Sentiment Analysis

Natural Language Processing

Master in Business Analytics and Big Data

acastellanos@faculty.ie.edu

Examples

Resumen de reseñas

"Vienen con una guarnición de **patatas** con una **salsa** de **barbacoa** picante."

"De **entrantes** pedimos unos **nachos**: muy buenos y **ración** abundante."

"Quizas si el **cocinero** hubiese probado **la salsa** habria visto que teniamos **razon**."

Destacado

✓ Terraza

Accesibilidad

⦸ Acceso para sillas de ruedas

Ofertas

✓ Cerveza ✓ Vino

Opciones del local

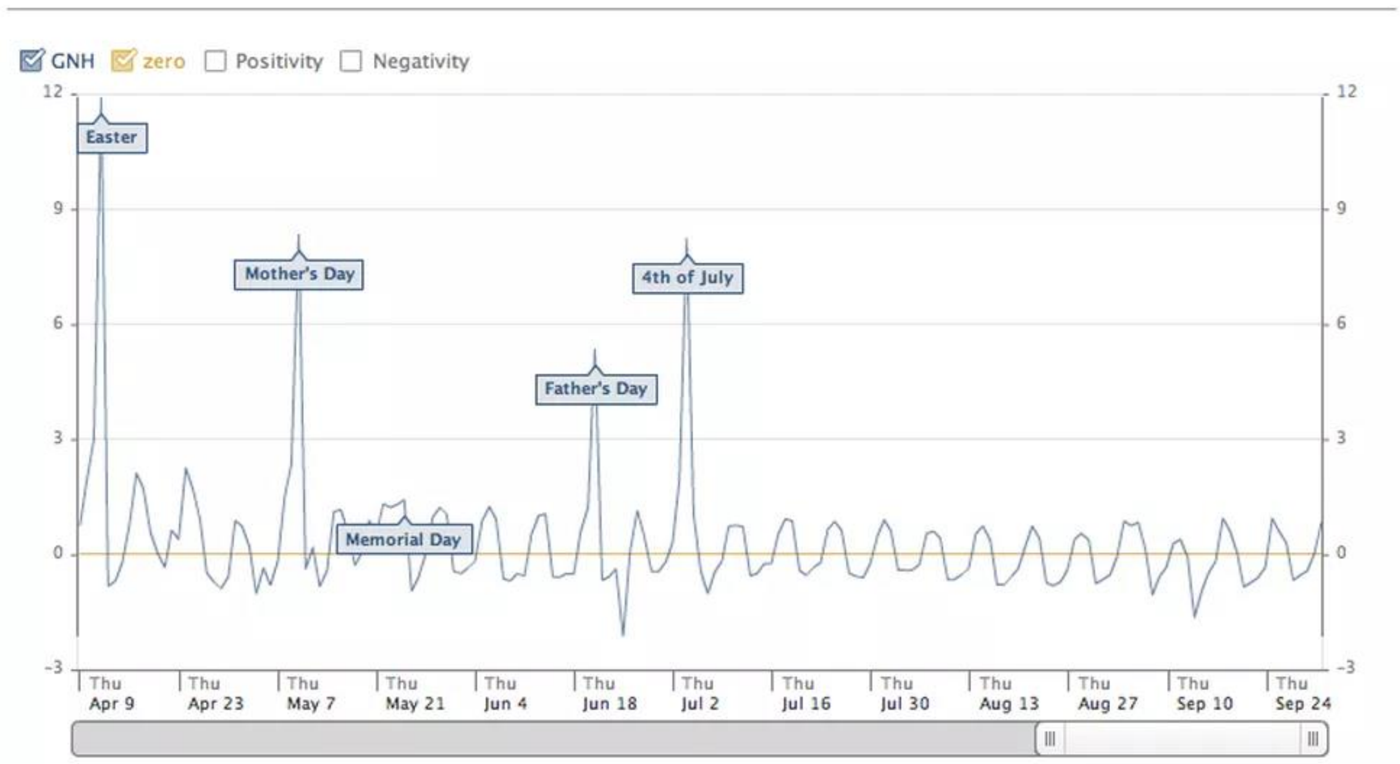
✓ Almuerzo ✓ Cena
✓ Reparto a domicilio ✓ Postres

Servicios

✓ Aseos ✓ Ideal para niños

Examples

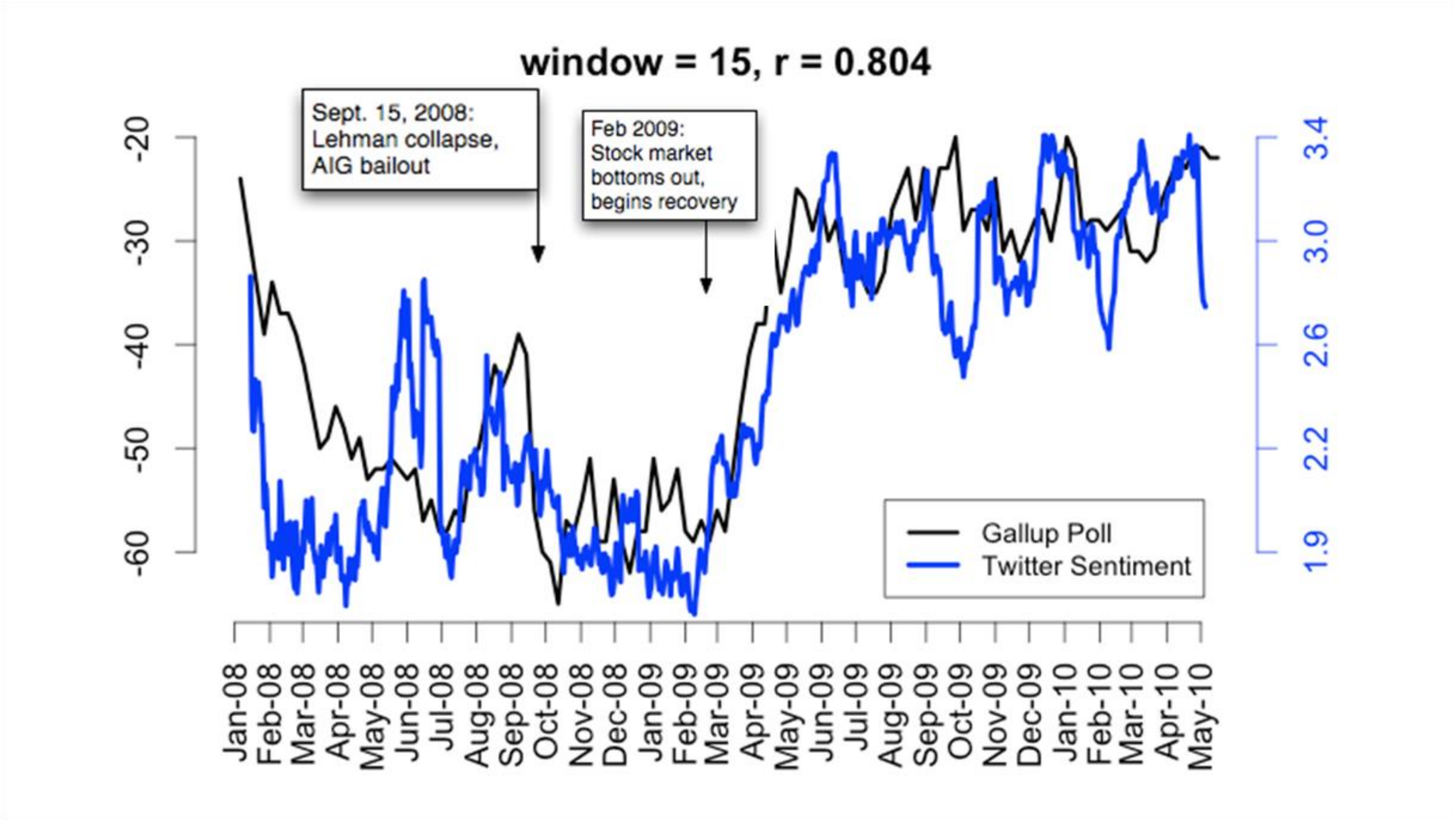
Facebook's "Gross National Happiness Index"



Examples

Twitter sentiment versus Gallup Poll of Consumer Confidence

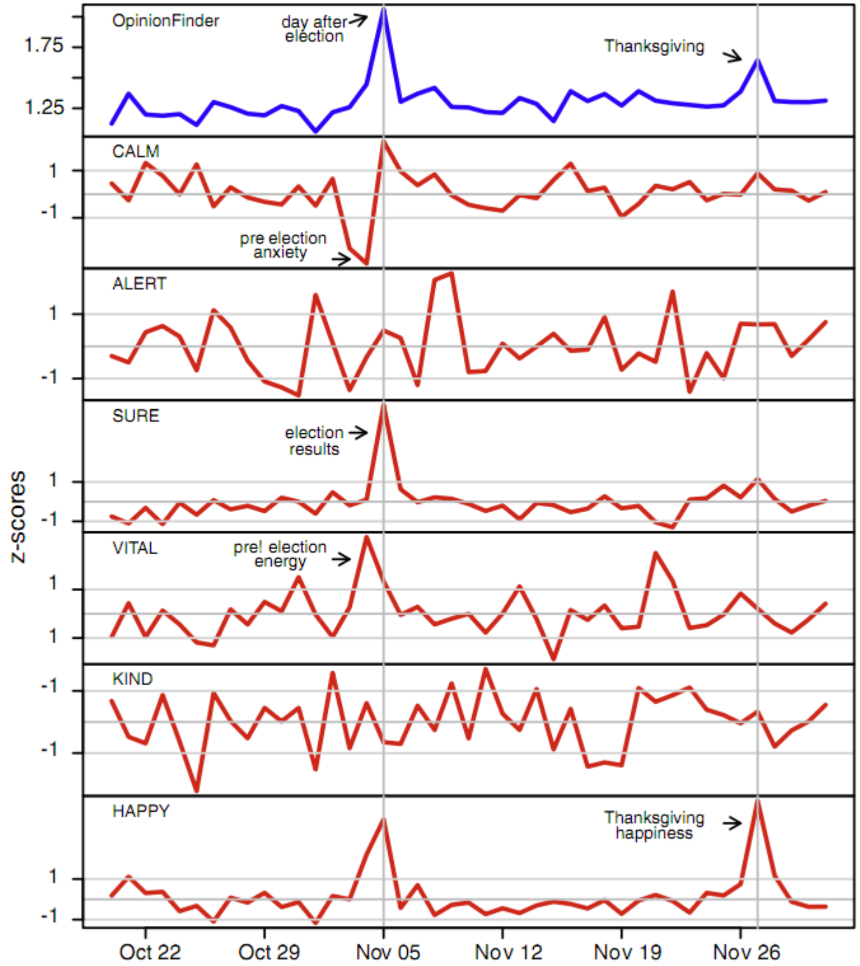
Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010



Examples

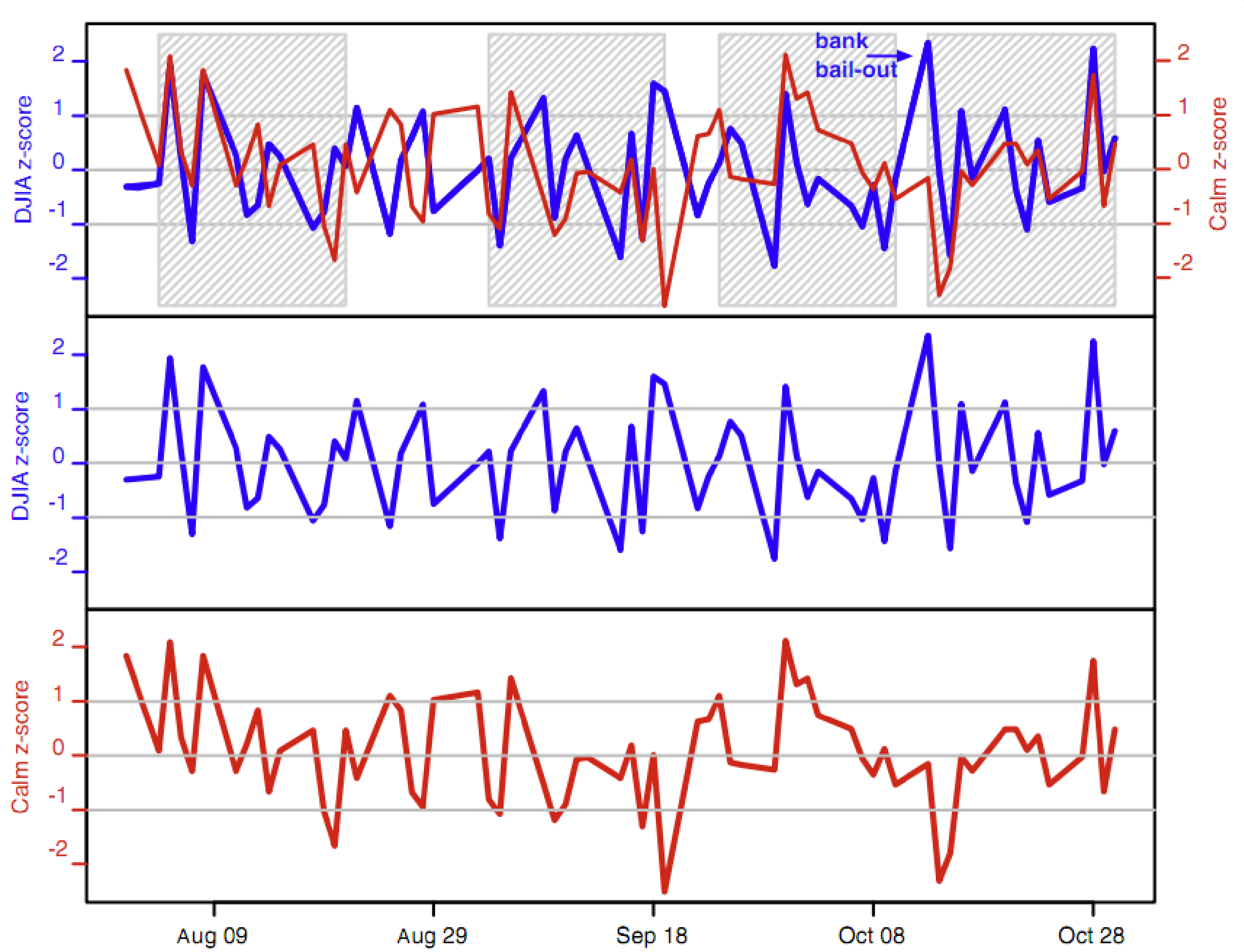
Twitter mood predicts the stock market

Johan Bollen, Huina Mao, Xiaojun Zeng. 2011.
Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.



Examples

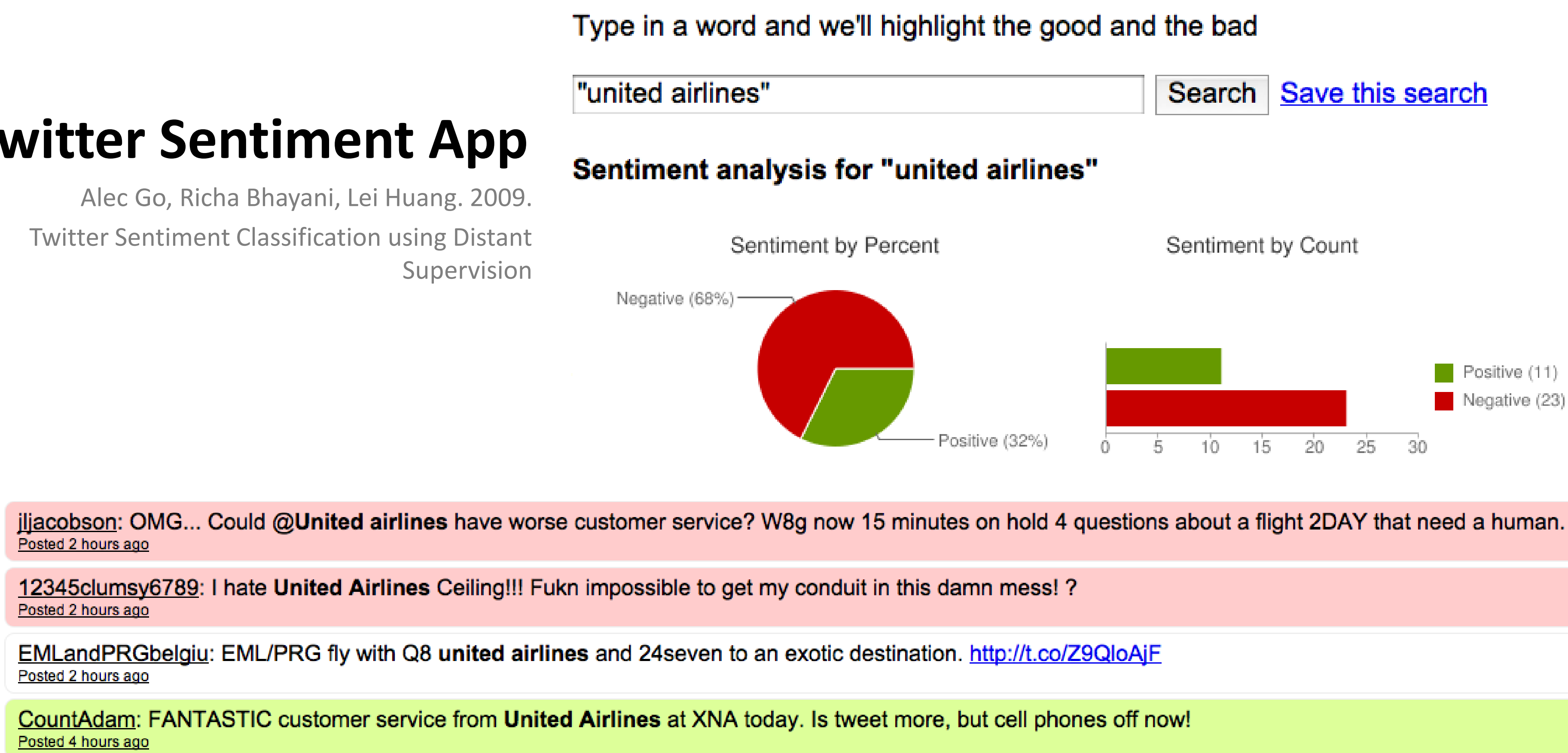
- CALM predicts DJIA 3 days later
- At least one current hedge fund uses this algorithm



Examples

Twitter Sentiment App

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



Challenges

- **Opinions expressed in complex ways**
- **Lexical content alone can be misleading**
- **Rhetorical such as sarcasm, irony, implication, etc.**

Why?

- **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

Why?

- Binary Decision
  - Is the attitude of this text positive or negative?
- Ranking
  - The attitude of this text from 1 to 5
- Faceted Analysis
  - Detect the target, source, or complex attitude types

Baseline Approach

- **Sentiment Classification in Movie Reviews**
  - 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

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

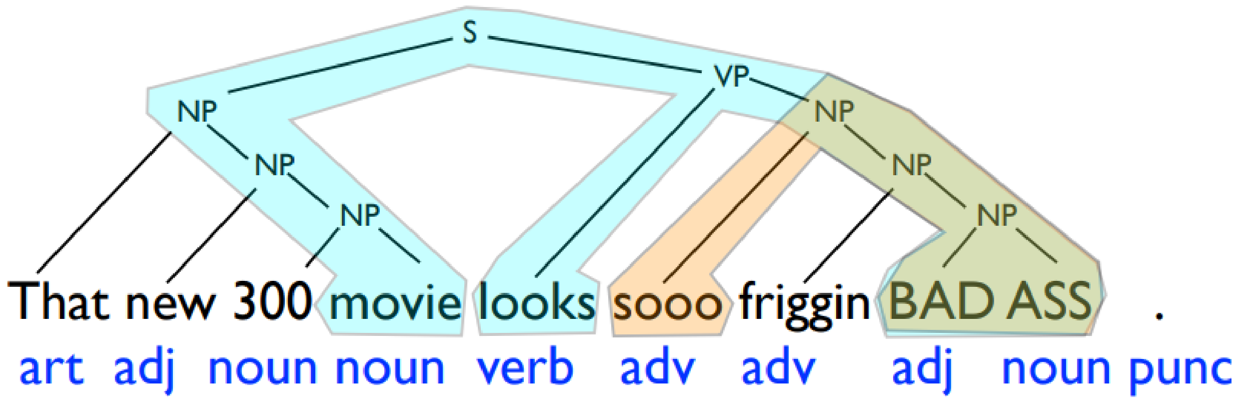
Baseline Algorithm

- Tokenization
- Feature Extraction
- Classification

Baseline Algorithm

- **Tokenization**
- Feature Extraction
- Classification

Tokenization




```
w=that
w=new
w=300
w=movie
w=looks
w=sooo
w=friggin
w=bad
w=ass
```

```
w=so
```

```
bi=<START>_that
bi=that_new
bi=new_300
bi=300_movie
bi=movie_looks
bi=looks_sooo
bi=sooo_friggin
bi=friggin_bad
bi=bad_ass
bi=ass_.
bi=._<END>
```

```
wt=that_art
wt=new_adj
wt=300_noun
wt=movie_noun
wt=looks_verb
wt=sooo_adv
wt=friggin_adv
wt=bad_adj
wt=ass_noun
```

```
subtree=S_NP_movie-S_VP_looks-S_VP_NP_bad_ass
```

```
subtree=NP_sooo_bad_ass
```

Tokenization

- **Twitter mark-up** (names, hashtags)
  - #love, #hate
- **Emoticons**

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

- **Useful code:**
  - Christopher Potts sentiment tokenizer
  - Brendan O’Connor twitter tokenizer

Baseline Algorithm

- Tokenization
- **Feature Extraction**
- Classification

Feature Extraction

- 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

Baseline Algorithm

- Tokenization
- Feature Extraction
- **Classification**

Classification

For sentiment (as well as 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
- More annotated data
  - Usual Suspects: SVM, RNN

Classification

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.

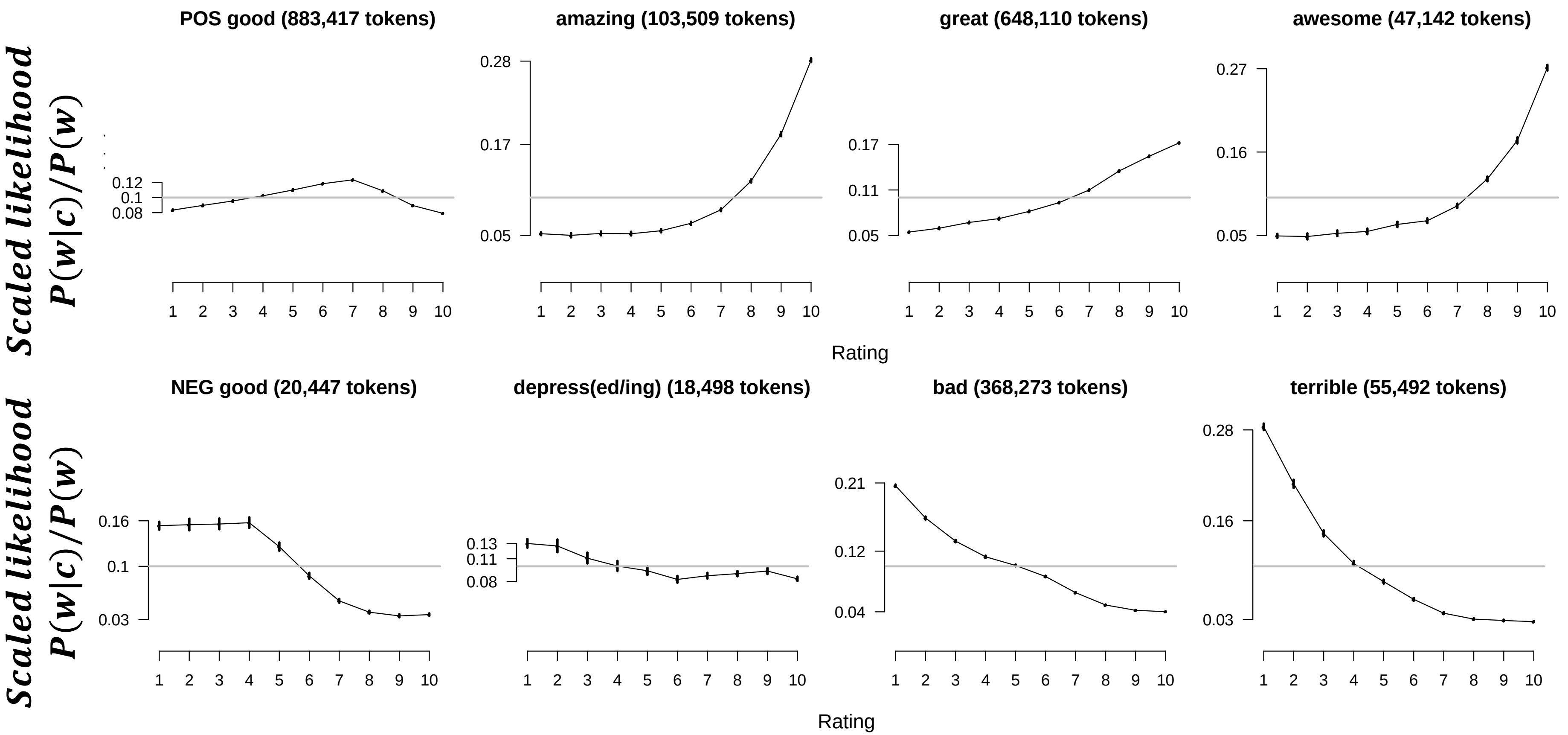
$$P(w|c) = \frac{f(w,c)}{\sum_{w \in c} f(w,c)}$$

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

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

Classification

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



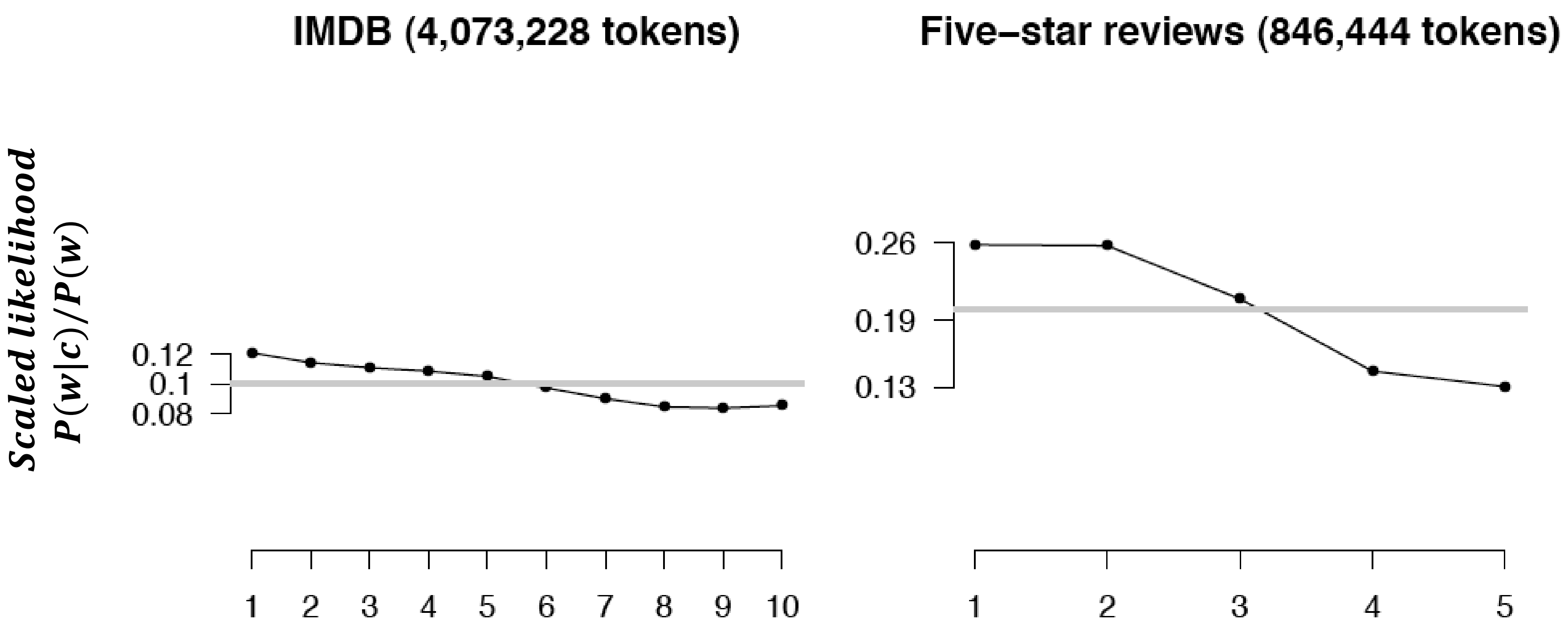
Classification

Logical Negation

- 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

Classification

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



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

- Adjectives conjoined by ***“and”*** have **same polarity**
  - Fair **and** legitimate, corrupt **and** brutal
  - *fair **and** brutal, *corrupt **and** legitimate
- Adjectives conjoined by ***“but”*** have **opposite polarity**
  - fair **but** brutal

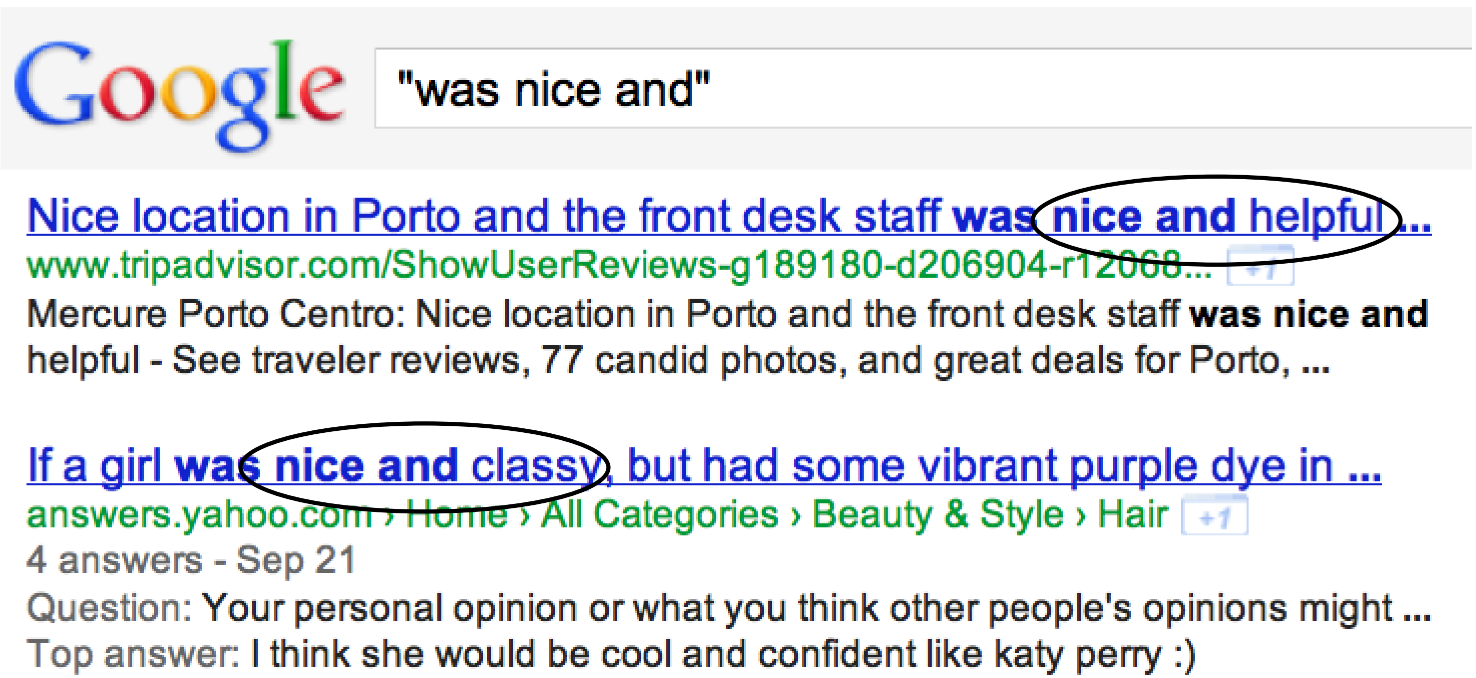
Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

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…

Step 2

- Expand seed set to conjoined adjectives

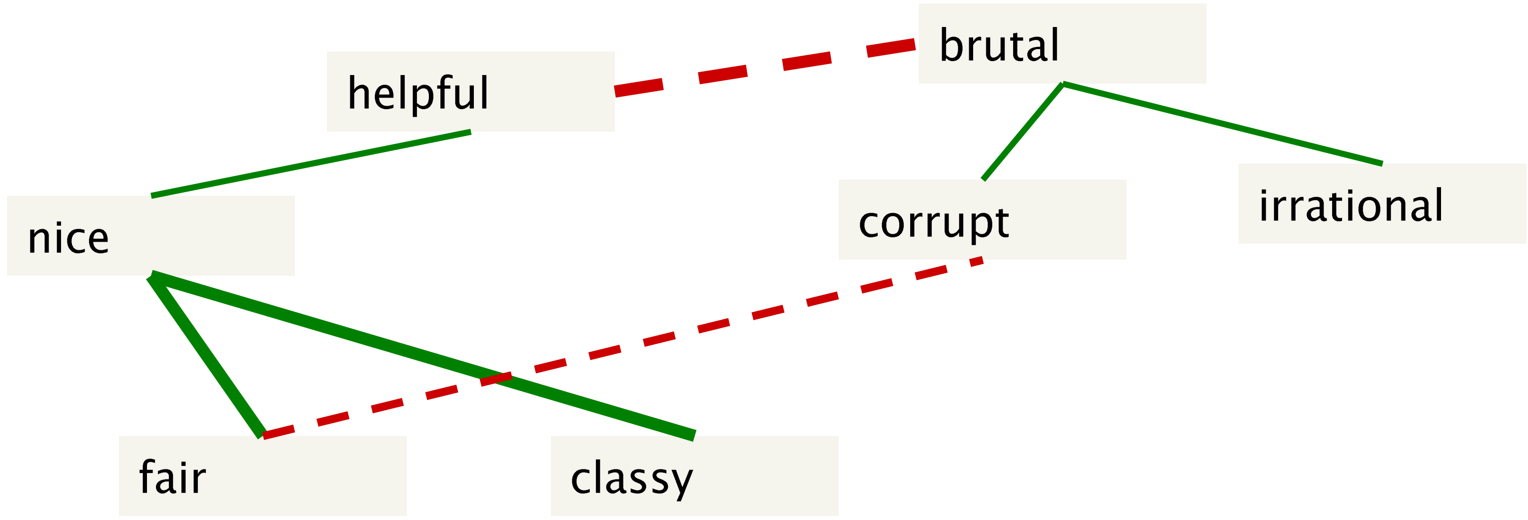


nice, helpful

nice, classy

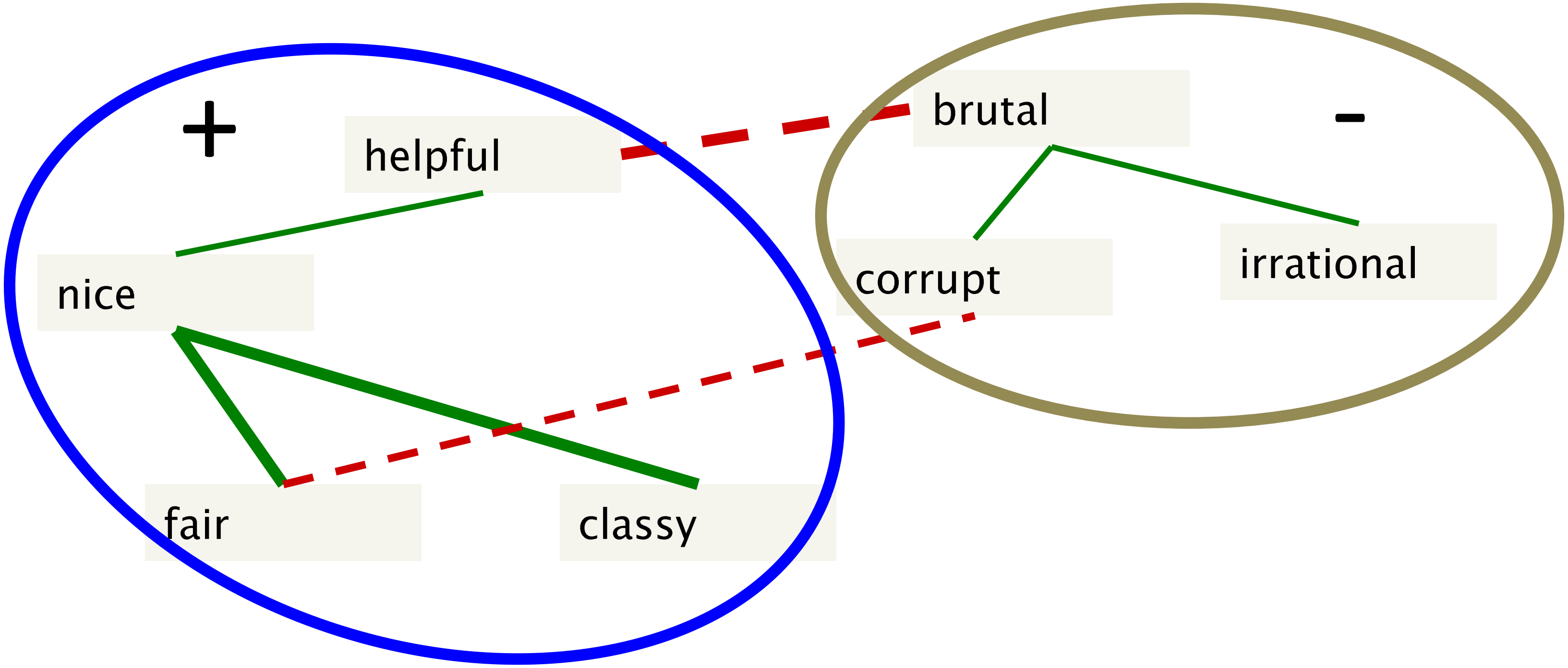
Step 3

- **Supervised classifier** assigns “polarity similarity” to each word pair, resulting in graph:



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

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

- Extract a **phrasal lexicon from reviews**
- **Learn polarity** of each phrase
- **Rate a review by the average polarity of its phrases**

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

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**?

Pointwise Mutual Information

- **Mutual information** between 2 random variables X and Y

$$\mathrm{I}(X,Y) = \sum_{x}\sum_{y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Pointwise Mutual Information

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **PMI between two words**:
  - How much more do two words co-occur than if they were independent?

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1) \cdot P(word_2)}$$

How to Estimate Pointwise Mutual Information

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

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

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

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

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

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

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

Results of Turney algorithm

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- **Turney algorithm: 74%**

Results of Turney algorithm

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 %

Phrases from a thumbs-up

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
Average		0.32

Phrases from a thumbs-up

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

Using WordNet to learn polarity

- **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

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

Using WordNet to learn polarity

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**

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

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
    - Use seeds and semi-supervised learning to induce lexicons

The General Inquirer

- Home page:
  - http://www.wjh.harvard.edu/~inquirer
- Categories:
  - Positive (1915 words) and Negative (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated versus Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for Research Use

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

LIWC (Linguistic Inquiry and Word Count)

- 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*)

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

MPQA Subjectivity Cues Lexicon

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

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.

Bing Liu Opinion Lexicon

- Bing Liu's Page on Opinion Mining

  http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

- 6786 words
  - 2006 positive
  - 4783 negative

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

SentiWordNet

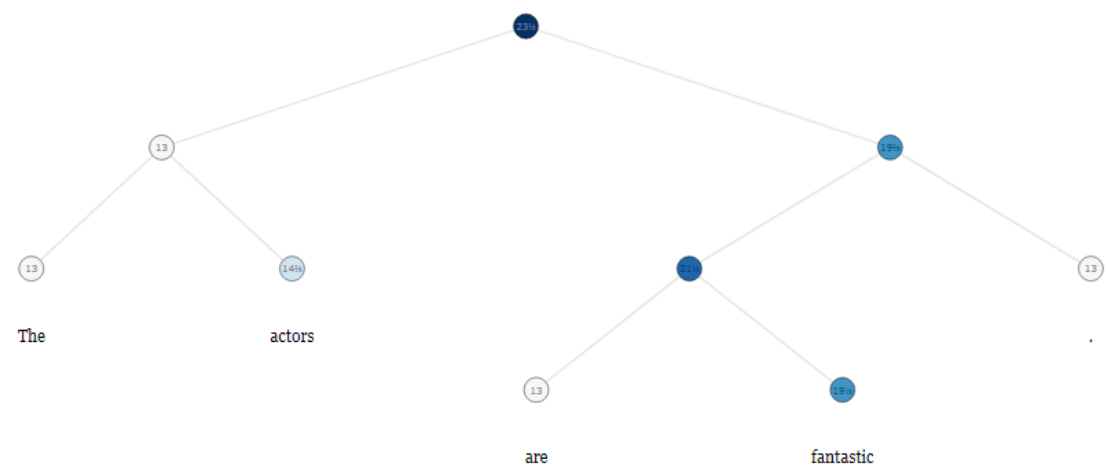
- Home page: http://sentiwordnet.isti.cnr.it/
- **WordNet synsets** automatically annotated for degrees of **positivity, negativity, and neutrality/ objectiveness**
- **NLTK Interface**

  http://www.nltk.org/howto/sentiwordnet.html

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

Sentiment TreeBank

- Home page:
  https://nlp.stanford.edu/sentiment/treebank.html
- **9645 parsed sentences**
  - Trees annotated with sentiment



Resources

- Deep Learning for Sentiment Analysis: A Survey
  - https://arxiv.org/pdf/1801.07883.pdf
- A Survey on Sentiment Analysis Methods and Approach
  - https://ieeexplore.ieee.org/document/7951748/
- SemEval Task on Sentiment Analysis
  - http://alt.qcri.org/semeval2017/task4/
- Sentiment Analysis with Spark
  - https://es.hortonworks.com/tutorial/sentiment-analysis-with-apache-spark/