room	ed morrissey
dana perino	track record
phil gramm	confirmed dead
senator mccain	american thinker
abu ghraib	illegal alien

Table 1: Top ten “sticky” partisan bigrams for each side.

Definite determiners

Acton (2010)

Using *the NP*, where *NP* denotes a group to which the speaker could belong, tends to indicate that the speaker thinks of himself as outside that group.

Definite determiners

Acton (2010)

Using *the NP*, where *NP* denotes a group to which the speaker could belong, tends to indicate that the speaker thinks of himself as outside that group.

Initial support

- *The Americans* by Robert Frank (Swiss emigre)
  "It was initially dismissed as the jaundiced work of an unpatriotic cynic who kept company with the similarly subversive and ragtag Beats." (NPR, Feb 13, 2009)
- *The Americans* by Gordon Sinclair (Canadian journalist)

Definite determiners

Data

Participants Taking at Least 1,000 Turns on *The McLaughlin Group*: May 23, 1998 - May 13, 2001		
Participant	**Turns Taken**	**Share of All Turns Taken on the Program***
John McLaughlin	12,279	39%
Eleanor Clift	4,643	15%
Tony Blankley	3,007	10%
Michael Barone	2,296	7%
Pat Buchanan	1,364	4%
Larry Kudlow	1,303	4%
Lawrence O'Donnell	1,267	4%
TOTALS:	26,159	84%

Source: Transcripts from www.mclaughlin.com. Calculations by Eric Acton.
*Includes *all* participants on the program during the time span. Accordingly, these percentages do not total to 100%.

Definite determiners

the Russians vs. *the Americans*

Tokens of *Russians* and *Americans* and Their Preceding Words
***The McLaughlin Group*:**
May 23, 1998 - May 13, 2001*

[a] Term	[b] Total # of Tokens	[c] # of Tokens Immediately Preceded by *the*	[d] % of Total Number of Tokens ([c] / [b])
Russians.....................	82	75	91%
Americans...................	186	10	5%
COMBINED.................	268	85	32%

Source: Transcripts from www.mclaughlin.com. Calculations by Eric Acton.
*Includes only those participants who took at least 1,000 turns talking over the specified time period.

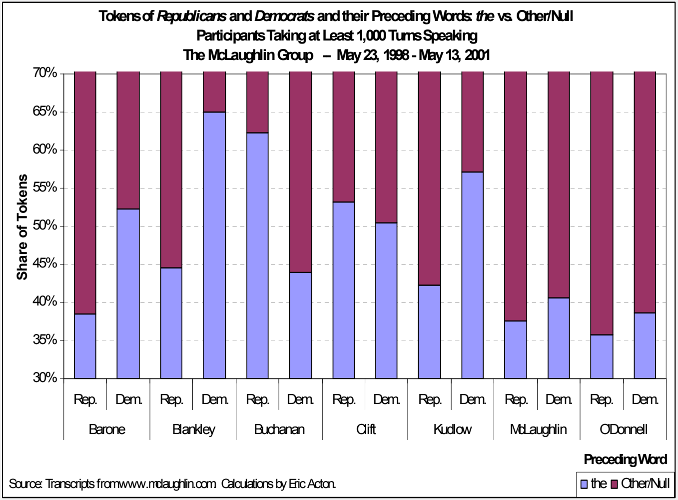
Definite determiners

Partisanship (quotes from Wikipedia, via Acton (2010))

John McLaughlin	"McLaughlin is a lifetime Republican. [...] His political views in general are diversified and often differ from the Republican Party."
Tony Blankley	"Blankley's political opinions are generally considered to fall within traditional conservatism."
Michael Barone	"generally conservative"
Pat Buchanan	"Republican (1960s-1999, 2004-present); Reform (1999-2000)."
Larry Kudlow	"Democrat turned conservative"
Eleanor Clift	"fierce defense of Hillary Rodham Clinton and Bill Clinton"
Lawrence O'Donnell	"He calls himself a 'practical European socialist.'"

Definite determiners

Results



Other features?

What else might we expect to carry information about bias/partisanship?

Classification

This section reviews the techniques and results of Thomas et al. (2006). Overview:

1. Data: snippets of Congressional speeches, labeled with how the speaker voted on the bill in question (along with party information and some other meta-data).
2. Standard unigrams-based text classifier predicting
3. Classifiers predicting speaker agreement.
4. Additional constraints across classification decisions.
5. A graph-based optimization method brings the pieces together to predict speaker votes.

The Convote corpus

Bill	052
Speaker	400011
Party	Democrat
Vote	No
Sample	the question is , what happens during those 45 days ? we will need to support elections . there is not a single member of this house who has not supported some form of general election , a special election , to replace the members at some point . but during that 45 days , what happens ?

(Thomas et al. 2006)

The Convote corpus

Bill	052
Speaker	400077
Party	Republican
Vote	Yes
Sample	i believe this is a fair rule that allows for a full discussion of the relevant points pertaining to the legislation before us . mr. speaker , h.r. 841 is an important step forward in addressing what are critical shortcomings in america 's plan for the continuity of this house in the event of an unexpected disaster or attack .

(Thomas et al. 2006)

The Convote corpus

	total	train	test	development
speech segments	3857	2740	860	257
debates	53	38	10	5
average number of speech segments per debate	72.8	72.1	86.0	51.4
average number of speakers per debate	32.1	30.9	41.1	22.6

Table 1: Corpus statistics.

Hierarchy of texts:

Debates (collections of speeches by different speakers)

$\Uparrow$

Speeches (collections of segments by the same speaker)

$\Uparrow$

Speech segments (documents in the corpus)

(Thomas et al. 2006)

Basic classification with same-speech links

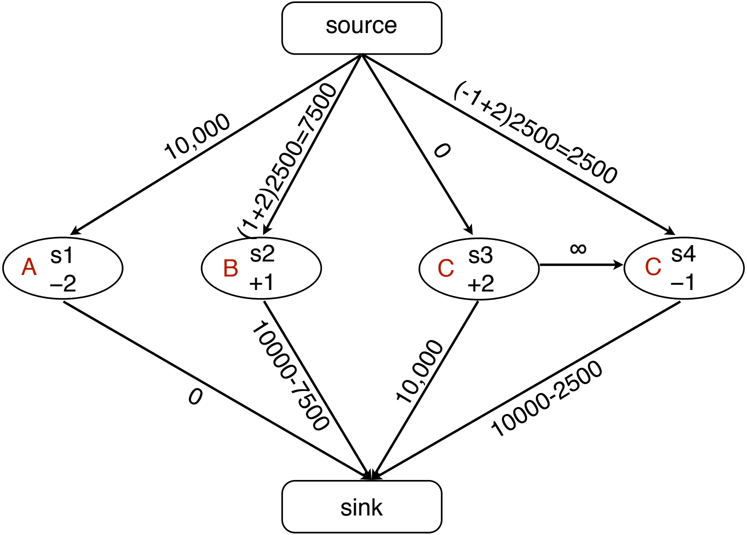
1. SVM classifier with unigram-presence features predicting, for each speech-segment, how the speaker voted (Y or N).
2. For each document s belonging to speech S, the SVM score for s is divided by the standard deviation for all $s' \in S$.
3. Debate-graph construction with minimal cuts:

$$\text{score}(s) \leqslant -2 \Rightarrow \left[\begin{array}{rcl} \text{source} & \overset{0}{\rightarrow} & s \\ s & \overset{10{,}000}{\rightarrow} & \text{sink} \end{array} \right.$$

$$\text{score}(s) \geqslant +2 \Rightarrow \left[\begin{array}{rcl} \text{source} & \overset{10{,}000}{\rightarrow} & s \\ s & \overset{0}{\rightarrow} & \text{sink} \end{array} \right.$$

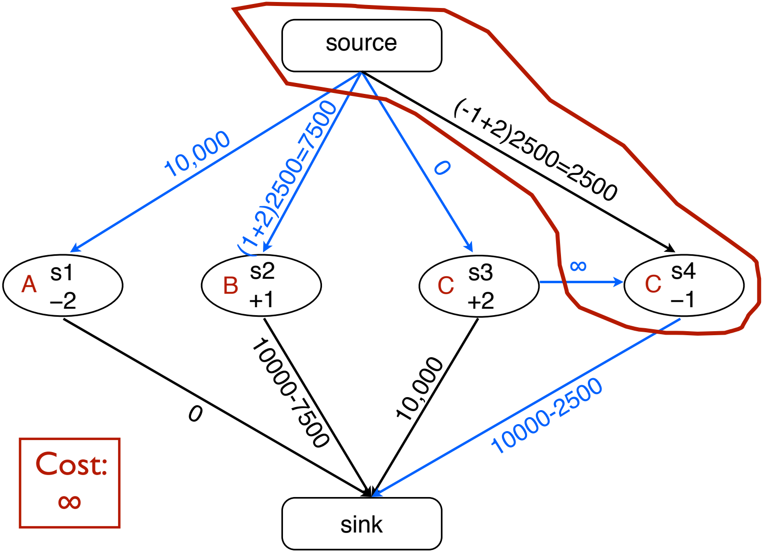
$$\text{else} \Rightarrow \left[\begin{array}{rcl} \text{source} & \overset{x=(\text{score}(s)+2)2500}{\rightarrow} & s \\ s & \overset{10{,}000-x}{\rightarrow} & \text{sink} \end{array} \right.$$

Graph construction and minimal cuts



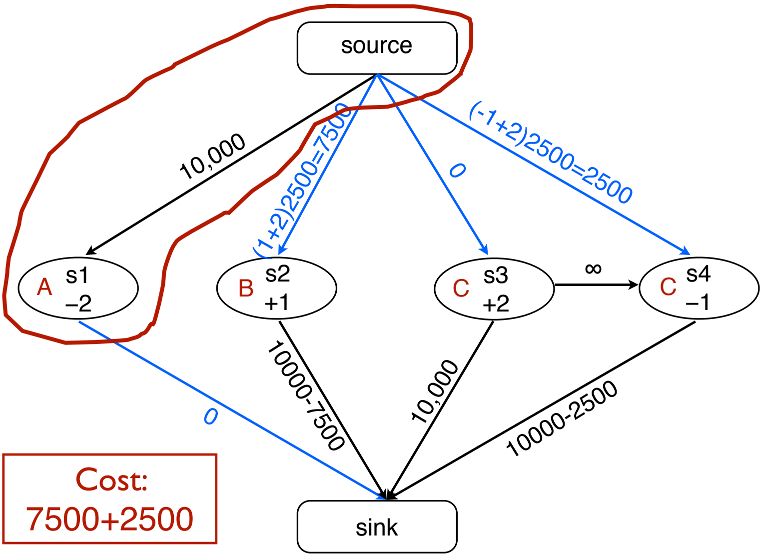
(source = No; sink = Yes)

Graph construction and minimal cuts



(source = No; sink = Yes)

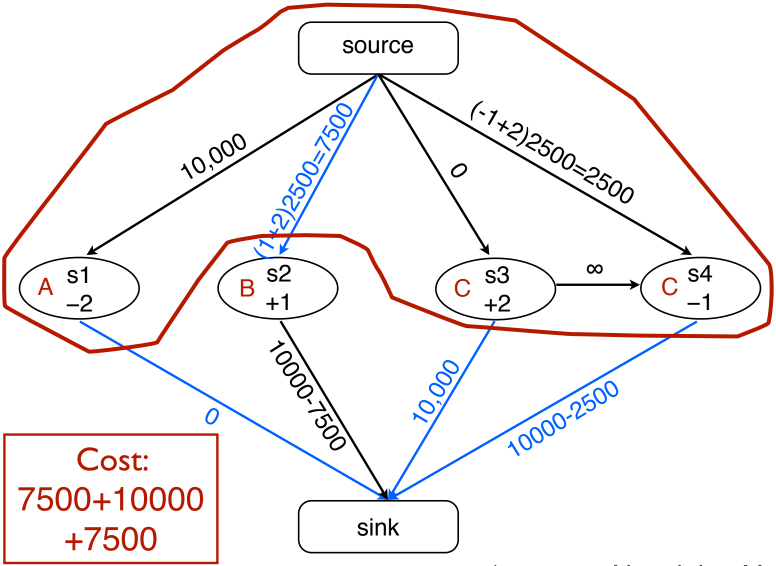
Graph construction and minimal cuts



Cost:
7500+2500

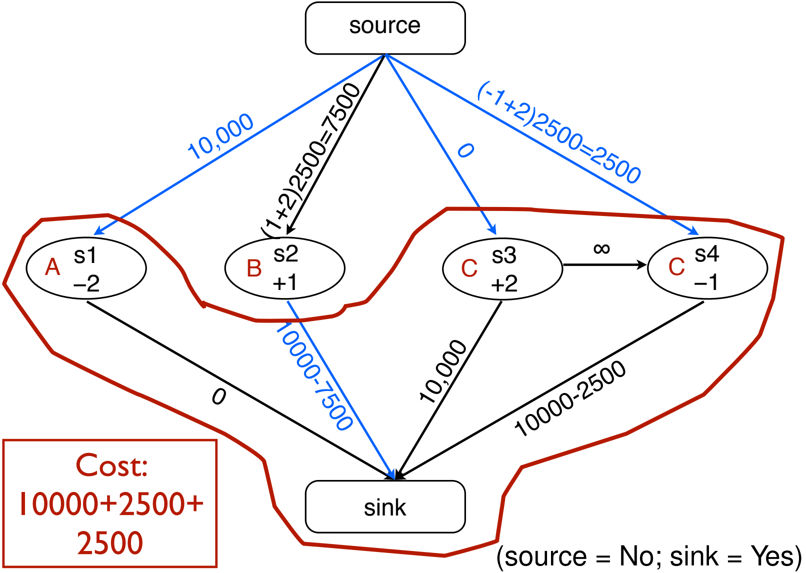
(source = No; sink = Yes)

Graph construction and minimal cuts



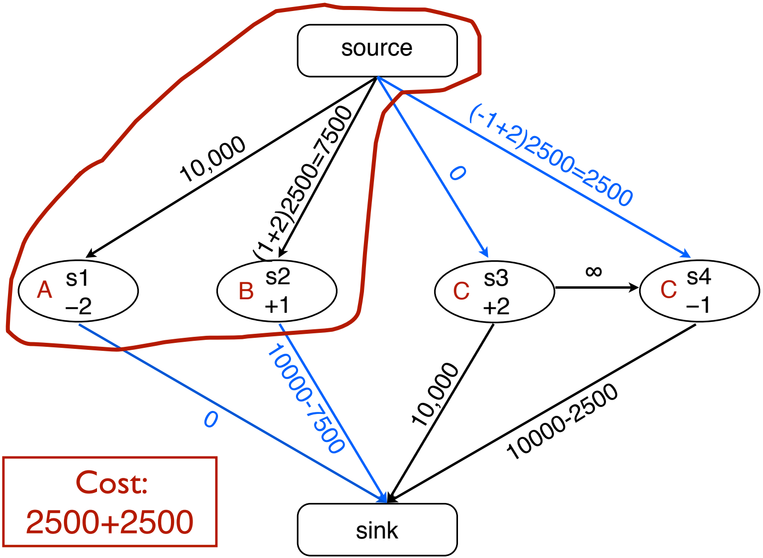
(source = No; sink = Yes)

Graph construction and minimal cuts



(source = No; sink = Yes)

Graph construction and minimal cuts



Cost:
2500+2500

(source = No; sink = Yes)

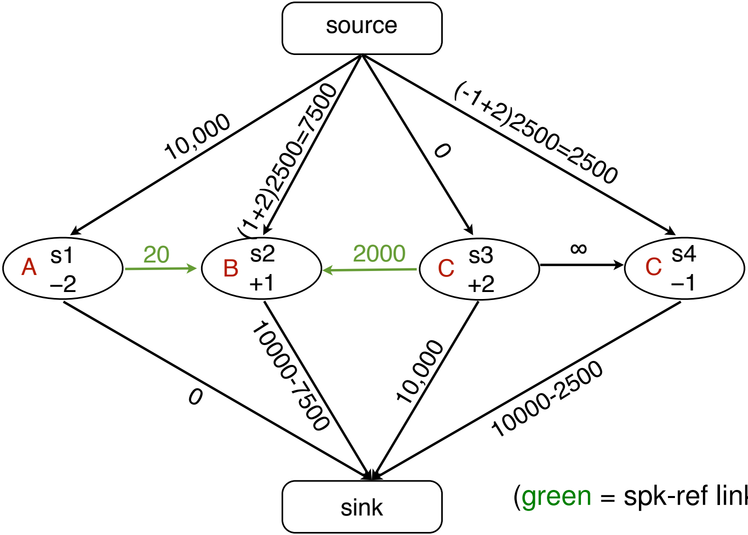
Speaker references

Bill	006
Speaker	400115
Party	Republican
Vote	Yes
Sample	mr. speaker , i am very happy to yield 3 minutes to the gentleman from new york (mr. boehlert) xz4000350 , the very distinguished chairman of the committee on science .
Bill	006
Speaker	400035
Party	Republican
Vote	Yes
Sample	mr. speaker , i rise in strong support of this balanced rules package . i want to speak particularly to the provisions regarding homeland security . [. . .]

Speaker reference classifier

1. Label a reference as Agree if the speaker and the Referent voted the same way, else Disagree.
2. Features: 30 unigrams before, the name, and 30 unigrams after
3. Normalized SVM scores from this classifier are then added to the debate graphs, at the level of speech segments. (Where a speaker has multiple speech segments, one is chosen at random; the infinite-weight links ensure that this information propagates to the others.)

Inter-text and inter-speaker links



(green = spk-ref links)

Results

Support/oppose classifer ("speech segment⇒yea?")	Devel. set	Test set
majority baseline	54.09	58.37
#("support") − #("oppos")	59.14	62.67
SVM [speech segment]	70.04	66.05
SVM + same-speaker links	79.77	67.21
SVM + same-speaker links ...		
+ agreement links, $\theta_{agr} = 0$	**89.11**	**70.81**
+ agreement links, $\theta_{agr} = \mu$	87.94	71.16

Table 4: Segment-based speech-segment classification accuracy, in percent.

θ_{agr} is a free-parameter in the scaling function for speaker agreement scores. The development results suggest that 0 is the better value than μ (a mean of all the debate's scores), but μ performs better in testing.

Twitter prognostication

Predicting political mood

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Central question: "Is text sentiment a leading indicator of polls?"

Data

- 1 billion messages from 2008 and 2009
- The ‘gardenhose’ feed: should be a uniform sample of public messages if run for long enough.
- 100,000 to 7 million messages per day, with the variation largely a function of Twitter’s overall size
- Very little preprocessing for these experiments.

Consumer confidence polls

Index of Consumer Sentiment (ICS)

Administered monthly by telephone to several hundred people in the U.S.. Participants respond to questions on the following topics:

- current and future personal financial situation
- current and future state of the economy
- confidence in the government to help the economy
- views on a few other economic indicators

Gallup Economic Confidence Index

Two questions asking participants to rate the overall economic health of the country. Administered daily.

Tweet analysis

Topic keywords

Consumer confidence: *economy*, *job*, and *jobs*.

Sentiment

- 1600 positive, 1200 negative words from OpinionFinder
- A message is positive (negative) if it contains a positive (negative) word according to the lexicon.
- Sentiment score for day t and topic words S:

$$x_{t,S} \stackrel{def}{=} \frac{\text{positive tweets with a word in } S \text{ on day } t}{\text{negative tweets with a word in } S \text{ on day } t}$$

- Smoothing over the past k days:

$$MA_{t,S} \stackrel{def}{=} \frac{1}{k} \sum_{i=t-k+1}^{t} x_{i,S}$$

Broad correlation between tweets and polls

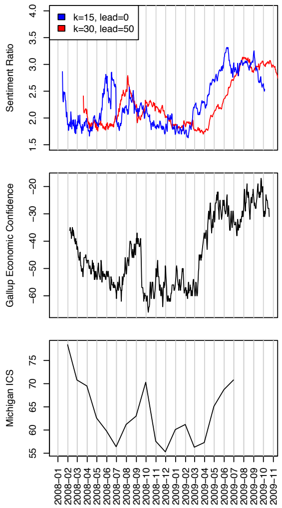


Figure 6: Sentiment ratio and consumer confidence surveys. Sentiment information captures broad trends in the survey data.

- Topic keyword: *jobs*
- Smoothing over 15 days
- Smoothing over 30 days with the tweets 50 days before the poll
- Twitter–Gallup correlation: 73.1%

- *job* and *economy* correlate very poorly with Gallup

Tweets as a leading indicator

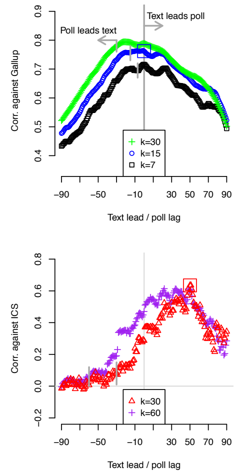


Figure 7: Cross-correlation plots: sensitivity to lead and lag for different smoothing windows. $L > 0$ means the text window completely precedes the poll, and $L < -k$ means the poll precedes the text. (The window straddles the poll for $L < -k < 0$.) The $L = -k$ positions are marked on each curve. The two parameter settings shown in Figure 6 are highlighted with boxes.

- The correlations are higher on the right side, where the tweets lead the polls.
- The correlations increase with increased smoothing.

Political polls

- O'Connor et al. (2010) also look at correlations between
  - $x_{t,\{obama,mccain\}}$ and the 2008 election polls
  - $x_{t,\{obama\}}$ and the 2009 presidential job approval polls.
- The results are pretty good for the job approval experiment, but that is easy to predict: it has been in decline since Obama took office.
- The results are poor for the election experiment.
- Tweet volume is a better predictor than sentiment, a result that is echoed in Asur and Huberman (2010), discussed next.

Predicting box office revenue

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. arXiv:1003.5699v1.

Central question: “Can Twitter data predict weekend box office revenue?”

- 2.89 million tweets, 1.2 million users, 24 movies released Dec. 2009 to Mar. 2010.
- Publicity, as measured by URLs and retweets: only modestly correlated with box office performance.

Predicting opening weekend

Linear model predicting opening box office. Features defined over the 7 days prior to release.

Model	Adjusted R^2
Mean tweets/hour	0.80
Mean tweets/hour, 1 feature per day	
without theater count	0.93
with theater count	0.973
Hollywood Stock Exchange index	
with theater count	0.965

- Rounding to 2 places brings the final two models together.
- All the data is used in training/estimation.

Beyond opening weekend

As before, features are defined for the seven days preceding the weekend, with predictions made about the weekend box office:

- Mean tweets/hour, 1 feature per day
- Theater count
- Number of weeks the movie has been playing

Week	Adjusted R^2
Jan 15-17	0.92
Jan 22-24	0.97
Jan 29-31	0.92
Feb 05-07	0.95

A dash of sentiment

Annotation with Mechanical Turk

Workers assigned sentiment (Positive, Negative, Neutral) to “a large random sample of tweets”. Three Turkers per tweet, keeping only tweets for which the response was unanimous.

Features

Removal of stopwords, all punctuation except ! and ?, and all URLs and user-ids. All movie titles were replaced with ‘MOV’.

Classifier

Results

98% accuracy! (This is either amazing or mundane, depending on how imbalanced the training set was.)

Predicting weekend #2 using sentiment

Subjective tweets about a movie more frequent after its release.

$$\text{PNratio} \stackrel{def}{=} \frac{\text{number of positive tweets as predicted by the classifier}}{\text{number of negative tweets as predicted by the classifier}}$$

Model	Adjusted R^2
Mean tweets/hour	0.79
with theater count	0.83
with PNratio	0.92
Mean tweets/hour, 1 feature per day	0.84
with theater count	0.863
with PNratio	0.94

Predicting Dow directions

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. arXiv:1010.3003v1.

Central question: “Can Twitter sentiment predict the day-to-day direction of Dow Jones Industrial Average movement?”

Watch for forthcoming data and tools (empty as of this writing):
`http://terramood.informatics.indiana.edu/data`

Data

Twitter

1. About 10 million tweets from 2.7 million users
2. Deb 28 to Dec 19, 2008
3. Group all tweets from the same day.
4. Keep only tweets that match:
   `(i feel|i am feeling|i'm feeling|i dont feel|i don't feel|i'm|im)`
   and don't match `(http|www)`.
5. Remove stopwords.

Financial

Daily Dow Jones Industrial Average closing values:

$$D_t \stackrel{def}{=} DJIA_t - DJIA_{t-1}$$

Mood assessment

OpinionFinder

2718 positive strings, and 4912 negative strings

GPOMS (Google Profile of Mood States)

1. Begin with the 72 strings in the Profile of Mood States, in 6 categories: Calm, Alert, Vital, Vital, Kind, Happy.
2. Expand to 964 strings using “word co-occurrences” in the Google Ngrams Corpus.

Definition (Daily GPOMS values)

For a day t and a sentiment category X:

$$X_t \stackrel{def}{=} \sum_{w \in X} \mathrm{count}(w) * \mathrm{weight}(w)$$

where weight(w) is the maximum co-occurrence weight for w and a word in the original category POMS category. (Note: I am not positive that this is the precise measure they use. See p. 3)

Correlations with known/expected public mood

The scores here and throughout are z-score normalized.

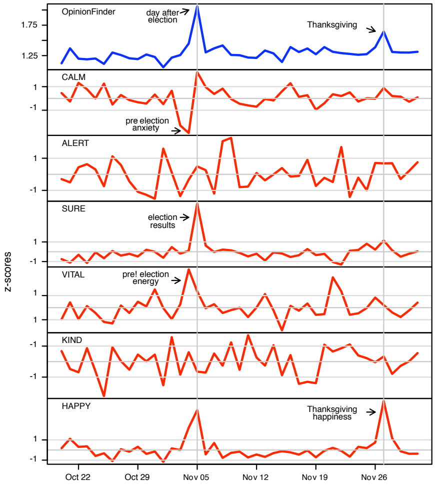


Fig. 2. Tracking public mood states from tweets posted between October 2008 to December 2008 shows public responses to presidential election and thanksgiving.

Linear models

Definition (Previous *n* Dow values only)

$$D_t \sim \text{intercept} + \sum_{i=1}^{n} \beta_i D_{t-i} + \varepsilon_i$$

Definition (Previous *n* Dow and *X* values.)

$$D_t \sim \text{intercept} + \sum_{i=1}^{n} \beta_i D_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \varepsilon_i$$

Results: Only $X = $ Calm is a reliable predictor.

TABLE II

STATISTICAL SIGNIFICANCE (P-VALUES) OF BIVARIATE GRANGER-CAUSALITY CORRELATION BETWEEN MOODS AND DJIA IN PERIOD FEBRUARY 28, 2008 TO NOVEMBER 3, 2008.

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 day	**0.085**⋆	0.272	0.952	0.648	0.120	0.848	0.388
2 days	0.268	**0.013**⋆⋆	0.973	0.811	0.369	0.991	0.7061
3 days	0.436	**0.022**⋆⋆	0.981	0.349	0.418	0.991	0.723
4 days	0.218	**0.030**⋆⋆	0.998	0.415	0.475	0.989	0.750
5 days	0.300	**0.036**⋆⋆	0.989	0.544	0.553	0.996	0.173
6 days	0.446	**0.065**⋆	0.996	0.691	0.682	0.994	**0.081**⋆
7 days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

(p-value < 0.05: ⋆⋆, p-value < 0.1: ⋆)

Linear models

TABLE II

STATISTICAL SIGNIFICANCE (P-VALUES) OF BIVARIATE GRANGER-CAUSALITY CORRELATION BETWEEN MOODS AND DJIA IN PERIOD FEBRUARY 28, 2008 TO NOVEMBER 3, 2008.

Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 day	**0.085***	0.272	0.952	0.648	0.120	0.848	0.388
2 days	0.268	**0.013****	0.973	0.811	0.369	0.991	0.7061
3 days	0.436	**0.022****	0.981	0.349	0.418	0.991	0.723
4 days	0.218	**0.030****	0.998	0.415	0.475	0.989	0.750
5 days	0.300	**0.036****	0.989	0.544	0.553	0.996	0.173
6 days	0.446	**0.065***	0.996	0.691	0.682	0.994	**0.081***
7 days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

(p-value < 0.05: $\star\star$, p-value < 0.1: $\star$)

Correlations between DJIA closings and Calm

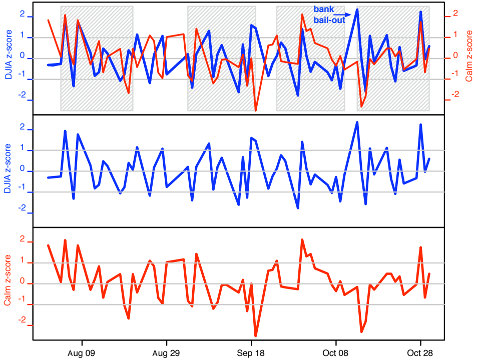


Fig. 3. A panel of three graphs. The top graph shows the overlap of the day-to-day difference of DJIA values (blue: $\mathbb{Z}_{D_t}$) with the GPOMS' Calm time series (red: $\mathbb{Z}_{X_t}$) that has been lagged by 3 days. Where the two graphs overlap the Calm time series predict changes in the DJIA closing values that occur 3 days later. Areas of significant congruence are marked by gray areas. The middle and bottom graphs show the separate DJIA and GPOMS' Calm time series.

Correlations between DJIA closings and Calm

Those were interesting times . . .

we point to a significant deviation between the two graphs on October 13th where the DJIA surges by more than 3 standard deviations trough-to-peak. The Calm curve however remains relatively flat at that time after which it starts to again track changes in the DJIA again. This discrepancy may be the result of the the Federal Reserve's announcement on October 13th of a major bank bailout initiative which unexpectedly increase DJIA values that day. The deviation between Calm values and the DJIA on that day illustrates that unexpected news is not anticipated by the public mood yet remains a significant factor in modeling the stock market. (p. 5)

Nonlinear models

Model

Self-organizing Fuzzy Neural Network

Training and testing

- Training: Feb. 28, 2008 to Nov. 28, 2008
- Testing: Dec. 1, 2008 to Dec. 19, 2008

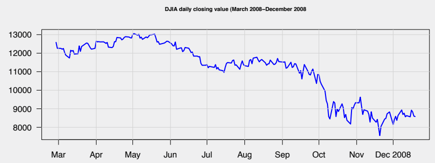


Fig. 4. Daily Dow Jones Industrial Average values between February 28, 2008 and December 19, 2008.

Results

All models use the previous three days of values. MAPE: mean absolute percentage error. Direction: up or down.

Model	MAPE (%)	Direction (%)
DJIA	1.94	73.3
DJIA + OpinionFinder	1.95	73.3
DJIA + Calm	1.83	$\boxed{86.7}$
DJIA + Calm + Alert	2.03	60.0
DJIA + Calm + Sure	2.13	46.7
DJIA + Calm + Vital	2.05	60.0
DJIA + Calm + Kind	1.85	73.3
DJIA + Calm + Happy	$\boxed{1.79}$	80.0

"The odds that the mentioned probability would hold by chance for a random period of 20 days within that period is then estimated to be [. . .] 3.4%." (p. 6)

Summary

- These papers are provocative in a good way: they strongly suggest that Twitter takes a measure of general public mood that correlates with real world events.
- A frustration: it's unclear, to me and to the authors of these papers, why their models work the way they do.
- (Interesting control experiment: try to use frequency of function words, or tokens of `Justin Bieber`, to predict things about the future.)

References I

Acton, Eric. 2010. The outsiders: ‘the’ as a marker of speaker non-membership. Ms., Stanford University.

Asur, Sitaram and Bernardo A. Huberman. 2010. Predicting the future with social media. ArXiv:1003.5699v1.

Lakoff, George. 2003. Simple framing. *The Rockridge Institite Website* URL `http://www.rockridgeinstitute.org/people/lakoff`.

Lakoff, George. 2004. *Don’t Think of an Elephant!* White River, VT: Chelsea Green.

Leskovec, Jure; Lars Backstrom; and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. New York: ACM.

Monroe, Burt L; Michael P. Colaresi; and Kevin M. Quinn. 2009. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16:372–403.

O’Connor, Brendan; Balasubramanyan; Bryan R. Routledge; and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Quinn, Kevin M.; Burt L Monroe; Michael P. Colaresi; and Michael H. Crespin. 2006. How to analyze political attention with minimal assumptioins and costs. Ms., Society for Political Methodology, Davis, CA.

Thomas, Matt; Bo Pang; and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, 327–335.

References II

Yano, Tae; William W. Cohen; and Noah A. Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of NAACL-HLT*. New York: ACL.

Yano, Tae; Philip Resnik; and Noah A. Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL-HLT Workshop on Creating Speech and Language Data With Mechanical Turk*. ACL.