Sentiment analysis and context dependence

Christopher Potts

CS 244U: Natural language understanding
Feb 23

Overview

1. Sentiment as blended and continuous (Experience Project data)
2. Topic-relative sentiment (review data)
3. Sentiment as social: congressional voting data (Thomas et al. 2006)
4. Sentiment as social: Twitter users (Tan et al. 2011)
5. Sentiment as social: Experience Project users and groups
6. Sentiment and morphosyntax

Sentiment as blended and continuous

This one is for the long-suffering fans, the bittersweet memories, the hilariously embarrassing moments, . . .

Sentiment as a classification problem

- Pioneered by Pang et al. (2002), who apply Naive Bayes, MaxEnt, and SVMs to the task of classifying movie reviews as positive or negative,
- and by Turney (2002), who developed vector-based unsupervised techniques (see also Turney and Littman 2003).
- Extended to different sentiment dimensions and different categories sets (Cabral and Hortaçsu 2006; Pang and Lee 2005; Goldberg and Zhu 2006; Snyder and Barzilay 2007; Bruce and Wiebe 1999; Wiebe et al. 1999; Hatzivassiloglou and Wiebe 2000; Riloff and Wiebe 2003; Riloff et al. 2005; Pang and Lee 2004; Thomas et al. 2006; Liu et al. 2003; Alm et al. 2005; Wiebe et al. 2005; Neviarouskaya et al. 2010).
- Fundamental assumption: each textual unit (at whatever level of analysis) either has or does not have each sentiment label — usually it has exactly one label.
- Fundamental assumption: while the set of all labels might be ranked, they are not continuous.

MaxEnt for sentiment classification

Definition (MaxEnt)

$$P(class|text, \lambda) = \frac{\exp\left(\sum_i \lambda_i f_i(class, text)\right)}{\sum_{class'} \exp\left(\sum_i \lambda_i f_i(class', text)\right)}$$

Minimize:

$$-\sum_{class,text} \log P(class|text, \lambda) + \log P(\lambda)$$

Gradient:

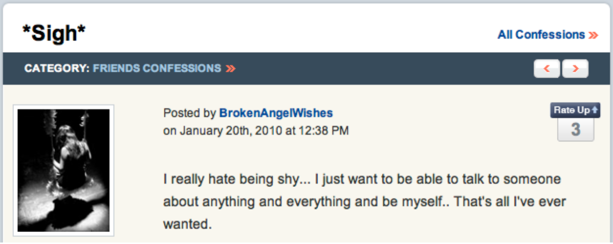
$$\mathit{empirical\ count}(f_i, c) - \mathit{predicted\ count}(f_i, \lambda)$$

- A powerful modeling idea for sentiment — can handle features of different type and feature sets with internal statistical dependencies.
- Output is a probability distribution, but classification is typically just based on the most probable class, with little attention to the full distribution.
- Uncertainty about the underlying labels in *empirical count*(f_i, c) is typically also supressed/ignored.

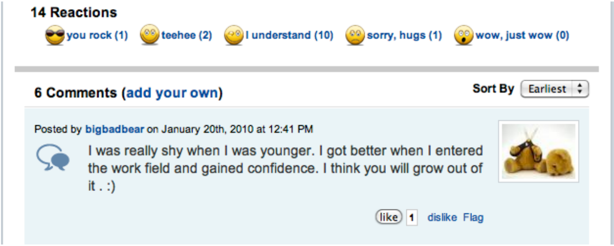
Objections to sentiment as classification

- The expression of emotion in language is nuanced, blended, and continuous Russell (1980); Ekman (1992); Wilson et al. (2006).
- Human reactions are equally complex and multi-dimensional.
- Insisting on a single label doesn't do justice to the author's intentions, and it leads to unreliable labels.
- Few attempts to address this at present (Potts and Schwarz 2010; Potts 2011; Maas et al. 2011; Socher et al. 2011), though that will definitely change soon:
    - New datasets emerging
    - Demands from industry
    - New statistical models

Experience Project confessions: blended, continuous sentiment reactions



[. . .]



Experience Project confessions: blended, continuous sentiment reactions

Confession:	I really hate being shy ... I just want to be able to talk to someone about anything and everything and be myself... That's all I've ever wanted.
Reactions:	*hugs*: 1; *rock*: 1; *teehee*: 2; *understand*: 10; *just wow*: 0;
Confession:	subconsciously, I constantly narrate my own life in my head. in third person. in a british accent. Insane? Probably
Reactions:	*hugs*: 0; *rock*: 7; *teehee*: 8; *understand*: 0; *just wow*: 1
Confession:	I have a crush on my boss! *blush* eeek *back to work*
Reactions:	*hugs*: 1; *rock*: 0; *teehee*: 4; *understand*: 1; *just wow*: 0
Confession:	I bought a case of beer, now I'm watching a South Park marathon while getting drunk :P
Reactions:	*hugs*: 2; *rock*: 3; *teehee*: 2, *understand*: 3, *just wow*: 0

Table: Sample Experience Project confessions with associated reaction data.

Experience Project confessions: blended, continuous sentiment reactions

	Texts	Words	Vocab	Mean words/text
Confessions	194,372	21,518,718	143,712	110.71
Comments	405,483	15,109,194	280,768	37.26

Table: The overall size of the corpus.

Reaction distributions

	Category	Reactions
exclamative, positive ←	sorry, hugs	91,222 (22%)
amused ←	you rock	80,798 (19%)
solidarity ←	teehee	59,597 (14%)
sympathy ←	I understand	125,026 (30%)
exclamative, negative (shocked) ←	wow, just wow	60,952 (15%)
	Total	417,595

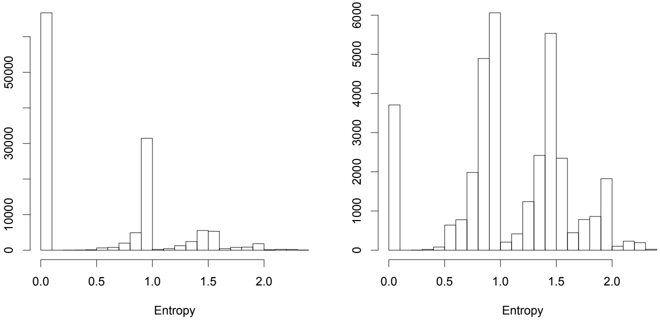
(a) All reactions.

	Texts
$\geqslant 1$	140,467
$\geqslant 2$	92,880
$\geqslant 3$	60,880
$\geqslant 4$	39,342
$\geqslant 5$	25,434

(b) Per text.

Table: In general, reader reactions are sympathetic and supportive.

Reaction distributions



(a) The full corpus.

(b) $\geqslant 4$ reactions.

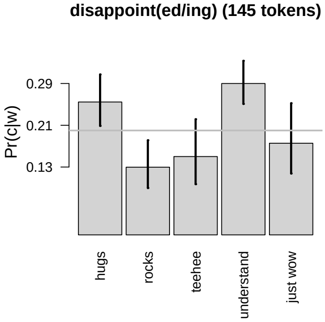
Figure: The entropy of the reaction distributions.

Counting and visualizing: Experience Project

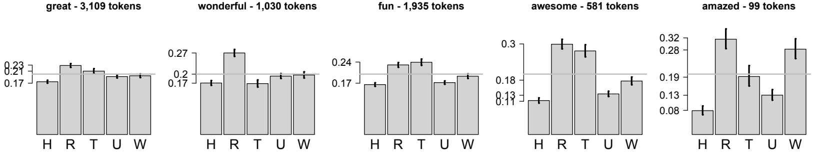
A Cat.	B Count	C Total	D $\Pr_{EP}(w\|c)$	E $\Pr_{EP}(c\|w)$
hugs	108	2,153,134	0.00005	0.25
rock	34	1,330,084	0.00002	0.13
teehee	25	845,397	0.00003	0.15
understand	197	3,447,377	0.00006	0.29
just wow	29	838,059	0.00004	0.18

$$\Pr_{EP}(w|c) \stackrel{def}{=} \mathrm{Count}(w, r)/\mathrm{Total}(r)$$

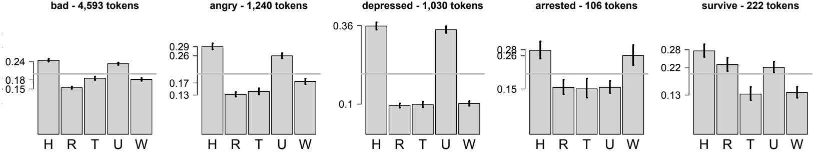
$$\Pr_{EP}(c|w) \stackrel{def}{=} \frac{\Pr_{EP}(w|c)}{\sum_{x \in \mathrm{Categories}} \Pr_{EP}(w|x)}$$



Word-level sentiment examples



(a) Words eliciting predominantly 'You rock' reactions. The data reveal other dimensions as well, including mixes of light-heartedness, negative exclamativity.



(b) Words eliciting sympathetic ('sorry, hugs', 'I understand') reactions. Other categories rise to prominence as well, depending on the lexical semantics and pragmatics of the word.

Figure: Word–category associations in the EP data.

A model for sentiment distributions

Definition (MaxEnt with distributional labels)

$$P(class|text, \lambda) = \frac{\exp\left(\sum_i \lambda_i f_i(class, text)\right)}{\sum_{class'} \exp\left(\sum_i \lambda_i f_i(class', text)\right)}$$

Minimize the KL divergence of the predicted distribution from the empirical one:

$$\sum_{class,text} empiricalProb(class|text) \log\left(\frac{empiricalProb(class|text)}{P(class|text, \lambda)}\right)$$

Gradient:

$$\sum_{text} empiricalProb(class|text) - P(class|text, \lambda)$$

Some results

Features	⩾ 5 reactions KL	⩾ 5 reactions Max Acc.	⩾ 1 reaction KL	⩾ 1 reaction Max Acc.
Uniform Reactions	0.861	20.2	1.275	20.4
Mean Training Reactions	0.763	43.0	1.133	46.7
Bag of Words (All unigrams)	0.637	56.0	1.000	53.4
Bag of Words (Top 5000 unigrams)	0.640	54.9	0.992	54.3
LSA	0.667	51.8	1.032	52.2
Our Method Laplacian Prior	0.621	55.7	0.991	54.7
Our Method Gaussian Prior	0.620	55.2	0.991	54.6

Table: Results from Maas et al. 2011. The first two are simple baselines. The 'Bag of words' models are MaxEnt/softmax. LSA and 'Our method' uses word vectors for predictions, by training on the average score in the vector. 'Our method' is distinguished primarily by combining an unsupervised VSM with a supervised component using star-ratings.

Topic-relative sentiment

- Sentiment is often topic relative
  (“We loved the food but hated the waiter.”)

- Sentiment vocabulary is topic dependent
  (*tasty*, *beautiful*, *melodious*, *plush*, . . .)

- Sentiment feature values can vary dramatically by topic
  (“The movie {*Scream*/*Love Story*} was totally gross!”)

Attribute-relative sentiment (Liu et al. 2005)

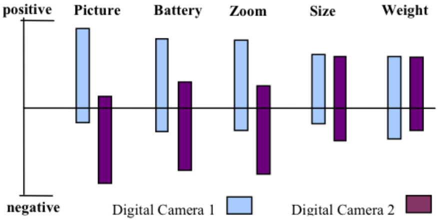


Figure 1: Visual comparison of consumer opinions on two products.

Associated datasets:
`http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`

OpenTable: attribute-level ratings

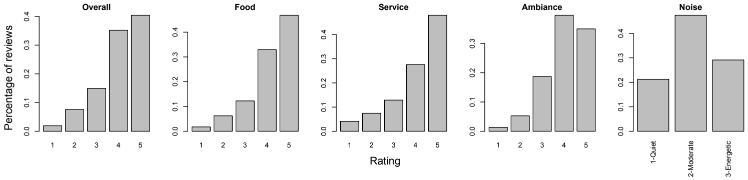
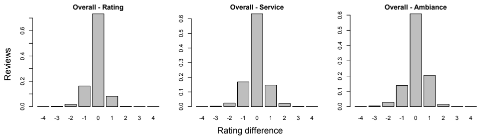


Figure: OpenTable rating distributions. Positive reviews dominate in all categories. ‘Noise’ is fundamentally different, since it doesn’t have a standard preference ordering.



(a) Comparisons with ‘Overall’. In each panel, the overall rating value is subtracted from the other rating value. Thus, a value of 0 indicates agreement between the two ratings for the review in question.

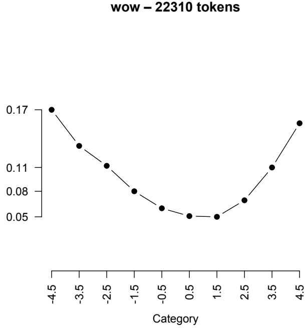
Overall, Food	0.82
Overall, Service	0.77
Overall, Ambiance	0.70
Food, Service	0.57
Food, Ambiance	0.56
Ambiance, Service	0.54

(b) Correlations.

Figure: OpenTable rating category comparisons. ‘Overall’ and ‘Food’ are highly correlated.

Counting and visualizing: IMDB

A Rating	B Count	C Total	D $\Pr(w\|r)$	E $\Pr(r\|w)$
−4.5	2983	28,962,201	0.00010	0.17
−3.5	1056	13,436,851	0.00008	0.13
−2.5	1041	15,987,151	0.00007	0.11
−1.5	819	17,095,212	0.00005	0.08
−0.5	848	23,293,790	0.00004	0.06
+0.5	975	31,317,918	0.00003	0.05
+1.5	1407	45,913,948	0.00003	0.05
+2.5	2326	55,634,817	0.00004	0.07
+3.5	2940	45,941,763	0.00006	0.11
+4.5	7915	84,294,625	0.00009	0.16



$$\Pr(w|r) \stackrel{def}{=} \text{Count}(w,r)/\text{Total}(r)$$

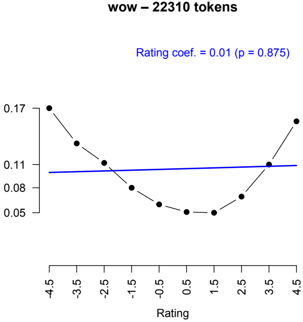
$$\Pr(r|w) \stackrel{def}{=} \frac{\Pr(w|r)}{\sum_{x\in\text{Rating}} \Pr(w|x)}$$

Counting and visualizing: IMDB

A Rating	B Count	C Total	D $\Pr(w \mid r)$	E $\Pr(r \mid w)$
−4.5	2983	28,962,201	0.00010	0.17
−3.5	1056	13,436,851	0.00008	0.13
−2.5	1041	15,987,151	0.00007	0.11
−1.5	819	17,095,212	0.00005	0.08
−0.5	848	23,293,790	0.00004	0.06
+0.5	975	31,317,918	0.00003	0.05
+1.5	1407	45,913,948	0.00003	0.05
+2.5	2326	55,634,817	0.00004	0.07
+3.5	2940	45,941,763	0.00006	0.11
+4.5	7915	84,294,625	0.00009	0.16

$$\Pr(w \mid r) \stackrel{def}{=} \text{Count}(w, r)/\text{Total}(r)$$

$$\boxed{\Pr(r \mid w) \stackrel{def}{=} \frac{\Pr(w \mid r)}{\sum_{x \in \text{Rating}} \Pr(w \mid x)}}$$



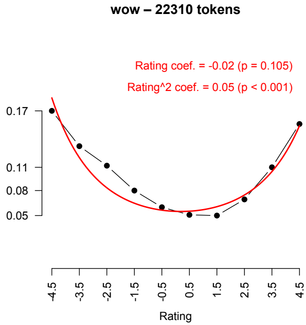
$$\Pr(wow) = \text{logit}^{-1}\left(\texttt{intercept} + \texttt{rating}\right)$$

Counting and visualizing: IMDB

A Rating	B Count	C Total	D $\Pr(w \mid r)$	E $\Pr(r \mid w)$
−4.5	2983	28,962,201	0.00010	0.17
−3.5	1056	13,436,851	0.00008	0.13
−2.5	1041	15,987,151	0.00007	0.11
−1.5	819	17,095,212	0.00005	0.08
−0.5	848	23,293,790	0.00004	0.06
+0.5	975	31,317,918	0.00003	0.05
+1.5	1407	45,913,948	0.00003	0.05
+2.5	2326	55,634,817	0.00004	0.07
+3.5	2940	45,941,763	0.00006	0.11
+4.5	7915	84,294,625	0.00009	0.16

$$\Pr(w \mid r) \stackrel{def}{=} \text{Count}(w, r)/\text{Total}(r)$$

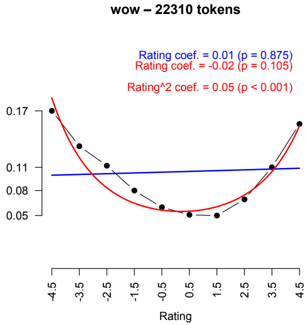
$$\Pr(r \mid w) \stackrel{def}{=} \frac{\Pr(w \mid r)}{\sum_{x \in \text{Rating}} \Pr(w \mid x)}$$



$$\Pr(wow) = \text{logit}^{-1}\begin{pmatrix} \texttt{intercept} & + \\ \texttt{rating} & + \\ \texttt{rating}^2 & \end{pmatrix}$$

Counting and visualizing: IMDB

A Rating	B Count	C Total	D $\Pr(w \mid r)$	E $\Pr(r \mid w)$
−4.5	2983	28,962,201	0.00010	0.17
−3.5	1056	13,436,851	0.00008	0.13
−2.5	1041	15,987,151	0.00007	0.11
−1.5	819	17,095,212	0.00005	0.08
−0.5	848	23,293,790	0.00004	0.06
+0.5	975	31,317,918	0.00003	0.05
+1.5	1407	45,913,948	0.00003	0.05
+2.5	2326	55,634,817	0.00004	0.07
+3.5	2940	45,941,763	0.00006	0.11
+4.5	7915	84,294,625	0.00009	0.16

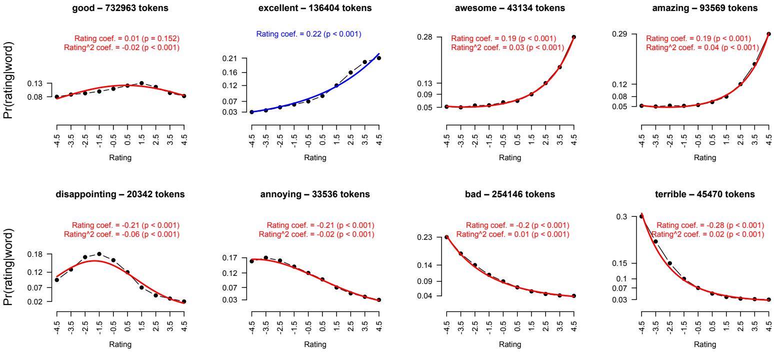


$$\Pr(w \mid r) \stackrel{def}{=} \text{Count}(w, r)/\text{Total}(r)$$

$$\Pr(r \mid w) \stackrel{def}{=} \frac{\Pr(w \mid r)}{\sum_{x \in \text{Rating}} \Pr(w \mid x)}$$

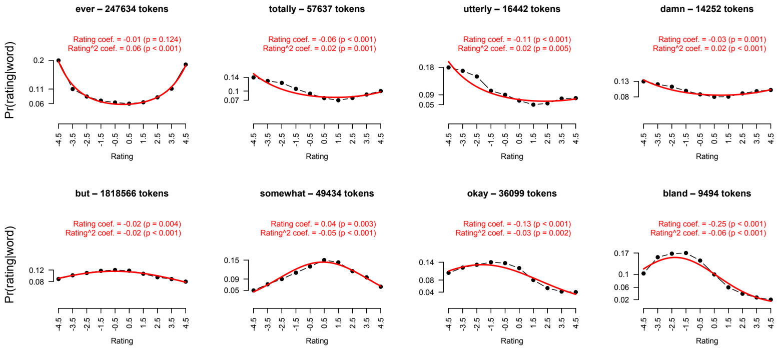
IMDB movie reviews: word-level distributional profiles

Positive and negative scalar terms



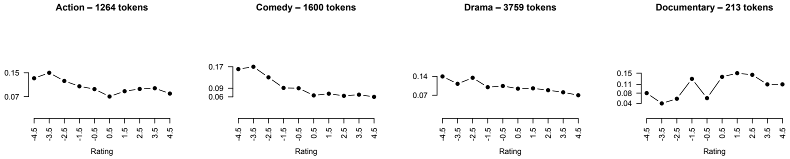
IMDB movie reviews: word-level distributional profiles

Emphasizing and attenuating terms

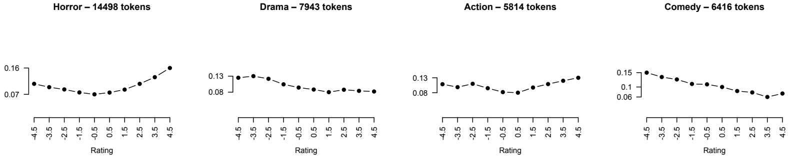


IMDB movie reviews: variation by genre

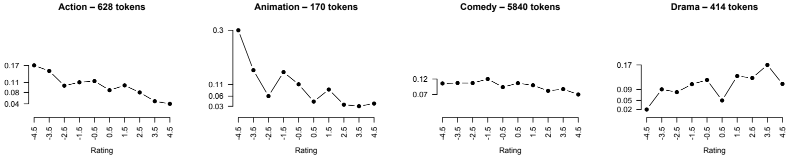
depressing



scary



Sandler



Other examples of topic/aspect relative sentiment

- user-level variation in both author + reader reactions
  - age
  - formality
  - in-group / public
- dialect features + regional differences in emphasis

Sentiment as social: Convote (Thomas et al. 2006)

- Using text and social ties to predict congressional voting.
- Adapts the hierarchical model of Pang and Lee (2004), where subjectivity scores are used to focus a subsequent polarity classifier.
- A pioneering attempt to treat sentiment (here, support/opposition) as a social phenomenon.

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 ?
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 .

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)

Basic classification with same-speech links

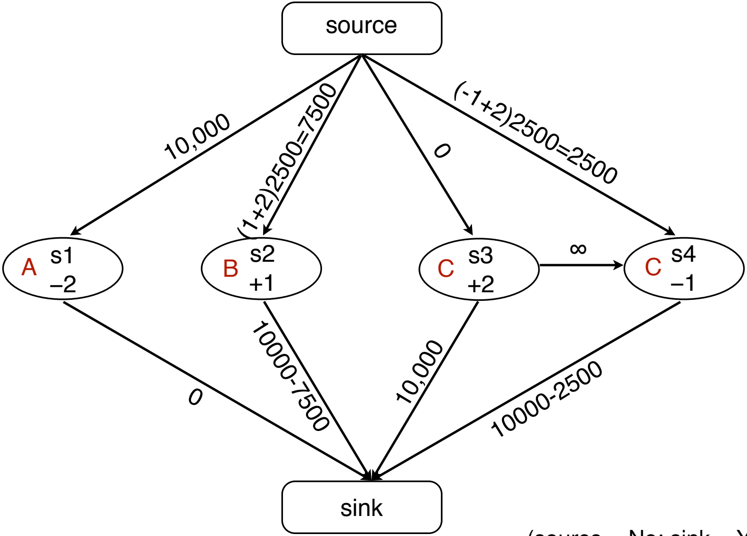
1. SVM classifier with unigram-presence features predicting, for each speech-segment, how the speaker voted (Y or N).
2. For each document s belonging to speech S, the SVM score for s is divided by the standard deviation for all $s' \in S$.
3. Debate-graph construction with minimal cuts:

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

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

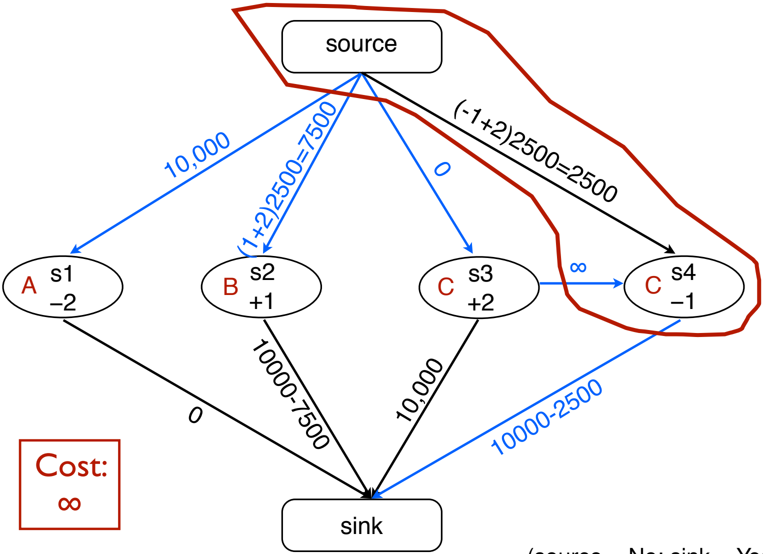
$$\text{else} \;\Rightarrow\; \left[\begin{array}{rcl} \text{source} & \overset{x=(\text{score}(s)+2)2500}{\rightarrow} & s \\ s & \overset{10{,}000-x}{\rightarrow} & \text{sink} \end{array}\right.$$

Graph construction and minimal cuts



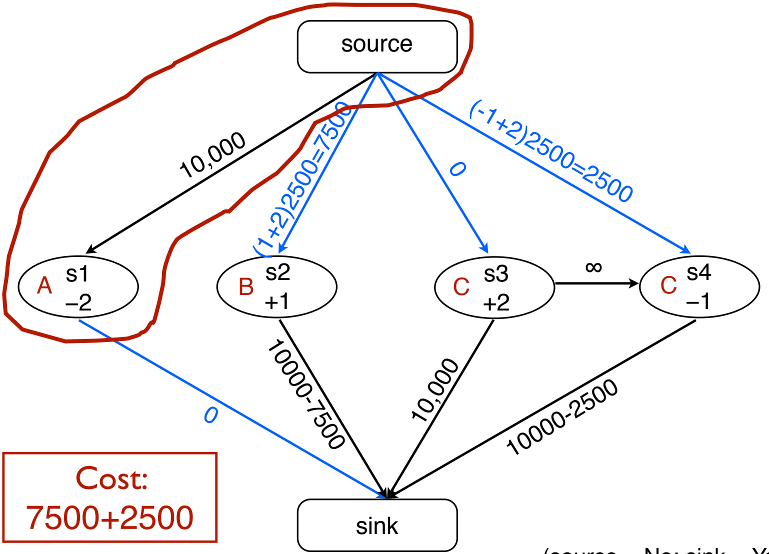
(source = No; sink = Yes)

Graph construction and minimal cuts



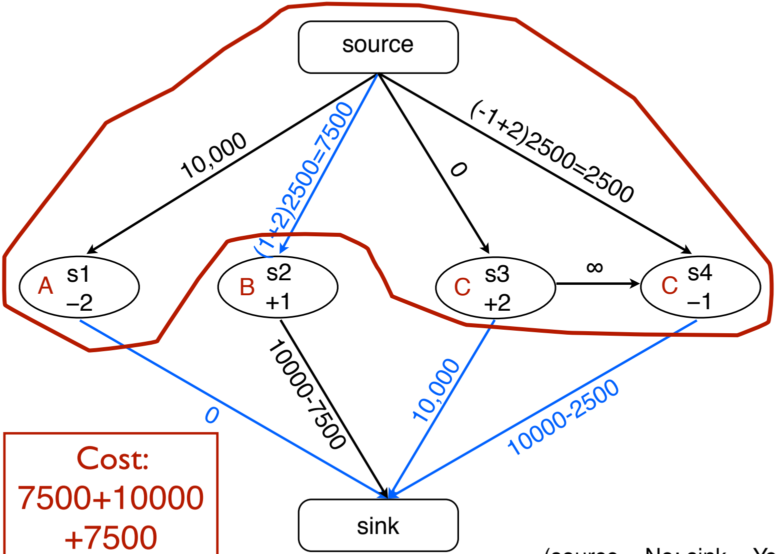
(source = No; sink = Yes)

Graph construction and minimal cuts



(source = No; sink = Yes)

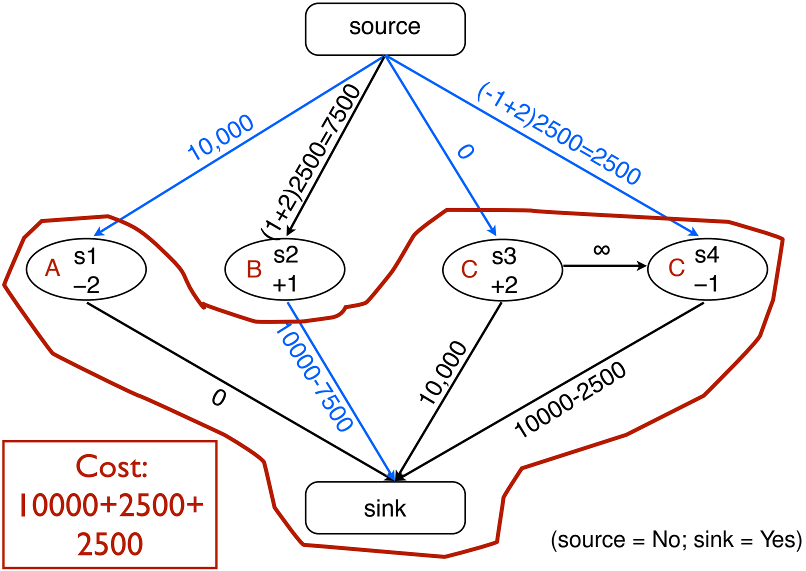
Graph construction and minimal cuts



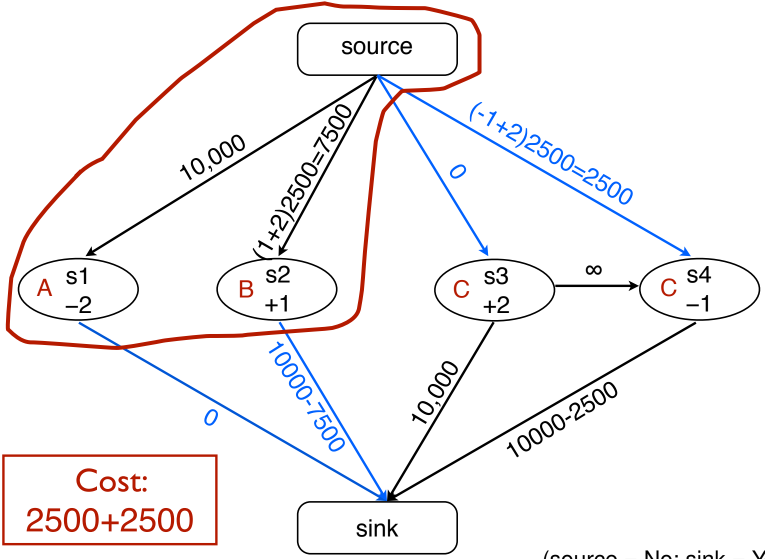
Cost:
7500+10000
+7500

(source = No; sink = Yes)

Graph construction and minimal cuts



Graph construction and minimal cuts



(source = No; sink = Yes)

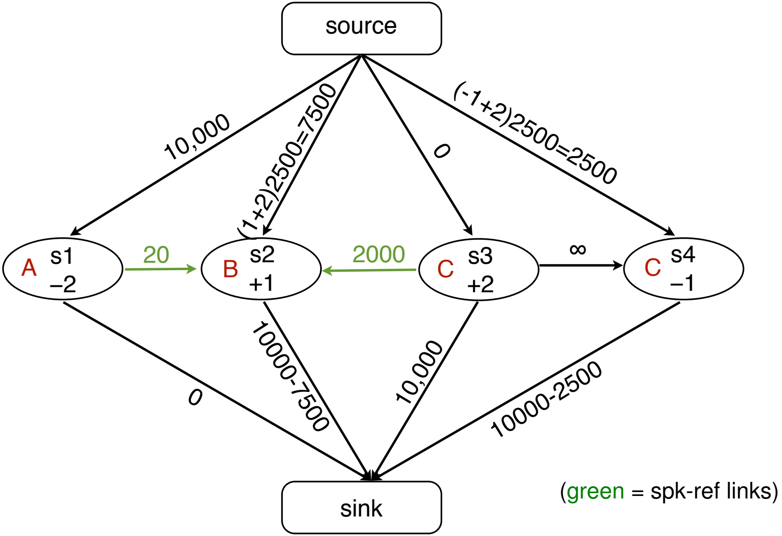
Speaker references

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



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.

Sentiment as social: Twitter users (Tan et al. 2011)

Goal

Given a topic q and a user v, predict whether v is positive or negative wrt topic q

Guiding idea (builds on Thomas et al. 2006)

Users in the same social network will tend to share sentiment, so bringing in these social ties will improve sentiment predictions.

Data

Topically-clustered tweets, with social network determined by the following relation or the connection user *a* makes with user *b* by tweeting "@*b* . . ."

Dataset (Tan et al. 2011)

Table 1: Statistics for our main datasets.

Topic	# users	#t-follow edges		#@ edges		# on-topic tweets
		dir.	mutual	dir.	mutual	
Obama	889	7,838	2,949	2,358	302	128,373
Sarah Palin	310	1,003	264	449	60	21,571
Glenn Beck	313	486	159	148	17	12,842
Lakers	640	2,297	353	1,167	127	35.250
Fox News	231	130	32	37	5	8,479

- Set of topics chosen by hand, explicitly favoring polarizing topics so that the classes could be balanced.
- For the following relations, ‘dir’ means that the following or @-link goes in at least one direction, whereas mutual means that it goes in both directions.
- User-level polarity was determined by inspecting biographies and in some cases their tweets and using that information to assign a label by hand.
- The dataset is only partially labeled.

Connected user tend to share topic-relative sentiment (Tan et al. 2011)

In keeping with the guiding intuition,

- connected users tend to share the same sentiment (left); and
- users who share sentiment are more likely to be connected.

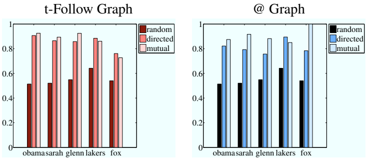


Figure 1: Shared sentiment conditioned on type of connection. Y-axis: probability of two users v_i and v_j having the same sentiment label, conditioned on relationship type. The left plot is for the t-follow graph, while the right one is for the @ graph. "random": pairs formed by randomly choosing users. "directed": at least one user in the pair links to the other. "mutual": both users in the pair link to each other. Note that the very last bar (a value of 1 for "Fox News", mutual @-graph) is based on only 5 edges (datapoints).

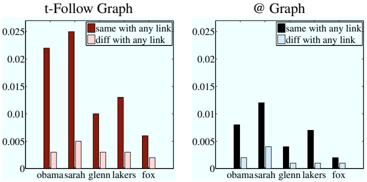
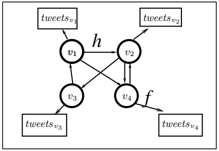


Figure 2: Connectedness conditioned on labels. Y-axis: probability that two users are connected, conditioned on whether or not the users have the same sentiment.

Graphical structure and model (Tan et al. 2011)

Heterogeneous graph



k = user sentiment label $\in \{0,1\}$
l = tweets sentiment label $\in \{0,1\}$

v_i sentiment label $\in \{0,1\}$

unknown label $\in \{0,1\}$ for v_i tweets

$$\log P(\mathbf{Y}) = \Big(\sum_{v_i \in V} \Big[\sum_{t \in tweets_{v_i}, k, \ell} \mu_{k,\ell} f_{k,\ell}(y_{v_i}, \hat{y}_t) + \sum_{v_j \in Neighbors_{v_i}, k, \ell} \lambda_{k,\ell} h_{k,\ell}(y_{v_i}, y_{v_j}) \Big] \Big) - \log Z,$$

vector of user labels for the topic in question

weights for impact of the feature functions; set by counting labels or SampleRank

= 1 in experiments = 0.125 in experiments

$$f_{k,\ell}(y_{v_i}, \hat{y}_t) = \begin{cases} \frac{w_{\text{labeled}}}{|tweets_{v_i}|} & y_{v_i} = k, \hat{y}_t = \ell, v_i \text{ labeled} \\ \frac{w_{\text{unlabeled}}}{|tweets_{v_i}|} & y_{v_i} = k, \hat{y}_t = \ell, v_i \text{ unlabeled} \\ 0 & \text{otherwise} \end{cases}$$

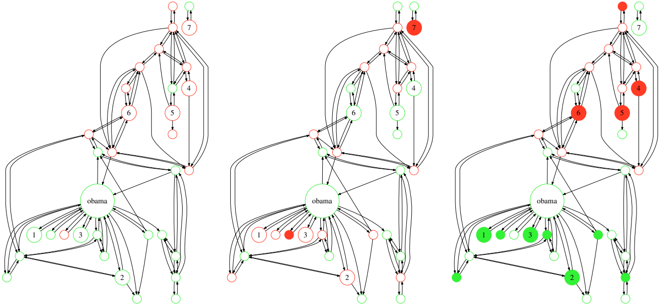
User–tweet factor

= 0.6 in experiments

$$h_{k,\ell}(y_{v_i}, y_{v_j}) = \begin{cases} \frac{w_{\text{relation}}}{|Neighbors_{v_i}|} & y_{v_i} = k, y_{v_j} = \ell \\ 0 & \text{otherwise} \end{cases}$$

User–user factor

Case study highlighting the value of social information



(a) Ground Truth (b) Text-Only Approach (c) Our algorithm

Figure 4: Case study: Portion of the t-follow graph for the topic "Obama", where derived labels on users are indicated by green (positive) and red (negative), respectively. Each node is a user, and the center one is "BarackObama". The numbers in the nodes are indices into the table below. (a): Ground truth (human annotation). (b) SVM Vote (baseline). (c) HGM-Learning in the directed t-follow graph. Filled nodes indicate cases where the indicated algorithm was right and the other algorithm was wrong; for instance, only our algorithm was correct on node 4.

By-topic results

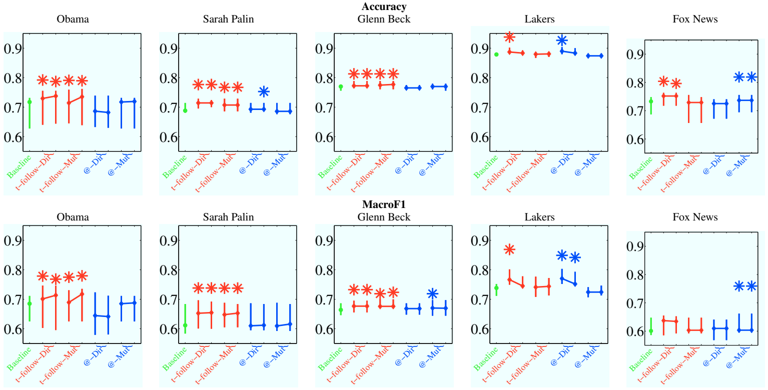


Figure 6: Performance Analysis in Different Topics. The x-axes are the same as in Figure 5. Bars summarize performance results for our "10-run" experiments: the bottom and top of a bar indicate the 25th and 75th percentiles, respectively. Dots indicate median results; in pairs connected by lines, the left is "NoLearning", while the right is "Learning". Green: SVM vote, our baseline. Red: network-based approaches applied to the t-follow graphs. Blue: results for the @ graphs. Stars (∗) indicate performance that is significantly better than the baseline, according to the paired t-test.

Possible extensions of the Convote/Twitter-graph approach

- newspapers as users, soft constraints on shared sentiment, network struc given by opinion and corp. structure
- Scientific network + shared views on controversial topics
- Convote 2012 — stronger social feats due to increased polarization?
- Sentiment in hyperlink structures

Sentiment as social: Experience Project

Confession:	I really hate being shy ... I just want to be able to talk to someone about anything and everything and be myself... That's all I've ever wanted.
Reactions:	*hugs*: 1; *rock*: 1; *teehee*: 2; *understand*: 10; *just wow*: 0;
Confession:	I bought a case of beer, now I'm watching a South Park marathon while getting drunk :P
Reactions:	*hugs*: 2; *rock*: 3; *teehee*: 2, *understand*: 3, *just wow*: 0

Table: Sample Experience Project confessions with associated reaction data, author demographics, and text groups.

Sentiment as social: Experience Project

Confession:	I really hate being shy ... I just want to be able to talk to someone about anything and everything and be myself... That's all I've ever wanted.
Reactions:	*hugs*: 1; *rock*: 1; *teehee*: 2; *understand*: 10; *just wow*: 0;
Author age	21
Author gender	female
Text group	friends
Confession:	I bought a case of beer, now I'm watching a South Park marathon while getting drunk :P
Reactions:	*hugs*: 2; *rock*: 3; *teehee*: 2, *understand*: 3, *just wow*: 0
Author age	25
Author gender	male
Text group	health

Table: Sample Experience Project confessions with associated reaction data, author demographics, and text groups.

Contextual variables

Age	Texts
teens	5,495
20s	26,564
30s	15,317
40s	7,413
50s	3,600
$\geqslant 60$	1130
unknown	80,948
Total	140,467

(a) Author ages.

Gender	Texts
female	34,921
male	15,333
unknown	90,213
Total	140,467

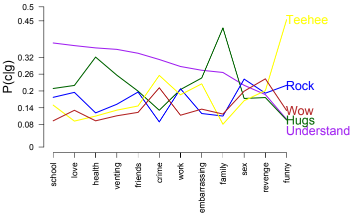
(b) Author genders.

Group	Texts
crime	312
embarrassing	5,349
family	5,114
friends	13,719
funny	3,692
health	6,467
love	36,242
revenge	1,406
school	1,698
sex	45,538
venting	19,090
work	1,840
Total	140,467

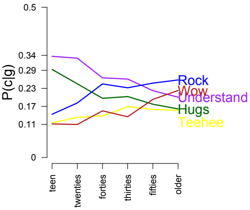
(c) Text groups.

Table: Contextual metadata. The EP's demographics seem to be skewed towards young women writing about issues concerning their interpersonal relationships.

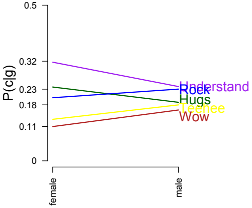
The influences of context



(a) Text groups.



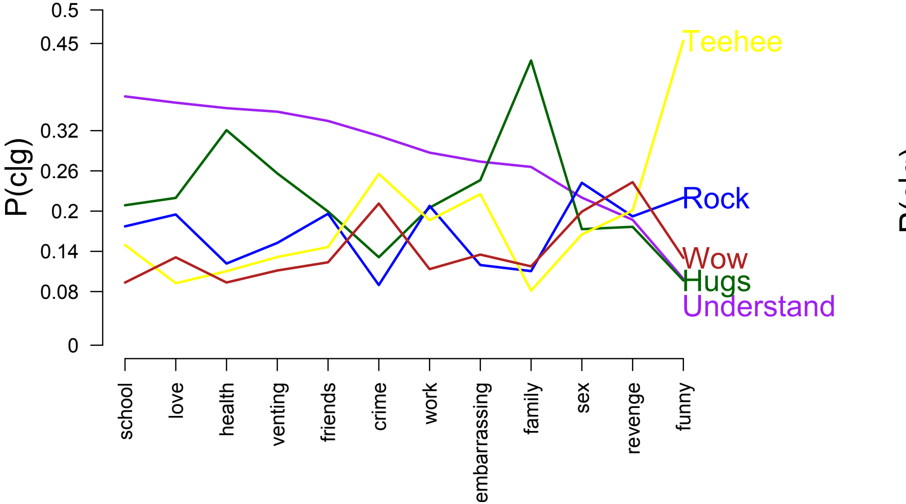
(b) Age.



(c) Gender.

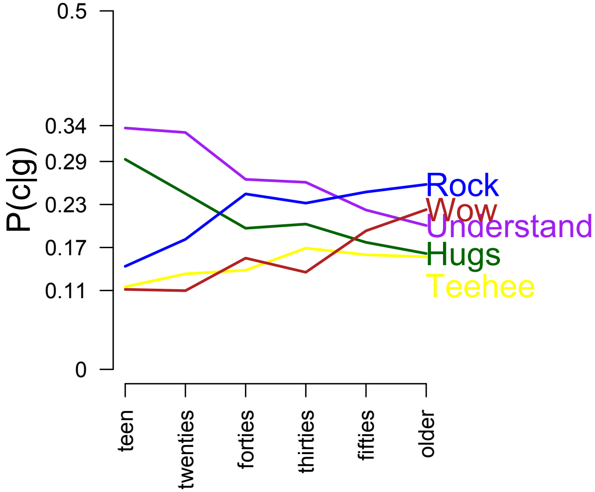
Figure: Text groups show the most variability. Age and gender are more stable by comparison, though the relationships remain interesting.

The influences of context



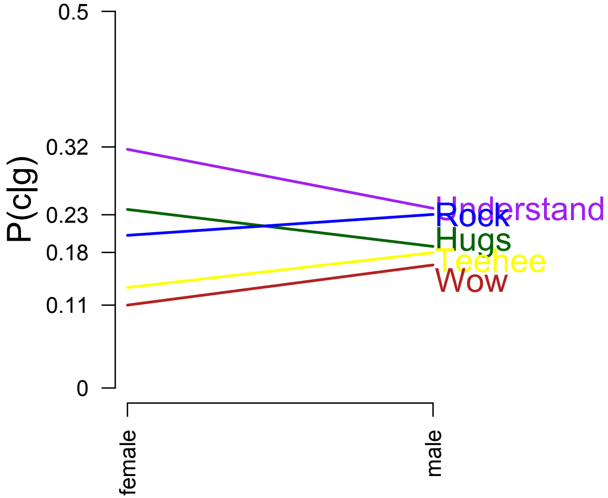
(a) Text groups.

The influences of context



(b) Age.

The influences of context



(c) Gender.

The influences of text groups

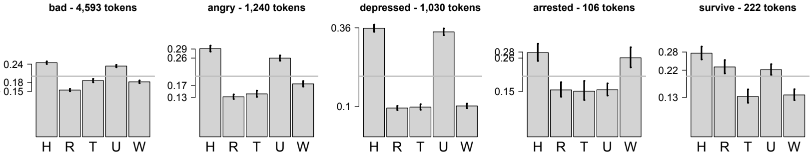


Figure: Words eliciting predominantly ‘You rock’ reactions. The data reveal other dimensions as well, including mixes of light-heartedness, negative exclamativity.

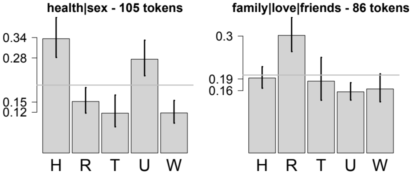


Figure: The bimodal distribution of *survive* seems to derive from an underlying distinction in text group.

The influences of age

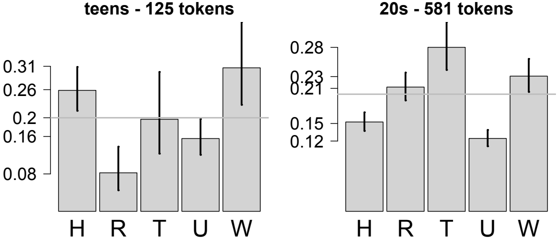


Figure: Age is a source of variation in responses to *drunk*.

Modeling ideas

- Demographic and text-group features can be treated on par with linguistic features.
- They could also be brought in as hierarchical effects in a multi-level generalized linear model (Gelman and Hill 2007; Baayen 2008).
- In ongoing work with Andrew Maas, Peter Pham, and Andrew Ng, we have been using Conditional Random Fields (Lafferty et al. 2001; Sutton and McCallum 2010) to define context-relative feature functions to directly model the distribution $P(\mathit{class}|\mathit{text}, \mathit{context}, \lambda)$.

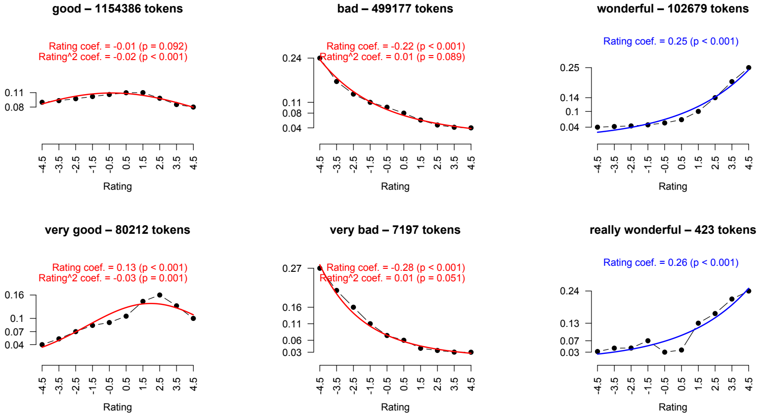
Sentiment and morphosyntax

I’ve so far concentrated on general features of the context of use. Sentiment is also profoundly influenced by the immediate linguistic context.

1. That was fun :)
2. That was miserable :(
3. I stubbed my damn toe
4. What’s with these friggin QR codes?
5. It was wonderful.
6. He knows it is wonderful.
7. It was not wonderful.
8. No one found it to be wonderful.
9. They said it would be wonderful, but they were wrong: it was awful!
10. This “wonderful” movie turned out to be boring.

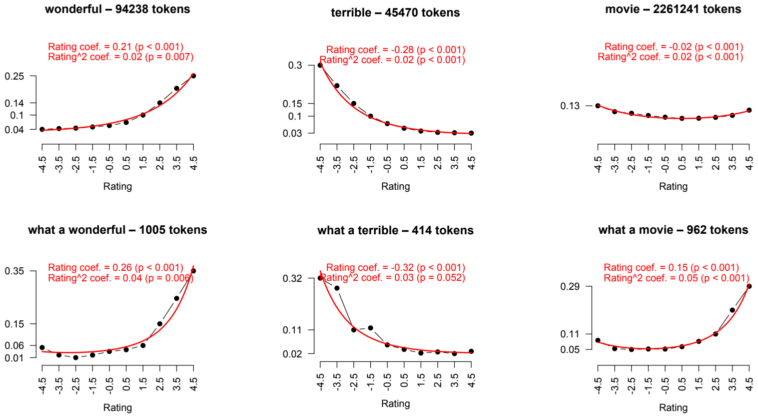
Degree modification

The intensifers *really* and *very* enhance sentiment:



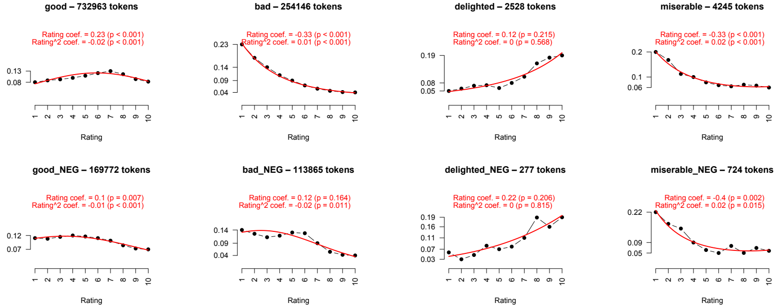
Exclamatives

Exclamatives (e.g., *what a view!*) both create and enhance sentiment):



Negation

Negating mid-scalar terms leads to polarity reversal. Negating high-scalar terms (positive or negative) leads to mere attenuation.

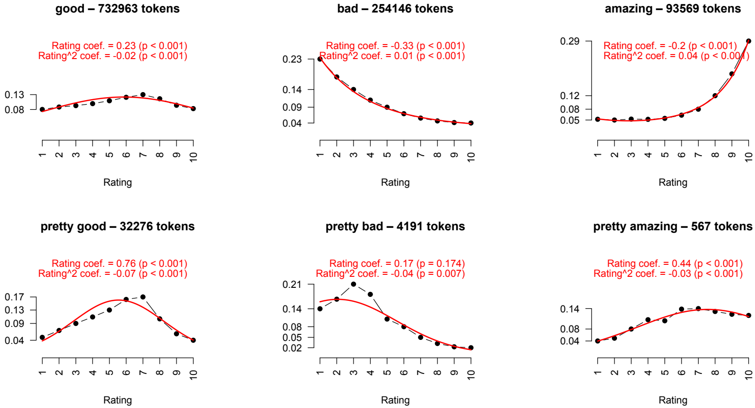


not good $\approx$ *bad*
not bad $\approx$ *good*

not delighted $\not\approx$ *miserable*
not miserable $\not\approx$ *delighted*

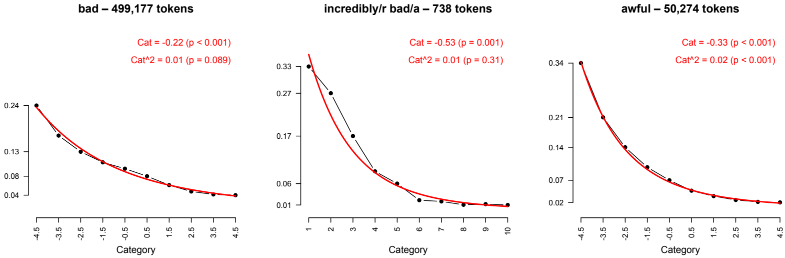
Attenuators

Adverbials like *pretty* weaken/attenuate sentiment:



Intensification: the weak overtake the strong

Low-scalar modifiers are likely to be intensified, which can confuse models into thinking that they are stronger than their high-scalar counterparts:



Attitude predictions and thwarted expectations

i had been looking forward to this film since i heard about it early last year , when matthew perry had just signed on . i'm big fan of perry's subtle sense of humor , and in addition , i think chris farley's on-edge , extreme acting was a riot . so naturally , when the trailer for " almost heroes " hit theaters , i almost jumped up and down . a soda in hand , the lights dimming , i was ready to be blown away by farley's final starring role and what was supposed to be matthew perry's big breakthrough . i was ready to be just amazed ; for this to be among farley's best , in spite of david spade's absence . i was ready to be laughing my head off the minute the credits ran . sadly , none of this came to pass . the humor is spotty at best , with good moments and laughable one-liners few and far between . perry and farley have no chemistry ; the role that perry was cast in seems obviously written for spade , for it's his type of humor , and not at all what perry is associated with . and the movie tries to be smart , a subject best left alone when it's a farley flick . the movie is a major dissapointment , with only a few scenes worth a first look , let alone a second . perry delivers not one humorous line the whole movie , and not surprisingly ; the only reason the movie made the top ten grossing list opening week was because it was advertised with farley . and farley's classic humor is widespread , too . almost heroes almost works , but misses the wagon-train by quite a longshot . guys , let's leave the exploring to lewis and clark , huh ? stick to " tommy boy " , and we'll all be " friends " .

Table: An example of thwarted expectations. This is a negative review. Inquirer positive terms are in blue, and Inquirer negative terms are red. There are 20 positive terms and six negative ones, for a Pos:Neg ratio of 3.33.

Attitude predictions and thwarted expectations

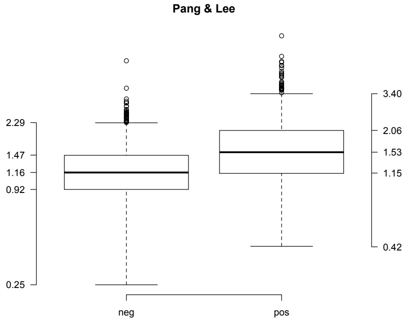


Figure: Inquirer Pos:Neg ratios obtained by counting the terms in the review that are classified as Positiv or Negativ in the Harvard Inquirer (Stone et al. 1966).

Proposed feature: the Pos:Neg ratio if that ratio is below 1 (lower quartile for the whole Pang & Lee data set) or above 1.76 (upper quartile), else 1.31 (the median). The goal is to single out 'imbalanced' reviews as potentially untrustworthy. (For a similar idea, see Pang et al. 2002.)

Looking ahead to Richard Socher's lecture

1. Sentiment-relevant semantic influences can come from
   - negation
   - adverbs and other modifiers
   - attitude predications, including modals and hedges
   - and combinations of all of the above.
2. This is just to say that all aspects of semantic composition are relevant.
3. Thus, rather than treating it as series of isolated and separate problems, we should approach it as part of a theory of semantic composition.
4. This is precisely what Richard Socher is seeking to do (Socher et al. 2011). Lots more about that on Tuesday!

Conclusion

Central insights

- Sentiment is blended and continuous.
- Sentiment is social and context-dependent.
- Sentiment is as hard as semantic composition.

Opportunities

- Increasingly, we have the rights dataset and models to honor the above insights.
- Careful, flexible sentiment analysis systems are in high demand.
- Extensions:
  - How does sentiment flow in a social network?
  - How does it affect the flow of other information?
  - What does sentiment reveal about social ties, media bias, polarization, . . .
  - . . .

References I

Alm, Cecilia Ovesdotter; Dan Roth; and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge University Press.

Bruce, Rebecca F. and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering* 5(2).

Cabral, Luís and Ali Hortaçsu. 2006. The dynamics of seller reputation: Theory and evidence from eBay. Working paper, downloaded version revised in March. URL `http://pages.stern.nyu.edu/~lcabral/workingpapers/CabralHortacsu_Mar06.pdf`.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Emotion,* 6(3/4):169–200.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Goldberg, Andrew B. and Jerry Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised leaarning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*.

Hatzivassiloglou, Vasileios and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Lafferty, John; Andrew McCallum; and Fernando Pereira. 2001. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, 282–289.

Liu, Bing; Minqing Hu; and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International World Wide Web Conference*, 342–351. ACM.

Liu, Hugo; Henry Lieberman; and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of Intelligent User Interfaces (IUI)*, 125–132.

References II

Maas, Andrew; Andrew Ng; and Christopher Potts. 2011. Multi-dimensional sentiment analysis with learned representations. Ms., Stanford University.

Neviarouskaya, Alena; Helmut Prendinger; and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 806–814. Beijing, China: COLING 2010 Organizing Committee.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 271–278. Barcelona, Spain. doi:\bibinfo{doi}{10.3115/1218955.1218990}. URL http://www.aclweb.org/anthology/P04-1035.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 115–124. Ann Arbor, Michigan: Association for Computational Linguistics.

Pang, Bo; Lillian Lee; and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86. Philadelphia: Association for Computational Linguistics.

Potts, Christopher. 2011. On the negativity of negation. In Nan Li and David Lutz, eds., *Proceedings of Semantics and Linguistic Theory 20*, 636–659. Ithaca, NY: CLC Publications.

Potts, Christopher and Florian Schwarz. 2010. Affective 'this'. *Linguistic Issues in Language Technology* 3(5):1–30.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Riloff, Ellen; Janyce Wiebe; and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of AAAI*, 1106–1111.

Russell, James A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161–1178.

References III

Snyder, Benjamin and Regina Barzilay. 2007. Multiple aspect ranking using the Good Grief algorithm. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, 300–307.

Socher, Richard; Jeffrey Pennington; Eric H. Huang; Andrew Y. Ng; and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 151–161. Edinburgh, Scotland, UK.: Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1014`.

Stone, Philip J; Dexter C Dunphry; Marshall S Smith; and Daniel M Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Sutton, Charles and Andrew McCallum. 2010. An introduction to conditional random fields. *Foundations and Trends in Machine Learning* .

Tan, Chenhao; Lillian Lee; Jie Tang; Long Jiang; Ming Zhou; and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1397–1405. San Diego, CA: ACM Digital Library.

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.

Turney, Peter D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 417–424. Philadelphia, PA: Association for Computational Linguistics. doi:\bibinfo{doi}{10.3115/1073083.1073153}. URL `http://www.aclweb.org/anthology/P02-1053`.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21:315–346. doi:\bibinfo{doi}{http://doi.acm.org/10.1145/944012.944013}. URL `http://doi.acm.org/10.1145/944012.944013`.

References IV

Wiebe, Janyce; Theresa Wilson; and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)* 39(2/3):164–210.

Wiebe, Janyce M.; Rebecca F. Bruce; and Thomas P. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the Association for Computational Linguistics (ACL)*, 246–253.

Wilson, Theresa; Janyce Wiebe; and Rebecca Hwa. 2006. Just how mad are you? Finding strong and weak opinion clauses. *Computational Intelligence* 2(22):73–99.