INSOFE
Inspire...Educate...Transform.

Inspire…Educate…Transform.

7. Sentiment Analysis

Dr. Manish Gupta
Sr. Mentor – Academics, INSOFE

Adapted from Prof. Dan Jurafsky’s slides at
http://spark-public.s3.amazonaws.com/nlp/slides/sentiment.pptx

Course Content

- Collection of three main topics of high recent interest.
  - Search engines (Crawling, Indexing, Ranking)
    - Language Modeling
    - Text Indexing and Crawling
    - Relevance Ranking
    - Link Analysis Algorithms
  - Text Processing (NLP, NER, Sentiments)
    - Natural Language Processing
    - Named Entity Recognition
    - **Sentiment Analysis**
    - Summarization
  - Social networks (Properties, Influence Propagation)
    - Social Network Analysis
    - Influence Propagation in Social Networks

Agenda

- **Applications of Sentiment Analysis**

Why Sentiment Analysis

- Also called as Opinion extraction, Opinion mining, Sentiment mining, Subjectivity analysis
- *Applications*
  - *Movie*: is this review positive or negative?
  - *Products*: what do people think about the new iPhone?
  - *Public sentiment*: how is consumer confidence? Is despair increasing?
  - *Politics*: what do people think about this candidate or issue?
  - *Prediction*: predict election outcomes or market trends from sentiment

Positive or Negative Movie Review?

- unbelievably disappointing 
- Full of zany characters and richly applied satire, and some great plot twists 
- this is the greatest screwball comedy ever filmed 
- It was pathetic. The worst part about it was the boxing scenes. 

Product Search

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, **$100** nearby ★★★★☆ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

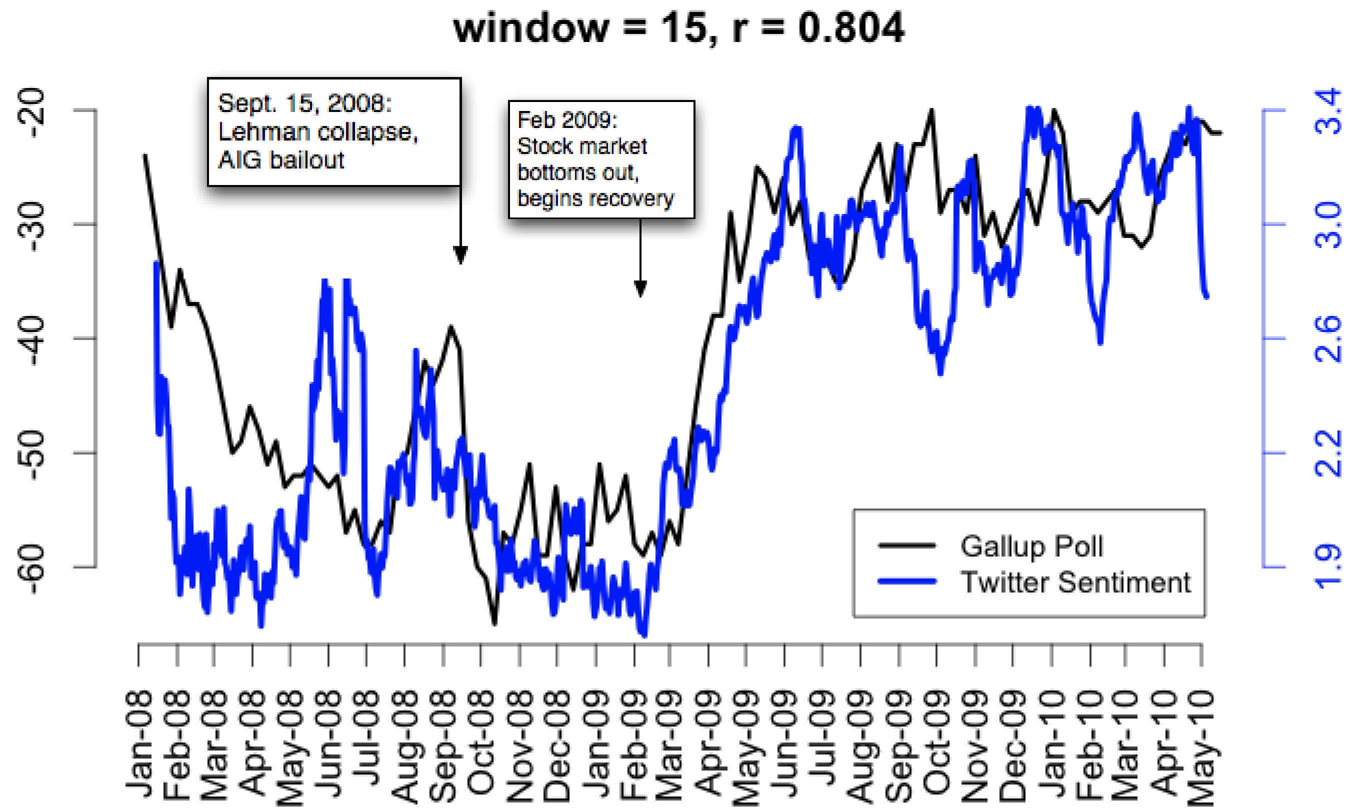
Summary - Based on 377 reviews

1 star	2	3	4 stars	5 stars

What people are saying

ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

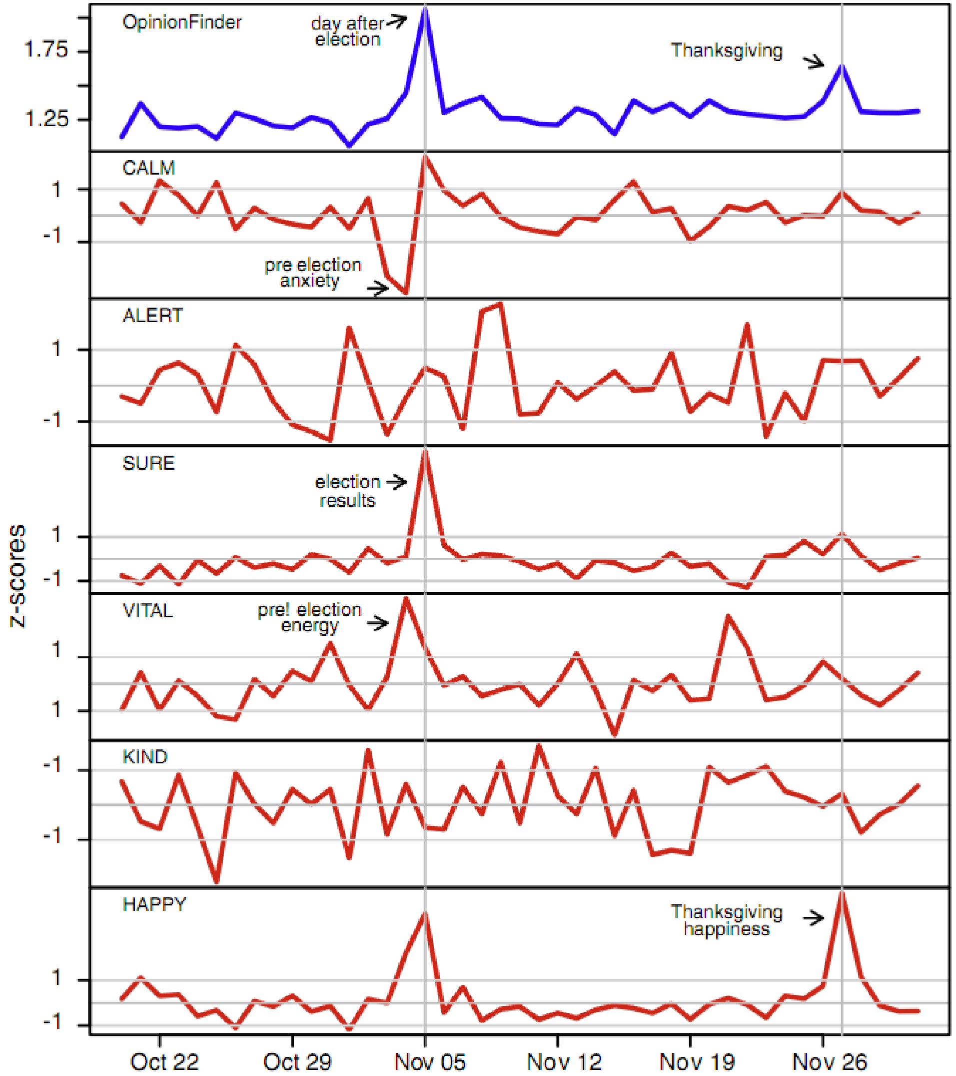
Twitter Sentiment versus Gallup Poll of Consumer Confidence



- 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 ICWSM-2010
- Gallup is a public opinion company.

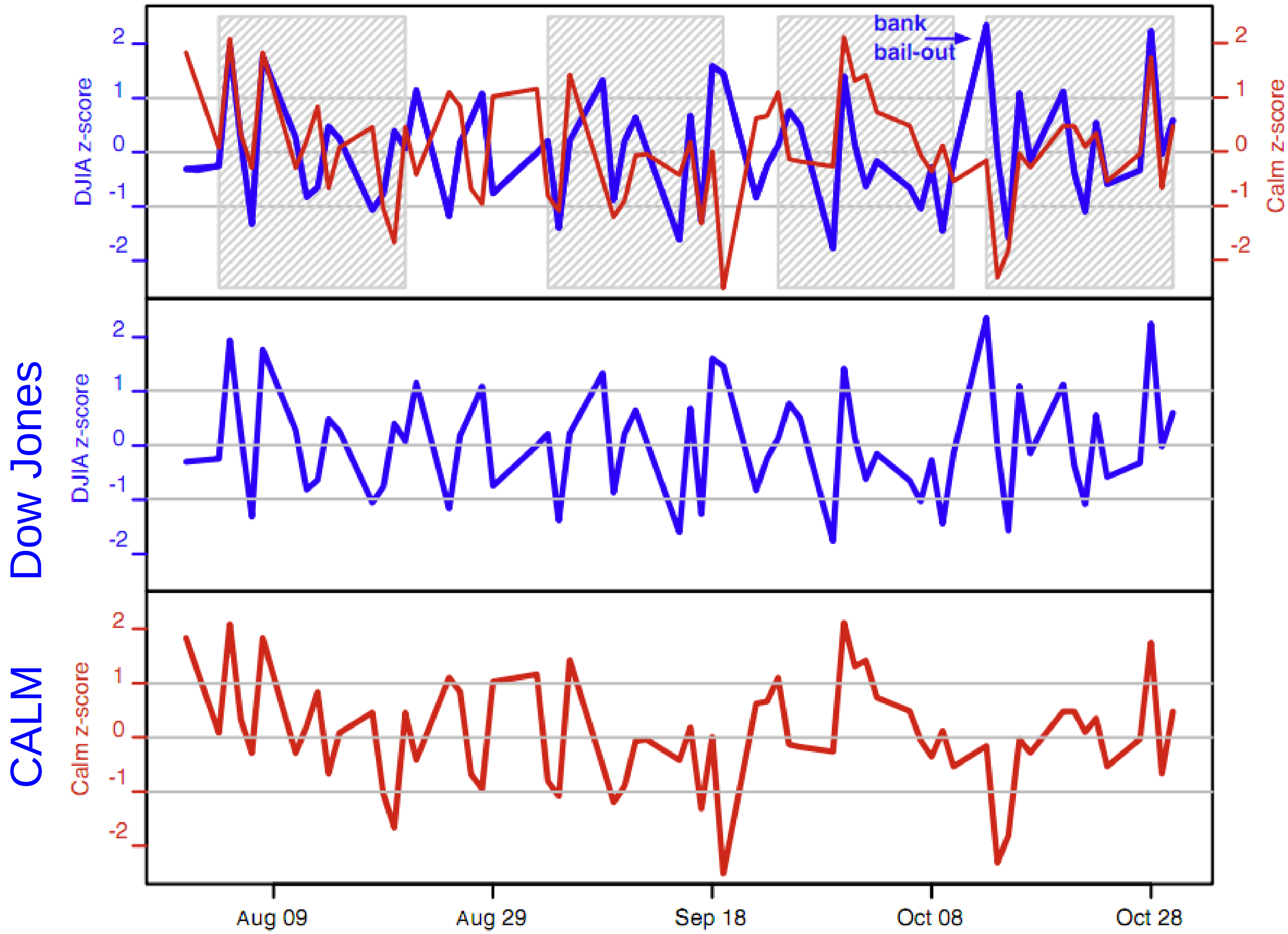
Twitter Sentiment

- Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. [Twitter mood predicts the stock market,]
- They analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy).
- They validate the ability of OF and GPOMS to capture various aspects of public mood, during the presidential elections and thanksgiving.



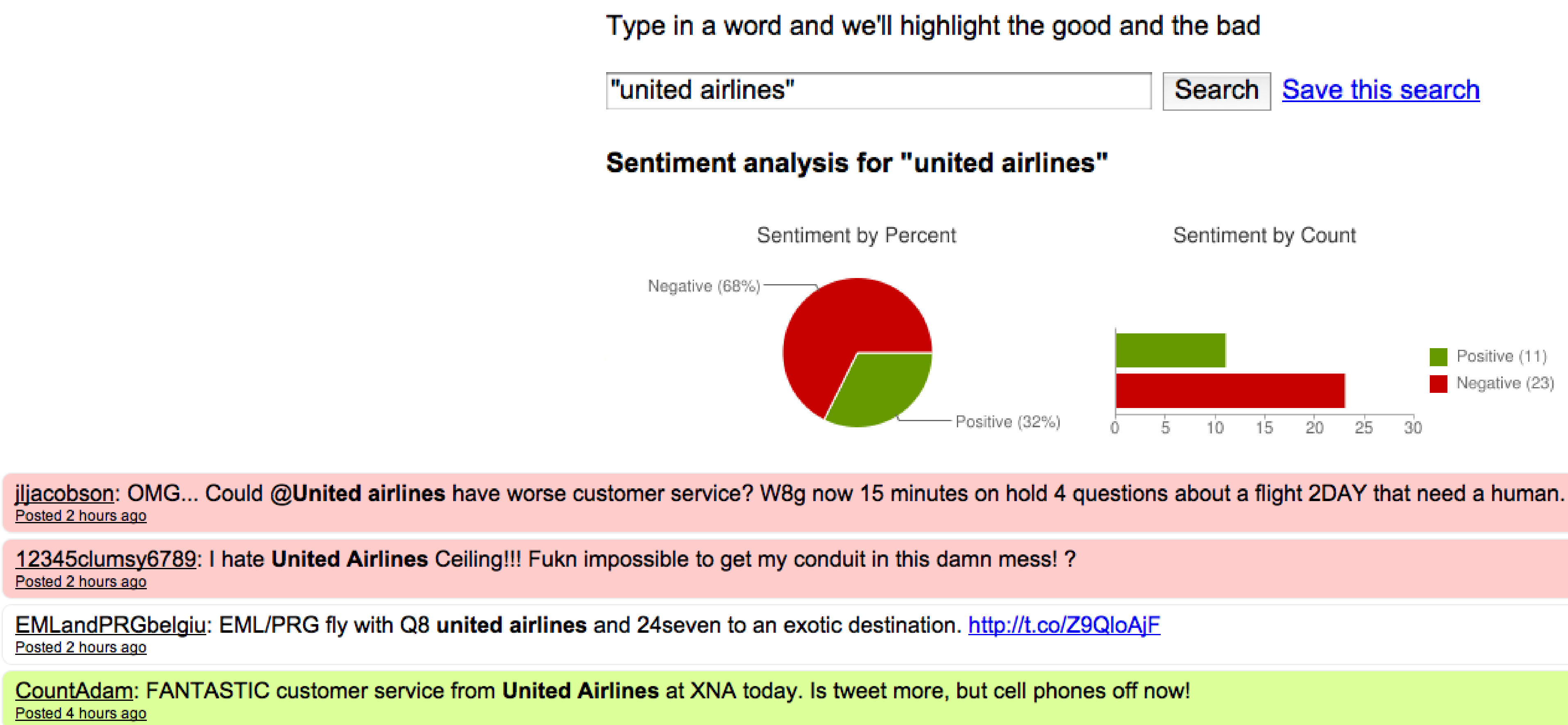
Twitter Sentiment and DJIA

- Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time.
- CALM predicts DJIA 3 days later



Target Sentiment on Twitter

- Twitter Sentiment App
- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision



Sentiment Analysis

- Simplest task:
  - Is the attitude of this text positive or negative?
- More complex:
  - Rank the attitude of this text from 1 to 5
- Advanced:
  - Detect the target, source, or complex attitude types

Agenda

- Applications of Sentiment Analysis
- **Word-based Classification Approach**

IMDB data in the Pang and Lee database (2002)

✓

when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . […]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

october sky offers a much simpler image–that of a single white dot , traveling horizontally across the night sky . [. . .]

✗

“ snake eyes ” is the most **aggravating** kind of movie : the kind that shows so much potential then becomes **unbelievably disappointing** .

it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare .

and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

Word-based Classification Approach

- Tokenization of words in review
- Feature Extraction
- Classification using different classifiers
  - Naïve Bayes – will use this one today
  - MaxEnt
  - SVM
  - MaxEnt and SVM tend to do better than Naïve Bayes

Sentiment Tokenization Issues

- Deal with HTML and XML markup
- Twitter mark-up (user names, hash tags)
- Capitalization (preserve words in all caps)
- Phone numbers, dates
- Emoticons
- Useful code:
  - Christopher Potts sentiment tokenizer
  - Brendan O'Connor twitter tokenizer

Extracting Features for Sentiment Classification

- How to handle negation
  - “I didn’t like this movie” vs “I really like this movie”
  - Add NOT_ to every word between negation and following punctuation:
    - didn’t like this movie , but I
    - didn’t NOT_like NOT_this NOT_movie but I
- Which words to use?
  - Only adjectives
  - All words
    - All words turns out to work better, at least on this data

Reminder: Naïve Bayes

In Naïve Bayes, prob of assigning a document to class c is computed as follows:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \left[P(c_j) \prod_{i \in positions} P(w_i \mid c_j) \right]$$

$$P(w_i|c_j) = \frac{count(w_i, c_j) + 1}{count(c) + |V|}$$

CSE 7306c

Binarized (Boolean feature) Multinomial Naïve Bayes

- Intuition:
  - For sentiment
  - Word occurrence may matter more than word frequency
    - The occurrence of the word *fantastic* tells us a lot
    - The fact that it occurs 5 times may not tell us much more.
  - Boolean Multinomial Naïve Bayes
    - Clips all the word counts in each document at 1

Boolean Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $= c_j$

$$P(c_j) \neg \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - Remove duplicates in each doc:
    - For each word type w in doc_j
      - Retain only a single instance of w
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$$

Boolean Multinomial Naïve Bayes on test doc *d*

- First remove all duplicate words from *d*
- Then compute NB using the same equation:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \left[P(c_j) \prod_{i \in positions} P(w_i \mid c_j) \right]$$

Normal vs. Boolean Multinomial NB

Normal	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Boolean	Doc	Words	Class
Training	1	Chinese Beijing	c
	2	Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Tokyo Japan	?

Binary seems to work better than full word counts

Problems: What makes reviews hard to classify?

- Perfume review in *Perfumes: the Guide*: “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
  - Negative review but very difficult to figure out using just positive or negative words.
- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”
  - Seems like a positive review but it is not.
- Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is **not so good** either, I was surprised.
  - Ordering effect
  - Multiple sentiments within same sentence.

Agenda

- Applications of Sentiment Analysis
- Word-based Classification Approach
- **Sentiment Lexicons and Negation Words**

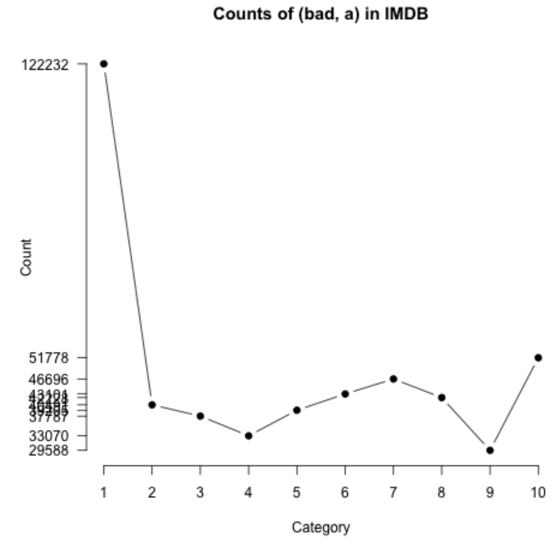
Sentiment Lexicons

- Positive words: abide, ability, able, abound, absolve, absorbent, …
- Negative words: abandon, abandonment, abate, abdicate, abhor, abject, …
- Available lexicons
  - General Inquirer: 1915 positive words and 2291 negative words
  - LIWC (Linguistic Inquiry and Word Count): 2300 words, >70 classes
  - MPQA Subjectivity Cues Lexicon: 2718 positive words and 4912 negative words. Each word annotated for intensity (strong, weak)
  - Bing Liu Opinion Lexicon: 2006 positive words and 4783 negative words
  - SentiWordNet: All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
    - [estimable(J,3)] "may be computed or estimated": Pos=0, Neg=0, Obj=1
    - [estimable(J,1)] "deserving of respect or high regard": Pos=0.75, Neg=0, Obj=0.25

Analyzing the Polarity of each Word in IMDB

- How likely is each word to appear in each sentiment class?
- Count("bad") in 1-star, 2-star, 3-star, etc.
- But can't use raw counts: count(10)>count(2) because there are many 10 star reviews and very few 2 star ones
- Instead, **likelihood:**

$$P(w|c) = \frac{f(w,c)}{\sum_{w \in c} f(w,c)}$$



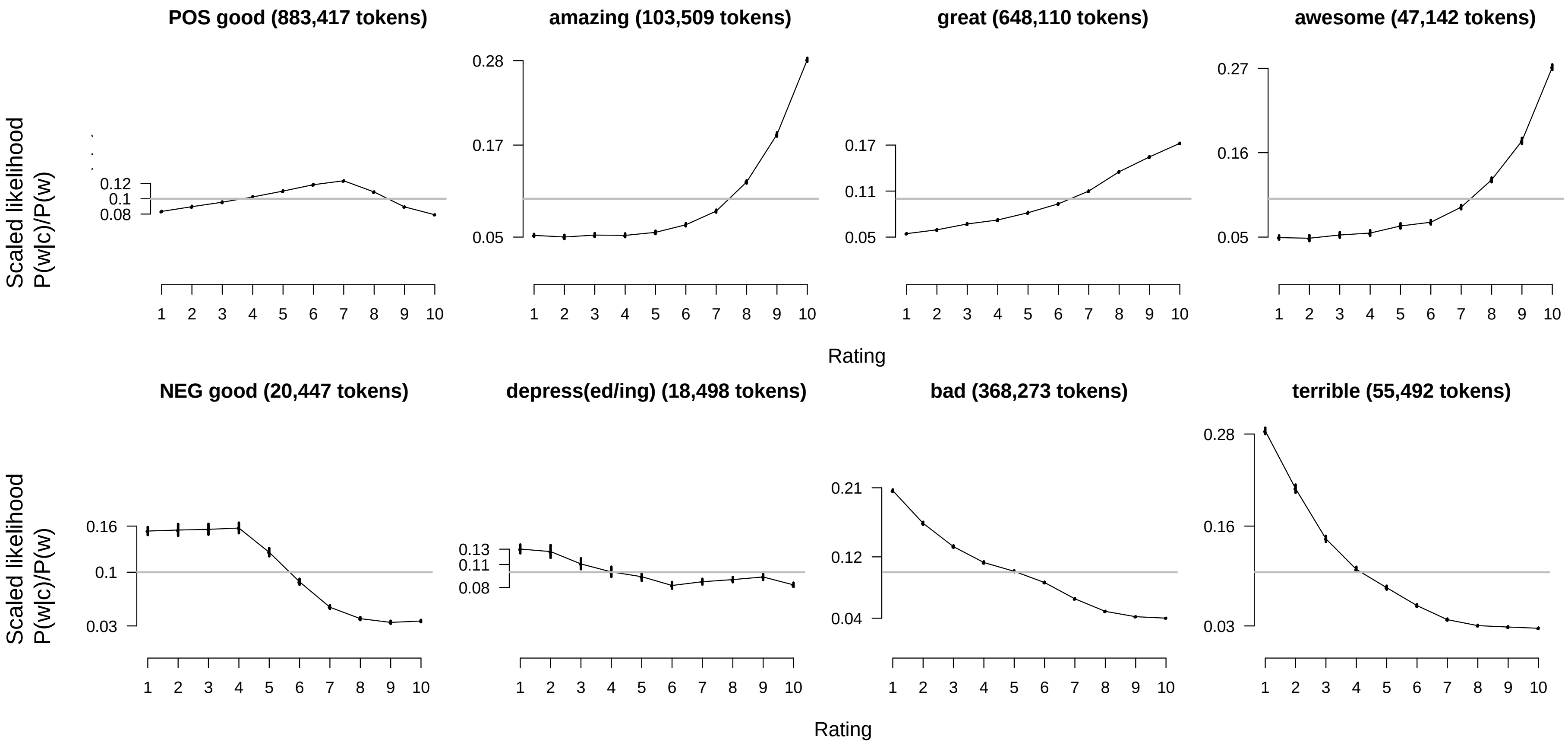
- Make them comparable between words
  - **Scaled likelihood:**

$$\frac{P(w|c)}{P(w)}$$

Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

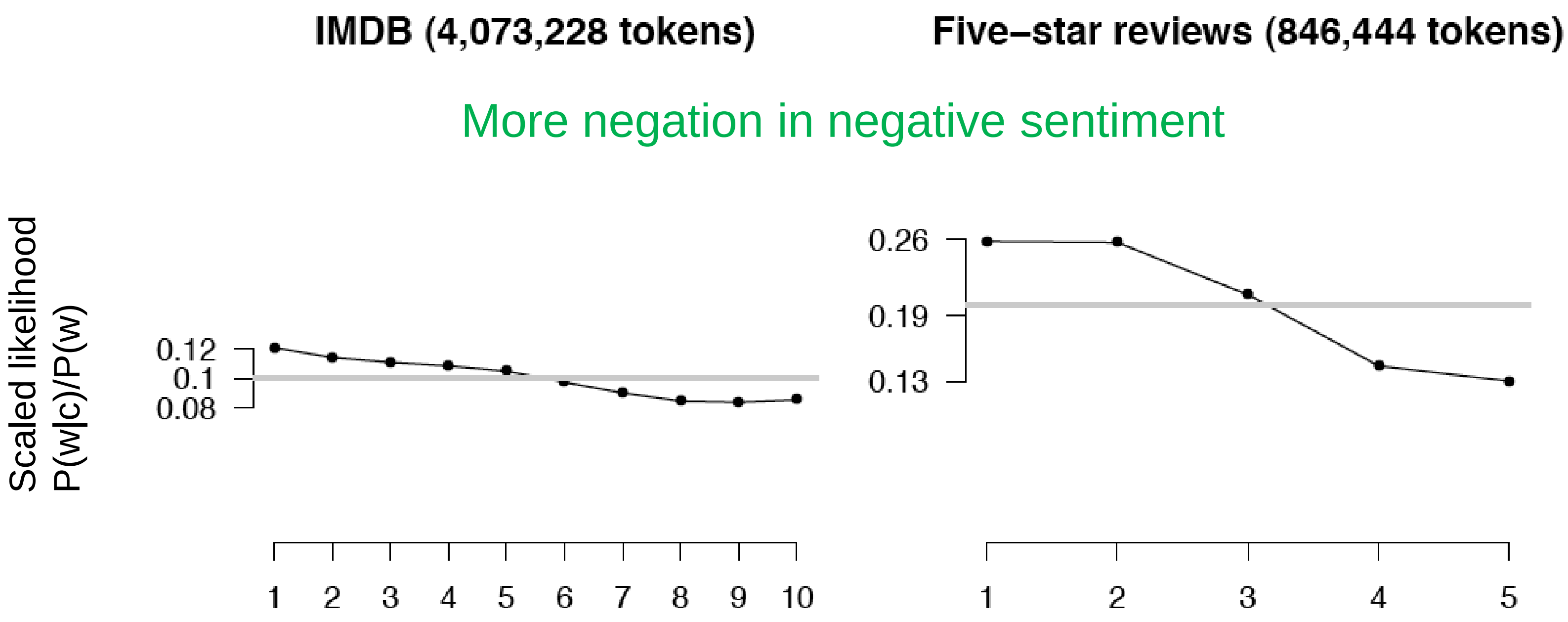
Analyzing the Polarity of each Word in IMDB

Potts, Christopher. 2011. On the negativity of negation, 636-659.



Other Sentiment Feature: Logical Negation

- Is logical negation (*no, not, never*) associated with negative sentiment?
  - Count negation (*not, n't, no, never*) in online reviews
  - Regress against the review rating



Agenda

- Applications of Sentiment Analysis
- Word-based Classification Approach
- Sentiment Lexicons and Negation Words
- **Learning Sentiment Lexicons**

Semi-supervised Learning of Lexicons

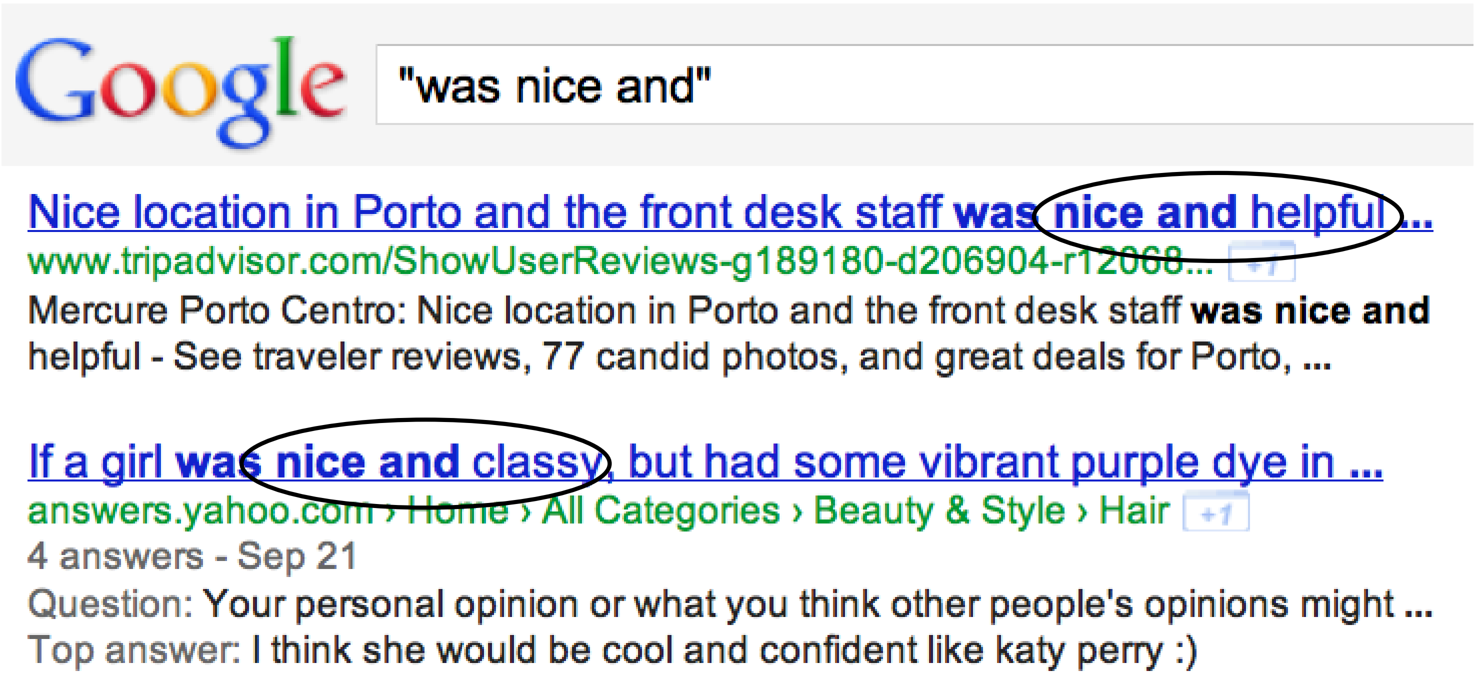
- Use a small amount of information
  - A few labeled examples
  - A few hand-built patterns
- To bootstrap a lexicon using this small amount of information
  - Just increase lexicon size
  - Adapt lexicon to another domain

Using “and” and “but”

- Adjectives conjoined by “*and*” have same polarity
  - Fair **and** legitimate, corrupt **and** brutal – high frequency
  - *fair **and** brutal, *corrupt **and** legitimate – low frequency
- Adjectives conjoined by “*but*” do not
  - fair **but** brutal – high frequency
- Step 1: Label seed set of 1336 adjectives (all >20 in 21 million word WSJ corpus): 657 positive and 679 negative

Using “and” and “but”: Step 2

- Expand seed set to conjoined adjectives

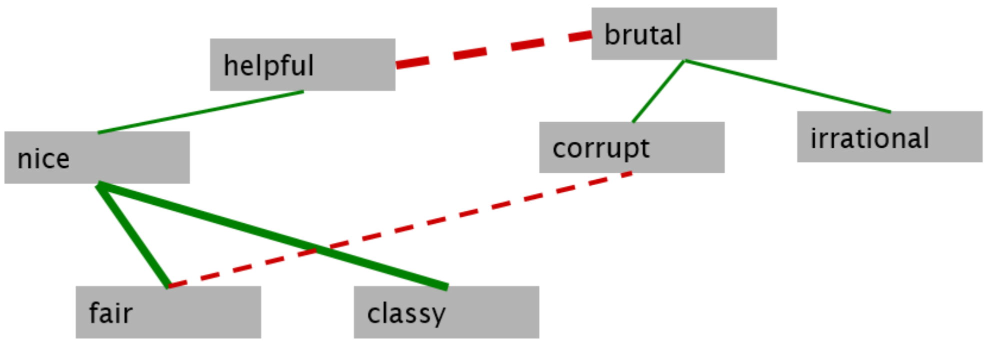


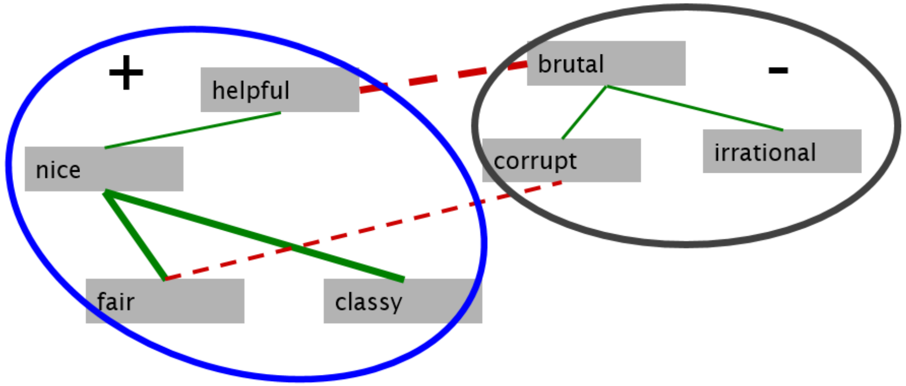
nice, helpful

nice, classy

Using “and” and “but”

- Step 3: Supervised classifier assigns “polarity similarity” to each word pair, resulting in graph:
  - Features are count(and), count(but), …
- Step 4: Clustering for partitioning the graph into two





Turney Algorithm

1. Extract a *phrasal lexicon* from reviews
2. Learn polarity of each phrase
3. Rate a review by the average polarity of its phrases

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

JJ=adjective, NN=noun, NNS=plural noun, RB=adverb, VB=verb

How to Measure Polarity of a Phrase?

- Positive phrases co-occur (on the web) more with *"excellent"*
- Negative phrases co-occur (on the web) more with *"poor"*
- But how to measure co-occurrence?
  - PMI between two words measures how much more do two words co-occur than if they were independent.
    - $PMI(w_1, w_2) = \log_2 \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$
    - Estimate P(w) by issuing "w" as a query to search engine and looking at #results
    - Estimate P($w_1 w_2$) by issuing "w_1 w_2" as a query to search engine and looking at #results
- Polarity of phrase = PMI(phrase, "excellent") − PMI(phrase, "poor")

Phrases from a Thumbs-Up Review

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
…		
low fees	JJ NNS	0.33
other bank	JJ NN	−0.85
inconveniently located	JJ NN	−1.5
Average		0.32

Phrases from a Thumbs-Down Review

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
…		
virtual monopoly	JJ NN	−2.0
other problems	JJ NNS	−2.8
low funds	JJ NNS	−6.8
unethical practices	JJ NNS	−8.5
Average		−1.2

Results of Turney Algorithm

- 410 reviews from Epinions (http://www.epinions.com/ review website)
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- Turney algorithm: 74%

- Phrases rather than words
- Semi-supervised algorithms like Turney's help us learn domain-specific information

Using WordNet to Learn Polarity

- WordNet: online thesaurus.
- Create positive ("good") and negative seed-words ("terrible")
- Find Synonyms and Antonyms
  - Positive Set: Add synonyms of positive words ("well") and antonyms of negative words
  - Negative Set: Add synonyms of negative words ("awful") and antonyms of positive words ("evil")
- Repeat, following chains of synonyms
- Filter out bad examples or wrong word senses.

Summary on Learning Lexicons

- Advantages
  - Can be domain-specific
  - Can be more robust (more words)
- Intuition
  - Start with a seed set of words (‘good’, ‘poor’)
  - Find other words that have similar polarity:
    - Using “and” and “but”
    - Using words that occur nearby in the same document (“poor” or “excellent”)
    - Using WordNet synonyms and antonyms

Take-aways

- Sentiment analysis is very useful for many web portals.
- Word-based classification works reasonably well.
- There are standard sentiment lexicons for word-based sentiment classification algorithms
- Using semi-supervised learning one can grow a small sentiment lexicon into larger ones.

Further Reading

- Book on Sentiment Analysis and Opinion Mining
  - https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
- Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press
- Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in
- Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.
- Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

Further Reading

- Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010
- Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.
- Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181
- Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews
- S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. COLING 2004
- M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, 2004

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals:	+91-9502334561/63 or 040-65743991
For Corporates:	+91-9618483483
Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering