Introduction to Sentiment Analysis

Overview

- What is sentiment analysis (SA)?
- Why is it worth doing?
- What are the challenges?
- (Very broadly) how is it done?

What is Sentiment?

- Sentiment = feelings
  - Attitudes
  - Emotions
  - Opinions
- Subjective impressions, not facts

What is Sentiment?

- Generally, a binary opposition in opinions is assumed
- For/against, like/dislike, good/bad, etc.
- Some sentiment analysis jargon:
  - “Semantic orientation”
  - “Polarity”

What is Sentiment Analysis?

- Using NLP, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit
- Sometimes refered to as *opinion mining,* although the emphasis in this case is on extraction

Questions SA might ask

- Is this product review positive or negative?
- Is this customer email satisfied or dis-satisfied?
- Based on a sample of tweets, how are people responding to this ad campaign/product release/news item?
- How have bloggers' attitudes about the president changed since the election?

Other related tasks

- Information extraction (discarding subjective information)
- Question answering (recognizing opinion-oriented questions)
- Summarization (accounting for multiple viewpoints)

Other related tasks

- “Flame” detection
- Identifying child-suitability of videos based on comments
- Bias identification in news sources
- Identifying (in)appropriate content for ad placement

Applications in Business Intelligence

- Question: “Why aren't consumers buying our laptop?”
- We know the concrete data: price, specs, competition, etc.
- We want to know subjective data: “the design is tacky,” “customer service was condescending”
- Misperceptions are also important, e.g. “updated drivers aren't available” (even though they are)

Applications in Business Intelligence

- It is very difficult to survey customers who *didn't* buy the company's laptop
- Instead, you could use SA to

  A) search the web for opinions and reviews of this and competing laptops. Blogs, Epinions, amazon, tweets, etc.

  B) create condensed versions or a digest of consensus points

Cross domain applications

- Insights and applications from SA have been useful in other areas
  - Politics/political science
  - Law/policy making
  - Sociology
  - Psychology

Political SA

- Numerous applications and possibilities
- Analyzing trends, identifying ideological bias, targeting advertising/messages, gauging reactions, etc.
- Evaluation of public/voters' opinions
- Views/discussions of policy
- More on this in lecture 3

SA and Sociology

- Idea propagation through groups is an important concept in sociology (cf. Rogers 1962, *Diffusion of Innovations*)
- Opinions and reactions to ideas are relevant to adoption of new ideas
- Analyzing sentiment reactions on blogs can give insight to this process
- E.g. Kale et al (2007), *Modeling trust and influence in the blogosphere using link polarity*

SA and Psychology

- Potential to augment psychological investigations/experiments with data extracted from NL text
- Dream sentiment analysis (Nadeau et al., 2006)

In general,

- Humans are subjective creatures and opinions are important. Being able to interact with people on that level has many advantages for information systems.

How SA is different

- Comparatively few categories (positive/negative, 3 stars, etc) compared to text categorization
- Crosses domains, topics, and users
- Categories not independent (opposing or regression-like)
- Characteristics of answers to opinion-based questions are different from fact-based questions, so opinion-based IE differs from trad IE

Challenges in SA

- People express opinions in complex ways
- In opinion texts, lexical content alone can be misleading
- Intra-textual and sub-sentential reversals, negation, topic change common
- Rhetorical devices/modes such as sarcasm, irony, implication, etc.

A letter to a hardware store*

“Dear <hardware store>

Yesterday I had occasion to visit <your competitor>. The had an excellent selection, friendly and helpful salespeople, and the lowest prices in town.

You guys suck.

Sincerely,”

*an apocryphal example

What to classify

- There are many possibilities for what we might want to classify:
  - Users
  - Texts
  - Sentences (paragraphs, chunks of text?)
  - Predetermined descriptive phrases (<ADJ N>, <N N>, <ADV ADJ>, etc)
  - Words
  - Tweets/updates

Classifying words/short phrases

- The building blocks of sentiment expression
- Short phrases may be just as important (or moreso) as words:
  - "lowest prices"
  - "high quality"
- We need an approach to deal with these before moving on to other classification tasks

Polarity keywords

- There seems to be *some* relation between positive words and postive reviews
- Can we come up with a set of keywords by hand to identify polarity?

Pang et al. (2002)

- Two human subjects were asked to pick keywords that would be good indicators of sentiment polarity

	Proposed word list	Accu-racy	Ties
Human 1	Pos: dazzling, brilliant, phenomenal, excellent, fantastic Neg: suck, terrible, awful, unwatchable, hideous	58%	75%
Human 2	Pos: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting Neg: bad, cliched, sucks, boring, stupid, slow	64%	39%
Statistics-based	Pos: love, wonderful, best, great, superb, still, beautiful Neg: bad, worst, stupid, waste, boring, ?, !	69%	16%

Key-word methods

- Data-driven methods can be used to generate keyword lists that model better than human-generated keyword lists
- Unigram methods on similar data have reached 80% accuracy (Pang et al, 2002)
- Not bad, but lower than you'd usually see in topic-based binary text classification

Smileys

- A common approach for working with tweets and short text updates
- Very little text to work with
- Sentiment most succinctly represented with emoticons/smileys

Some actual examples of sentiment text

Amazon (5 star)

“The characters are so real and handled so carefully, that being trapped inside the Overlook is no longer just a freaky experience. You run along with them, filled with dread, from all the horrible personifications of evil inside the hotel's awful walls. There were several times where I actually dropped the book and was too scared to pick it back up. Intellectually, you know it's not real. It's just a bunch of letters and words grouped together on pages. Still, whenever I go into the bathroom late at night, I have to pull back the shower curtain just to make sure.”

Amazon.com (1 star)

“The original Star Wars trilogy was a defining part of my childhood. Born as I was in 1971, I was just the right age to fall headlong into this amazing new world Lucas created. I was one of those kids that showed up early at toy stores [...] anxiously awaiting each subsequent installment of the series.

I'm so glad that by my late 20s, the old thrill had faded, or else I would have been EXTREMELY upset over *Episode I: The Phantom Menace*... perhaps the biggest let-down in film history.”

Pitchfork.com (0.0 out of 10)

“Ten years on from *Exile*, Liz has finally managed to achieve what seems to have been her goal ever since the possibility of commercial success first presented itself to her: to release an album that could have just as easily been made by anybody else.”

Amazon.com (1 star)

“It took a couple of goes to get into it, but once the story hooked me, I found it difficult to put the book down -- except for those moments when I had to stop and shriek at my friends, "SPARKLY VAMPIRES!" or "VAMPIRE BASEBALL!" or "WHY IS BELLA SO STUPID?" These moments came increasingly often as I reached the climactic chapters, until I simply reached the point where I had to stop and flail around laughing.”

Tools and Resources

- Heuristic/hand made references
  - Inadequate in practice on their own
  - Can be useful for augmenting ML approaches
- Sentiment-oriented data sets
  - Highly domain sensitive
  - Difficult to create/collect

Heuristic/manual references

General Inquirer

- Content analysis tool
- Created in 1966
- Database of words and manually created semantic and cognitive categories, including positive and negative connotations
- Used to generate counts of words in categories

http://www.wjh.harvard.edu/~inquirer/

LIWC

- Linguistic Inquiry and Word Count
- Similar to GI
- Counts words belonging to categories, including positive and negative

  http://www.liwc.net/

Wordnet

- A lexical database for English with emphasis on synonymy
- Nouns, verbs, adjectives and adjectives are grouped into synonym sets
- Words are linked according to lexical and conceptual relations (creating a “net”)
- Not specifically sentiment oriented, but has been used to help derive sentiment related information (Hu & Liu)

http://wordnet.princeton.edu/

SentiWordNet

- A lexical resource for opinion mining
- Based on Wordnet synsets
- Each synset is assigned three sentiment scores: positivity, negativity, and objectivity

http://sentiwordnet.isti.cnr.it/

Whissell's Dictionary of Affective Language

- About 9000 words rated in terms of their Pleasantness, Activation, and Imagery (concreteness)
- App:

  http://sail.usc.edu/~kazemzad/emotion_in_text_cgi/DAL_app/

The steak was tough and tasteless but the wine was wonderful

Datasets for SA learning

Pang & Lee data sets

- Movie review polarity datasets
- Sentiment scale datasets
- Subjectivity datasets
- http://www.cs.cornell.edu/People/pabo/movie-review-data/

Blitzer et al Multi-domain sentiment dataset

- Reviews from Amazon.com from many product types (domains)
- Include star ratings
- Also divided into positive/negative

http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

MPQA Opinion Corpus

- Multi-Perspective Question Answering (MPQA) (Stoyanov et al, 2005)
- News articles and other text documents manually annotated for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, etc.).
- 692 documents (15,802 sentences)

http://www.cs.pitt.edu/mpqa/

Data for PMI-IR-based polarity identification

- The Web (for unsupervised training via PMI-IR)
- Waterloo-Multitext (alternate support database for PMI-IR method of assigning semantic orientation to phrases. Private access)

Thomas, Pang, & Lee, 2006

- Congressional speech data
- Transcripts of floor debates on policy

  http://www.cs.cornell.edu/home/llee/data/convote.html

Creating Sentiment-oriented Data sets

- Self-annotated data
  - Data has “built in” ordinal or binary labeling of some kind to complement NL text, ideally by the author of the text.
  - E.g. Amazon reviews (1-5 stars)
  - Pitchfork.com record reviews (0.0-10.0 range)
- Hand-annotated data
  - Annotated independently of the author
  - Usually labor intensive

Inter-annotator agreement

- Hand annotated sentiment data can vary in reliability
- Inter-annotator agreement is the degree to which multiple human annotators arrive at the same annotations when confronted with the same NL text
- Represents theoretical upper bound for sentiment classification

Mechanical turk

- Snow et al (2008) analyzed Amazon's mturk service for NLP annotation
- Roughly $1 for 1000 labels
- 5 non-expert annotators achieve equivalent accuracy to 1 expert annotator

Things to consider

- What elements do you want to classify, rank, or score?
- What classification/scale do you want to use?
- Is domain-appropriate annotated data available?
- If not, can it be created? Is inter-annotator agreement acceptable?

Techniques in Sentiment Analysis

Overview

- Semantic orientation and polarity of words
- Text-based sentiment classification
- Incorporating shallow linguistics
- Other approaches

Semantic Orientation

- Hatzivassiloglou & McKeown, 2002
- A real-number measure of positive or negative sentiment in a phrase
- *Polarity* is a binary value either positive or negative

Where to start?

- Texts are made up of words
- Words are in dictionaries
- Let's look up the words in the text, see what they mean, and be done with it!
- This (slightly more sophisticated) is what we do when we use heuristic tools

Heuristic methods

- “Heuristic” means applying what we know
- Dictionaries, thesauruses, word lists, etc
- General Inquirer (1966) groups words into 180 categories (like a dictionary with more categories)
- Wordnet creates a network of synonymy (like an extended, souped-up thesaurus with richer semantic organization)

HASSLE **Neg** Noun Hostile Work |

HASTE Noun Travl Actv |

HASTEN IAV SUPV Travl Actv |

HAT Noun Object Tool | noun: A shaped covering for the head

HATE#1 SV **Neg** SUPV Hostile Ngtv Psv Arousal | 80% verb: To dislike passionately, to detest
HATE#2 **Neg** Noun EMOT Hostile Ngtv Psv Arousal | 19% noun: Intense dislike, aversion, hostility
HATE#3 **Neg** Modif EVAL EMOT Hostile Ngtv Psv Arousal | 0% adj: 'hated'-loathed--'the hated dictator'
5 TOR(K+0,K+0,,10,ROOT.S.
6 TOR(K-1,K-1,APLY(2),,DET.PREP.
7 TOR(K-1,K-1,APLY(1),,TO.MOD.LY.HU.DO.DEF.
8 TOR(K+1,K+1,APLY(1),APLY(2),DET.PRON.
10 TOR(K+0,K+0,,APLY(1),ED.
11 TOR(K-1,K-1,APLY(3),APLY(1),DET.PREP.

HATER **Neg** Noun HU Ngtv Psv Hostile Role |

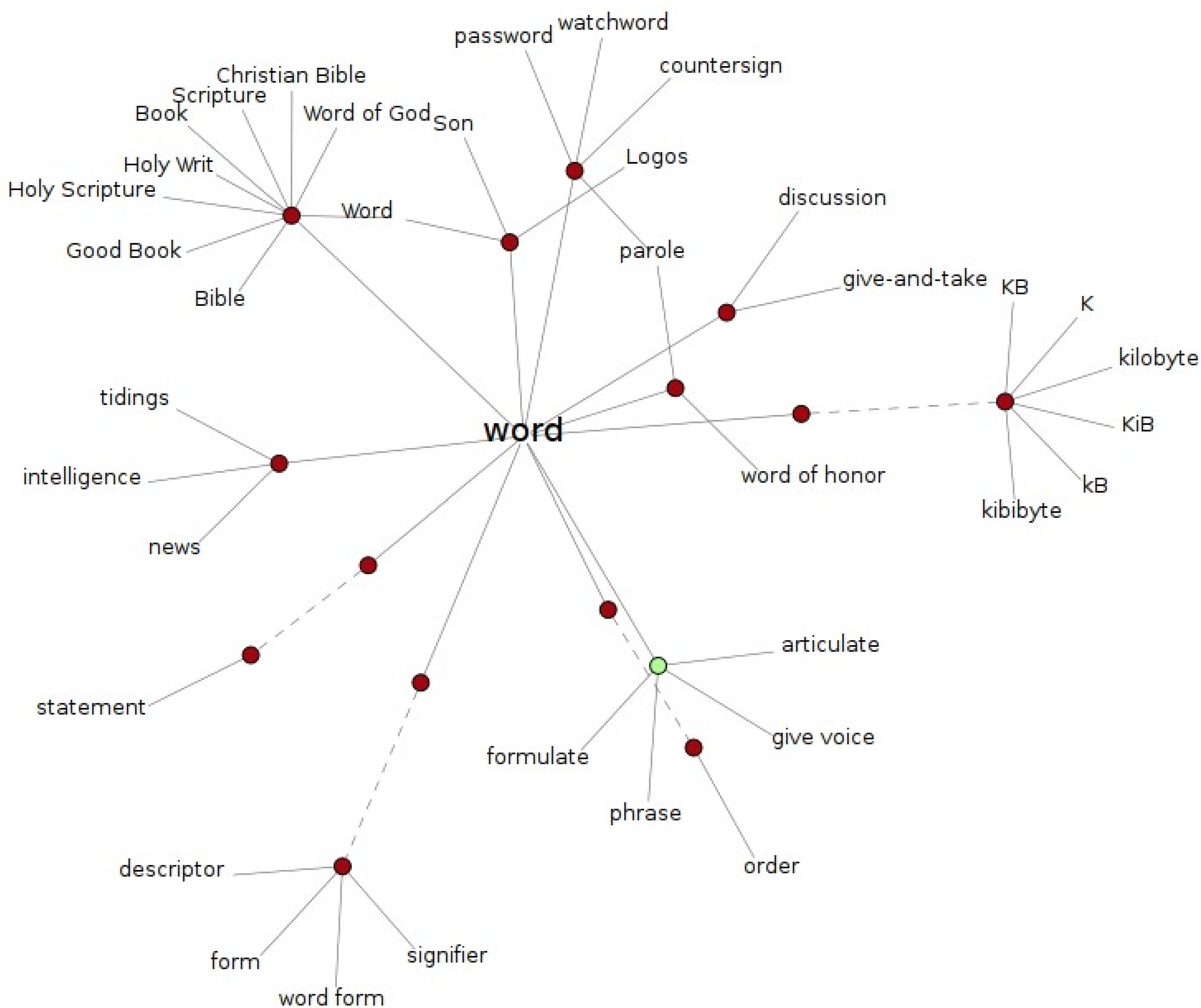
HATRED **Neg** Noun EMOT Hostile Ngtv Psv Arousal |

HAUGHTY IndAdj **Neg** Modif Emot Strng Power |

General Inquirer and polarity

- For identifying word polarity, we can use Neg and Pos categories
- Some problems
  - Binary, no gradations/weighting
  - Manually classed (intuitions are not always reliable)
  - Single word level only
  - Blind to context
- You cannot accurately classify texts as positive or negative using only lexical GI values

Wordnet



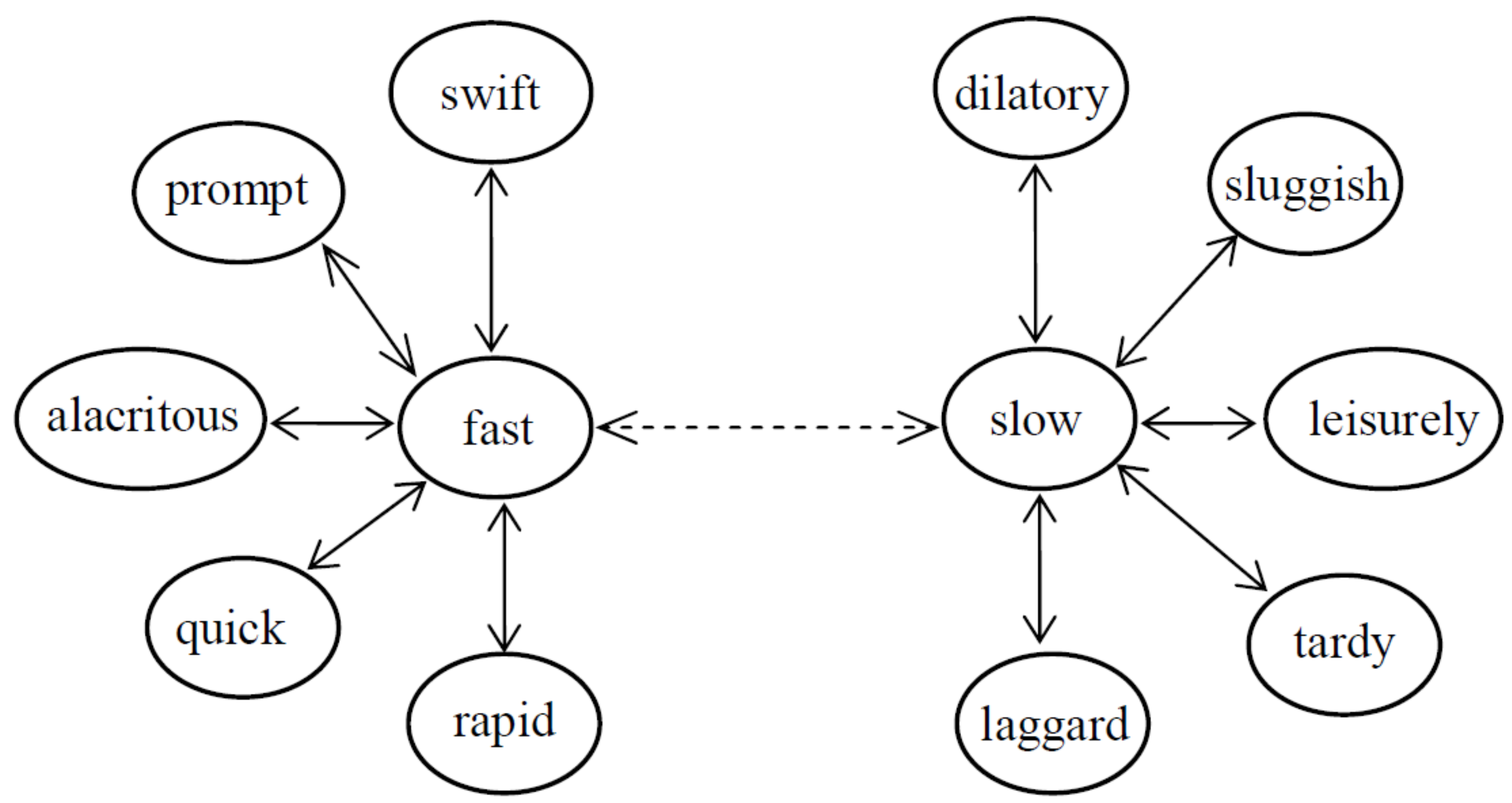
password
watchword
countersign
Christian Bible
Scripture
Book
Word of God
Son
Logos
Holy Writ
Holy Scripture
Word
discussion
Good Book
parole
Bible
give-and-take
KB
K
kilobyte
tidings
word
KiB
intelligence
word of honor
kB
kibibyte
news
articulate
statement
give voice
formulate
phrase
order
descriptor
signifier
form
word form

Wordnet

- Synonyms grouped in synsets
- Relationships between synsets:
  - HYPONYM: “type-of” relationship
  - HYPERNYM: {oak} -> {tree}
  - HAS-MEMBER: {family, family unit} -> {child, kid}
  - HAS-STUFF: {tank, army tank} -> {steel}
  - ENTAIL: {snore, saw wood} -> {sleep, slumber}
  - CAUSE-TO: {develop} -> {grow, become larger}
  - ATTRIBUTE: {hypocritical} -> {insincerity}

Wordnet

- Relationships between words:
  - PERTAINYM: academic -> academia
  - ANTONYM: presence -> absence
  - SIMILAR-TO: abridge -> shorten
  - SEE-ALSO: touch -> touch down


swift
prompt
alacritous
fast
quick
rapid
dilatory
sluggish
slow
leisurely
tardy
laggard

Polarity identification with Wordnet

- Hu & Liu (2004) identify polarity for adjectives using Wordnet
  - Begin with a set of “seed” adjectives of known orientation: “good”, “fantastic”, “wonderful”, “awful”, “terrible”, “bad”, etc.
  - For unknown adjectives, measure proximity via synonymy/antonymy relations to seed adjectives
  - If an adjective is close in synonymy to positive words, or close in antonymy to negative words, it's positive
  - Add newly labeled words to seed set

Evaluating sentence polarity

- Extract "opinion sentences" based on the presence of a predetermined list of product features and adjectives
  - e.g. "The **lens** is **excellent**"
- Evaluate the sentences based on counts of positive vs negative polarity words (as determined by the Wordnet algorithm)

Results (Hu & Liu, 2004)

- Predicting sentence polarity based on constituent word orientations
- Lowish extraction recall and precision due to disagreement with human annotators on what constitutes an "opinion sentence"

Product name	Opinion sentence extraction		Sentence orientation accuracy
	Recall	Precision	
Digital camera1	0.719	0.643	0.927
Digital camera2	0.634	0.554	0.946
Cellular phone	0.675	0.815	0.764
Mp3 player	0.784	0.589	0.842
DVD player	0.653	0.607	0.730
Average	**0.693**	**0.642**	**0.842**

Polarity identification with Wordnet

- Advantages
  - Very fast
  - No training data necessary
  - Good predictive accuracy
- Disadvantages
  - Does not deal with multiple word sense, context issues
  - Does not work for multiple word phrases (or non-adjective words)

Osgood values for words

- Theory of Semantic Differentiation (Osgood, 1957)
- Three values pertinent to the emotive meaning of adjectives
  - Potency (strong or weak)
  - Activity (active or passive)
  - Evaluative (good or bad)

Deriving Osgood values with Wordnet

- Kamps and Marx (2002) used Wordnet to assign scores to words based on Osgood factors
- For each Osgood factor, compared the minimal path length (MPL) in Wordnet between two words representing the factor's range.
- E.g., for Evaluative factor (EVA), compare MPLs for word between “good” and “bad”

Deriving Osgood values with Wordnet

- Only adjectives connected by synonymy to both opposites receive scores (i.e., an adjective must have a synonymy path to both “good” and “bad” to receive an EVA score)
- Yields a list of adjectives with EVA, POT and ACT scores

Semantic orientation of phrases

- Words may not be enough
  - *unpredictable plot – unpredictable steering*
  - *flakey crust - flakey politician*
  - *ridiculous comedy – ridiculous drama*
  - *cheap construction – cheap deal*
- We might want to assign SO scores to certain kinds of phrases
- Binary polarity judgments don't capture nuance

The PMI-IR method

- Turney (2002)
- Using Pointwise Mutual Information (PMI) on data gathered using Information Retrieval (IR) techniques
- Yields real-numbered positive and negative scores for potentially any combination of words
- Requires WWW-sized unstructured training data resources

The PMI-IR method

- Extract descriptive 2-word phrases based on POS

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN or NNS
5.	RB, RBR, or RBS	VB, VBD, VBN or VBG	anything

The PMI-IR method

- For each phrase, conducted Altavista searches using the NEAR operator, one with the word *excellent* and one with the word *poor.*
- NEAR operator (now discontinued) searched for the phrase occurring within ten words of the value word.
- Derive a score based on returned hit counts for each search and hit counts of the words and phrases on their own

The PMI-IR method

- Calculating PMI
- *word1* is the descriptive phrase, *word2* is the value word
- p() is Altavista hit count (& is NEAR operator)

$$\mathrm{PMI}(word_1, word_2) = \log_2\left(\frac{\mathrm{p}(word_1 \,\&\, word_2)}{\mathrm{p}(word_1)\mathrm{p}(word_2)}\right)$$

The PMI-IR method

- Deriving semantic orientation from PMI

$$\mathrm{SO}(\textit{phrase}) = \mathrm{PMI}(\textit{phrase}, \text{"excellent"}) - \mathrm{PMI}(\textit{phrase}, \text{"poor"})$$

Classifying whole documents

- Based on average SO of phrases in the review

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		0.322

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

Results (Turney, 2002)

Domain of Review	Accuracy
Automobiles	84.00 %
Honda Accord	83.78 %
Volkswagen Jetta	84.21 %
Banks	80.00 %
Bank of America	78.33 %
Washington Mutual	81.67 %
Movies	65.83 %
The Matrix	66.67 %
Pearl Harbor	65.00 %
Travel Destinations	70.53 %
Cancun	64.41 %
Puerto Vallarta	80.56 %
All	74.39 %

Incorporating diverse information sources

- We might want to combine information sources
- Words, phrases, other methods of evaluation, topic information, sentence position, etc...
- To do this involves building more sophisticated models

Pang et al. 2002

- Compared a variety of well-known text classification techniques and feature sets (IMDB dataset)

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Support Vector Machines

- SVMs are a widely used ML technique for creating feature-vector-based classifiers
- Each instance to be classified is represented by a vector of real-numbered features
- Training data is used to generate a high-dimensional space that can be divided by a hyperplane between positive and negative instances
- New instances are classified by finding their position in the space with respect to the hyperplane

Support Vector Machines

- Very good at combining diverse information sources
- Does not assume feature independence; overlapping information sources OK
- Supervised learning; requires annotated training data
- Like statistical methods, sensitive to sparse and insufficient data

SA with diverse information sources

- Mullen & Collier (2004)
- Incorporate a variety of overlapping information sources based on Turney scores and Osgood values
- Primary motivation was to incorporate topic information

The data

- 100 record reviews from Pitchfork.com
- Author-assigned rank from 0.0 to 10.0
- 50 reviews selected from >8.0 score
- 50 reviews selected from <3.0 score
- Hand-annotated with `THIS_WORK` and `THIS_ARTIST` tags for all references (including co-references) to the title of the album and the artist, respectively.

Features (traditional)

- Word token unigrams
- Lemmatized unigrams
  - lemmatized using Conexor FDG parser

Features (Turney-based)

- **Turney value**: Average value of all phrases' SO values
- **In sentence with THIS_WORK**: Average value of all SO scores for phrases in the same sentence as a reference to the work being reviewd
- **Following THIS_WORK**: Average value of SO scores for phrases which follow a reference to the work being reviewed directly or separated by the copula or a preposition
- **Preceding THIS_WORK**:Average value of SO scores for phrases which precede a reference to the work being reviewed directly or separated by the copula or a preposition
- **In sentence with THIS_ARTIST**: Similar to above, but for artist
- **Following THIS_ARTIST**: Similar to above, but for artist
- **Preceding THIS_ARTIST**: Similar to above, but for artist

Features (Osgood-based)

- **Text-wide EVA**: Average ETA of all adjectives in document
- **Text-wide POT**: Average POT of all adjectives in document
- **Text-wide ACT**: Average ACT of all adjectives in document
- **Topic-sentence EVA**: Average ETA of all adjectives that share a sentence with the topic (artist or work) of the review
- **Topic-sentence POT**: Average POT of all adjectives that share a sentence with the topic (artist or work) of the review
- **Topic-sentence ACT**: Average ACT of all adjectives that share a sentence with the topic (artist or work) of the review

Results (IMDB)

Model	**3 folds**	**10 folds**
Pang et al. 2002	82.9%	NA
Turney Values only	68.4%	68.3%
Osgood only	56.2%	56.4%
Turney Values and Osgood	69.0%	68.7%
Unigrams	82.8%	83.5%
Unigrams and Osgood	82.8%	83.5%
Unigrams and Turney	83.2%	85.1%
Unigrams, Turney, Osgood	82.8%	85.1%
Lemmas	84.1%	85.7%
Lemmas and Osgood	83.1 %	84.7%
Lemmas and Turney	84.2%	84.9%
Lemmas, Turney, Osgood	83.8%	84.5%
Hybrid SVM (Turney and Lemmas)	84.4%	**86.0%**
Hybrid SVM (Turney/Osgood and Lemmas)	**84.6%**	**86.0%**

Results (pitchfork)

Model	**5 folds**	**10 folds**	**20 folds**	**100 folds**
Turney Values only	72%	73%	72%	72%
All (THIS_WORK and THIS_ARTIST) PMI	70%	70%	68%	69%
THIS_WORK PMI	72%	69%	70%	71%
All Osgood	64%	64%	65%	64%
All PMI and Osgood	74%	71%	74%	72%
Unigrams	79%	80%	78%	82%
Unigrams, PMI, Osgood	81%	80%	82%	82%
Lemmas	83%	85%	84%	84%
Lemmas and Osgood	83%	84%	84%	84%
Lemmas and Turney	84%	85%	84%	84%
Lemmas, Turney, text-wide Osgood	84%	85%	84%	84%
Lemmas, PMI, Osgood	84%	85%	84%	86%
Lemmas and PMI	84%	85%	**85%**	86%
Hybrid SVM (PMI/Osgood and Lemmas)	**86%**	**87%**	84%	**89%**

Some conclusions

- Various good word and phrase classification methods exist
- Topic information is very useful when known
- SVM good for bringing different information sources together
- Using diverse overlapping word and phrase-based features with topic information can yield good results

Sentiment Analysis of Political Content

Overview

- Definitions and varieties of political content
- Motivations and goals
- Some pertinent research

Political Sentiment Analysis

- Public opinion
  - Attitudes to policies, parties, government agencies, politicians
- Policy-making and government
  - Arguments and beliefs informing discussions between lawmakers or representatives
- Informal or formal environments

Analyzing political opinion

- Possible applications:
  - Analyzing political trends/Augmenting opinion polling data
  - Targeting advertising and communications such as notices, donation requests, or petitions
  - Identifying political bias, e.g. in news texts
  - Evaluating lawmakers positions, arguments, or biases

Sentiment Analysis of Informal Political Texts

- What is informal political discourse?
- Why try to analyze it?
- Successes and failures

What is informal political discourse?

- Informal political discourse can be found in
  - Newsgroups
  - Blogs
  - Online publications reader feedback sections
  - Social Networking Services
- Generally organized as linear threads by topic
- Discourse is “informal”; written quickly, as thought
- Overall discourse not real-time, but individual exchanges often near real-time.

Idiosyncrasies of informal political discourse

- Informal
  - Rampant spelling errors
  - Casual usage (sentence fragments, etc.)
- Political
  - Jargon, names, non-dictionary terms
- Informal and political
  - Specific jargon, terms of abuse (“wingnuts”, “moonbats”)
  - Satirical re-spellings of known words (“Raygun”, Repugnicans”, “Dumbocrats”)

Sentiment analysis of informal political discourse

- What is “political opinion?”
  - SA often considers a binary “thumbs up” vs “thumbs down” classification
  - This is too simple to represent political opinion.
- Political attitudes encompass a variety of favorability judgments
- Relations between judgments are not always clear; e.g., in the US political domain anti-abortion judgment often corresponds to pro-death penalty judgment.

Possible goals

- Aside from binary judgments about a specific issue, candidate, or proposal, we might want to:
  - Identify political party affiliation
  - Classify according to some more general taxonomy, e.g. right vs left
  - Gauge the “extremeness” or distance from a politically centrist position of the writer’s views
  - Evaluate the degree of confidence with which the writer expresses views
  - Evaluate the degree of agreeability/argumentativeness with which the writer communicates
  - Identify particular issues of special importance to the writer

Sentiment analysis of informal political discourse (Mullen & Malouf 2008)

- Goal: to automatically classify participants in an online political discussion forum according to political viewpoint

Classifying political attitudes

- As a preliminary task, we opted for the simplest classification scheme we could think of:
  - right
  - left
  - other
- Many viewpoints do not fit tidily on the left/right line, and “other” is so general as to be essentially noise

The data

- Data from the (now defunct) www.politics.com discussion site
- 77,854 posts organized by topic thread
- 408 individual posters
- Number of posts follows a Zipf-like distribution, with 19% of posters logging only a single post.
- Greatest number of posts by a single poster is 6885, second is 3801

Identifying quotes

- Each post broken into “chunks” based upon typographical cues such as new lines, quotes, boldface, and italics, to identify sections of the post which are quoted from previous posts.
- Chunks of three words or greater which are complete substrings of previous posts are considered quotes.
- The database is broken into 229,482 individual chunks, of which 22,391 are identified as quotes from other posts.

Supplementary data

- Additional data from the web was used to support spelling correction
  - 6481 politically oriented syndicated columns from right and left leaning websites, to provide professionally edited spellings of domain specific terms
  - A wordlist of email, chat, and text message slang, including such terms as “lol” meaning “laugh out loud”

Political affiliation in the data

- Posters have a self-described political affiliation.
- After some hand-editing, nine modified labels were identified:
  - Republican
  - Conservative
  - R-fringe
  - Democrat
  - Liberal
  - L-fringe
  - Centrist
  - Independent
  - Libertarian

Classes to stated affiliation

Right	34%	Republican	53
		Conservative	30
		R-fringe	5
Left	37%	Democrat	62
		Liberal	28
		L-fringe	6
Other	28%	Centrist	7
		Independent	33
		Libertarian	22
Unknown			151

Naïve Bayes lexical model

- First, we used naïve Bayes to classify posts lexically as Left or Right
- “Other” users were disregarded
- Total number of users were 96 left, and 89 right, so the baseline was 51.9%
- Lexical model performed at 60.4%

Observations on the lexical model

- Unlike with topic identification, arguments from both sides of an issue use many of the same terms.
- Irregular spellings are harmful to lexical models, necessitating far more training data.
- Skewed distribution of posting frequency means that frequent posters are better modeled than infrequent posters

Some adjustments

- Restricting experiments to frequent posters (500+ words)
  - Baseline 50%
  - Naïve Bayes: 61.38%
  - With spelling correction: 64.48%
- Human gold standard 87.5% for all users, 91% for frequent posters

Quote patterns

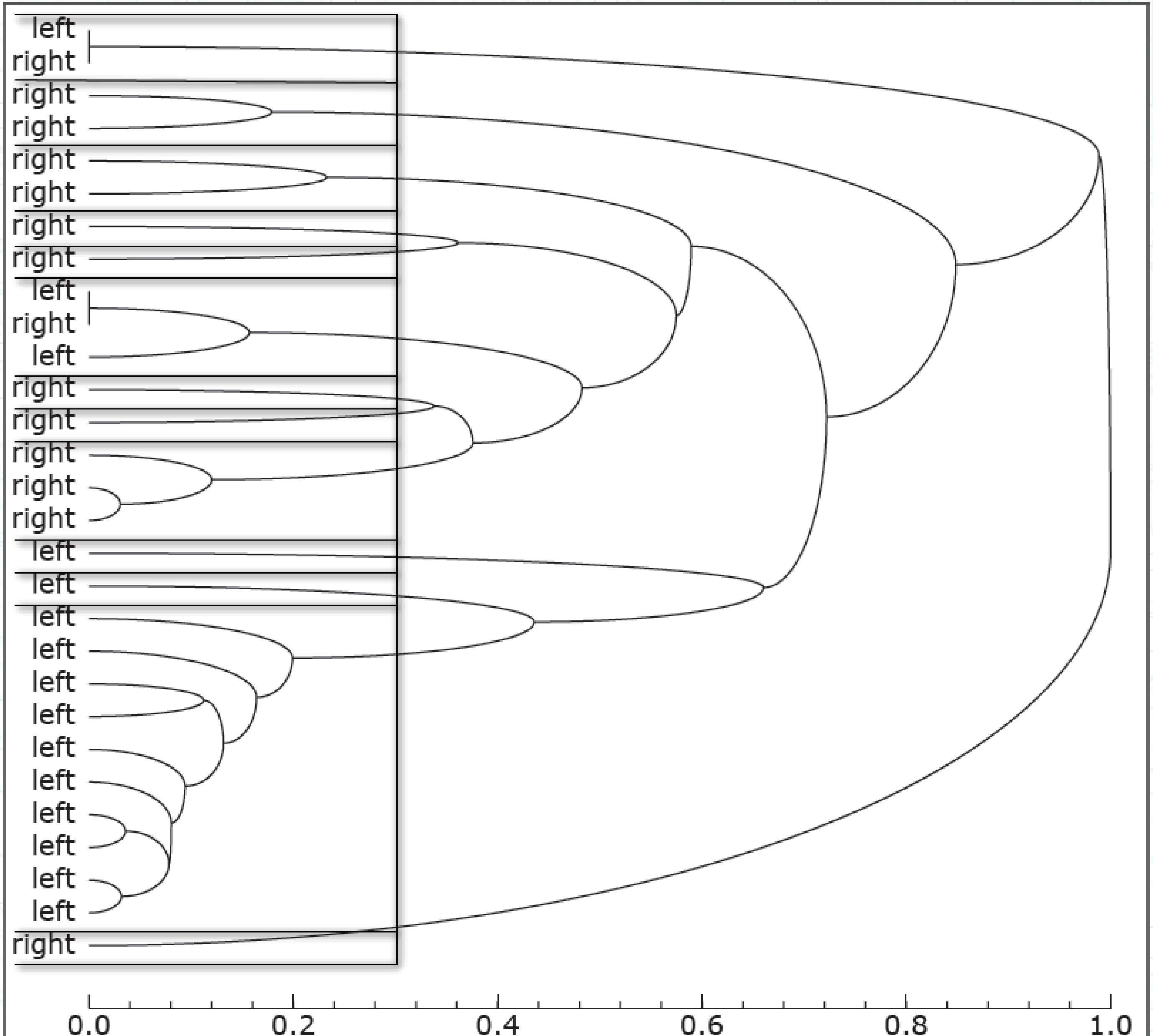
- Of 41,605 posts 4,583 contained quoted material
- Strong tendency to quote users from opposite end of political spectrum
  - Left quoted right: 62.2%
  - Right quoted left: 77.5%

Classification by quote

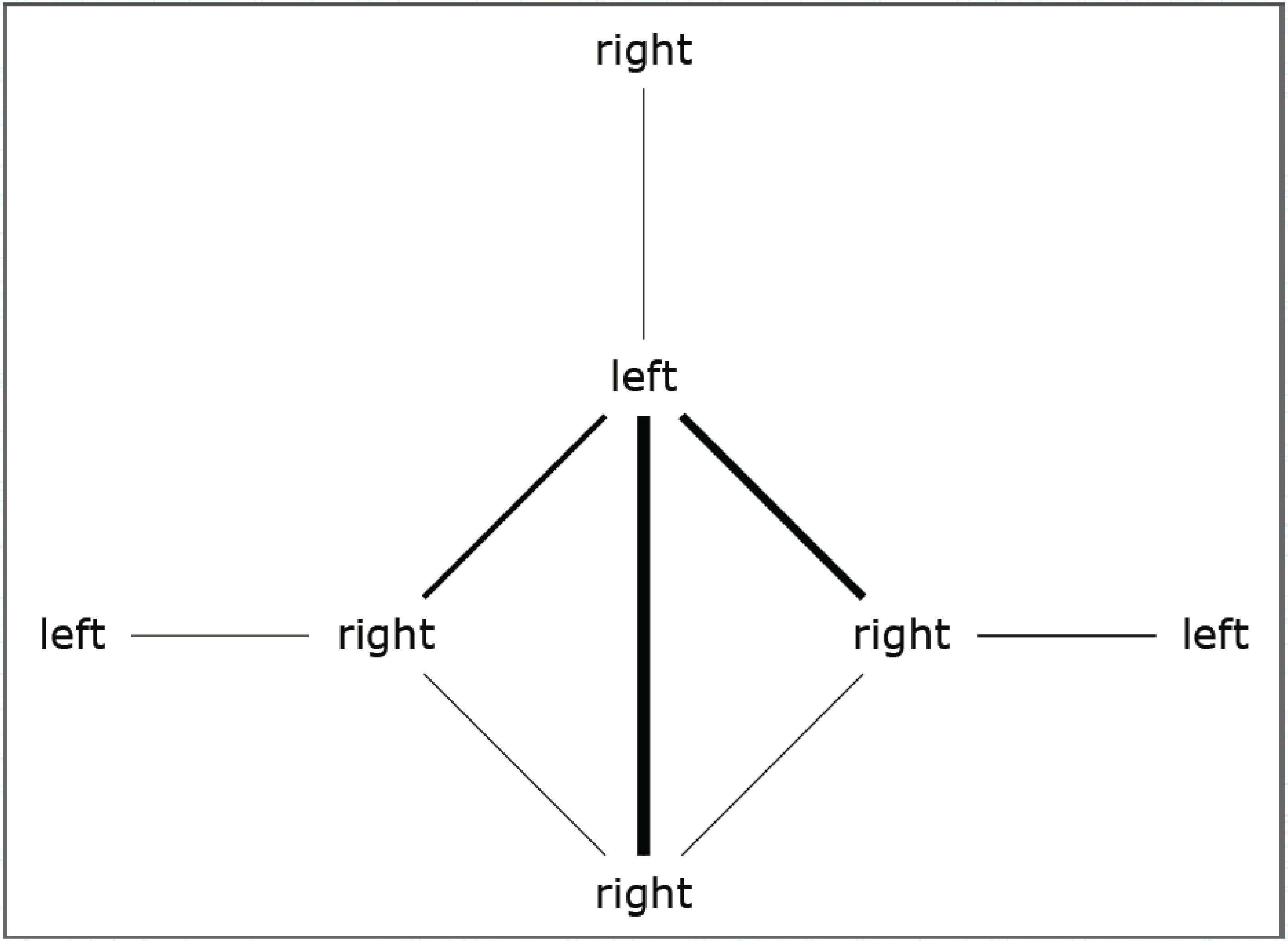
- For frequent posters:
  - For those who quote/are quoted: 83.53
  - Overall: 79.38
- However, this assumes that we know the class of the quoted poster

Using user citation graph information

- Created a graph with each user as a node and each quote an edge
- Singular value decomposition on graph’s adjacency matrix to compute a “citation space” in which distances between users could be measured
- Derived equivalence classes via alliance/agreement patterns


left
right
right
right
right
right
right
right
left
right
left
right
right
right
right
right
left
left
left
left
left
left
left
left
left
left
left
left
right
0.0
0.2
0.4
0.6
0.8
1.0

Equivalence classes



Using user citation graph information

- Graph-based clustering + NB yielded 68.48% accuracy for all users, 73% for frequent posters

Sentiment Analysis of Texts

- Assumptions
  - Political attitudes are (the same as| analogous to|composed of) the kind of opinions found in reviews
  - Political discussion is rhetorically similar in some significant respect to opinion/review writing

Simple PMI-IR inspired political classifcation

$$\mathrm{SO}(\mathit{phrase}) = \mathrm{PMI}(\mathit{phrase}, \text{"liberal"}) - \mathrm{PMI}(\mathit{phrase}, \text{"conservative"})$$

- Derived SO values from Reuters corpus
- Results considerably below baseline

Simple PMI-IR inspired political classifcation

- Possible reasons for poor performance
  - Wrong choice of contrast terms?
  - Inadequate training data?
  - Deeper assumptions mistaken?

Single-Issue PMI-IR feature vectors

- Assume that political attitudes are collections of positive/negative judgements on single, hot-button issues
- Draw up a list of politically contentious words/terms/names
- From each poster, select all sentences containing each of these terms
- Evaluate using PMI-IR to get an SO score for each concept
- SVM model with resulting feature vectors

Single-Issue PMI-IR feature vectors

- Created approximately 100 contentious concepts by hand, intuitively likely to distinguish right from left in American political discussion.
- Turney’s Waterloo Multitext system to derive SO values
- Used various sets of opposing keywords (for PMI-IR)
- No deviation from the baseline

What are the problems?

- As usual, data is sparse
- Political opinions expressed more obliquely than, e.g. movie reviews?
- Rhetorical goals different?
  - Reviews are written to express/describe/justify opinions
  - Political discussion posts treat underlying opinions as given and focus on convincing and/or attacking

Some conclusions

- Patterns of agreement/disagreement more salient than actual opinion content
- Political discussion more than just a description of opinions on various topics
- PMI-IR based methods not promising for informal political text analysis

Sentiment analysis and the policy-making process

- Determining support or opposition from Congressional floor-debate transcripts
- Thomas et al, 2006
- Evaluate a formal speech on policy to determine whether the speaker supports or opposes the policy

Sentiment and eRulemaking

- Electronic rulemaking, or *eRulemaking* initiatives seek to involve the public more closely in policy-making through "electronic collection, distribution, synthesis, and analysis of public commentary in the regulatory rulemaking process" (Shulman & Schlosberg, 2002)
- Analysis of NL policy discussion would benefit from SA

Why this is difficult

- Congressional debates contain very rich language and cover a wide variety of topics
- Subject to potentially wide digressions (e.g. “Why are we discussing this bill when the plight of my constituents regarding this other issue is being ignored?”
- Speakers spend more time presenting evidence in support of their position than stating their opinions explicitly

The data

- Congressional floor debate data
- Speeches labeled by the speaker's eventual “yea” or “nay” vote on the proposed bill

	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

Models

- Unigram-based SVM for classifying individual segments as yea or nay
- Identify instances of inter-speaker agreement based on by-name reference and predetermined words and phrases indicating agreement
- An agreement threshold is adjusted to control precision vs accuracy of agreement

Results (Thomas, et al 2006)

Support/oppose classifer ("speech segment⇒yea?")	Devel. set	Test set
SVM [speaker]	71.60	70.00
SVM + agreement links . . .		
with $\theta_{\text{agr}} = 0$	**88.72**	**71.28**
with $\theta_{\text{agr}} = \mu$	84.44	76.05

Some more conclusions

- Political sentiment analysis is difficult even when restricted to straight support vs opposition judgments in formal environments
- Thomas et al (2006) gives some insight into why our attempted single-issue feature-based SVM failed
- If yea or nay accuracy is low, an SVM based on these features would have high levels of noise for each feature

Some more conclusions

- Most political discussions do occur in some wider discourse context
- Agreement/disagreement/alliance information should be regarded as a crucial component for political sentiment analysis
- Traditional classification approaches may provide a starting point

To sum up

- Sentiment analysis is a difficult task
- The difficulty increases with the nuance and complexity of opinions expressed
- Product reviews, etc are relatively easy
- Books, movies, art, music are more difficult
- Policy discussions, indirect expressions of opinion more difficult still
- Non-binary sentiment (political leanings etc) is extremely difficult
- Patterns of alliance and opposition between individuals become central