A Taste of Sentiment Analysis

Rob Zinkov

May 26th, 2011

Outline

1. Introduction

2. Basics of NLP

3. Basic Techniques for Sentiment Analysis

4. Advanced Techniques for Sentiment Analysis

5. Further Questions

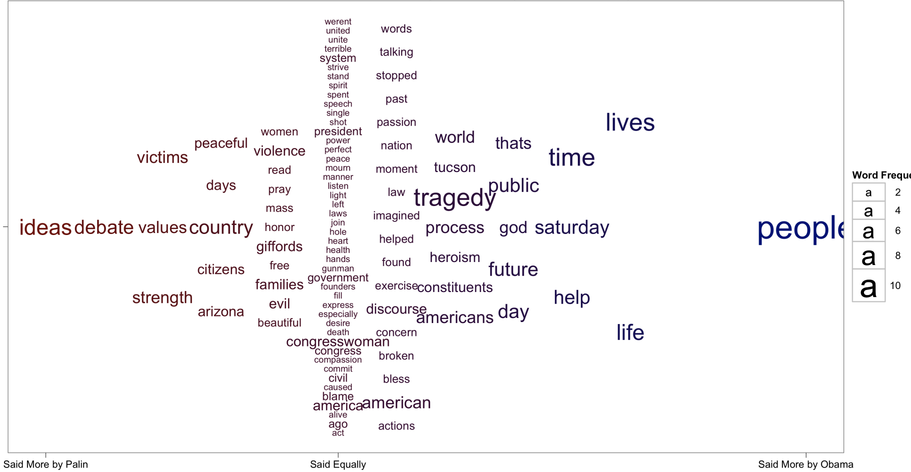
What is Sentiment Analysis?

Sentiment Analysis is a subfield of Computational Linguistics concerned with extracting emotions from text

Applications

Applications - Political Blogs

Word Cloud 2.0, Tuscon Shooting Speeches (Obama vs. Palin)



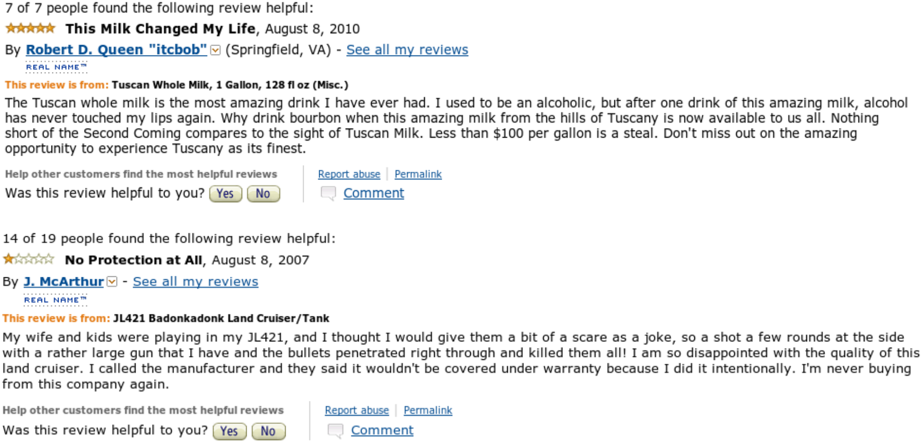
Applications - Political Blogs

- Tracking opinions on issues
- Tracking which issues are held emotionally
- Tracking subjectivity of bloggers

Political Blogs - Challenges

- Identifying opinion holder
- Associating opinions with issue
- Identifying public figures and legislation

Applications - Product Reviews



7 of 7 people found the following review helpful:

★★★★★ **This Milk Changed My Life**, August 8, 2010

By **Robert D. Queen "itcbob"** (Springfield, VA) - See all my reviews

This review is from: Tuscan Whole Milk, 1 Gallon, 128 fl oz (Misc.)

The Tuscan whole milk is the most amazing drink I have ever had. I used to be an alcoholic, but after one drink of this amazing milk, alcohol has never touched my lips again. Why drink bourbon when this amazing milk from the hills of Tuscany is now available to us all. Nothing short of the Second Coming compares to the sight of Tuscan Milk. Less than $100 per gallon is a steal. Don't miss out on the amazing opportunity to experience Tuscany as its finest.

Help other customers find the most helpful reviews
Was this review helpful to you? Yes No
Report abuse | Permalink
Comment

14 of 19 people found the following review helpful:

★☆☆☆☆ **No Protection at All**, August 8, 2007

By **J. McArthur** - See all my reviews

This review is from: JL421 Badonkadonk Land Cruiser/Tank

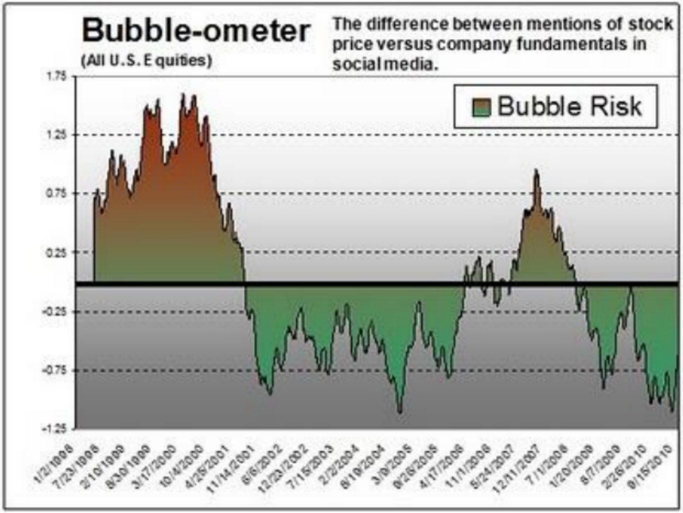
My wife and kids were playing in my JL421, and I thought I would give them a bit of a scare as a joke, so a shot a few rounds at the side with a rather large gun that I have and the bullets penetrated right through and killed them all! I am so disappointed with the quality of this land cruiser. I called the manufacturer and they said it wouldn't be covered under warranty because I did it intentionally. I'm never buying from this company again.

Help other customers find the most helpful reviews
Was this review helpful to you? Yes No
Report abuse | Permalink
Comment

Product Reviews - Challenges

- Identifying aspects of product
- Associating opinions with aspects of product
- Identifying Fake Reviews
- No canonical form

Applications - Financial News



Financial News - Challenges

- Identifying the equity in the article (think commodities)
- Associating entities with market symbols
- Specialized financial terms with distinct sentiment
- Articles rarely only about one equity

Applications - Brand Tracking

Brand Tracking - Challenges

- Text likely to be unstructured
- Identifying Brand
- Identifying Opinion Holder/Demographic

Goals

- Give a broad overview of the field
- Showcase the best current tools and approaches

Caveats

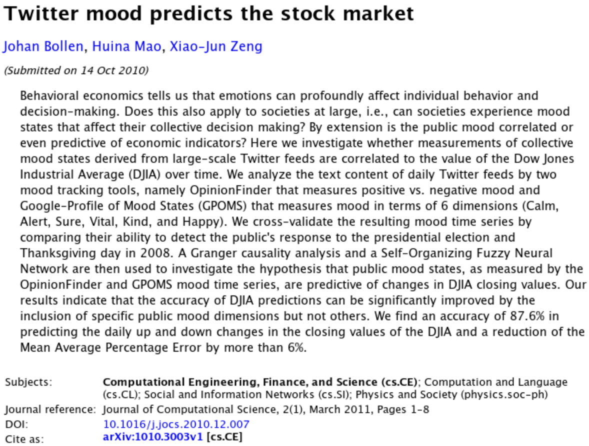
- There are no good R code/libraries to do this (yet)
- This talk is biased towards my domains
- No one in this area really knows what they are doing

History

History

- Grew out of Web integration Field
- Started as extension of knowledge extraction
- This is why field sometimes called Opinion Mining
- Also why papers as likely to occur in ACL as in WWW
- Many early algorithms are extraction patterns
- Field was still largely academic

Then something happened



Twitter mood predicts the stock market

Johan Bollen, Huina Mao, Xiao-Jun Zeng

(Submitted on 14 Oct 2010)

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e., can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We 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). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.

Subjects:	**Computational Engineering, Finance, and Science (cs.CE)**; Computation and Language (cs.CL); Social and Information Networks (cs.SI); Physics and Society (physics.soc-ph)
Journal reference:	Journal of Computational Science, 2(1), March 2011, Pages 1-8
DOI:	10.1016/j.jocs.2010.12.007
Cite as:	**arXiv:1010.3003v1 [cs.CE]**

Unique Challenges in Sentiment Analysis

Opinions are not Facts

Order Matters

- Sentences at end of article have stronger influence on sentiment
- Sentences at beginning of article have stronger influence on sentiment
- Irrelevant sentences influence sentiment of document.

Order Matters - Valience Shifts

The camera is reasonable,but there are far better ones at this price

The meal could have been better,though still tasty.

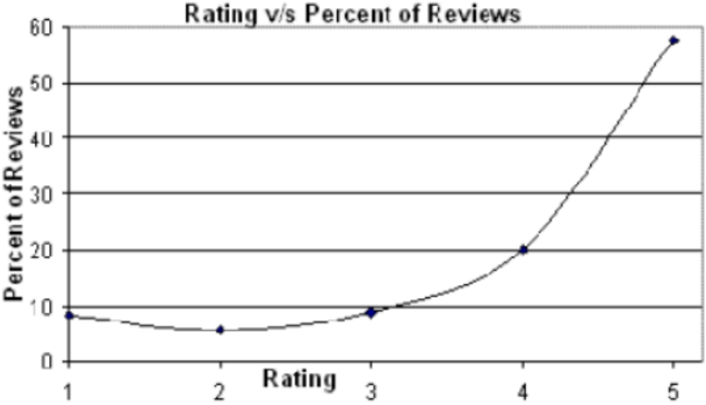
Sentiment Orientation

- shifts in sentiment noted by special words
- special words usually have no sentiment of their own
- sentiment though consistent in each phrase

Sentiment Orientation - continued

- Naive method misses these shifts
- Bag of Words model fails here

Opinions polarize


Rating v/s Percent of Reviews
Percent of Reviews
60
50
40
30
20
10
0
Rating
1
2
3
4
5

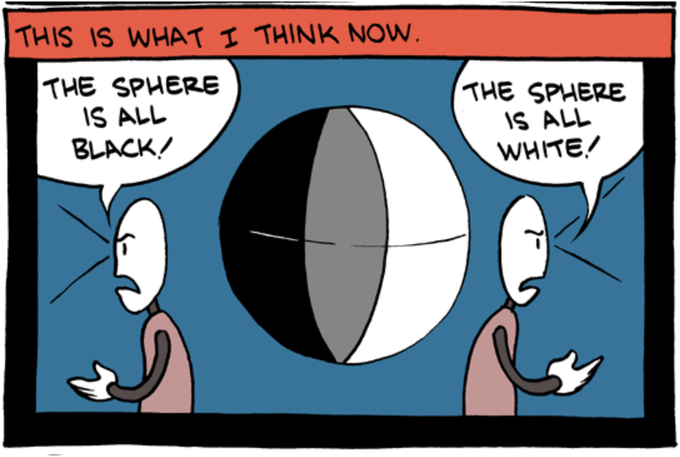
Opinions have context

Small screen

Small carbon footprint

Opinions need to be normalized

People disagree on what words mean



Basics of Natural Language Processing

Introduction to NLP

- Computational Linguistics in centered in Frequency Counts
- Frequency Counts become statistic through which we reason
- This statistic has flaws but still useful

Stemming

It is useful to combine words with a common root.

When counting terms this groups words that denote the same term

This is done by dropping the end

$$\left.\begin{array}{l}\text{sleeping}\\ \text{sleeper}\\ \text{sleeps}\end{array}\right\}\text{sleep}$$

Stopwords

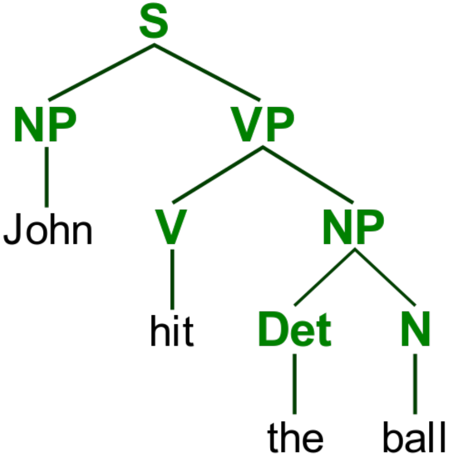
It is important to remove common words as they dominate all counts
Common words in English:

a, the, an, is, be, could, there

Most NLP libraries packaged with a list of stopwords

Sometimes words will need to more finely processed
The following tools exist in most NLP packages
I prefer the Stanford NLP software suite
http://nlp.stanford.edu/software/index.shtml

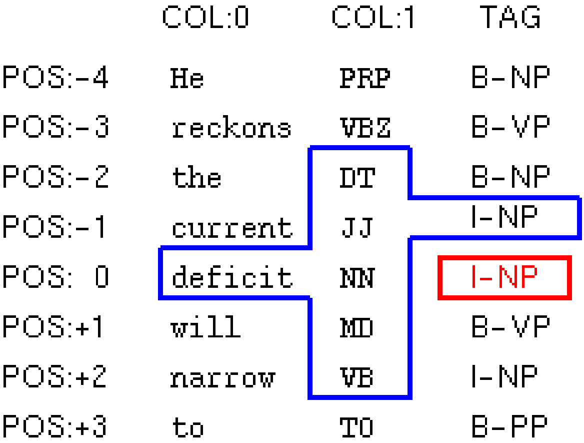
Parsing



Parsing

- Structure also derivable by parsing sentences
- Treat text like programming language
- Algorithms can then convert text into Tree
- Algorithms exist to learn grammar
- Very Heavyweight

Shallow Parsing

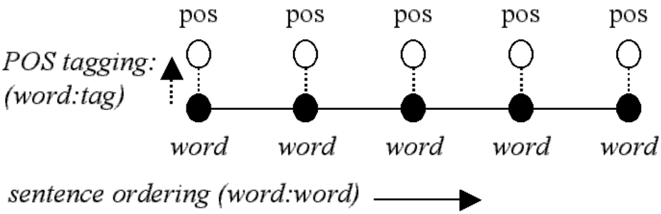


	COL:0	COL:1	TAG
POS:-4	He	PRP	B-NP
POS:-3	reckons	VBZ	B-VP
POS:-2	the	DT	B-NP
POS:-1	current	JJ	I-NP
POS: 0	deficit	NN	I-NP
POS:+1	will	MD	B-VP
POS:+2	narrow	VB	I-NP
POS:+3	to	TO	B-PP

Shallow Parsing

- Less heavy to use than a full parser
- Processes words into phrases
- Training Chunking parser significantly easier/faster
- Requires having words tagged with their part of speech

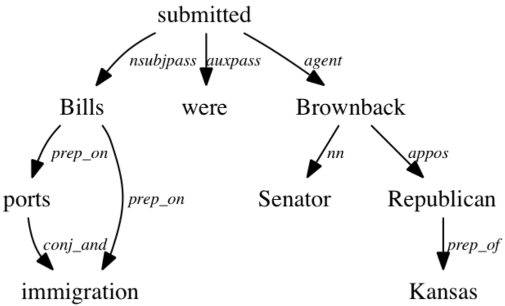
Part of Speech tagging



POS tagging

- Simplest operation to perform on words
- All NLP libraries support this operation
- Provides lightweight metadata
- Very common word feature
- Used by nearly all more complex NLP techniques

Dependency Parsing



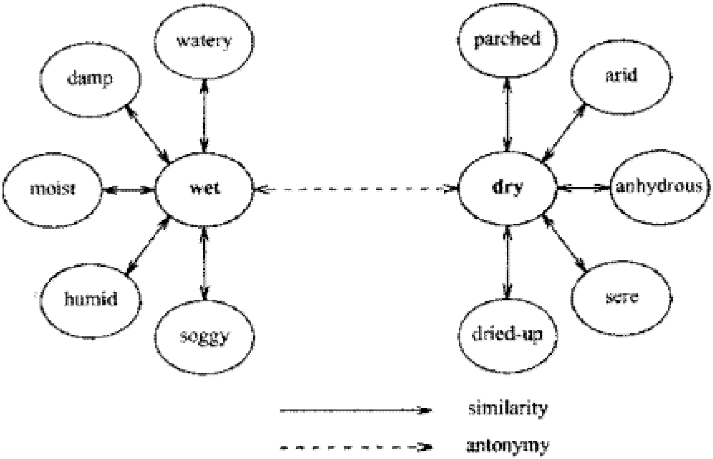
Dependency Parsing

- Traditional Treebank Parsing is a bit bureaucratic
- Hides relations words have with each in sentence
- Dependency Parsing provides a lightweight alternative
- Alternative has looser representation, more language agnostic
- More readily captures which words modify each other

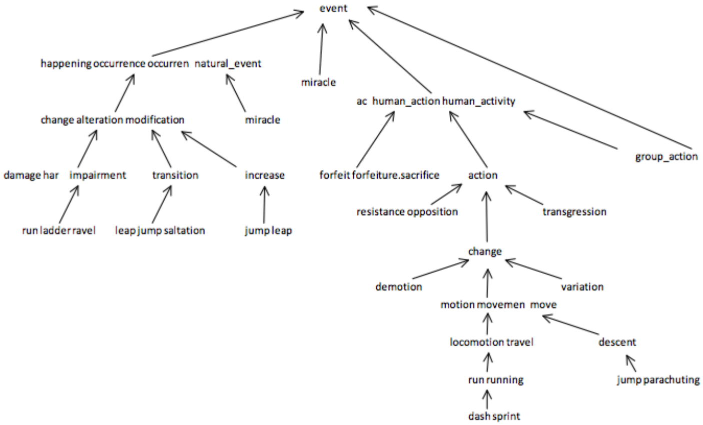
Wordnet

- Words can be related by how similar they are
- Words are similar if they mean similar things
- Words are similar is they are a type of another word
- Words can have many meanings
- Wordnet is a hand curated ontology that annotates these relations

Wordnet synsets



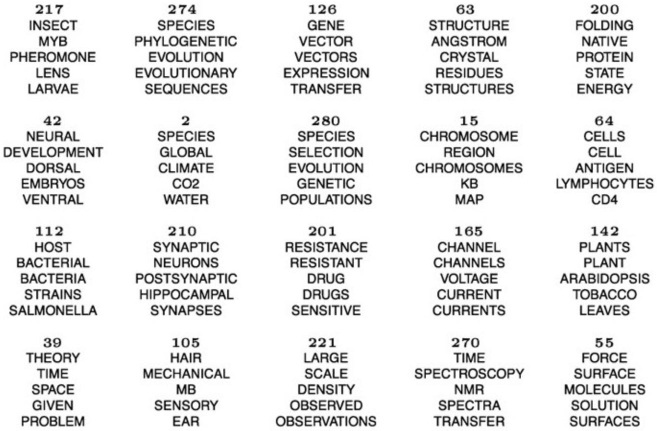
Wordnet concept network



Topic Modeling

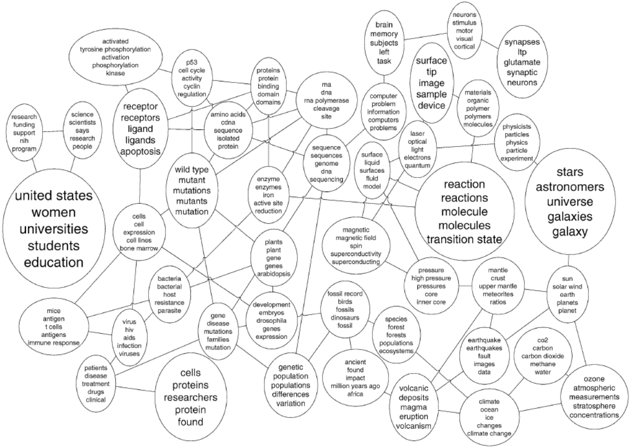
Topic Modeling is a way to group and categorize documents

Usually unsupervised approach



217 INSECT MYB PHEROMONE LENS LARVAE	274 SPECIES PHYLOGENETIC EVOLUTION EVOLUTIONARY SEQUENCES	126 GENE VECTOR VECTORS EXPRESSION TRANSFER	63 STRUCTURE ANGSTROM CRYSTAL RESIDUES STRUCTURES	200 FOLDING NATIVE PROTEIN STATE ENERGY
42 NEURAL DEVELOPMENT DORSAL EMBRYOS VENTRAL	2 SPECIES GLOBAL CLIMATE CO2 WATER	280 SPECIES SELECTION EVOLUTION GENETIC POPULATIONS	15 CHROMOSOME REGION CHROMOSOMES KB MAP	64 CELLS CELL ANTIGEN LYMPHOCYTES CD4
112 HOST BACTERIAL BACTERIA STRAINS SALMONELLA	210 SYNAPTIC NEURONS POSTSYNAPTIC HIPPOCAMPAL SYNAPSES	201 RESISTANCE RESISTANT DRUG DRUGS SENSITIVE	165 CHANNEL CHANNELS VOLTAGE CURRENT CURRENTS	142 PLANTS PLANT ARABIDOPSIS TOBACCO LEAVES
39 THEORY TIME SPACE GIVEN PROBLEM	105 HAIR MECHANICAL MB SENSORY EAR	221 LARGE SCALE DENSITY OBSERVED OBSERVATIONS	270 TIME SPECTROSCOPY NMR SPECTRA TRANSFER	55 FORCE SURFACE MOLECULES SOLUTION SURFACES

CTM - Coorelated Topic Models



CTM - Coorelated Topic Models

- CTMs model the underlying topics within a document
- They differ from earlier approaches in capturing correlations between topics
- Give superior performance compared to other unsupervised models
- Available for use as an R package in CRAN (topicmodels)

Named Entity Recognition

The purpose of NER is to extract out and label phrases in a sentence

Bill Clinton arrived at the United Nations Building in Manhattan.

Sentiment Definitions

Opinion

A vector denoting representing an opinion
with values positive, negative, or neutral gradings

Opinion Holder

The agent an opinion belongs to.
This mostly relevant in political blogs

Item Features

Facets of the object that are readily available

Sentiment Features

Facets of the object that an opinion may be subscribed.
These are usually hard to tease out of the text

1. Gather a Seed set

Opinion corpora available at:

- Wiebe's corpora http://www.cs.pitt.edu/mpqa/
- Sentiwordnet: http://sentiwordnet.isti.cnr.it/
- Personal dictionaries (available on request)

Gathering initial seed words

- Wiebe's work comes with subjectivity scores in addition to sentiment
- Sentiwordnet was autogenerated, quality could be better
- Personal dictionaries hand generated, small but good quality

2. Learn sentiment of unknown words

Learn sentiment - Supervised

Learn sentiment - Supervised

- Get a large collection of them labeled
- Use this collection as is

Learn sentiment - Unsupervised - Turney

- Use Turney's Method
- Calculate Pointwise Mutual Information between every word and the seed words 'excellent' 'poor'

$$SO(w) = lg\left(\frac{hits(\mathrm{w\ NEAR\ excellent})hits(excellent)}{hits(\mathrm{w\ NEAR\ poor})hits(poor)}\right)$$

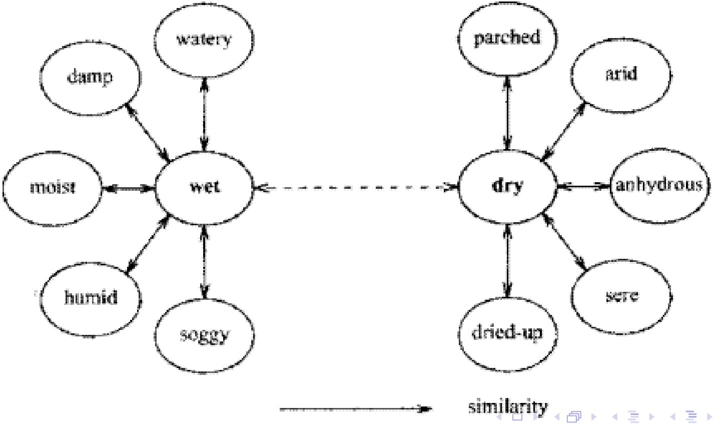
where hits($\mathrm{w\ NEAR\ y}$) = number of times w is within 10 words of the y

Learn sentiment - Unsupervised - Twitter

- Use Turney's Method with Twitter
- Calculate Pointwise Mutual Information between every word and whenever it appears with $\ddot\frown$ or $\ddot\smile$ within a tweet
- This method has the advantage of being multilingual, other kinds of smiles aside

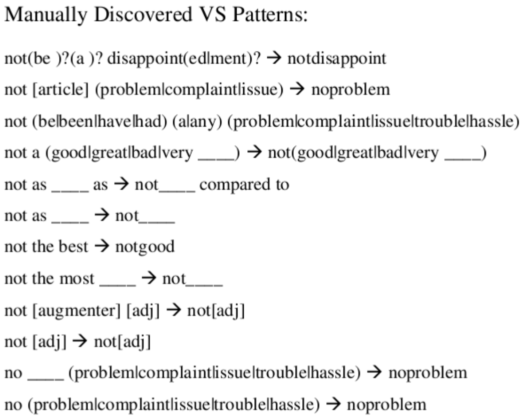
Learn sentiment - Unsupervised - Wordnet

- Use wordnet to walk random paths from start word until arriving at a seed word
- Average across sentiments of all seed words arrived at
- This method is the fastest and most accurate



3. Apply rules to simplify document

- Rules make words more independent
- Rewrites make it less likely to misclassify a phrase



Manually Discovered VS Patterns:

not(be)?(a)? disappoint(ed|ment)? → notdisappoint

not [article] (problem|complaint|issue) → noproblem

not (be|been|have|had) (a|any) (problem|complaint|issue|trouble|hassle)

not a (good|great|bad|very ____) → not(good|great|bad|very ____)

not as ____ as → not____ compared to

not as ____ → not____

not the best → notgood

not the most ____ → not____

not [augmenter] [adj] → not[adj]

not [adj] → not[adj]

no ____ (problem|complaint|issue|trouble|hassle) → noproblem

no (problem|complaint|issue|trouble|hassle) → noproblem

4. Identify opinion phrases

- Shallow Parse the document into chunks
- Remove chunks with mostly neutral words

Alternatively, extract with some rules

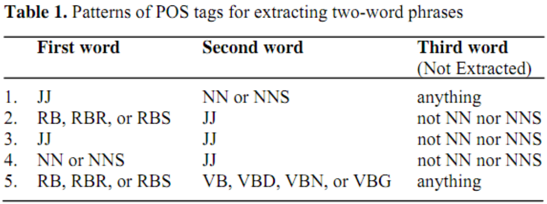


Table 1. Patterns of POS tags for extracting two-word phrases

	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

5. Extend sentiment to phrases and sentences

- Ultimately, sentiment is for phrases and sentences
- Use sentiment on individual words as priors
- Sentiment is based on joint probability across words in phrase
- Use Naive Bayes or a Markov Model as needed

6. Aggregate sentiments for display

Group phrases based on what you want the sentiment

- Entities
- Topics
- Sentiment Features
- Item Features
- Users

8. Generating Summary

Generating Summary

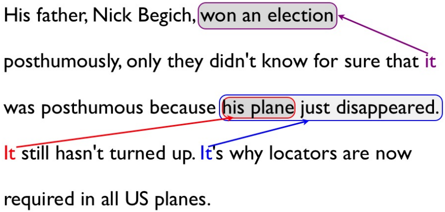
- Largely only relevant when you returning text
- Rate all sentences based on readability
- Return snippet of text for each group with sentiment vector attached

Summary

1. Gather a seed set
2. Learn sentiment of unknown words
3. Apply rules to simplify document
4. Identify opinion phrases
5. Extend sentiment to phrases and document
6. Aggregate sentiments for display
7. Generate summary

Anaphora Resolution

- Many articles refer entities by their name only a few times
- Opinions will usually co-occur with an anaphora of the entity



His father, Nick Begich, won an election posthumously, only they didn't know for sure that it was posthumous because his plane just disappeared. It still hasn't turned up. It's why locators are now required in all US planes.

Anaphora Resolution

- Simplest solution, replace all anaphora with their referent
- Trickier solution, aggregate all opinions associated with anaphora later
- Other options?

Sentiment Analysis is fundamentally a Discriminative Learning Task

Conditional Random Fields

- Sentiment is clearly affected by its surrounding context
- Sentiment is also affected by orientation shifting words
- Why not make these connections explicit in our model?
- Conditional Random Fields (CRFs) are a flexible way of representing these connections.

Conditional Random Fields

In a CRF, we represent posterior probability of a set of sentiments given the underlying text. A is a collection of cliques in the graph of connections.

$$p(y|x) = \frac{1}{Z} \prod_{A} \Psi_A(x_A, y_A)$$

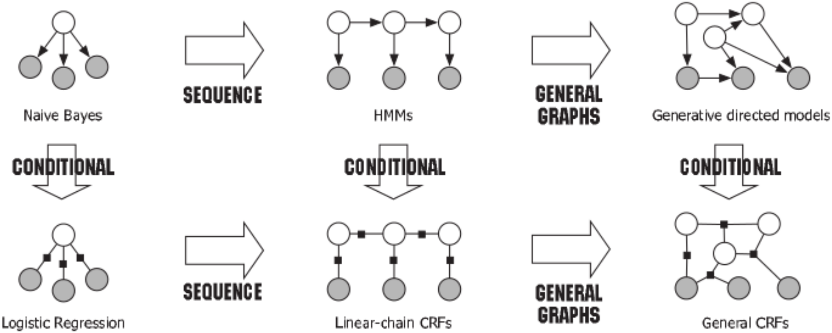
$$\Psi_A(x_A, y_A) = exp \left\{ \sum_{k} \theta_{Ak} f_{Ak}(x_A, y_A) \right\}$$

Linear Chain CRFs

If we assume the sentiment of any given word only depends on the previous, the formula simplifies to

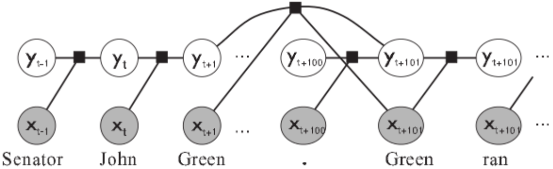
$$p(y|x) = \frac{1}{Z} \prod^{t} exp\left\{ \sum_{k} \theta_k f_k(x_t, y_t, y_{t-1}) \right\}$$

Linear Chain CRFs are best understood as a discriminative version of Hidden Markov Models



Skip-chain CRFs

But we can assume sentiment depends on words much further away



We can now connect entities to each other and connect phrases explicitly separated by a sentiment shifting word.

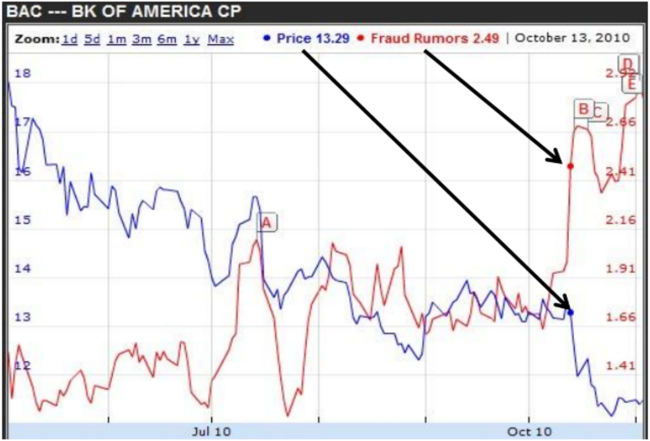
CRFs - Conclusions

- CRFs allow us to add context to opinion
- Properly used they can handle the connections between sentiments on phrases as well as words
- CRFs allow us to link arbitrary features of words and labels to each other

Extensions

Extensions - Time Series

Just order your documents in time, and can plot changes in sentiment

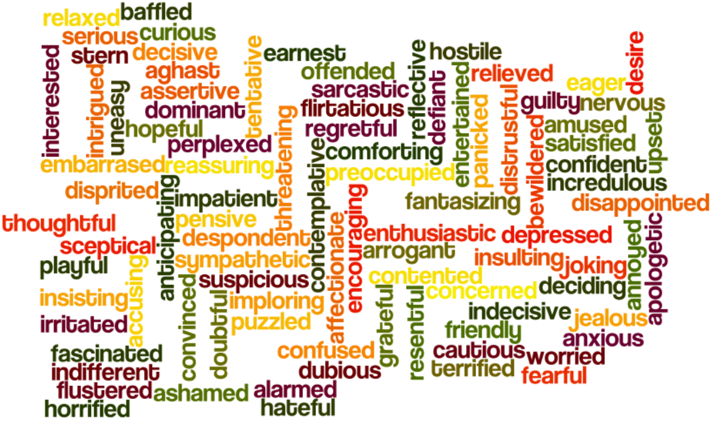


Extensions - Time Series

- This one tends to get used with financial data and monitoring brands
- Requires having access to lots of articles to make sense
- There can be sparsity issues so apply proper shrinkage

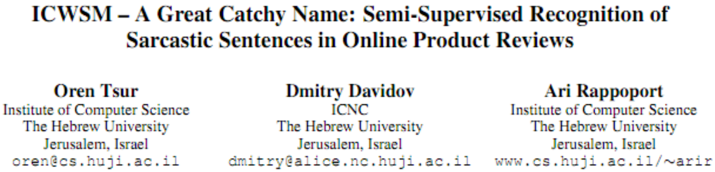
Beyond Positive and Negative

We can be more subtle



Sarcasm

If you deal with Product Review this is helpful



ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews

Oren Tsur
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
oren@cs.huji.ac.il

Dmitry Davidov
ICNC
The Hebrew University
Jerusalem, Israel
dmitry@alice.nc.huji.ac.il

Ari Rappoport
Institute of Computer Science
The Hebrew University
Jerusalem, Israel
www.cs.huji.ac.il/~arir

Sarcasm is best detected through punctuation and capitalization features

Detecting Fake Reviews

- Fake Reviews are best treated as a classification task
- Collect enough and use frequency counts for features
- This is useful in production deployments and simple to implement

Multilingual Sentiment Analysis

- Sentiment does not translate well
- Words that mean the same thing can not correspond wrt sentiment
- Retrain for each new language you wish to support

Word-sense disambiguation

- This is largely not worth the effort
- Using the first sense of the word gives comparable performance to more sophisticated approaches
- Exception: domain specific corpus where word is unlikely to be the first sense. Use specialized dictionaries for this case

Comparisons

- Sometimes opinions are stated relevant two separate entities
- Superlatives are a special case of this
- Treat these as a ranking problem and handle as a separate problem
- Merge sentiments during aggregation

R is much better than SPSS

Lingering Questions

What keeps me from doing this in R?

Further Questions - Large Data

- Text analysis is hard to do in R
- R has memory limits
- Using Hadoop or BigMemory usually means giving up many libraries
- tm.plugins.distributed helps a bit
- snow and OpenMPI gives mixed results

Further Questions - Metadata

Is there a lightweight metadata format?

Index	Offset	Property	Value
2	10	POS	NP
35	5	Sentiment	Positive
17	7	POS	JJ
51	20	Chunk	NULL
20	8	Entity	Person
2	45	Sentence	NULL

Further Questions - Model Files

- Not enough of the tools take model files
- Model files are needed for tokenization,sentence splitting, pos tagging, chunking
- Without easy support for model files, multilingual support is difficult
- Without easy support, impossible to train better models as data becomes available

Further Questions - Rule Files

- No standard on preprocessing rules
- DSL required for them
- Is this something we need to provide?
- Until better techniques come around, essential for any performance

Theoretical Formulation

- Can these techniques be made less hacky?
- Dependency Parses provide much of the structure for tracking sentiment orientation
- Can structure be handled in a more unsupervised manner?

References

Best starting point:
Sentiment Analysis and Subjectivity by Bing Liu
http://www.cs.uic.edu/ liub/FBS/NLP-handbook-sentiment-analysis.pdf

References (More)

- Joint Extraction of Entities and Relations for Opinion Recognition (Choi 2006)
- Mining Opinion Features in Customer Reviews (Liu 2004)
- A Holistic Lexicon-Based Approach to Opinion Mining (Deng 2008)
- I Cant Recommend This Paper Highly Enough (Dillard thesis)
- Entity Discovery and Assignment for Opinion Mining Applications (Deng 2009)
- Extracting Product Features and Opinions from Reviews (Popescu 2005)

Conclusions

- Sentiment Analysis is a relatively young area
- Still plenty of ideas to be explored
- Widely applicable
- Really fun

Questions?