Sentiment Lexicon

What is sentiment lexicons?

How to extract or build them ?

Recent Research

Sentiment Lexicons Introduction

- English Sentiment Lexicons (from Christopher Potts of Dept. of Linguistics, and from Dan in Dept. Computer Science of Stanford University)
- Chinese 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**
  - 不确定性Tentative (*maybe, perhaps, guess*), 压抑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/
- 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
SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining in LREC 2010.

- Home Page: http://sentiwordnet.isti.cnr.it
- An enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications
- the result of the automatic annotation of all the synsets of WORDNET according to the notions of "positivity", "negativity", and "neutrality".

Estimable (for example)

1. Effects and two-factor interactions are ***estimable.***
2. It is an ***estimable*** model, fun to drive and easy to handle

1. the synset [estimable(J,3)], corresponding to the sense "may be computed or estimated" of the adjective estimable, has an Obj score of 1.0 (and Pos and Neg scores of 0.0)
2. the synset [estimable(J,1)] corresponding to the sense "deserving of respect or high regard" has a Pos score of 0.75, a Neg score of 0.0, and an Obj score of 0.25.

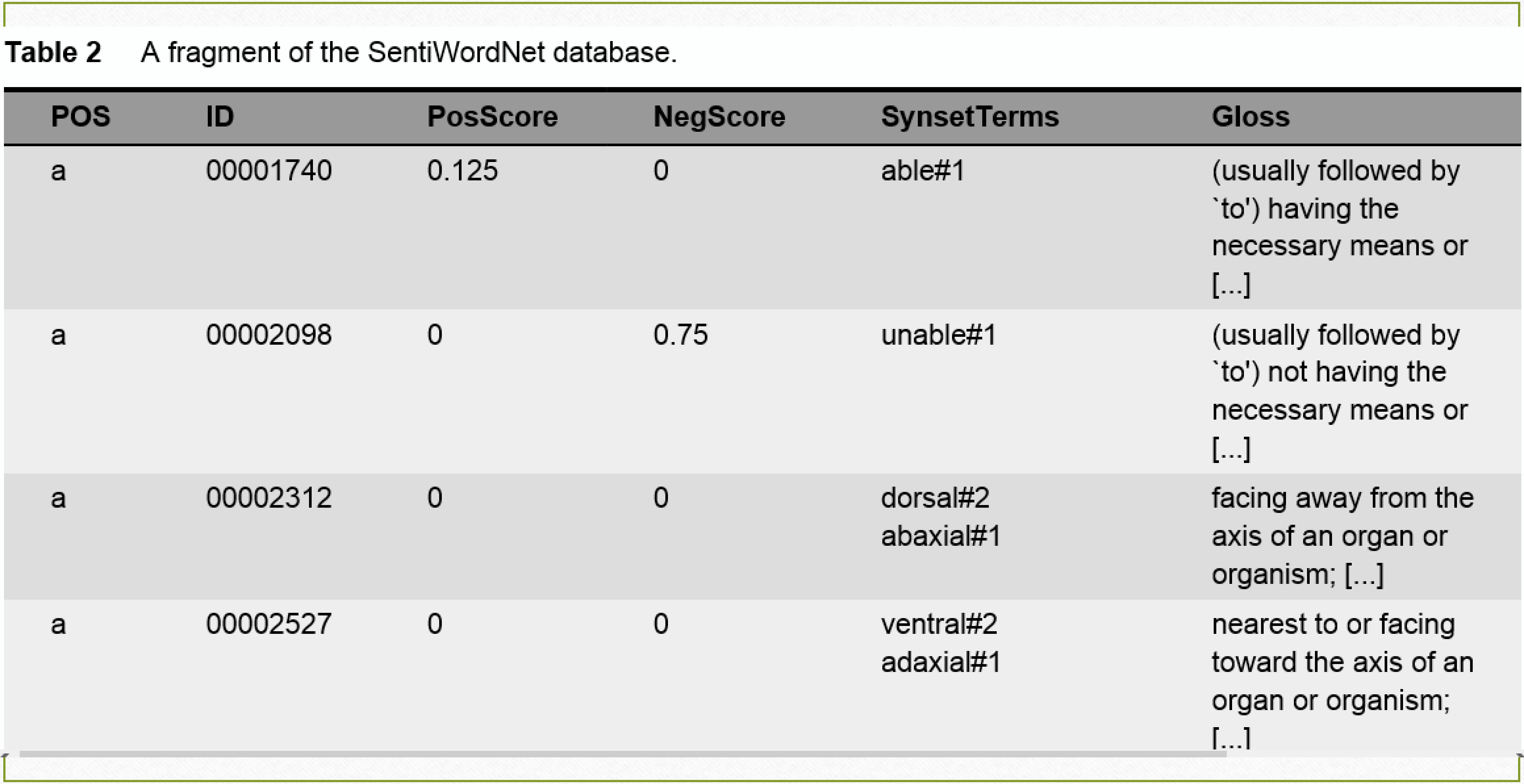
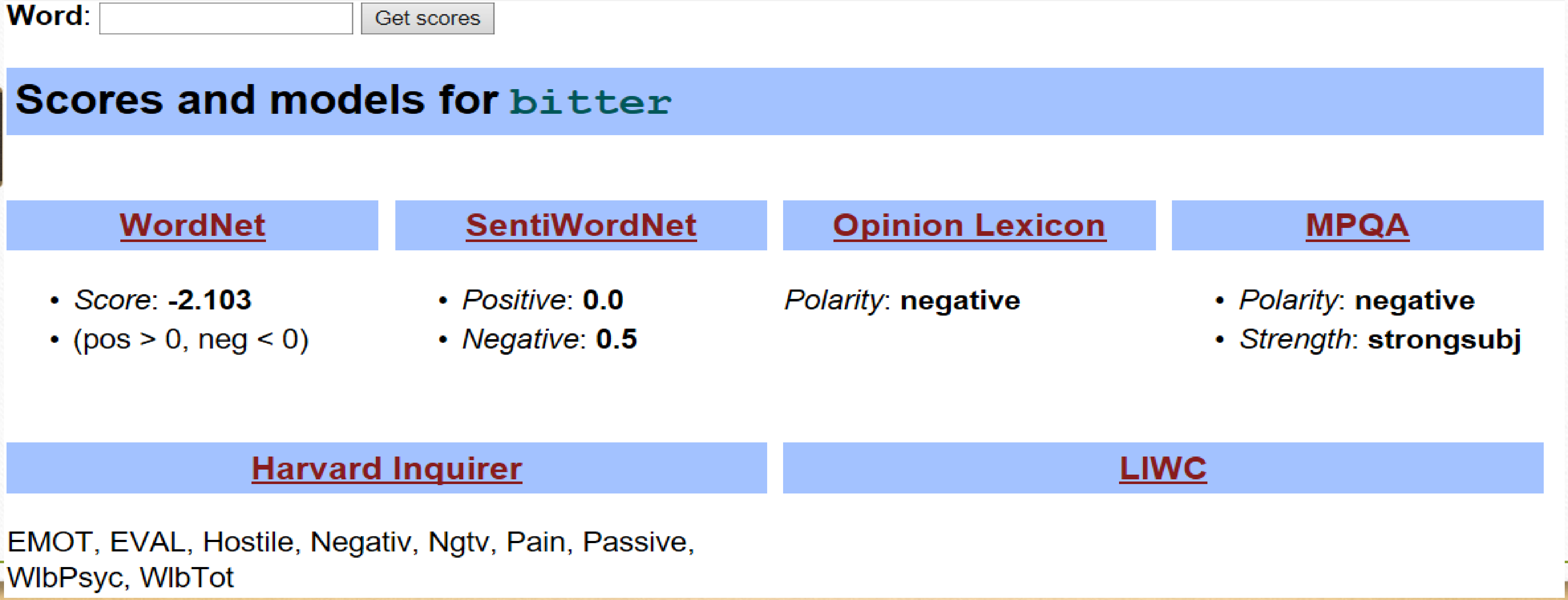


Table 2 A fragment of the SentiWordNet database.

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by \`to') having the necessary means or [...]
a	00002098	0	0.75	unable#1	(usually followed by \`to') not having the necessary means or [...]
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; [...]
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; [...]

http://sentiment.christopherpotts.net/lexicons.html



Scores and models for bitter

WordNet

- *Score*: **-2.103**
- (pos > 0, neg < 0)

SentiWordNet

- *Positive*: **0.0**
- *Negative*: **0.5**

Opinion Lexicon

Polarity: **negative**

MPQA

- *Polarity*: **negative**
- *Strength*: **strongsubj**

Harvard Inquirer

EMOT, EVAL, Hostile, Negativ, Ngtv, Pain, Passive, WlbPsyc, WlbTot

LIWC

Chinese Sentiment Lexicons

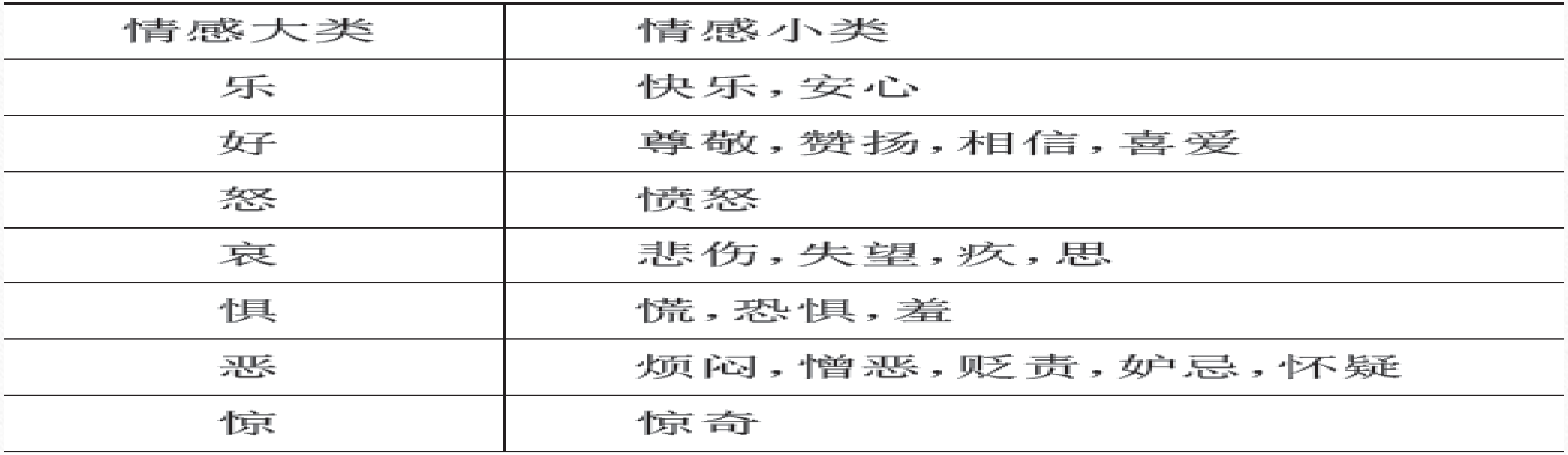
- Qinghua university: positive 5567, negative 4468
- Beijing university: positive 95, negative 420
- Dalian technology university: positive 11043, negative: 10646
- Hownet: positive 4528, negative 4320

Chinese sentimental lexicons (cont.)

大连理工大学的情感词汇库（徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185）

- Sentiment category has 7 types and 21 subtypes

(词汇本体中的情感共分为7大类21小类)



情感大类	情感小类
乐	快乐，安心
好	尊敬，赞扬，相信，喜爱
怒	愤怒
哀	悲伤，失望，疚，思
惧	慌，恐惧，羞
恶	烦闷，憎恶，贬责，妒忌，怀疑
惊	惊奇

Chinese sentimental lexicons (cont.)

词语	*词性种类*	*词义数*	*词义序号*	*情感分类*	*强度*	*极性*	*辅助情感分类*	*强度*	*极性*
无所畏惧	idiom	1	1	PH	7	1			
手头紧	idiom	1	1	NE	7	0			
周到	adj	1	1	PH	5	1			
言过其实	idiom	1	1	NN	5	2			

Sentimental Lexicons from QingHua University

Jun Li and Maosong Sun, Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques, in Proceding of IEEE NLPKE 2007

李军 中文评论的褒贬义分类实验研究 硕士论文 清华大学 2008

- **Positive words**

遂意, 得救, 稳帖, 谦诚, 赞成, 谦虚谨慎, 清淡, 佳境, 患得患失, 不惑， 宰相肚里好撑船

- **Negative words**

下流，挑刺儿，憾事，日暮途穷, 日暮途穷，散漫，谗言

How to Build Sentiment Lexicons

- **How to extract opinion words?**

✓Semi-supervised method (use a small amount of information)

  ✓A few labeled examples

  ✓A few hand-built patterns

✓ Classification-based method

- **How to identify their polarity?**
- **How to score opinion words?**

Using an Ontology

- Begins with **n small, hand-crafted seed-sets** and then follows WordNet relations from them, thereby expanding their size. The expanded sets of iteration i are used as seed-sets for iteration i+1, generally after pruning any pairwise overlap between them.
- Positive: excellent, good, nice, fortunate,..
- Negative: nasty, bad, poor, unfortunate,…
- Objective: administrative, financial, department, measurement, public,..

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

- 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

Classification based method to identify sentimental words

- Training phrase:

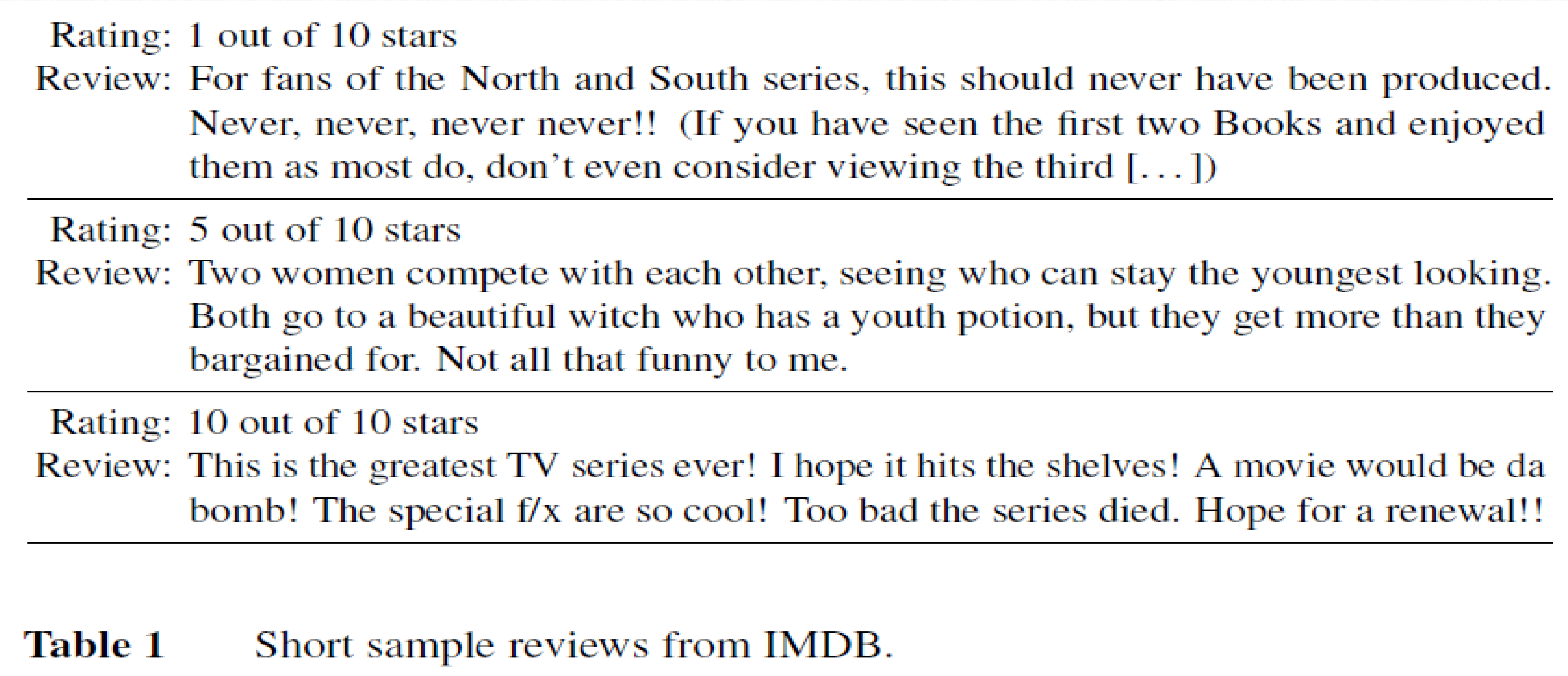
1. Construct training corpus: data set **D**, sentimental lexicon **S,** non sentimental words O, any word d belongs to D, if they appear in S, +1 or -1. if they appear in O, assign 0.
2. Feature selection: **POS**, **context of the word**, **how many words are pos, neg**, **and how many punctuations such as ！**

- Test phrase

1. Candidate sentimental words extraction (all the other words in D, tf, POS meets some requirements)
2. Use Model to identify when it is a sentimental words or not

Analyzing the polarity of each word in IMDB

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



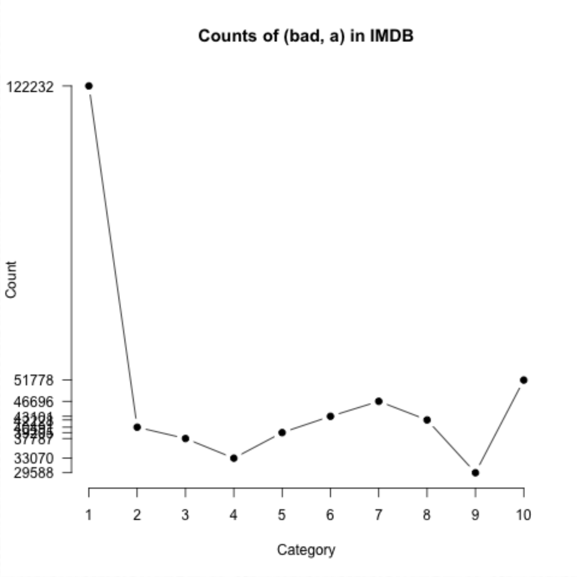
Rating:	1 out of 10 stars
Review:	For fans of the North and South series, this should never have been produced. Never, never, never never!! (If you have seen the first two Books and enjoyed them as most do, don't even consider viewing the third [. . .])
Rating:	5 out of 10 stars
Review:	Two women compete with each other, seeing who can stay the youngest looking. Both go to a beautiful witch who has a youth potion, but they get more than they bargained for. Not all that funny to me.
Rating:	10 out of 10 stars
Review:	This is the greatest TV series ever! I hope it hits the shelves! A movie would be da bomb! The special f/x are so cool! Too bad the series died. Hope for a renewal!!

Table 1 Short sample reviews from IMDB.

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:
- Instead, **likelihood:**

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

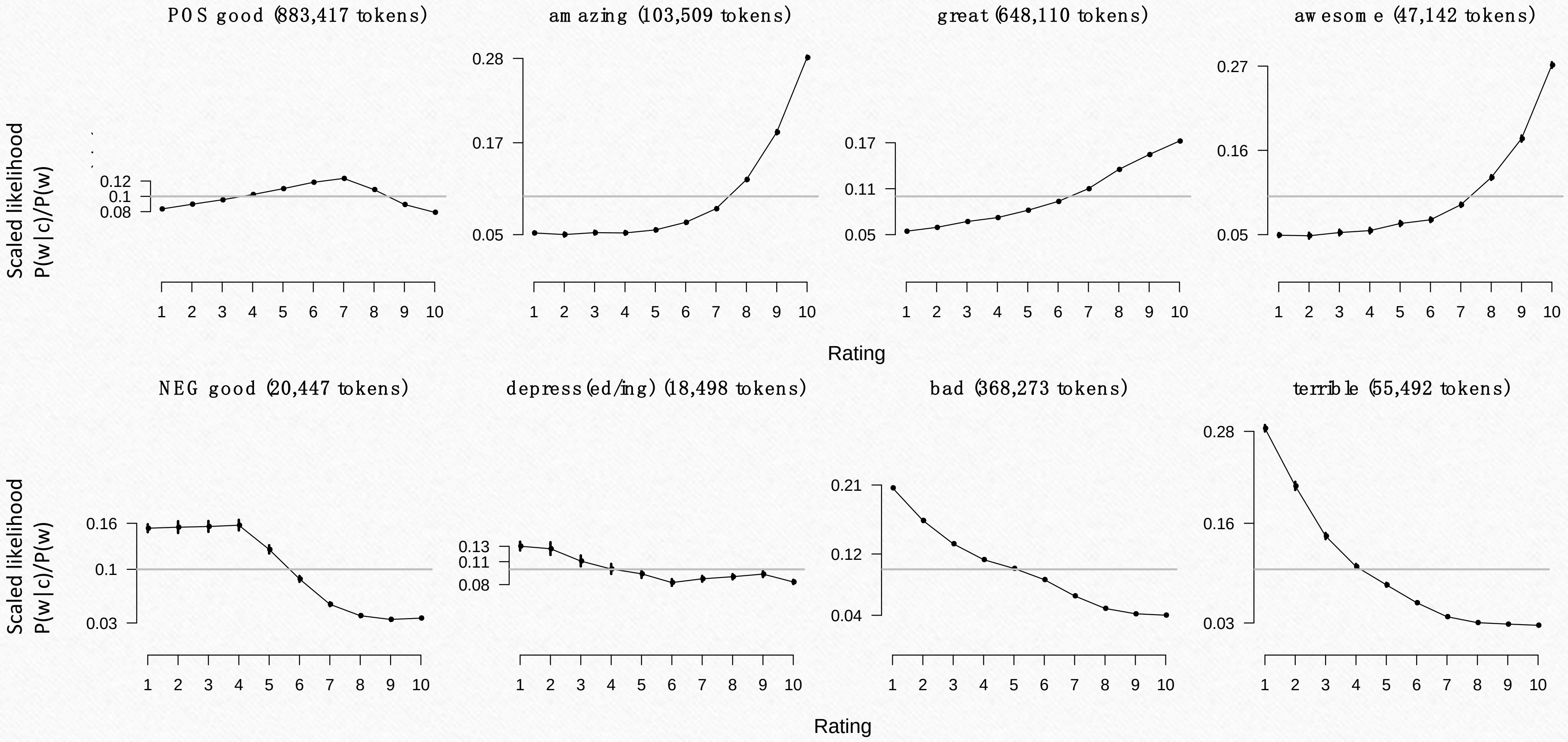


- Make them comparable between words
  - **Scaled likelihood:**

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

Analyzing the polarity of each word in IMDB

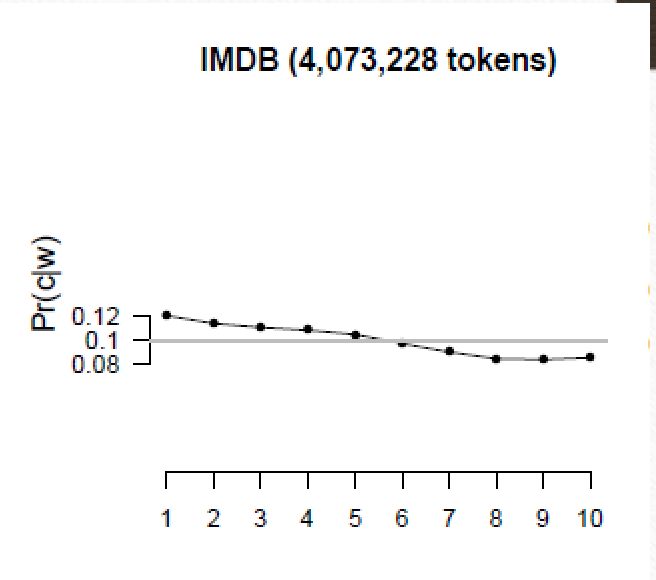
Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.



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



Hatzivassiloglou and McKeown **intuition for identifying word polarity**

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

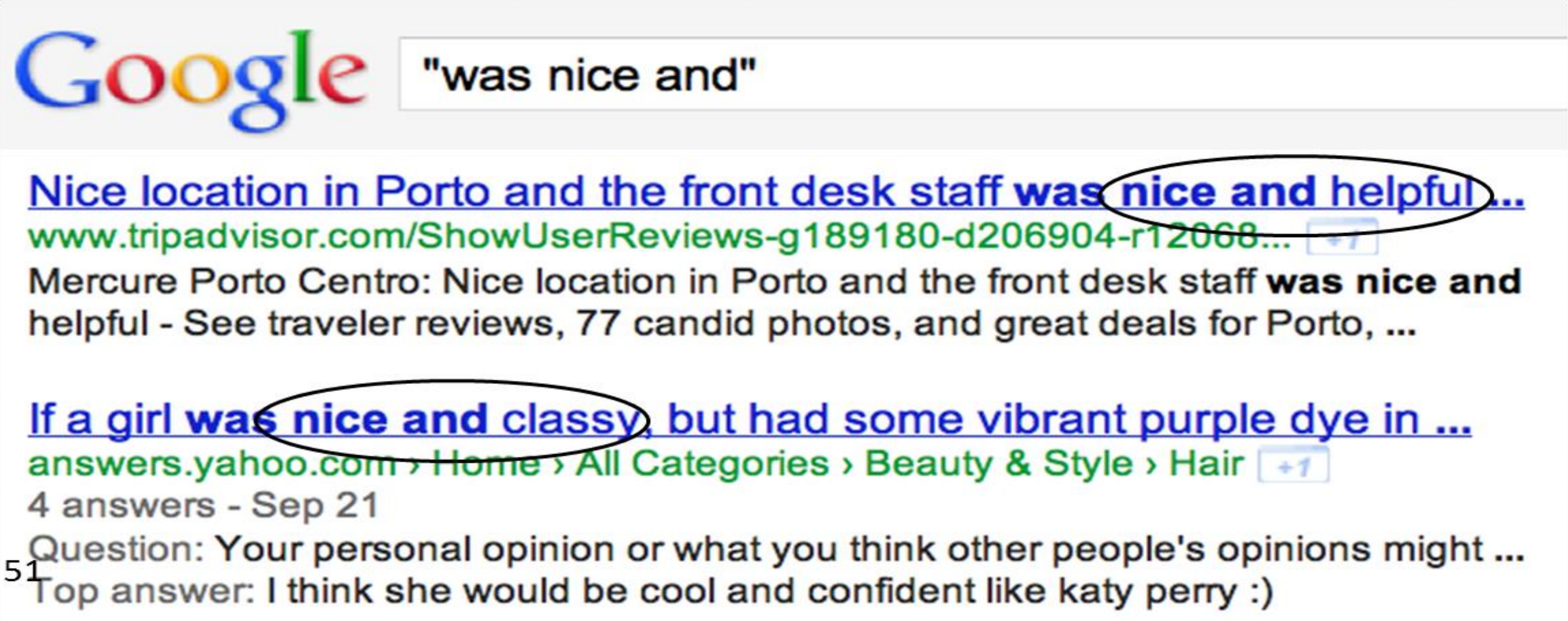
- Adjectives conjoined by "*and*" have same polarity
  - Fair **and** legitimate, corrupt **and** brutal
- Adjectives conjoined by "*but*" do not
  - fair **but** brutal

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



"was nice and"

Nice location in Porto and the front desk staff **was nice and** helpful ...
www.tripadvisor.com/ShowUserReviews-g189180-d206904-r12068...
Mercure Porto Centro: Nice location in Porto and the front desk staff **was nice and** helpful - See traveler reviews, 77 candid photos, and great deals for Porto, ...

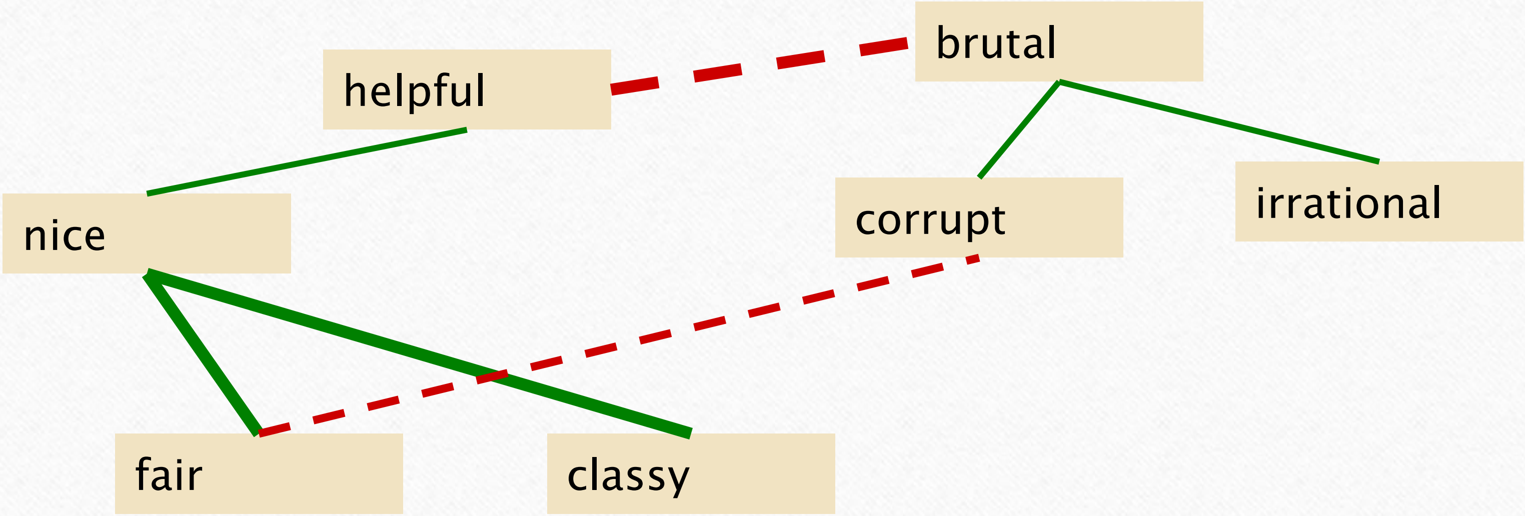
If a girl **was nice and** classy, but had some vibrant purple dye in ...
answers.yahoo.com › Home › All Categories › Beauty & Style › Hair
4 answers - Sep 21
Question: Your personal opinion or what you think other people's opinions might ...
Top answer: I think she would be cool and confident like katy perry :)

nice, helpful

nice, classy

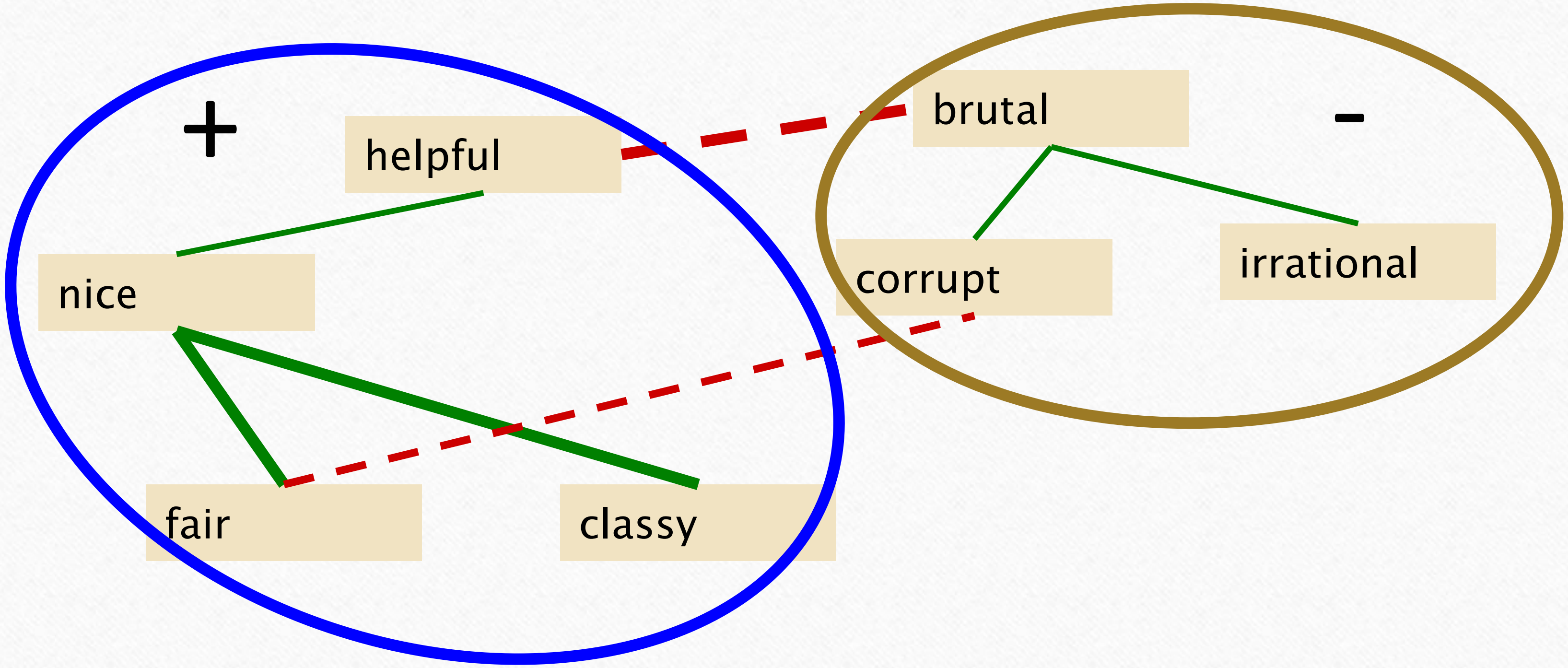
Hatzivassiloglou & McKeown 1997
Step 3

- 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…

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

1) Extract a *phrasal lexicon* from reviews

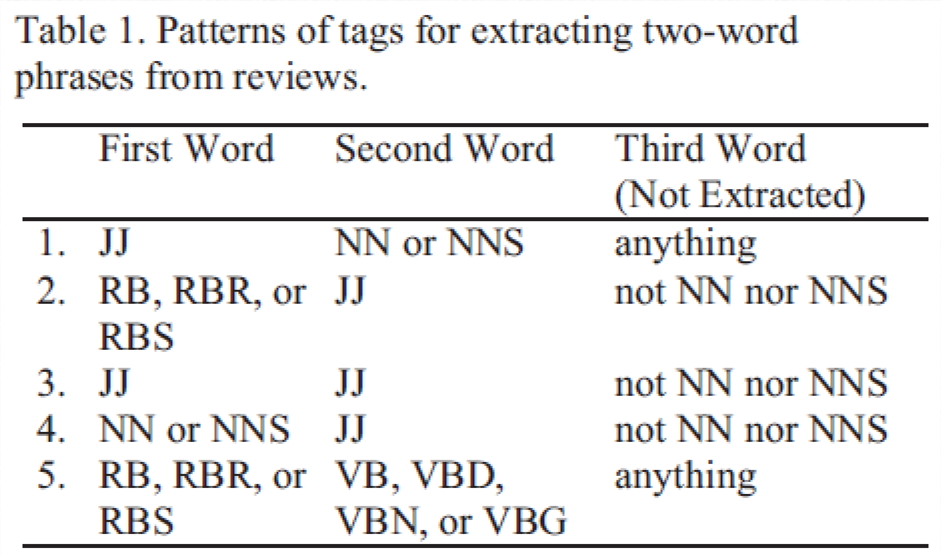


Table 1. Patterns of tags for extracting two-word phrases from reviews.

	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 nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

- Two consecutive words are extracted from the review if their tags conform to any of the patterns in Table 1.

2) 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?

MI and PMI

- **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?

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

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}

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

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

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

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

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

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

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
Average		0.32

The average value is 0.32, is positive

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

Summarization

- What are sentiment lexicons?
- How to extract them?
- How to identify their polarity?

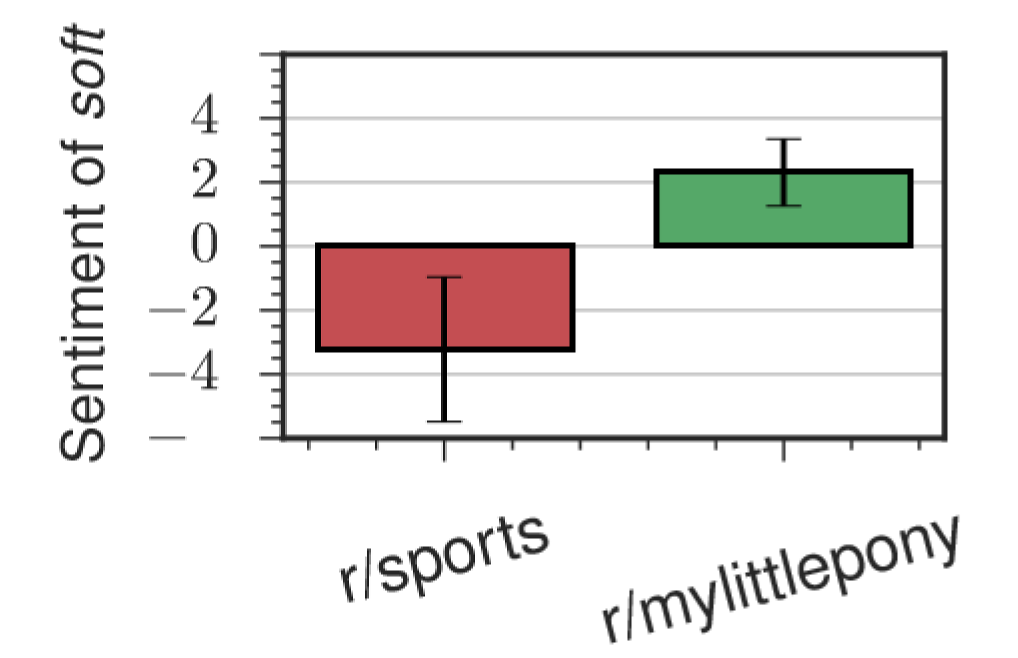
- Start with a seed set of words (‘good’, ‘poor’)
- Find other words that have similar polarity:
  - Using “and” and “but”
  - Using words that occur nearby in the same document
  - Using WordNet synonyms and antonyms
  - Using sentiment corpus

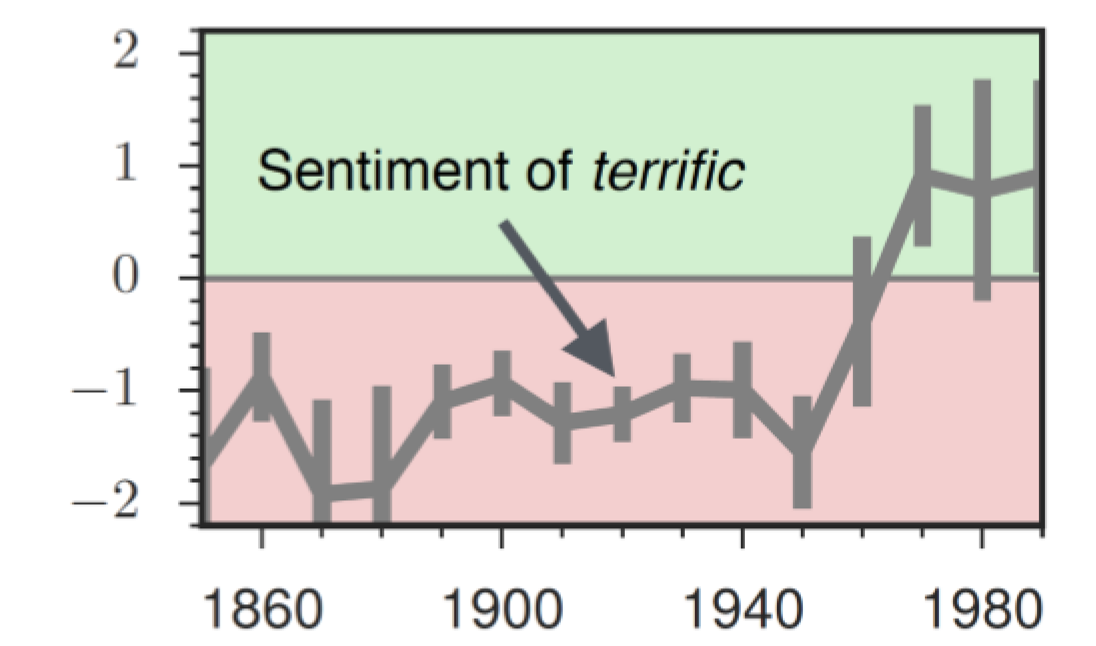
Recent Research

- Inducing Domain-specific Sentiment Lexicons from Unlabeled Corpora from nlp stanford university in 2016 EMNLP.

- Main idea: combine domain-specific word embeddings with a label propagation framework
- Two large scale Experiments: word sentiment varies across time and between communities.

Research Aim





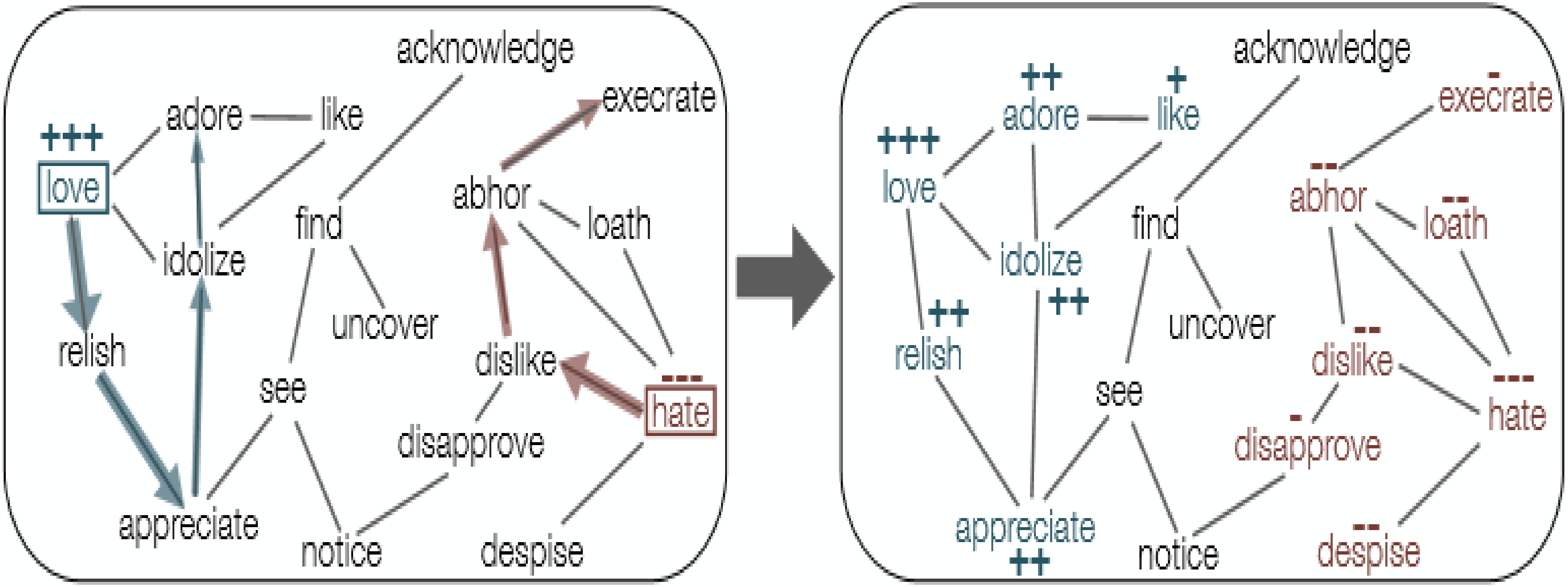
Method

- constructing a lexical graph

① Are constructed from distributional word embeddings, using PPMI(w1,w2)

② Define the graph edges

- propagating sentiment labels from a seed set

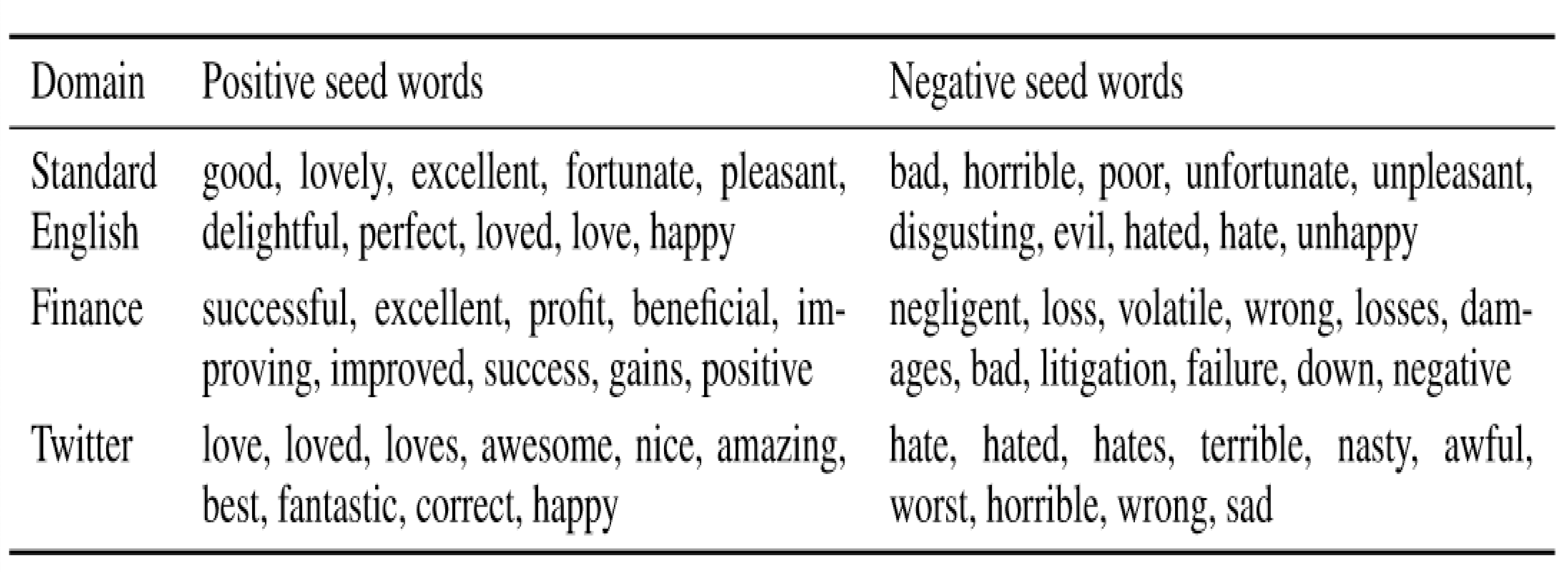


a. Run random walks from seed words.

b. Assign polarity scores based on frequency of random walk visits.

Figure 3: Visual summary of the SENTPROP algorithm.

Seed sets

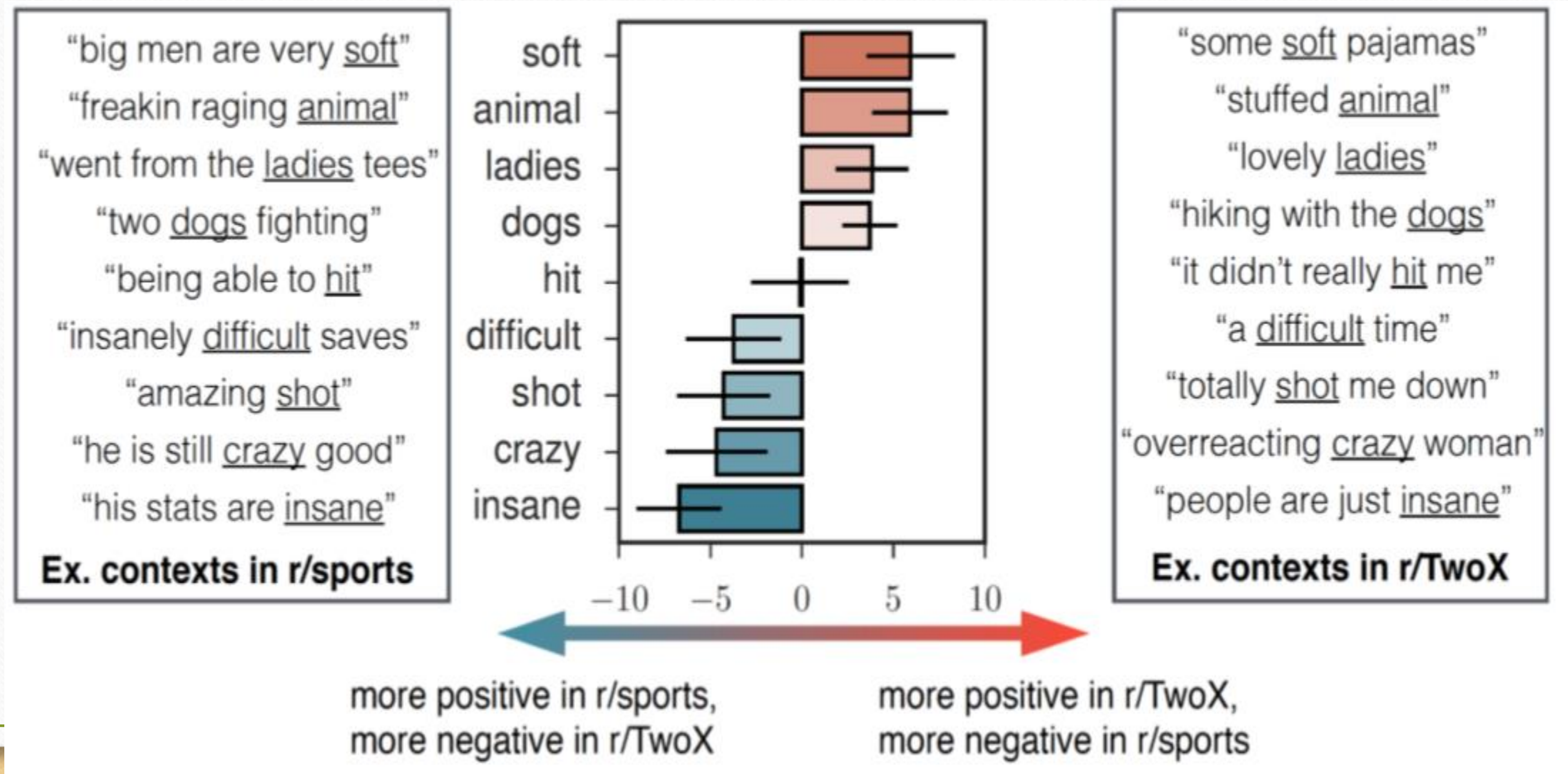


Domain	Positive seed words	Negative seed words
Standard English	good, lovely, excellent, fortunate, pleasant, delightful, perfect, loved, love, happy	bad, horrible, poor, unfortunate, unpleasant, disgusting, evil, hated, hate, unhappy
Finance	successful, excellent, profit, beneficial, improving, improved, success, gains, positive	negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative
Twitter	love, loved, loves, awesome, nice, amazing, best, fantastic, correct, happy	hate, hated, hates, terrible, nasty, awful, worst, horrible, wrong, sad

Corpus: user generated topic-specific forum

- Induce sentiment lexicons for the top-250.
- Sentiment was induced for the top-5000 non-stop words in each subset.

Words in different forum have different sentiments



"big men are very soft"
"freakin raging animal"
"went from the ladies tees"
"two dogs fighting"
"being able to hit"
"insanely difficult saves"
"amazing shot"
"he is still crazy good"
"his stats are insane"

Ex. contexts in r/sports

"some soft pajamas"
"stuffed animal"
"lovely ladies"
"hiking with the dogs"
"it didn't really hit me"
"a difficult time"
"totally shot me down"
"overreacting crazy woman"
"people are just insane"

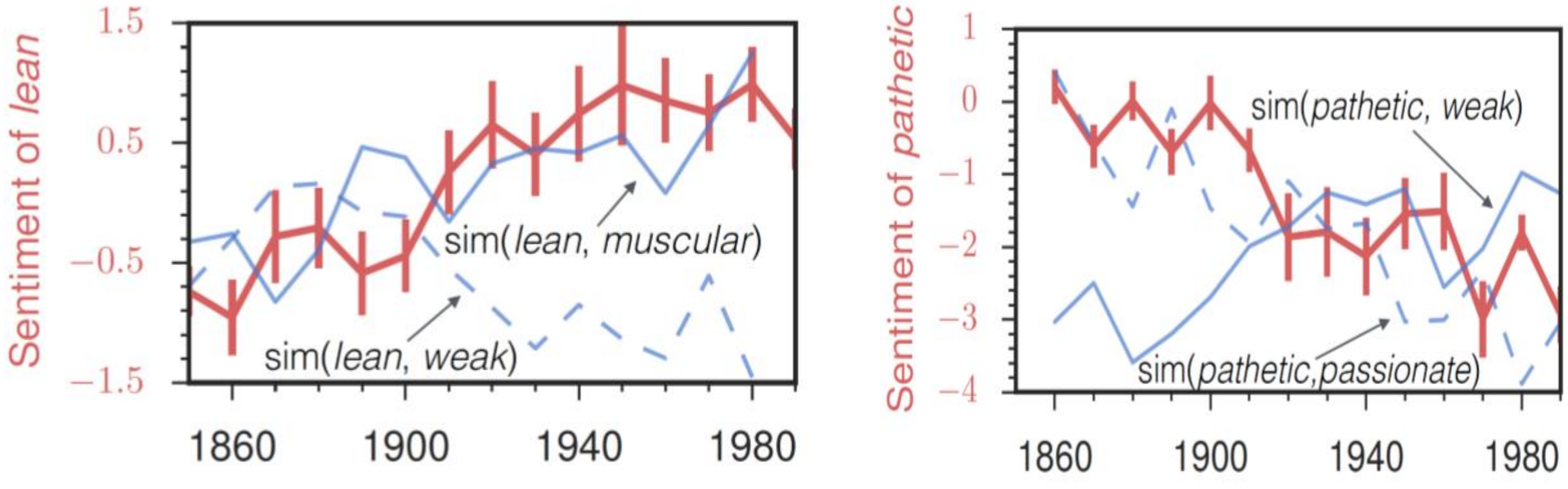
Ex. contexts in r/TwoX

soft
animal
ladies
dogs
hit
difficult
shot
crazy
insane
−10
−5
0
5
10
more positive in r/sports, more negative in r/TwoX
more positive in r/TwoX, more negative in r/sports

Inducing diachronic sentiment Lexicons

- built lexicons for all adjectives with counts above 100 in a given decade and also for the top-5000 non-stop words within each year.
- t>5% of sentiment-bearing (positive/negative) words completely switched polarity during this 150-year time-period
- >25% of all words changed their sentiment label (including switches to/from neutral)

Words sentiments varies across time



(a) *Lean* becomes more positive. *Lean* underwent amelioratioı, becoming more similar to *muscular* and less similar to *weak*.

(b) *Pathetic* becomes more negative. *Pathetic* underwent pejoration, becoming similar to *weak* and less similar to *passionate*.

Advantages of their method

- **Resource-light**: Accurate performance without massive corpora or hand-curated resources.
- **Interpretable**: Uses small seed sets of “paradigm” words to maintain interpretability and avoid ambiguity in sentiment values.
- **Robust**: Bootstrap-sampled standard deviations provide a measure of confidence.
- **Out-of-the-box**: does not rely on signals that are specific to only certain domains.

References

- http://sentiment.christopherpotts.net/lexicons.html From Christopher Potts of Dept. of Linguistics, Dan from Science of Stanford University
- Andrea Esuli et al. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining
- 大连理工大学的情感词汇库（徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185）