THE UNIVERSITY OF CHICAGO
GRAHAM SCHOOL
CONTINUING LIBERAL AND PROFESSIONAL STUDIES

Natural Language Processing
Session 4

Nick Kadochnikov

Session 4 Agenda

- Introduction to text classification
- Sentiment analysis
- Maximum entropy classifiers

Text Classification and Naïve Bayes

The Task of Text Classification

Is this spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

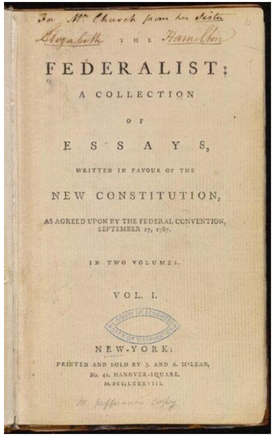
Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

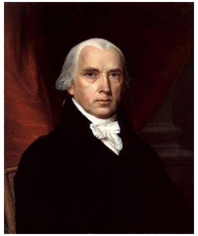
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

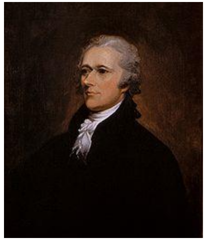
Who wrote which Federalist papers?



- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



James Madison



Alexander Hamilton

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, volume 23, number 3, pp. 321–346

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.

What is the subject of this article?

MEDLINE Article


Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Text Classification: definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- *Input:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2,\ldots, c_J\}$
  - A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma:d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
  - …

Text Classification and Naïve Bayes

Naïve Bayes (I)

Naïve Bayes Intuition

- Simple ("naïve") classification method based on Bayes rule
- Relies on very simple representation of document
  - Bag of words

The bag of words representation

$\gamma($

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

$) = c$

The bag of words representation

$\gamma($

I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun**… It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet.

$) = c$

The bag of words representation: using a subset of words

$\gamma($

```
x love xxxxxxxxxxxxxxxx sweet
xxxxxxx satirical xxxxxxxxxx
xxxxxxxxxxx great xxxxxxx
xxxxxxxxxxxxxxxxxxx fun xxxx
xxxxxxxxxxxxx whimsical xxxx
romantic xxxx laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx recommend xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx several xxxxxxxxxxxxxxxxx
xxxxx happy xxxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
```

$)=c$

The bag of words representation

$\gamma($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

$)=c$

Multinomial Naïve Bayes Independence Assumptions

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Bag of Words assumption**: Assume position doesn't matter
- **Conditional Independence**: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) \cdot \ldots \cdot P(x_n \mid c)$$

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

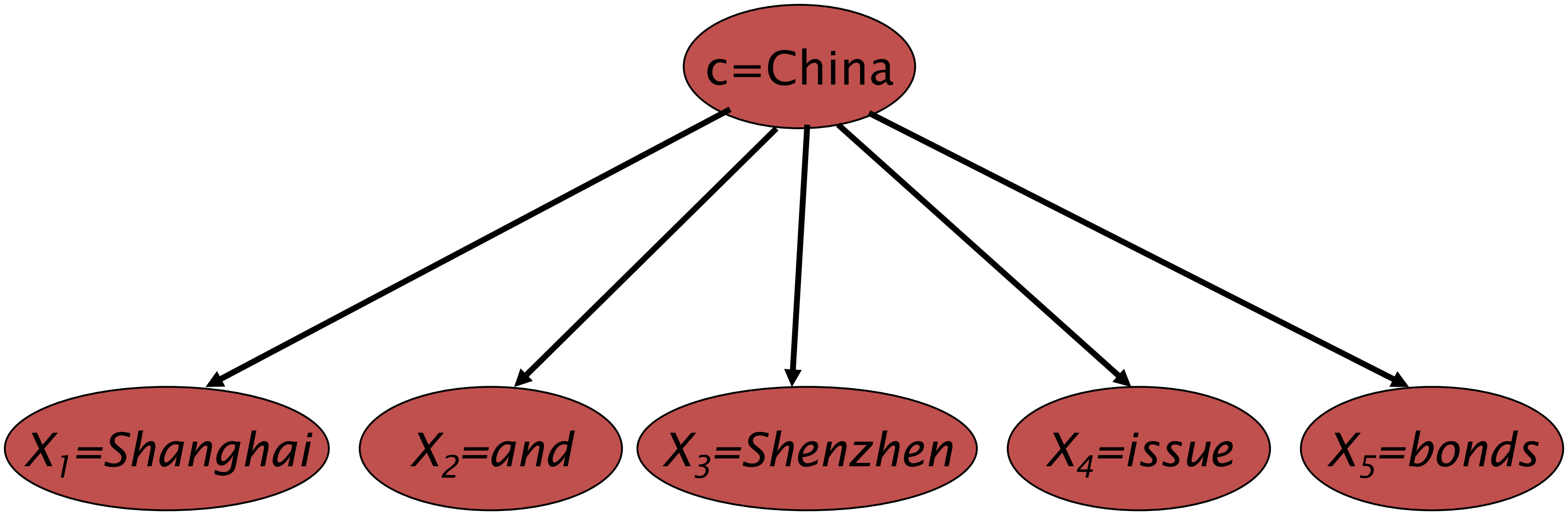
Parameter estimation

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic *j* by concatenating all docs in this topic
  - Use frequency of *w* in mega-document

Generative Model for Multinomial Naïve Bayes



Naïve Bayes and Language Modeling

- Naïve bayes classifiers can use any sort of feature
  - URL, email address, dictionaries, network features
- But if, as in the previous slides
  - We use **only** word features
  - we use **all** of the words in the text (not a subset)
- Then
  - Naïve bayes has an important similarity to language modeling.

Each class = a unigram language model

- Assigning each word: P(word | c)
- Assigning each sentence: P(s|c)=∠P(word|c)

Class *pos*

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

I	love	this	fun	film
0.1	0.1	.05	0.01	0.1

P(s | pos) = 0.0000005

Naïve Bayes as a Language Model

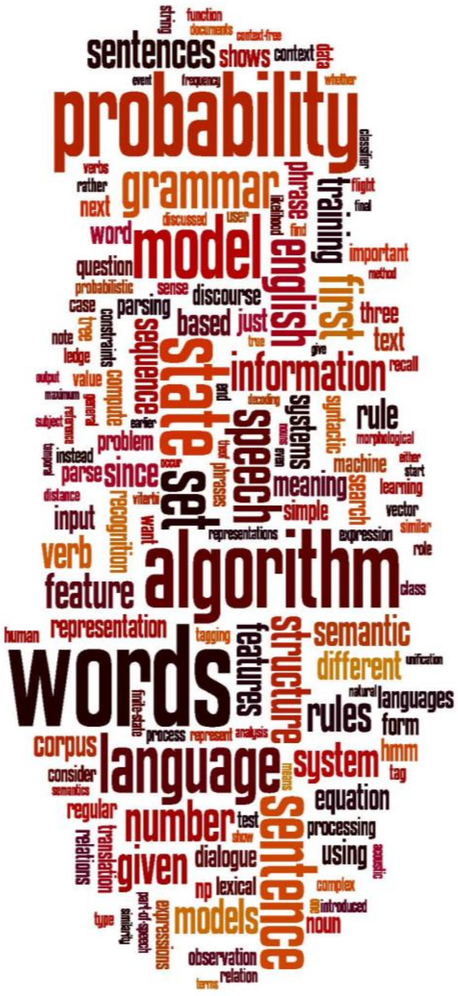
- Which class assigns the higher probability to s?

Model pos	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$P(s|pos) > P(s|neg)$



Text Classification and Naïve Bayes

Multinomial Naïve Bayes: A Worked Example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	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	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Choosing a class:

P(c|d5) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14 $\approx$ 0.0003

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

P(j|d5) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9 $\approx$ 0.0001

Naïve Bayes in Spam Filtering

- SpamAssassin Features:
  - Mentions Generic Viagra
  - Online Pharmacy
  - Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - Phrase: impress ... girl
  - From: starts with many numbers
  - Subject is all capitals
  - HTML has a low ratio of text to image area
  - One hundred percent guaranteed
  - Claims you can be removed from the list
  - 'Prestigious Non-Accredited Universities'
  - http://spamassassin.apache.org/tests_3_3_x.html

Summary: Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Robust to Irrelevant Features

  Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features

  Decision Trees suffer from *fragmentation* in such cases – especially if little data
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
  - **But we will see other classifiers that give better accuracy**

Text Classification and Naïve Bayes

Text Classification: Evaluation and Practical Issues

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Precision and recall

- **Precision**: % of selected items that are correct
  Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

The Real World

- Gee, I’m building a text classifier for real, now!
- What should I do?

No training data? Manually written rules

If (wheat or grain) and not (whole or bread) then

Categorize as grain

- Need careful crafting
  - Human tuning on development data
  - Time-consuming: 2 days per class

Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a "high-bias" algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …

A reasonable amount of data?

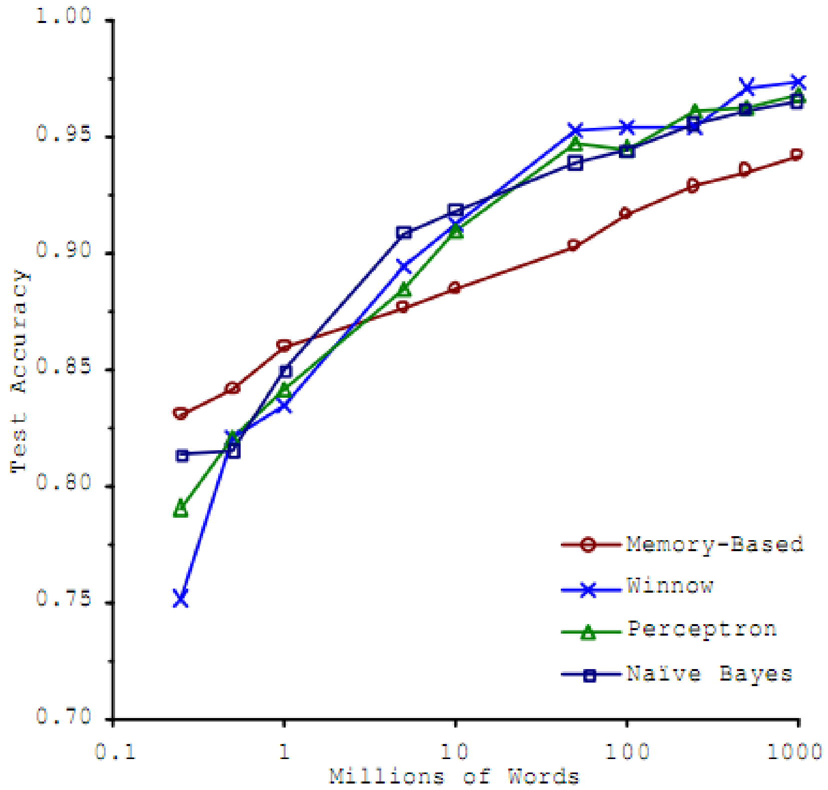
- Perfect for all the clever classifiers
  - SVM
  - Regularized Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
  - Management likes quick fixes

A huge amount of data?

- Can achieve high accuracy!
- At a cost:
  - SVMs (train time) or kNN (test time) can be too slow
  - Regularized logistic regression can be somewhat better
- So Naïve Bayes can come back into its own again!

Accuracy as a function of data size

- With enough data
  - Classifier may not matter



Brill and Banko on spelling correction

Real-world systems generally combine:

- Automatic classification
- Manual review of uncertain/difficult/"new” cases

How to tweak performance

- Domain-specific features and weights: *very* important in real performance
- Sometimes need to collapse terms:
  - Part numbers, chemical formulas, …
  - But stemming generally doesn't help
- Upweighting: Counting a word as if it occurred twice:
  - title words (Cohen & Singer 1996)
  - first sentence of each paragraph (Murata, 1999)
  - In sentences that contain title words (Ko *et al*, 2002)

Text Classification in Python

Sentiment Analysis

What is Sentiment Analysis?

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.

Google 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

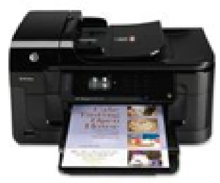
What people are saying

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."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

Product summary   Find best price   **Customer reviews**   Specifications   Related items



$121.53 - $242.39 (14 stores)

☐ Compare

Average rating ★★★½☆ (144)

★★★★★	(55)
★★★★☆	(54)
★★★☆☆	(10)
★★☆☆☆	(6)
★☆☆☆☆	(23)
☆☆☆☆☆	(0)

Most mentioned

Performance	(57)
Ease of Use	(43)
Print Speed	(39)
Connectivity	(31)

More ▼

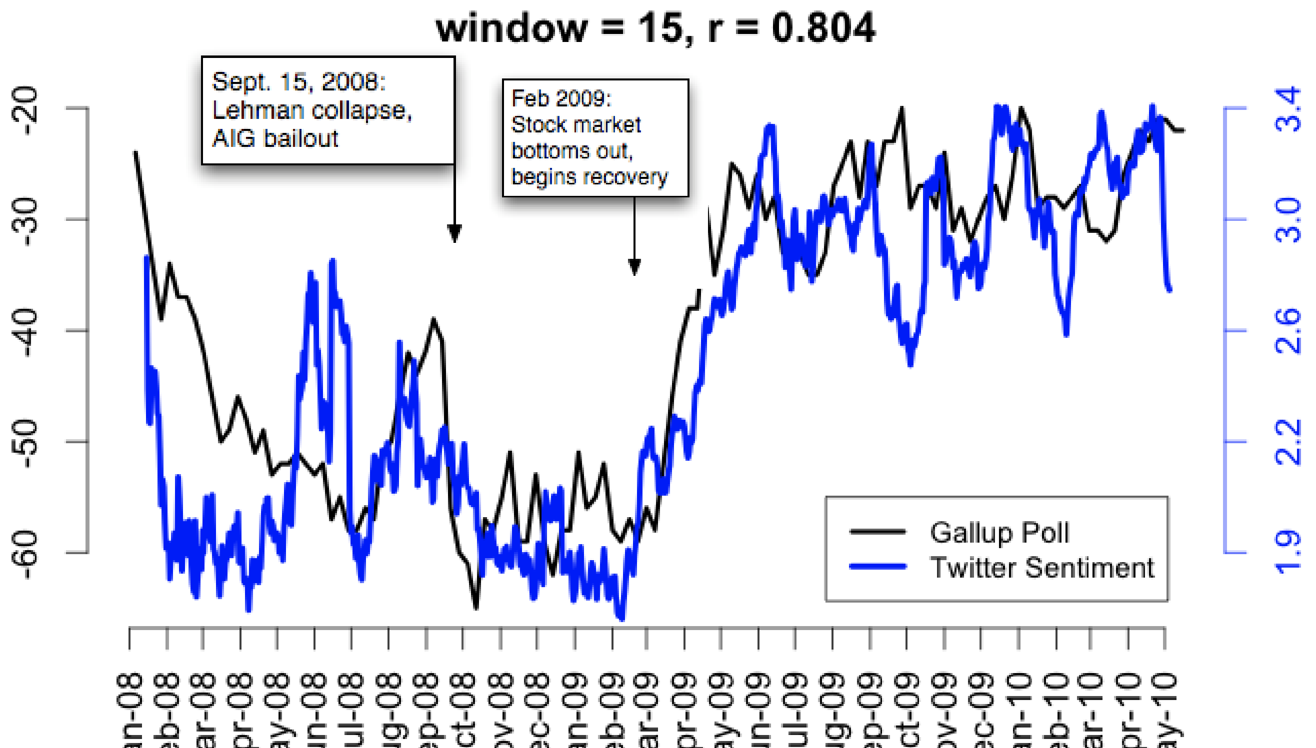
Show reviews by source

Best Buy (140)
CNET (5)
Amazon.com (3)

Dan Jurafsky

Twitter sentiment versus Gallup Poll of Consumer Confidence

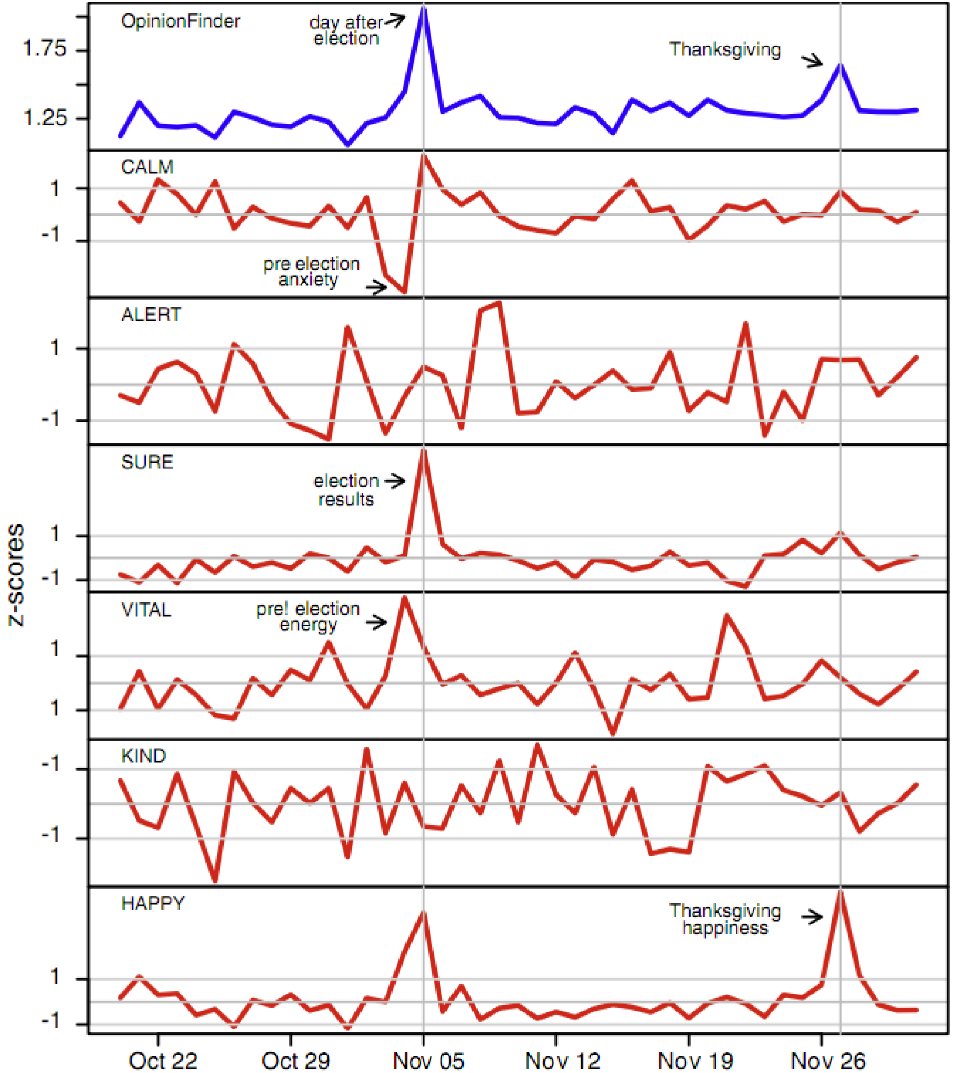
Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010



Twitter sentiment:

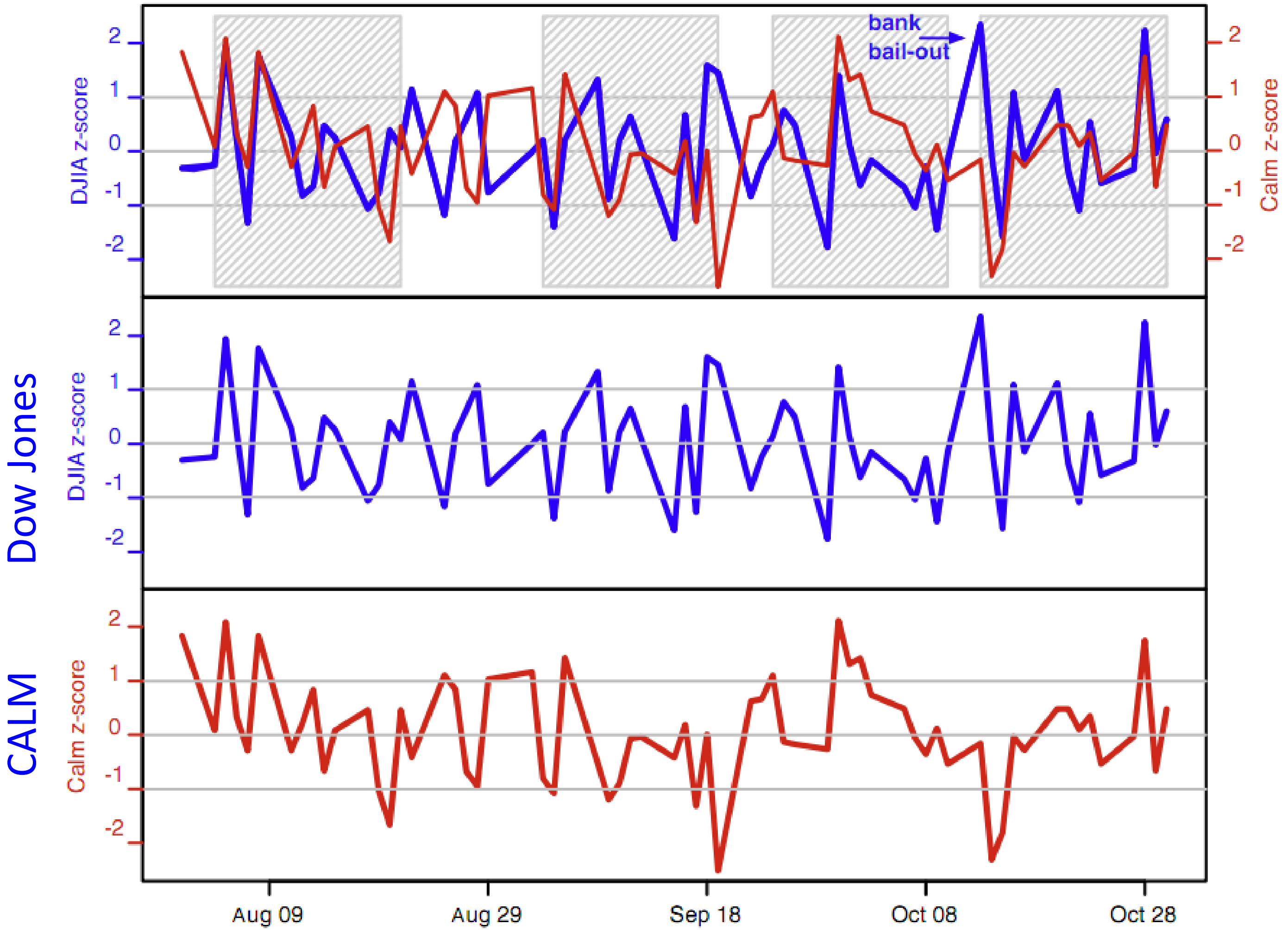
Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. [Twitter mood predicts the stock market,]

Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.



Bollen et al. (2011)

- CALM predicts DJIA 3 days later
- At least one current hedge fund uses this algorithm



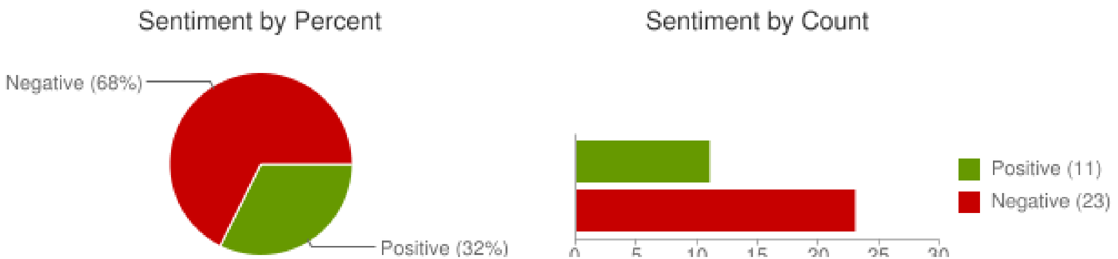
Target Sentiment on Twitter

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

Type in a word and we'll highlight the good and the bad

"united airlines" Search Save this search

Sentiment analysis for "united airlines"



jljacobson: OMG... Could **@United airlines** have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate **United Airlines** Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 **united airlines** and 24seven to an exotic destination. http://t.co/Z9QloAjF
Posted 2 hours ago

CountAdam: FANTASTIC customer service from **United Airlines** at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago

Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

Why sentiment analysis?

- *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

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - *liking, loving, hating, valuing, desiring*
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - *nervous, anxious, reckless, morose, hostile, jealous*

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons**
  - *liking, loving, hating, valuing, desiring*
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - *nervous, anxious, reckless, morose, hostile, jealous*

Sentiment Analysis

- Sentiment analysis is the detection of **attitudes**

  "enduring, affectively colored beliefs, dispositions towards objects or persons"

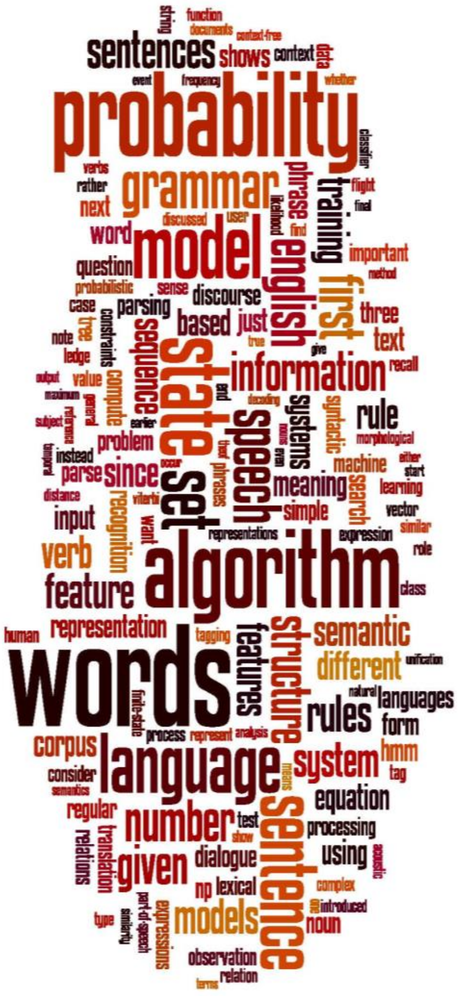
  1. **Holder (source)** of attitude
  2. **Target (aspect)** of attitude
  3. **Type** of attitude
     - From a set of types
       - *Like, love, hate, value, desire,* etc.
     - Or (more commonly) simple weighted **polarity**:
       - *positive, negative, neutral,* together with *strength*
  4. **Text** containing the attitude
     - Sentence or entire document

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

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



Sentiment Analysis

A Baseline Algorithm

Sentiment Classification in Movie Reviews

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

- Polarity detection:
  - Is an IMDB movie review positive or negative?
- Data: *Polarity Data 2.0:*
  - http://www.cs.cornell.edu/people/pabo/movie-review-data

IMDB data in the Pang and Lee database

✓

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 .

Baseline Algorithm (adapted from Pang and Lee)

- Tokenization
- Feature Extraction
- Classification using different classifiers
  - Naïve Bayes
  - MaxEnt
  - SVM

Sentiment Tokenization Issues

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

Potts emoticons

```
[<>]?                          # optional hat/brow
[:;=8]                         # eyes
[\-o\*\']?                     # optional nose
[\)\]\(\[dDpP/\:\}\{@\|\\]     # mouth
|                              #### reverse orientation
[\)\]\(\[dDpP/\:\}\{@\|\\]     # mouth
[\-o\*\']?                     # optional nose
[:;=8]                         # eyes
[<>]?                          # optional hat/brow
```

Extracting Features for Sentiment Classification

- How to handle negation
  - `I **didn't** like this movie`

    vs
  - `I really like this movie`
- Which words to use?
  - Only adjectives
  - All words
    - All words turns out to work better, at least on this data

Negation

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

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
```

Reminder: Naïve Bayes

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

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Binarized (Boolean feature) Multinomial Naïve Bayes

- Intuition:
  - For sentiment (and probably for other text classification domains)
  - 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
      - 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 $Text_j$

$$P(w_k | c_j) \neg \frac{n_k + a}{n + a|Vocabulary|}$$

Boolean Multinomial Naïve Bayes on a test document *d*

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

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

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	?

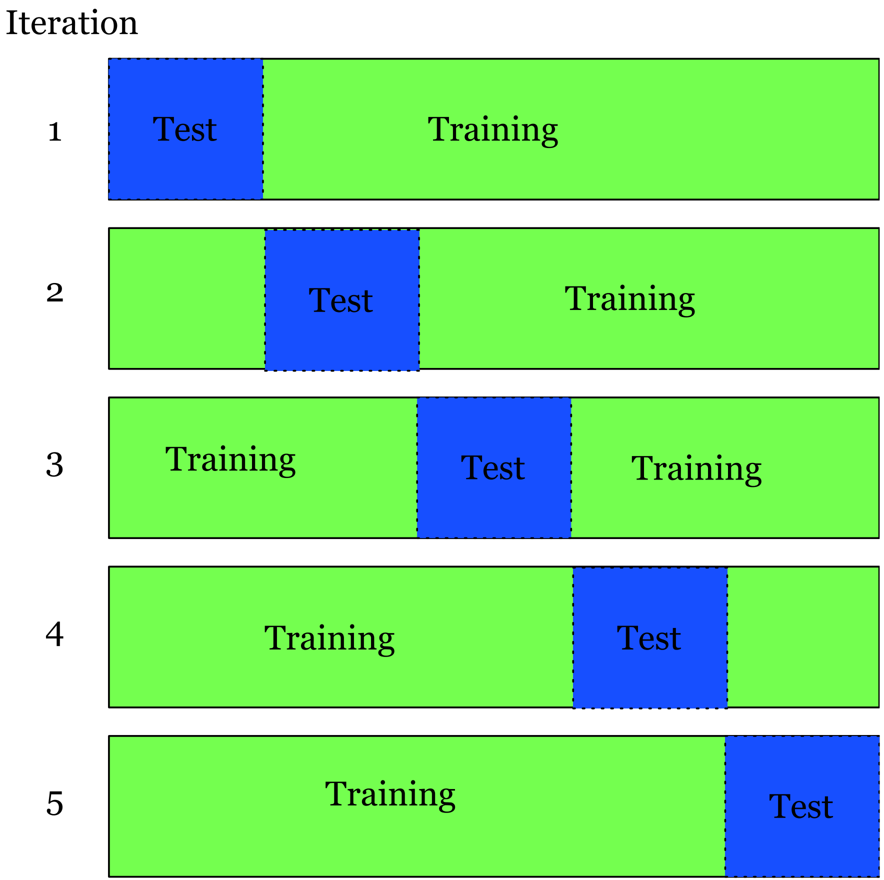
Binarized (Boolean feature) Multinomial Naïve Bayes

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
V. Metsis, I. Androutsopoulos, G. Paliouras. 2006. Spam Filtering with Naive Bayes – Which Naive Bayes? CEAS 2006 - Third Conference on Email and Anti-Spam.
K.-M. Schneider. 2004. On word frequency information and negative evidence in Naive Bayes text classification. ICANLP, 474-485.
JD Rennie, L Shih, J Teevan. 2003. Tackling the poor assumptions of naive bayes text classifiers. ICML 2003

- Binary seems to work better than full word counts
  - This is **not** the same as Multivariate Bernoulli Naïve Bayes
    - MBNB doesn't work well for sentiment or other text tasks
- Other possibility: log(freq(*w*))

Cross-Validation

- Break up data into 10 folds
  - (Equal positive and negative inside each fold?)
- For each fold
  - Choose the fold as a temporary test set
  - Train on 9 folds, compute performance on the test fold
- Report average performance of the 10 runs



Other issues in Classification

- MaxEnt and SVM tend to do better than Naïve Bayes

Problems: What makes reviews hard to classify?

- Subtlety:
  - 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."
  - Dorothy Parker on Katherine Hepburn
    - "She runs the gamut of emotions from A to B"

Thwarted Expectations and Ordering Effects

- "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**."
- Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is **not so good** either, I was surprised.

Sentiment Analysis

Sentiment Lexicons

The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Home page: http://www.wjh.harvard.edu/~inquirer
- List of Categories: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Spreadsheet: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls
- Categories:
  - Positiv (1915 words) and Negativ (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated versus Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for Research Use

LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

- Home page: http://www.liwc.net/
- 2300 words, >70 classes
- **Affective Processes**
  - negative emotion (*bad, weird, hate, problem, tough*)
  - positive emotion (*love, nice, sweet*)
- **Cognitive Processes**
  - Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)
- **Pronouns, Negation** (*no, never*), **Quantifiers** (*few, many*)
- $30 or $90 fee

MPQA Subjectivity Cues Lexicon

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.

- Home page: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- 6885 words from 8221 lemmas
  - 2718 positive
  - 4912 negative
- Each word annotated for intensity (strong, weak)
- GNU GPL

Bing Liu Opinion Lexicon

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.

- Bing Liu's Page on Opinion Mining
- http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

- 6786 words
  - 2006 positive
  - 4783 negative

SentiWordNet

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

- Home page: http://sentiwordnet.isti.cnr.it/
- 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 .75  Neg 0    Obj .25
```

Disagreements between polarity lexicons

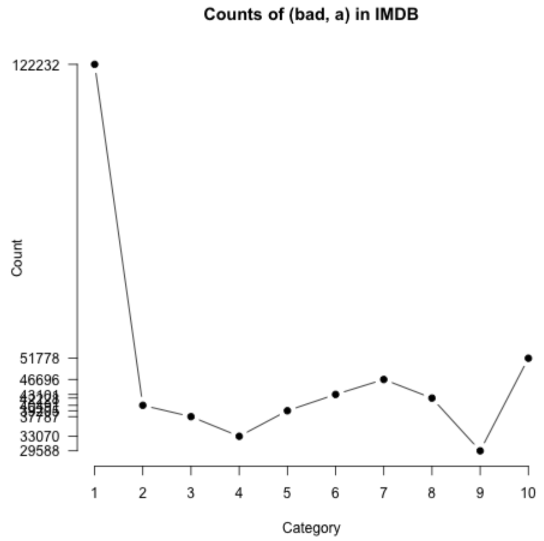
Christopher Potts, Sentiment Tutorial, 2011

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**	12/363 **(3%)**
Opinion Lexicon		32/2411 **(1%)**	1004/3994 **(25%)**	9/403 **(2%)**
General Inquirer			520/2306 **(23%)**	1/204 **(0.5%)**
SentiWordNet				174/694 **(25%)**
LIWC				

Analyzing the polarity of each word in IMDB

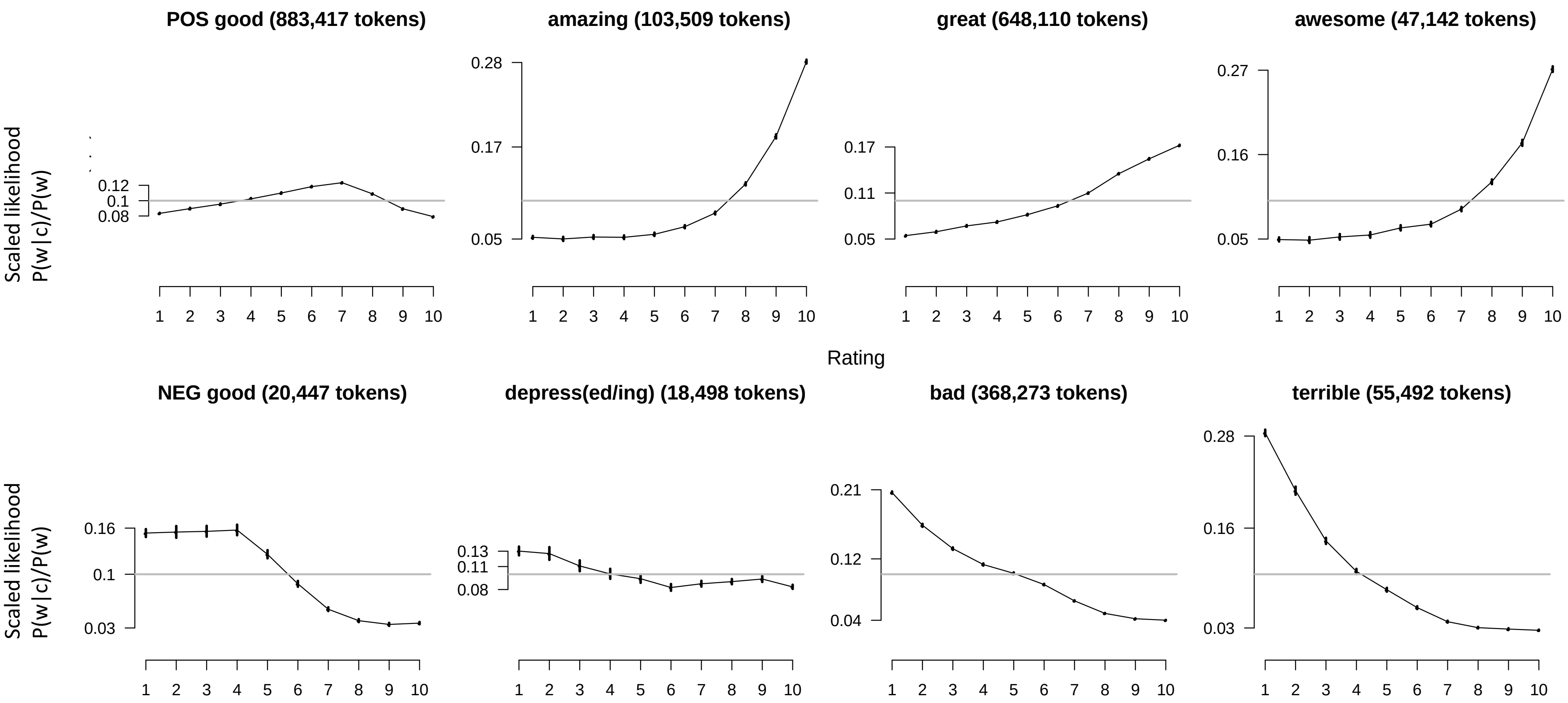
Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

- 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:
- Instead, **likelihood:** $$P(w|c) = \frac{f(w,c)}{\sum_{w \in c} f(w,c)}$$
- Make them comparable between words
  - **Scaled likelihood:** $$\frac{P(w|c)}{P(w)}$$



Analyzing the polarity of each word in IMDB

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



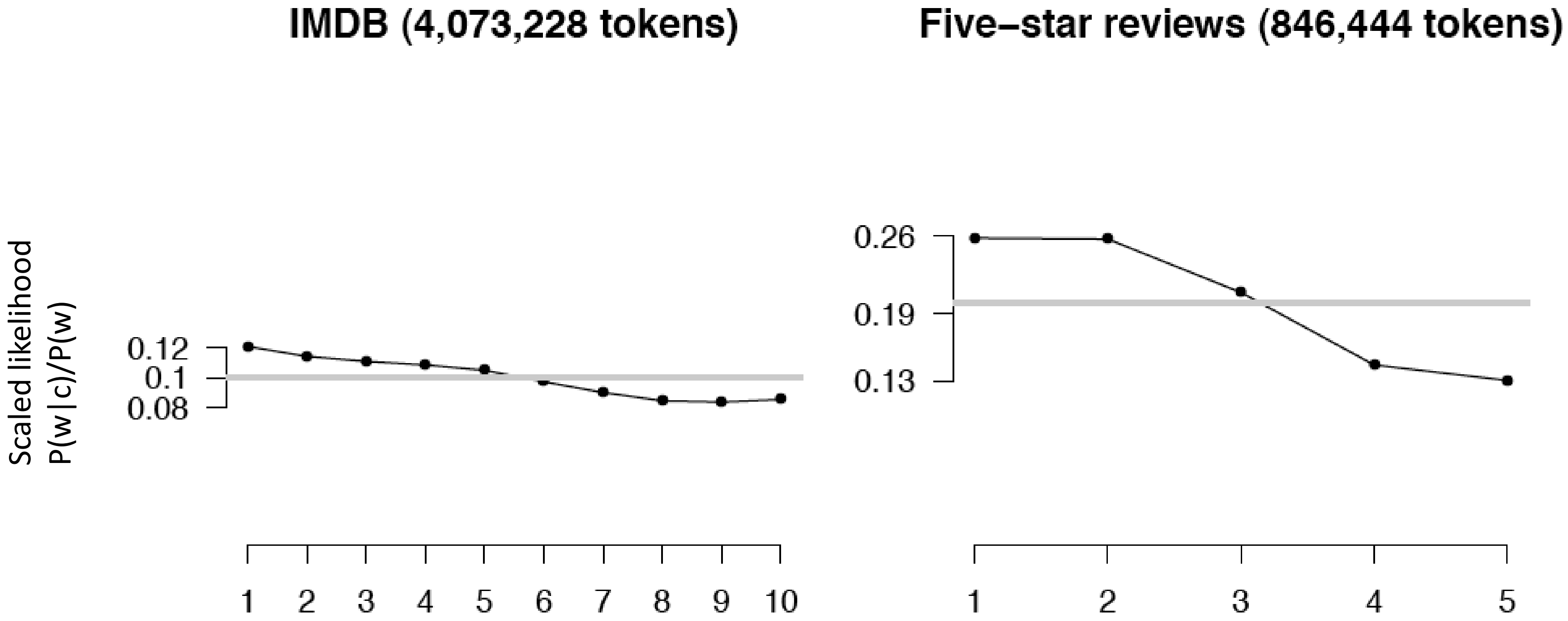
Other sentiment feature: Logical negation

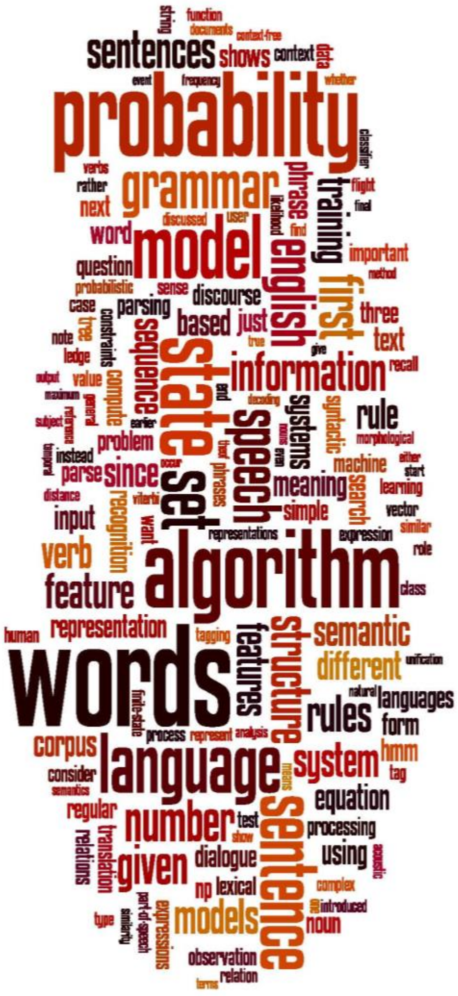
Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

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

Dan Jurafsky

Potts 2011 Results: More negation in negative sentiment





Sentiment Analysis

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

Hatzivassiloglou and McKeown intuition for identifying word polarity

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

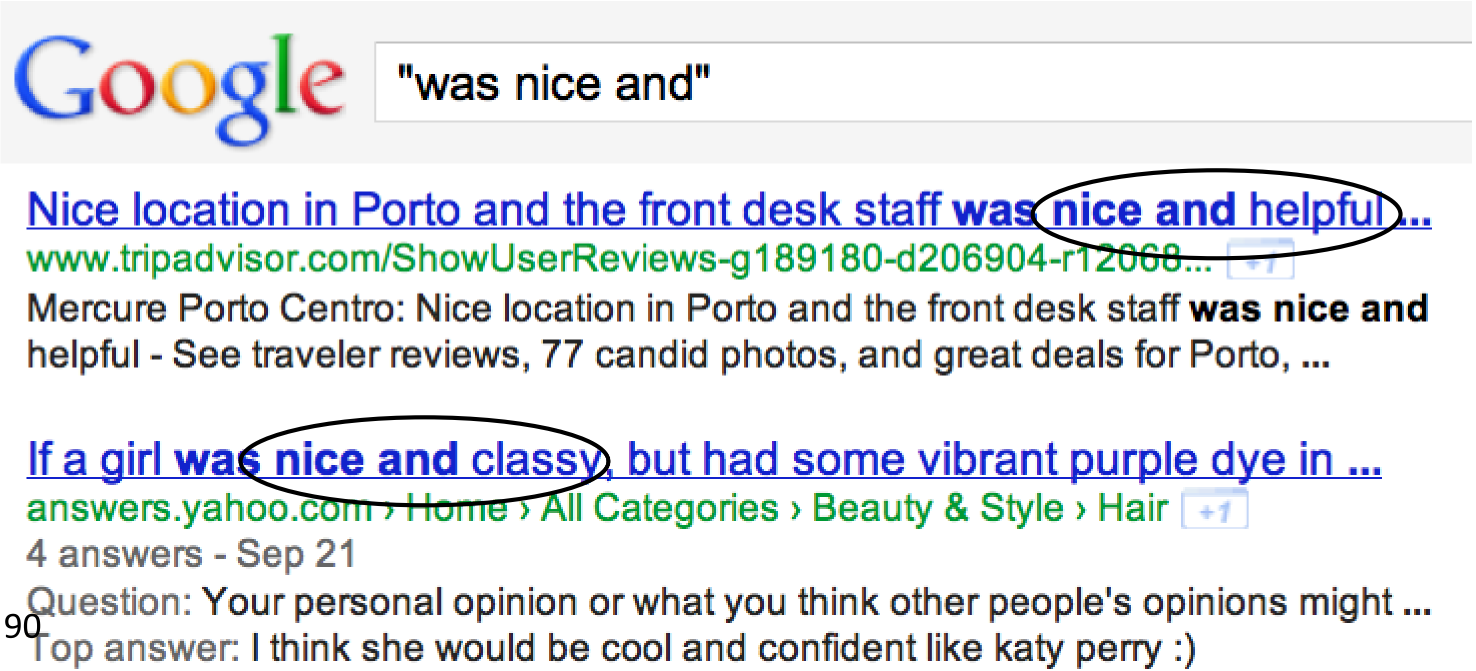
- Adjectives conjoined by "*and*" have same polarity
  - Fair **and** legitimate, corrupt **and** brutal
  - *fair **and** brutal, *corrupt **and** legitimate
- Adjectives conjoined by "*but*" do not
  - fair **but** brutal

Hatzivassiloglou & McKeown 1997 Step 1

- Label **seed set** of 1336 adjectives (all >20 in 21 million word WSJ corpus)
  - 657 positive
    - adequate central clever famous intelligent remarkable reputed sensitive slender thriving…
  - 679 negative
    - contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting…

Hatzivassiloglou & McKeown 1997
Step 2

- Expand seed set to conjoined adjectives

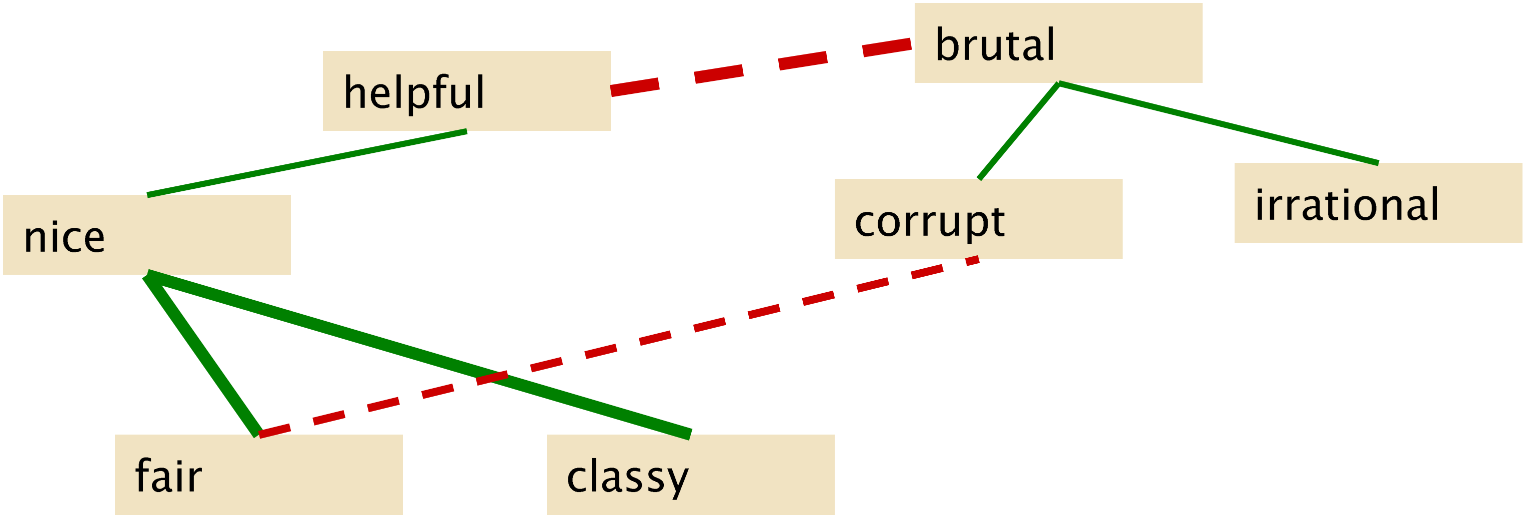


nice, helpful

nice, classy

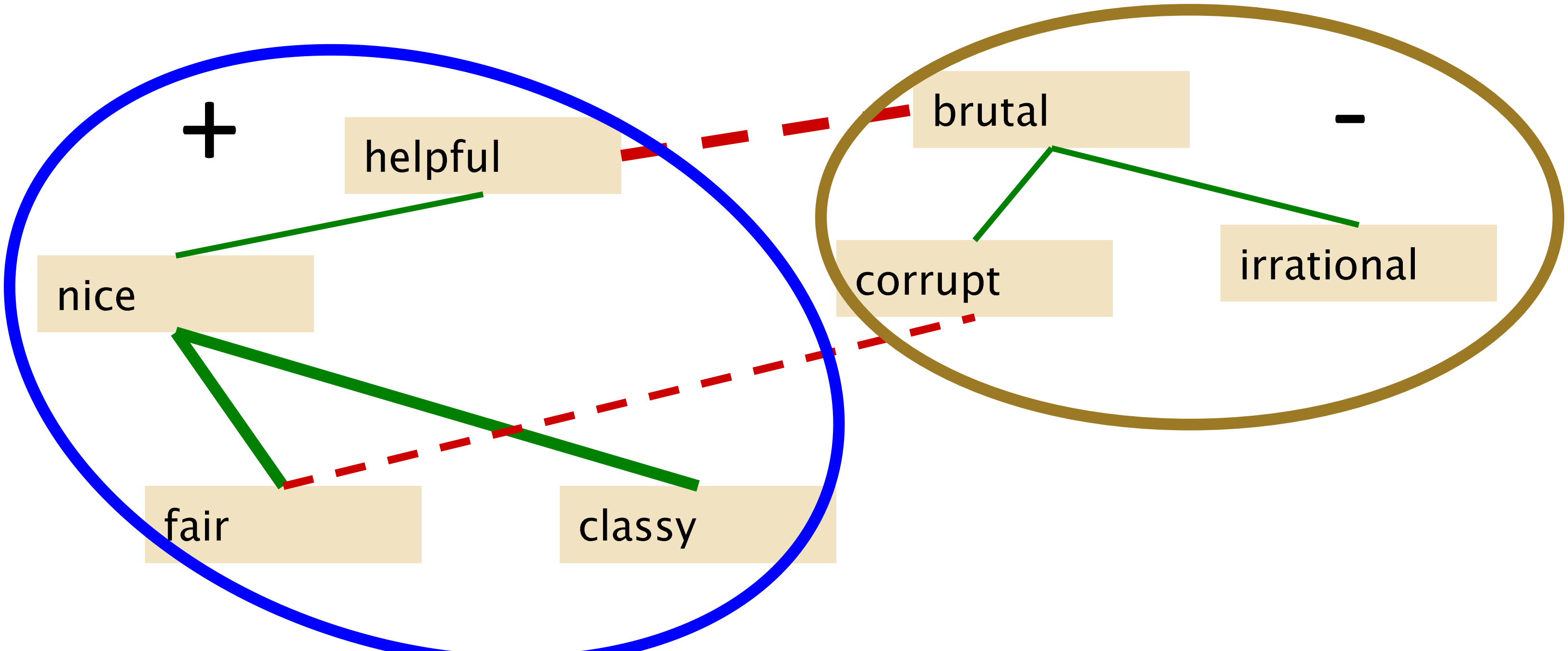
Hatzivassiloglou & McKeown 1997
Step 3

- Supervised classifier assigns "polarity similarity" to each word pair, resulting in graph:



Hatzivassiloglou & McKeown 1997 Step 4

- Clustering for partitioning the graph into two



Output polarity lexicon

- Positive
  - bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty…
- Negative
  - ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful…

Output polarity lexicon

- Positive
  - bold decisive **disturbing** generous good honest important large mature patient peaceful positive proud sound stimulating straightforward **strange** talented vigorous witty…
- Negative
  - ambiguous **cautious** cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor **outspoken pleasant** reckless risky selfish tedious unsupported vulnerable wasteful…

Turney Algorithm

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

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

Extract two-word phrases with adjectives

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

How to measure polarity of a phrase?

- Positive phrases co-occur more with *"excellent"*
- Negative phrases co-occur more with *"poor"*
- But how to measure co-occurrence?

How to Estimate Pointwise Mutual Information

- Query search engine (Altavista)
  - P(word) estimated by `hits(word)/N`
  - P($word_1$,$word_2$) by `hits(word1 NEAR word2)/N`^{2}

$$\text{PMI}(word_1, word_2) = \log_2 \frac{hits(word_1 \text{ NEAR } word_2)}{hits(word_1)hits(word_2)}$$

Does phrase appear more with “poor” or “excellent”?

$$\text{Polarity}(\mathit{phrase}) = \text{PMI}(\mathit{phrase}, \text{"excellent"}) - \text{PMI}(\mathit{phrase}, \text{"poor"})$$

$$= \log_2 \frac{\text{hits}(\mathit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\mathit{phrase})\text{hits}(\text{"excellent"})} - \log_2 \frac{\text{hits}(\mathit{phrase}\ \text{NEAR "poor"})}{\text{hits}(\mathit{phrase})\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(\mathit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\mathit{phrase})\text{hits}(\text{"excellent"})} \frac{\text{hits}(\mathit{phrase})\text{hits}(\text{"poor"})}{\text{hits}(\mathit{phrase}\ \text{NEAR "poor"})}$$

$$= \log_2 \left(\frac{\text{hits}(\mathit{phrase}\ \text{NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(\mathit{phrase}\ \text{NEAR "poor"})\text{hits}(\text{"excellent"})} \right)$$

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
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5

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
lesser evil	RBR JJ	-2.3
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
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- Turney algorithm: 74%

- Phrases rather than words
- Learns domain-specific information

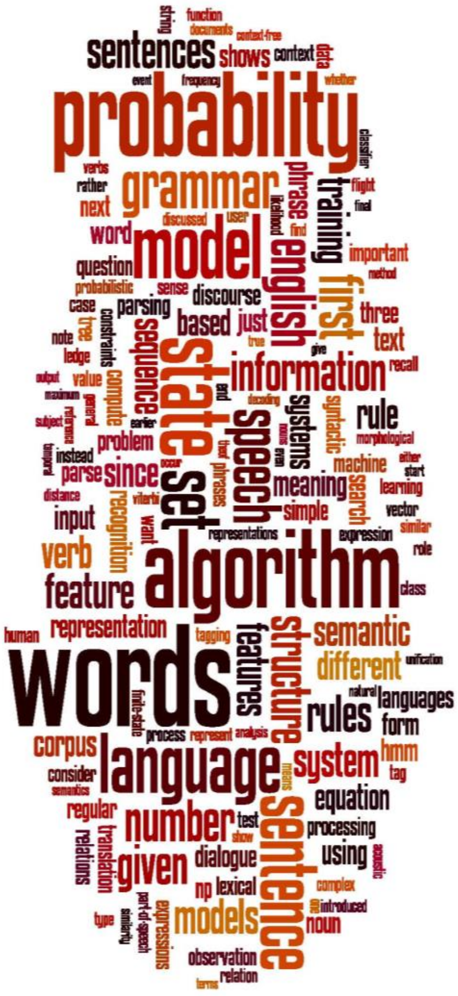
Using WordNet to learn polarity

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

- WordNet: online thesaurus (covered in later lecture).
- 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

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
    - Using WordNet synonyms and antonyms



Sentiment Analysis

Other Sentiment Tasks

Finding sentiment of a sentence

- Important for finding aspects or attributes
  - Target of sentiment

- `The food was great but the service was awful`

Finding aspect/attribute/target of sentiment

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In Proceedings of KDD.
S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop.

- Frequent phrases + rules
  - Find all highly frequent phrases across reviews (“`fish tacos`”)
  - Filter by rules like “occurs right after sentiment word”
    - “`...great fish tacos`” means `fish tacos` a likely aspect

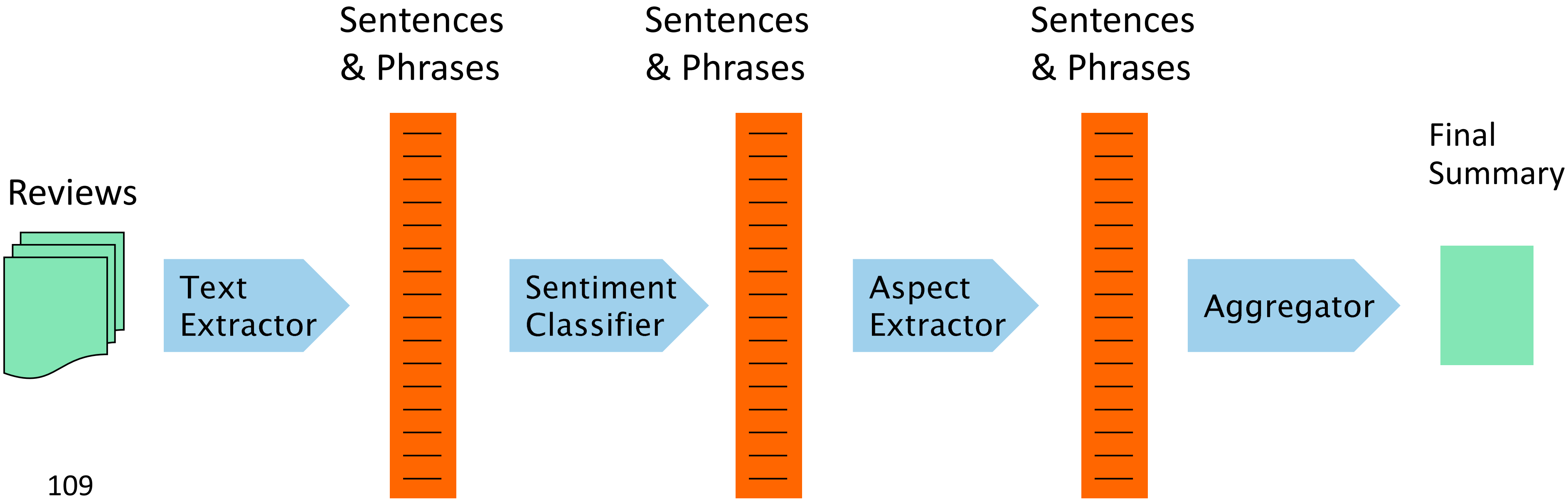
Casino	casino, buffet, pool, resort, beds
Children’s Barber	haircut, job, experience, kids
Greek Restaurant	food, wine, service, appetizer, lamb
Department Store	selection, department, sales, shop, clothing

Finding aspect/attribute/target of sentiment

- The aspect name may not be in the sentence
- For restaurants/hotels, aspects are well-understood
- Supervised classification
  - Hand-label a small corpus of restaurant review sentences with aspect
    - food, décor, service, value, NONE
  - Train a classifier to assign an aspect to asentence
    - "Given this sentence, is the aspect *food, décor, service, value,* or *NONE*"

Putting it all together: Finding sentiment for aspects

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop



Results of Blair-Goldensohn et al. method

Rooms (3/5 stars, 41 comments)

- (+) The room was clean and everything worked fine – even the water pressure ...
- (+) We went because of the free room and was pleasantly pleased ...
- (-) …the worst hotel I had ever stayed at ...

Service (3/5 stars, 31 comments)

- (+) Upon checking out another couple was checking early due to a problem ...
- (+) Every single hotel staff member treated us great and answered every ...
- (-) The food is cold and the service gives new meaning to SLOW.

Dining (3/5 stars, 18 comments)

- (+) our favorite place to stay in biloxi.the food is great also the service ...
- (+) Offer of free buffet for joining the Play

Baseline methods assume classes have equal frequencies!

- If not balanced (common in the real world)
  - can’t use accuracies as an evaluation
  - need to use F-scores
- Severe imbalancing also can degrade classifier performance
- Two common solutions:
  1. Resampling in training
     - Random undersampling
  2. Cost-sensitive learning
     - Penalize SVM more for misclassification of the rare thing

Summary on Sentiment

- Generally modeled as classification or regression task
  - predict a binary or ordinal label
- Features:
  - Negation is important
  - Using all words (in naïve bayes) works well for some tasks
  - Finding subsets of words may help in other tasks
    - Hand-built polarity lexicons
    - Use seeds and semi-supervised learning to induce lexicons

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - *liking, loving, hating, valuing, desiring*
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - *nervous, anxious, reckless, morose, hostile, jealous*

Computational work on other affective states

- **Emotion**:
  - Detecting annoyed callers to dialogue system
  - Detecting confused/frustrated versus confident students
- **Mood**:
  - Finding traumatized or depressed writers
- **Interpersonal stances**:
  - Detection of flirtation or friendliness in conversations
- **Personality traits**:
  - Detection of extroverts

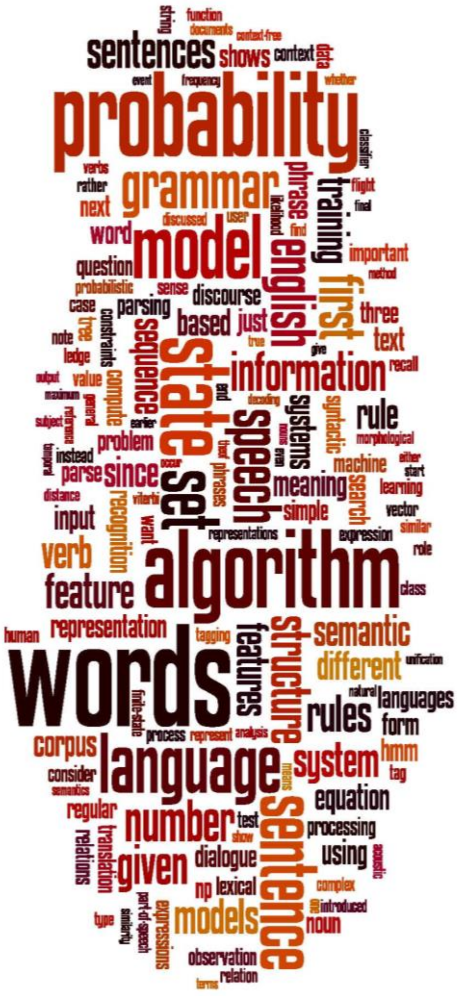
Detection of Friendliness

Ranganath, Jurafsky, McFarland

- Friendly speakers use collaborative conversational style
  - Laughter
  - Less use of negative emotional words
  - More sympathy
    - `That's too bad`   `I'm sorry to hear that`
  - More agreement
    - `I think so too`
  - Less hedges
    - `kind of`   `sort of`   `a little …`

THE UNIVERSITY OF CHICAGO
GRAHAM SCHOOL
CONTINUING LIBERAL AND PROFESSIONAL STUDIES

Sentiment Analysis in Python



Maxent Models and Discriminative Estimation

Generative vs. Discriminative models

Christopher Manning

Introduction

- So far we've looked at "generative models"
  - Language models, Naive Bayes
- But there is now much use of conditional or discriminative probabilistic models in NLP, Speech, IR (and ML generally)
- Because:
  - They give high accuracy performance
  - They make it easy to incorporate lots of linguistically important features
  - They allow automatic building of language independent, retargetable NLP modules

Joint vs. Conditional Models

- We have some data {(*d*, *c*)} of paired observations *d* and hidden classes *c*.
- Joint (generative) models place probabilities over both observed data and the hidden stuff (generate the observed data from hidden stuff):
  - All the classic StatNLP models:
    - *n*-gram models, Naive Bayes classifiers, hidden Markov models, probabilistic context-free grammars, IBM machine translation alignment models

$P(c,d)$

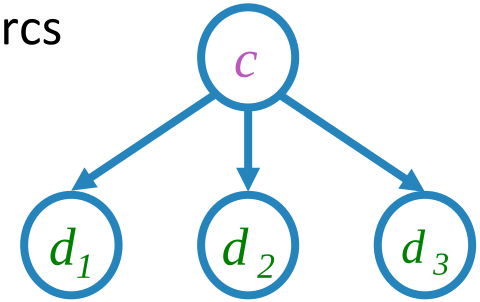
Joint vs. Conditional Models

- Broadly speaking, joint probability is the probability of two things* happening together: e.g., the probability that I wash my car, *and* it rains.
- Conditional probability is the probability of one thing happening, given that the other thing happens: e.g., the probability that, given that I wash my car, it rains.
- Discriminative (conditional) models take the data as given, and put a probability over hidden structure given the data:
  - Logistic regression, conditional loglinear or maximum entropy models, conditional random fields
  - Also, SVMs, (averaged) perceptron, etc. are discriminative classifiers (but not directly probabilistic)

$P(c|d)$

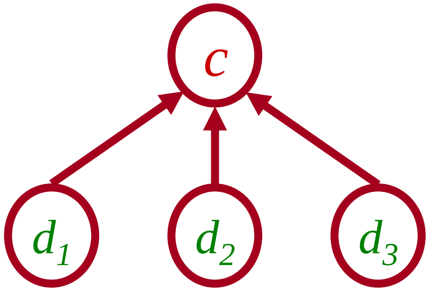
Bayes Net/Graphical Models

- Bayes net diagrams draw circles for random variables, and lines for direct dependencies
- Some variables are observed; some are hidden
- Each node is a little classifier (conditional probability table) based on incoming arcs



Naive Bayes

Generative



Logistic Regression

Discriminative

Conditional vs. Joint Likelihood

- A *joint* model gives probabilities $P(d,c)$ and tries to maximize this joint likelihood.
  - It turns out to be trivial to choose weights: just relative frequencies.
- A *conditional* model gives probabilities $P(c|d)$. It takes the data as given and models only the conditional probability of the class.
  - We seek to maximize conditional likelihood.
  - Harder to do (as we'll see…)
  - More closely related to classification error.

Features

- In these slides and most maxent work: *features* f are elementary pieces of evidence that link aspects of what we observe *d* with a category *c* that we want to predict
- A feature is a function with a bounded real value: $f\colon C \times D \rightarrow \mathbb{R}$

Example features

- $f_1(c, d) \equiv [c = \text{LOCATION} \wedge w_{-1} = \text{“in”} \wedge \text{isCapitalized}(w)]$
- $f_2(c, d) \equiv [c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$
- $f_3(c, d) \equiv [c = \text{DRUG} \wedge \text{ends}(w, \text{“c”})]$

LOCATION	LOCATION	DRUG	PERSON
in Arcadia	*in Québec*	*taking Zantac*	*saw Sue*

- Models will assign to each feature a *weight:*
  - A positive weight votes that this configuration is likely correct
  - A negative weight votes that this configuration is likely incorrect

Features

- In NLP uses, usually a feature specifies
  1. an indicator function – a yes/no boolean matching function – of properties of the input and
  2. a particular class

$$f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$$

[Value is 0 or 1]

- Each feature picks out a data subset and suggests a label for it

Feature-Based Models

- The decision about a data point is based only on the features active at that point.

Data	Data	Data
BUSINESS: Stocks hit a yearly low …	… to restructure bank:MONEY debt.	DT JJ NN … The previous fall …
Label: BUSINESS Features {…, stocks, hit, a, yearly, low, …}	Label: MONEY Features {…, w_{-1}=restructure, w_{+1}=debt, L=12, …}	Label: NN Features {w=fall, t_{-1}=JJ w_{-1}=previous}
Text Categorization	Word-Sense Disambiguation	POS Tagging

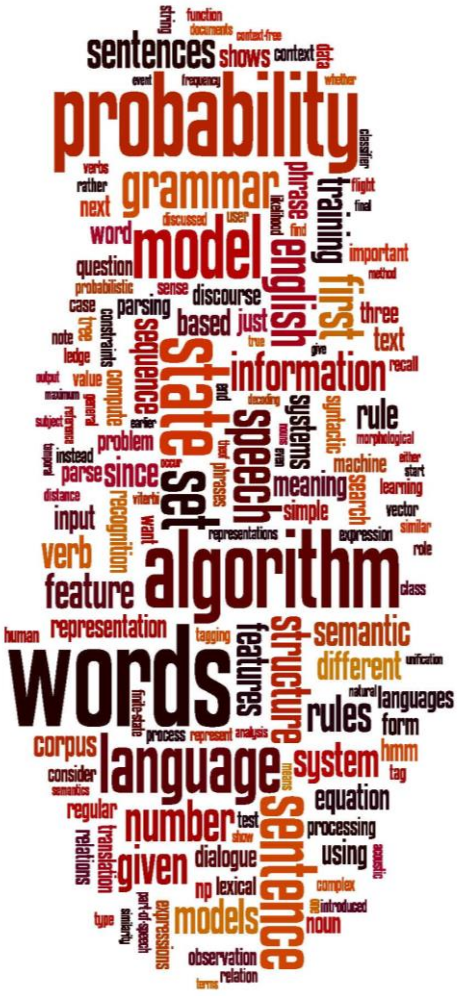
Example: Text Categorization

(Zhang and Oles 2001)

- Features are presence of each word in a document and the document class (they do feature selection to use reliable indicator words)
- Tests on classic Reuters data set (and others)
  - Naïve Bayes: 77.0% F_1
  - Linear regression: 86.0%
  - Logistic regression: 86.4%
  - Support vector machine: 86.5%
- Paper emphasizes the importance of *regularization* (smoothing) for successful use of discriminative methods (not used in much early NLP/IR work)

Other Maxent Classifier Examples

- You can use a maxent classifier whenever you want to assign data points to one of a number of classes:
  - Sentence boundary detection (Mikheev 2000)
    - Is a period end of sentence or abbreviation?
  - Sentiment analysis (Pang and Lee 2002)
    - Word unigrams, bigrams, POS counts, …
  - PP attachment (Ratnaparkhi 1998)
    - Attach to verb or noun? Features of head noun, preposition, etc.
  - Parsing decisions in general (Ratnaparkhi 1997; Johnson et al. 1999, etc.)



Discriminative Model Features

Making features from text for discriminative NLP models

Christopher Manning

Feature-based Linear Classifiers

How to put features into a classifier

Feature-Based Linear Classifiers

- Linear classifiers at classification time:
  - Linear function from feature sets $\{f_i\}$ to classes $\{c\}$.
  - Assign a weight λ_i to each feature f_i.
  - We consider each class for an observed datum d
  - For a pair (c,d), features vote with their weights:
    - $\text{vote}(c) = \Sigma\lambda_i f_i(c,d)$

PERSON *in Québec* | LOCATION *in Québec* | DRUG *in Québec*

  - Choose the class c which maximizes $\Sigma\lambda_i f_i(c,d)$

Feature-Based Linear Classifiers

There are many ways to chose weights for features

- Perceptron: find a currently misclassified example, and nudge weights in the direction of its correct classification
- Margin-based methods (Support Vector Machines)

Feature-Based Linear Classifiers

- Exponential (log-linear, maxent, logistic, Gibbs) models:
  - Make a probabilistic model from the linear combination $\Sigma\lambda_i f_i(c,d)$

$$P(c \mid d,\lambda)=\frac{\exp\sum_i \lambda_i f_i(c,d)}{\sum_{c'}\exp\sum_i \lambda_i f_i(c',d)}$$

Makes votes positive

Normalizes votes

  - P(LOCATION|*in Québec*) = $e^{1.8}e^{-0.6}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.586$
  - P(DRUG|*in Québec*) = $e^{0.3}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.238$
  - P(PERSON|*in Québec*) = $e^{0}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.176$
- The weights are the parameters of the probability model, combined via a “soft max” function

Aside: logistic regression

- Maxent models in NLP are essentially the same as multiclass logistic regression models in statistics (or machine learning)
  - If you haven't seen these before, don't worry, this presentation is self-contained!
  - If you have seen these before you might think about:
    - The parameterization is slightly different in a way that is advantageous for NLP-style models with tons of sparse features (but statistically inelegant)
    - The key role of feature functions in NLP and in this presentation
      - The features are more general, with *f* also being a function of the class – when might this be useful?

Thank You!