Sentiment Mining
from user generated content

Lyle Ungar

University of Pennsylvania

With thanks to Ronen Feldman a

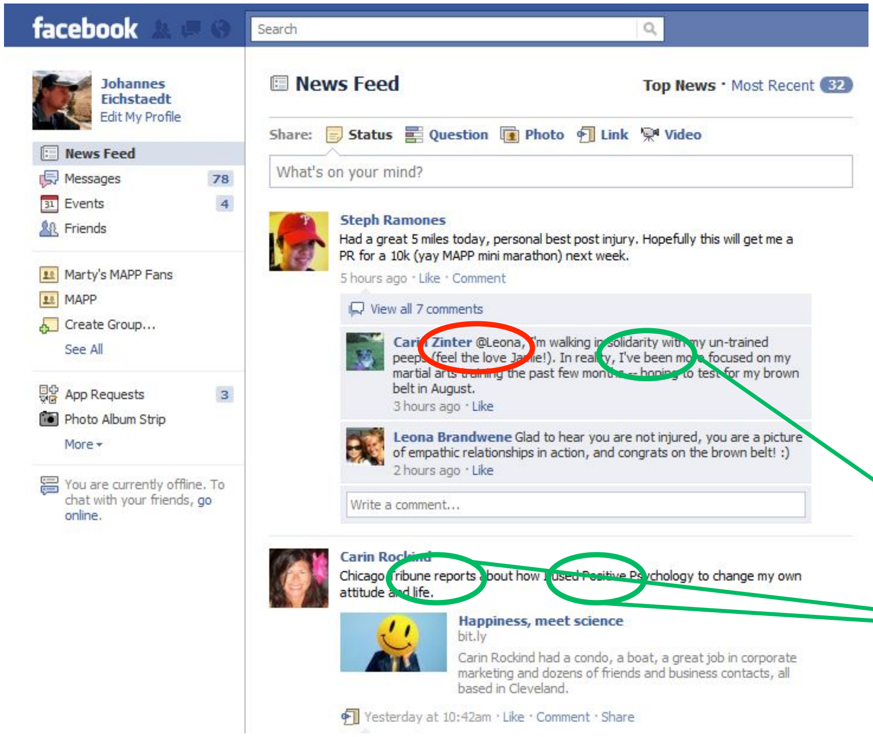
Tutorial outline

- Intro
- Core sentiment analysis (SA) methods
  - **Simple:** using lexica (dictionaries)
  - **Aspect-based:** using information extraction
- Machine learning for SA
  - Unsupervised: open language SA (LDA)
  - Supervised: regression and deep learning
- SA extensions
  - Post, person and community
  - Multi-media

Introductions

- **How many of you have done sentiment mining?**
  - Lexicon-based?
  - Information extraction-based?
- **NLP?**
- **Machine learning**
  - LDA?

What is sentiment analysis?



You are currently offline. To chat with your friends, go online.

News Feed

Steph Ramones Had a great 5 miles today, personal best post injury. Hopefully this will get me a PR for a 10k (yay MAPP mini marathon) next week.
5 hours ago · Like · Comment

Carin Zinter @Leona, I'm walking in solidarity with my un-trained peeps (feel the love Jamie!). In reality, I've been more focused on my martial arts training the past few months -- hoping to test for my brown belt in August.
3 hours ago · Like

Leona Brandwene Glad to hear you are not injured, you are a picture of empathic relationships in action, and congrats on the brown belt! :)
2 hours ago · Like

Carin Rockind Chicago Tribune reports about how I used Positive Psychology to change my own attitude and life.

Happiness, meet science
bit.ly
Carin Rockind had a condo, a boat, a great job in corporate marketing and dozens of friends and business contacts, all based in Cleveland.
Yesterday at 10:42am · Like · Comment · Share

- **What are people saying about my product ?**
  - What do they like or dislike about it?
  - What products do they compare it against?
    - On what dimensions?
  - How are their opinions or priorities changing?

What is sentiment analysis

- **Sentiment analysis**
  - Computational study of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in text.
    - Text = Reviews, blogs, discussions, news, comments, feedback ….
- **Typically: extract from text how people feel about different products**
  - But a huge variety of applications
- **Sometimes called *opinion mining***

Opinions are widely stated

- **Organization internal data**
  - Customer feedback from emails, call centers, etc.
- **News and reports**
  - Opinions in news articles and commentaries
- **Word-of-mouth on the Web**
  - Personal experiences and opinions about anything in reviews, forums, blogs, Twitter, micro-blogs, etc
  - Comments about articles, issues, topics, reviews, etc.
  - Postings at social networking sites, e.g., Facebook.

User-Generated Content

- **Social media**
  - Web pages, blogs, newsgroups
  - Yelp, Zagats …
  - Email, whatsapp, Kik, SMS
  - Facebook, Linkedin
  - Twitter, weibo
  - Pintrest, Instagram, Tumblr

- **Different from newswire and scientific literature**
  - Same general approach, but information extraction is harder in social media

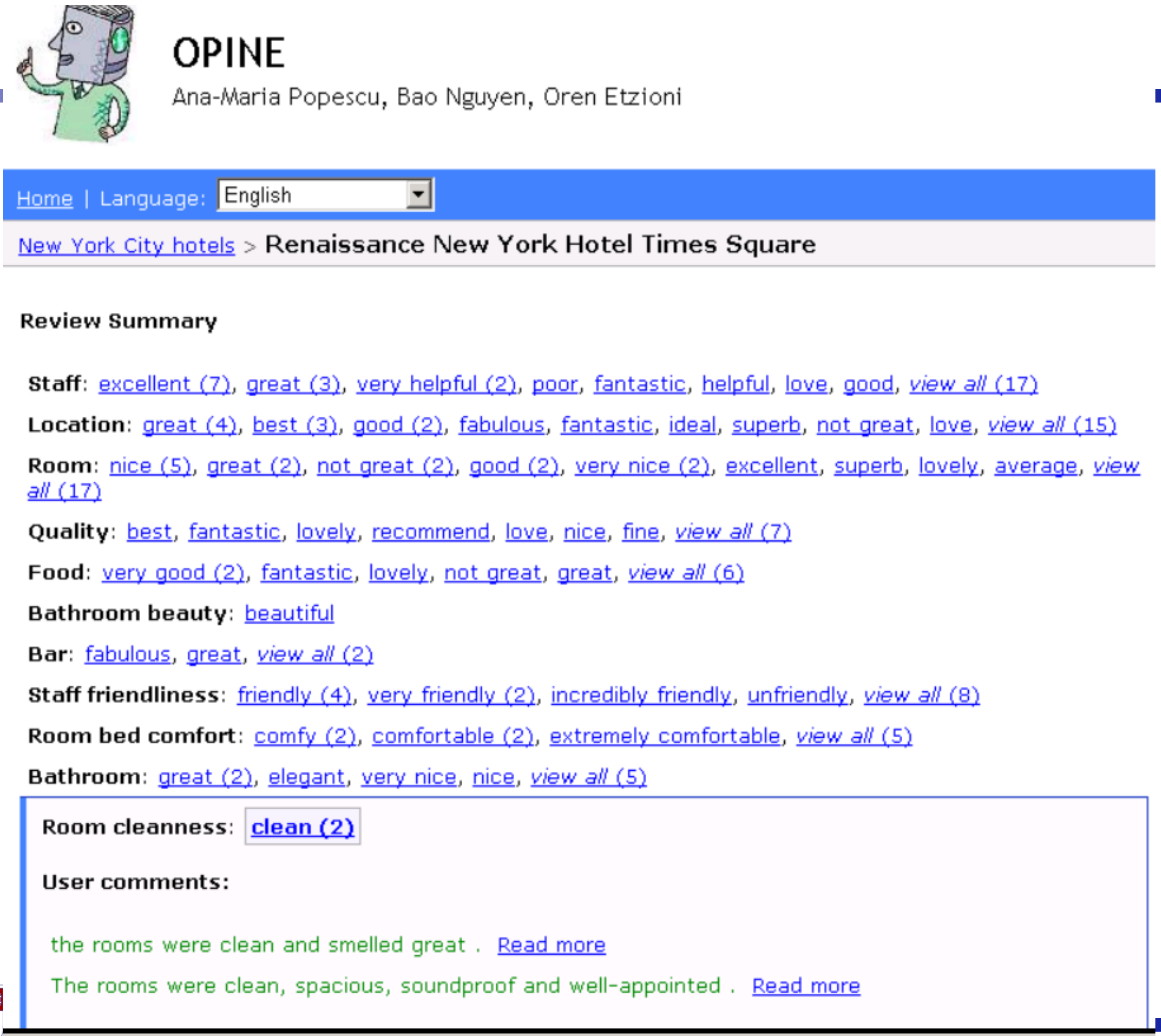
Sentiment Analysis Uses

- **To guide consumer product search**
  - Hotels, restaurants, presidents, barbers, ….
- **To guide product design and marketing**
  - What features do people care about?
  - How many dollars is one point higher ranking worth?
  - What sort of people use my product?
- **Reputation management**
  - What do people think about my product, company, candidate, policy, ..
- **Psychology, political science ···**

Partial substitute for surveys and focus groups

Many different types of sentiment

- **Product rating (simple sentiment analysis)**
  - How many stars did you give the book?
    - Awesome, loved it, terrible, don’t read, give it a skip
- **Aspect (feature) rating**
  - What did you like or dislike about it?
    - Plot, writing, delivery ….
- **Comparative sentiment**
  - “The camry is roomier than the corolla”
- **Emotional**
  - Mood (happy, angry, excited)
  - Emotional connection to product



OPINE

Ana-Maria Popescu, Bao Nguyen, Oren Etzioni

New York City hotels > Renaissance New York Hotel Times Square

Review Summary

Staff: excellent (7), great (3), very helpful (2), poor, fantastic, helpful, love, good, *view all (17)*

Location: great (4), best (3), good (2), fabulous, fantastic, ideal, superb, not great, love, *view all (15)*

Room: nice (5), great (2), not great (2), good (2), very nice (2), excellent, superb, lovely, average, *view all (17)*

Quality: best, fantastic, lovely, recommend, love, nice, fine, *view all (7)*

Food: very good (2), fantastic, lovely, not great, great, *view all (6)*

Bathroom beauty: beautiful

Bar: fabulous, great, *view all (2)*

Staff friendliness: friendly (4), very friendly (2), incredibly friendly, unfriendly, *view all (8)*

Room bed comfort: comfy (2), comfortable (2), extremely comfortable, *view all (5)*

Bathroom: great (2), elegant, very nice, nice, *view all (5)*

Room cleanness: **clean (2)**

User comments:

the rooms were clean and smelled great . Read more

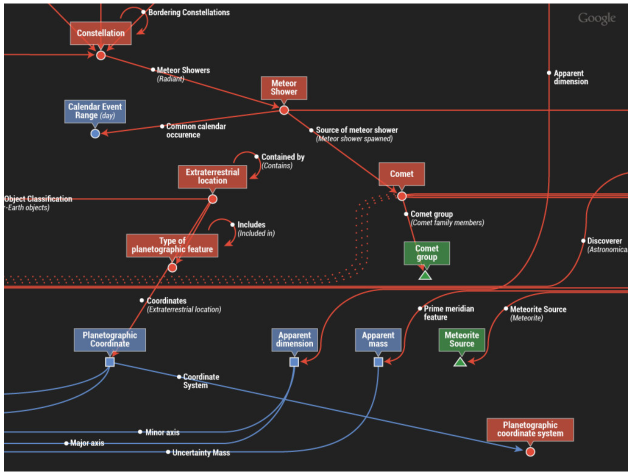
The rooms were clean, spacious, soundproof and well-appointed . Read more

Sentiment Analysis Methods

- **Simple sentiment analysis**
  - Uses *lexica*: collections of positive and negative terms
- **Aspect-based sentiment analysis**
  - Find sentiment about features (“aspects”) of products
  - High quality sentiment analysis builds on *information extraction* (IE)
  - Like most text mining, it works well on aggregate data, but less well on single instances

Sentiment mining builds on IE

- **Information extraction**
  - **Search enhancement**
  - **Question answering**
    - Watson
    - Siri
    - Wolfram's alpha



- **There are lots of "things" out there**
  - Google's *Universe of things* contains over ½ billion things plus attributes and connections

Sentiment Analysis can be tricky

- **Many forms of complexity**
  - Mostly stemming from the challenges of NLP
    - Entity recognition and normalization
    - Complexity and ambiguity of language

- **`Honda Accords and Toyota Camrys are nice sedans`**
- **`Honda Accords and Toyota Camrys are nice sedans, but hardly the best car on the road`**

Sentiment Analysis is Hot

- Over a thousand research papers
- Dozens of companies
  - BuzzMetrics, Reputica, Umbria, Cymfony, BuzzLogic, SentiMetrix ….

BUZZLOGIC

LingPipe

Nielsen BuzzMetrics

umbria

REUTERS

IVolatility.com®

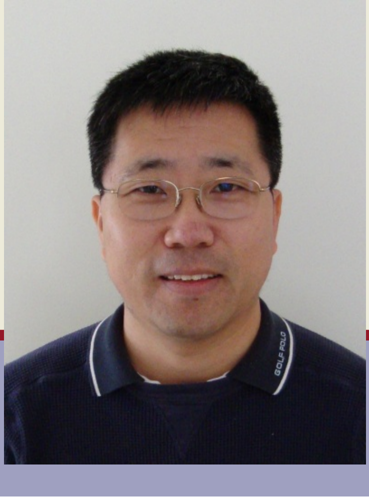
reputica

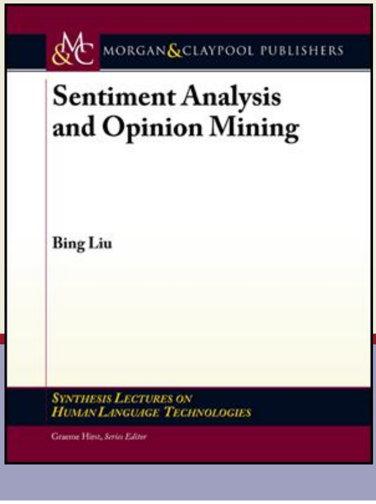
Tutorial outline

- Intro
- **Core sentiment analysis (SA) methods**
  - **Simple:** using lexica (dictionaries)
  - **Aspect-based:** using information extraction
  - variations and complications
- **Machine learning for SA**
  - Unsupervised: open language SA
  - Supervised: regression and deep learning
- **SA extensions**
  - Post, person and community
  - Multi-media

Sentiment Analysis

Based on in part on slides from Bing Liu





M&C MORGAN & CLAYPOOL PUBLISHERS

Sentiment Analysis and Opinion Mining

Bing Liu

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Graeme Hirst, Series Editor

Roadmap

- Problem Definition
  - Opinion identification and summarization
- Simple sentiment classification
- Aspect-based sentiment analysis
- Mining comparative opinions
- Some complications

Sentiment analysis

- **Consists of two parts**

(1) Opinion definition

What is an opinion?

(2) Opinion summarization

An opinion from a single person (unless a VIP) is often not sufficient for action.

- We to need to summarize opinions from many people

- **Can be done on documents, sentences, entities…**

What is an opinion?

- **Id: Abc123 on 5-1-2008 "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"**
- **One can look at this review/blog at the**
  - document level, i.e., is this review + or -?
  - sentence level, i.e., is each sentence + or -?
  - entity and feature/aspect level

Terminology variations

- Entity is also called object.
- Aspect is also called feature, attribute, facet, etc
- Opinion holder is also called opinion source

Entity and aspect/feature level

- **Id: Abc123 on 5-1-2008 "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"**
- **What do we see?**
  - Opinion targets: entities and their features/aspects
  - Sentiments: positive and negative
  - Opinion holders: persons who hold the opinions
  - Time: when opinions are expressed

Two main types of opinions
(Jindal and Liu 2006; Liu, 2010)

- **Regular opinions: Sentiment/opinion expressions on some target entities**
  - Direct opinions:
    - "The touch screen is really cool."
  - Indirect opinions:
    - "After taking the drug, my pain has gone."
- **Comparative opinions: Comparisons of more than one entity.**
  - E.g., "iPhone is better than Blackberry."
- **We focus on regular opinions, and just call them opinions.**

A (regular) opinion

- An *opinion has the following basic components*

  $(g_i, so_{ijkl}, h_i, t_l)$,

  where

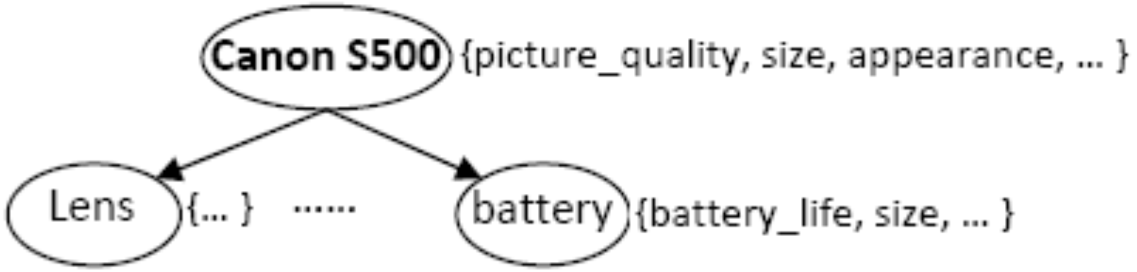
  - g_j is a target
  - so_{ijl} is the sentiment value of the opinion from opinion holder h_i on target g_j at time t_l. so_{ijl} is positive, negative or neutral, or a rating score
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Opinion target

- **In some cases, opinion target is a single entity or topic.**
  - *"I love iPhone"* and *"I support tax cut."*
- **But in many other cases, it is more complex.**
  - *"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool."*
    - Opinion target of the 3rd sentence is not just touch screen, but the "touch screen of iPhone".
  - *"I support tax cut for the middle class, but for the rich"*
- **We decompose the opinion target**

Entity and aspect (Hu and Liu, 2004; Liu, 2006)

- Definition (entity): An *entity e* is a product, person, event, organization, or topic. e is represented as
  - a hierarchy of components, sub-components, and so on.
  - Each node represents a component and is associated with a set of attributes of the component.


Canon S500
{picture_quality, size, appearance, ... }
Lens
{... }
......
battery
{battery_life, size, ... }

- An opinion can be expressed on any node or attribute of the node.
- For simplicity, we use the term *aspects* (features) to represent both components and attributes.

Opinion definition (Liu, Ch. in NLP handbook, 2010)

- An *opinion* is a quintuple

  $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

  where

  - e_j is a target entity.
  - a_{jk} is an aspect/feature of the entity e_j.
  - so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on aspect a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Our example blog in quintuples

- **Id: Abc123 on 5-1-2008 "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"**
- **In quintuples**

  (iPhone, GENERAL, +, Abc123, 5-1-2008)

  (iPhone, touch_screen, +, Abc123, 5-1-2008)

  ….

  - We will discuss comparative opinions later.

Document vs. entity/aspect sentiment

- **Goal: Given an opinionated document,**
  - classify the sentiment of the entire document or
  - Find all quintuples (e_j, a_{jk}, so_{ijkl}, h_i, t_l),
- **With the quintuples,**
  - Unstructured Text → Structured Data
    - Traditional data and visualization tools can be used to slice, dice and visualize the results.
    - Enables qualitative and quantitative analysis.

Subjectivity

- **Sentence subjectivity: An *objective sentence* presents some factual information, while a *subjective sentence* expresses some personal opinions, beliefs, views, feelings, or emotions.**
  - Not the same as emotion

Subjectivity

- Subjective expressions come in many forms, e.g., opinions, allegations, desires, beliefs, suspicions, and speculations (Wiebe, 2000; Riloff et al 2005).
  - A subjective sentence may contain a positive or negative opinion
- Most opinionated sentences are subjective, but objective sentences can imply opinions too (Liu, 2010)
  - “The machine stopped working in the second day”
  - “We brought the mattress yesterday, and a body impression has formed.”
  - “After taking the drug, there is no more pain”

Rational and emotional evaluations

- Rational evaluation: Many evaluation/opinion sentences express no emotion
  - e.g., “The voice of this phone is clear”
- Emotional evaluation
  - e.g., “I love this phone”
  - “The voice of this phone is crystal clear” (?)
- Some emotion sentences express no (positive or negative) opinion/sentiment
  - e.g., “I am so surprised to see you”.

Opinion summary

- **With many opinions, a summary is necessary.**
  - A multi-document summarization task
- **For factual texts, summarization selects the most important facts and avoids repetition**
  - 1 fact = any number of the same fact
- **Opinion documents are different: opinions have a quantitative side & have targets**
  - 1 opinion ≠ a number of opinions
  - Aspect-based summary is more suitable
    - Quintuples form the basis for opinion summarization

Feature-based opinion summary[1]

(Hu & Liu, 2004)

““I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …”

1.

….

Feature Based Summary of iPhone:

Feature1: Touch screen

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…

Feature2: voice quality

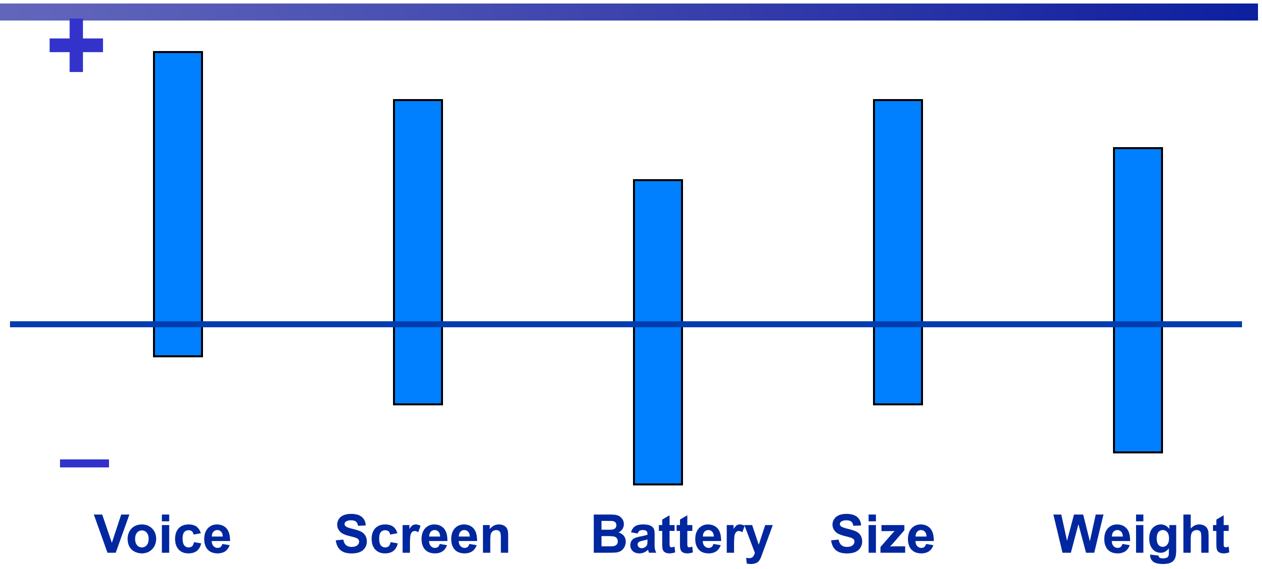
…

Note: We omit opinion holders

Opinion Observer (Liu et al. 2005)

Summary of reviews of Cell Phone 1

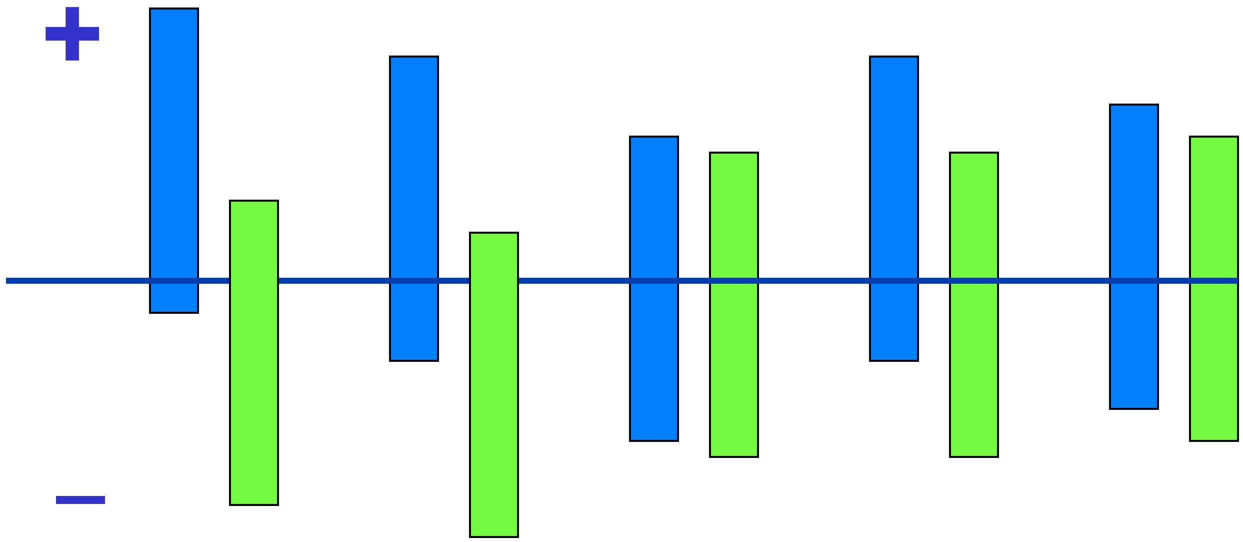
🟦



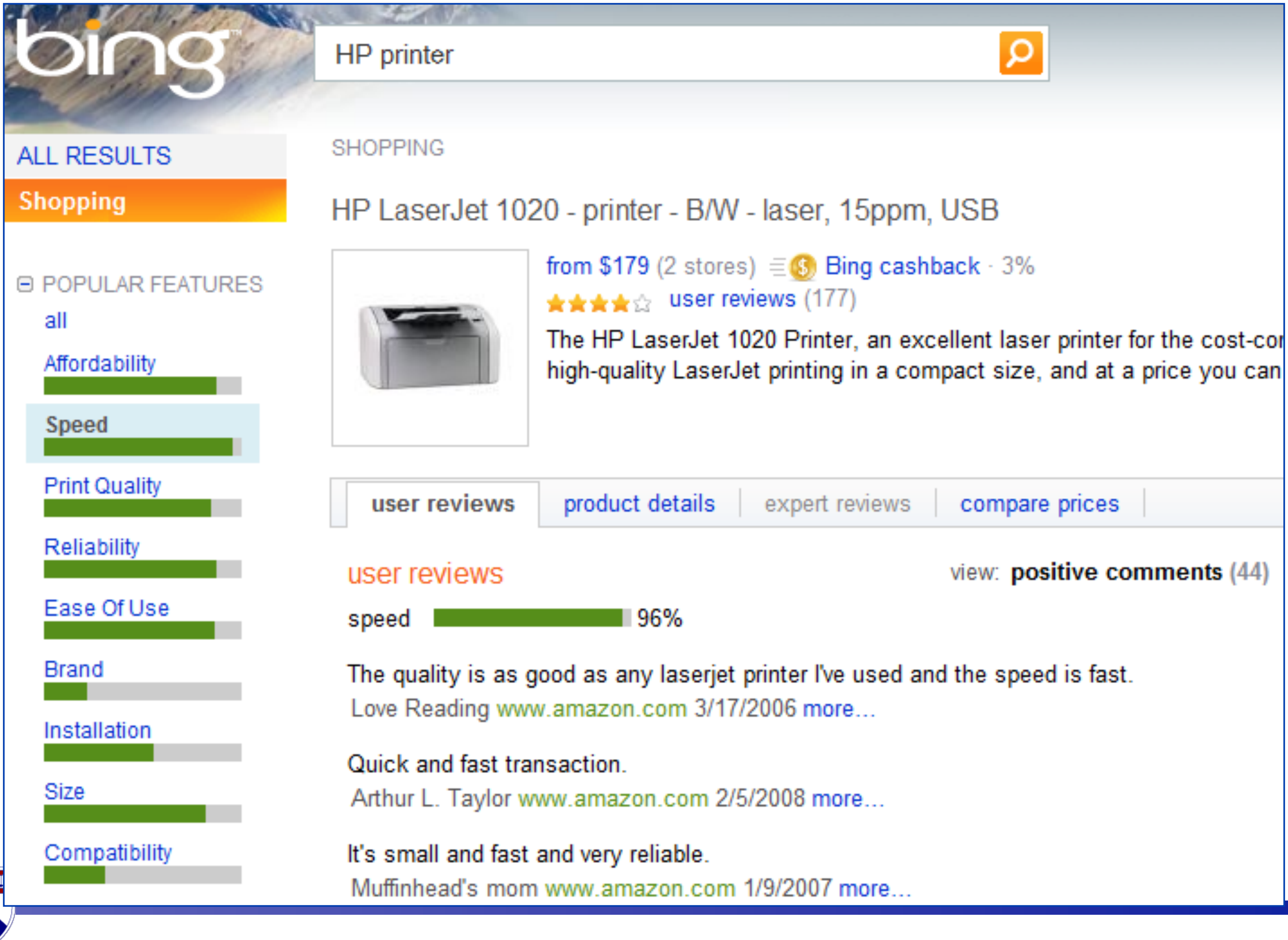
Comparison of reviews of

🟦 Cell Phone 1

🟩 Cell Phone 2



Aspect-based opinion summary



HP printer

ALL RESULTS

Shopping

POPULAR FEATURES

all

Affordability

Speed

Print Quality

Reliability

Ease Of Use

Brand

Installation

Size

Compatibility

SHOPPING

HP LaserJet 1020 - printer - B/W - laser, 15ppm, USB

from $179 (2 stores) Bing cashback · 3%

★★★★☆ user reviews (177)

The HP LaserJet 1020 Printer, an excellent laser printer for the cost-co high-quality LaserJet printing in a compact size, and at a price you can

user reviews | product details | expert reviews | compare prices

user reviews

view: **positive comments** (44)

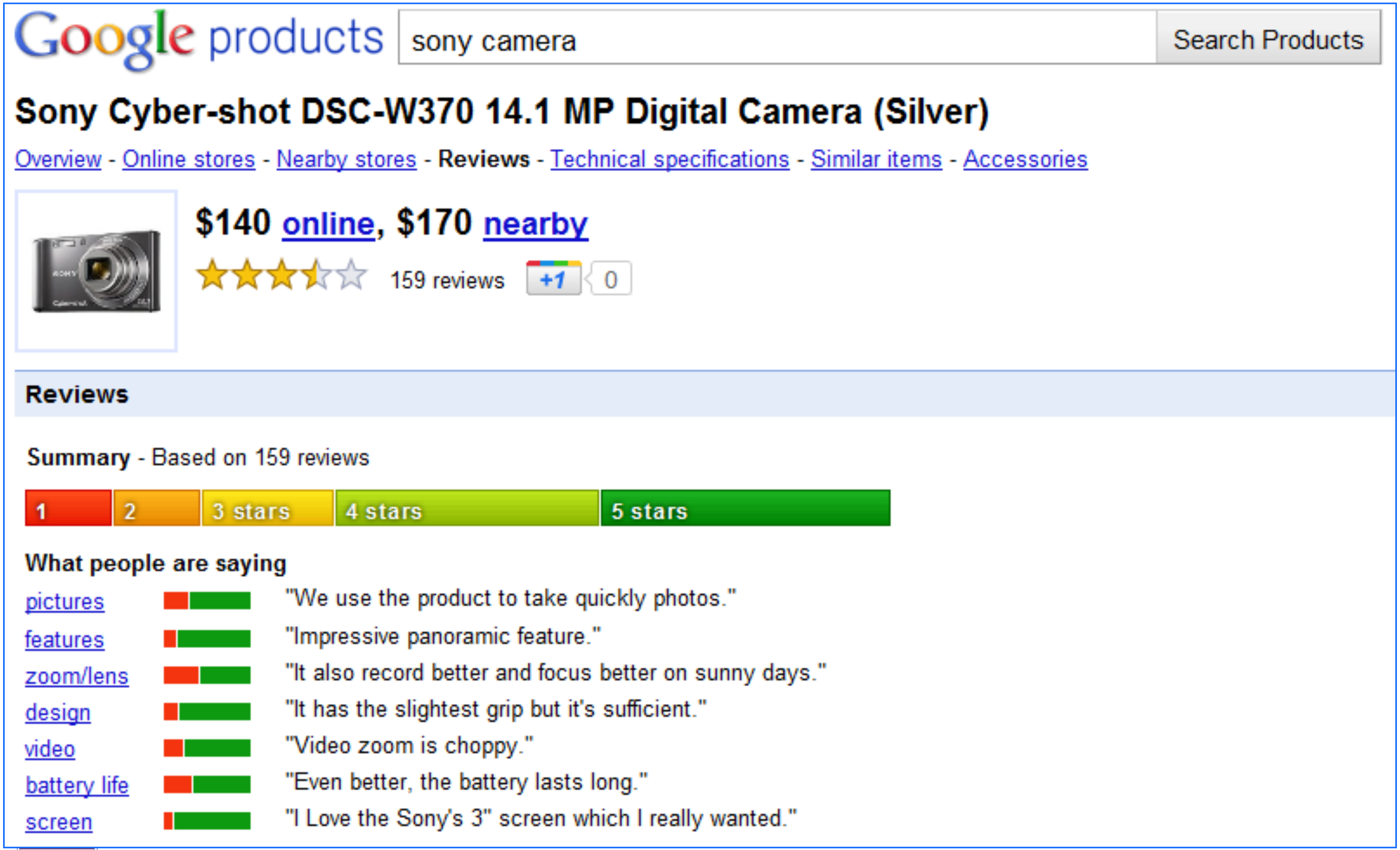
speed 96%

The quality is as good as any laserjet printer I've used and the speed is fast.
Love Reading www.amazon.com 3/17/2006 more...

Quick and fast transaction.
Arthur L. Taylor www.amazon.com 2/5/2008 more...

It's small and fast and very reliable.
Muffinhead's mom www.amazon.com 1/9/2007 more...

Google Product Search (Blair-Goldensohn et al 2008 ?)



Sony Cyber-shot DSC-W370 14.1 MP Digital Camera (Silver)

Overview - Online stores - Nearby stores - **Reviews** - Technical specifications - Similar items - Accessories

$140 online, $170 nearby

★★★⯪☆ 159 reviews +1 0

Reviews

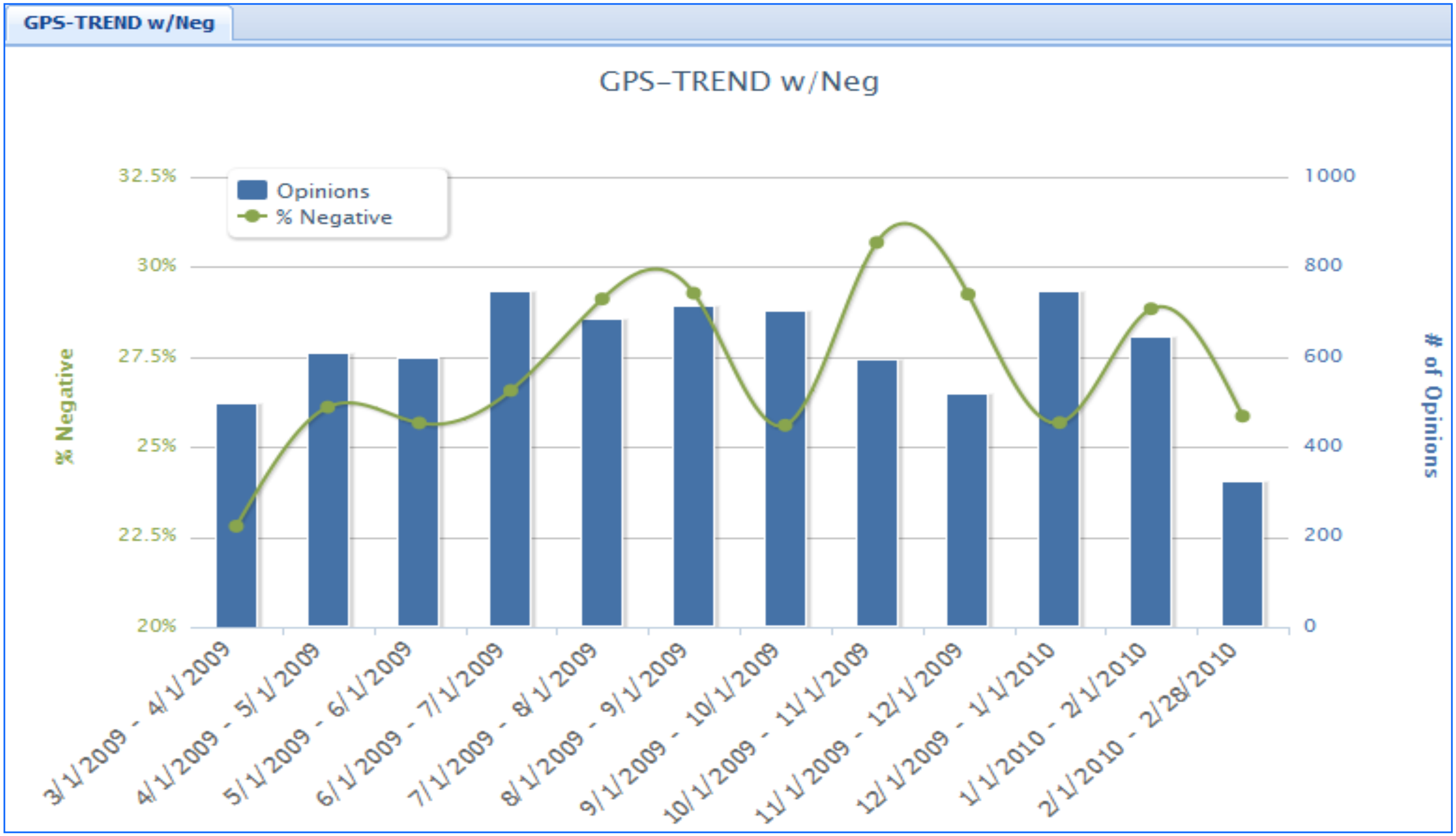
Summary - Based on 159 reviews

1	2	3 stars	4 stars	5 stars

What people are saying

pictures	🟥🟩	"We use the product to take quickly photos."
features	🟥🟩	"Impressive panoramic feature."
zoom/lens	🟥🟩	"It also record better and focus better on sunny days."
design	🟥🟩	"It has the slightest grip but it's sufficient."
video	🟥🟩	"Video zoom is choppy."
battery life	🟥🟩	"Even better, the battery lasts long."
screen	🟥🟩	"I Love the Sony's 3" screen which I really wanted."

Aggregate opinion trend



Sentiment mining requires solving coupled IE problems

- **(e_j, a_{jk}, so_{ijkl}, h_i, t_l),**
  - e_j - a target entity: Named Entity Extraction (more)
  - a_{jk} – an aspect/feature of e_j: Information Extraction
  - so_{ijkl} is sentiment: Sentiment Identification
  - h_i is an opinion holder: Information/Data Extraction
  - t_l is the time: Information/Data Extraction
  - These 5 pieces of information must match
- **Requires coreference and entity resolution**

Easier and harder problems

- **Tweets from Twitter are probably the easiest**
  - short and thus usually straight to the point
- **Reviews are next**
  - entities are given (almost) and there is little noise
- **Discussions, comments, and blogs are hard.**
  - Multiple entities, comparisons, noisy, sarcasm, etc
- **Extracting entities and aspects, and determining sentiments/opinions about them are hard.**
- **Combining them is harder.**

Sentiment Analysis Methods

- **Simple Sentiment Analysis**
  - Classify documents or sentences based on positive or negative sentiment words
  - Use standard lexicon or learn a custom one
- **Aspect-based Sentiment Analysis**
  - Use *information extraction* and *named entity recognition* to find (entity, aspect/feature, sentiment, holder, time)
  - Doing it “right” is very hard, requiring NER and IE
- **Visualization**

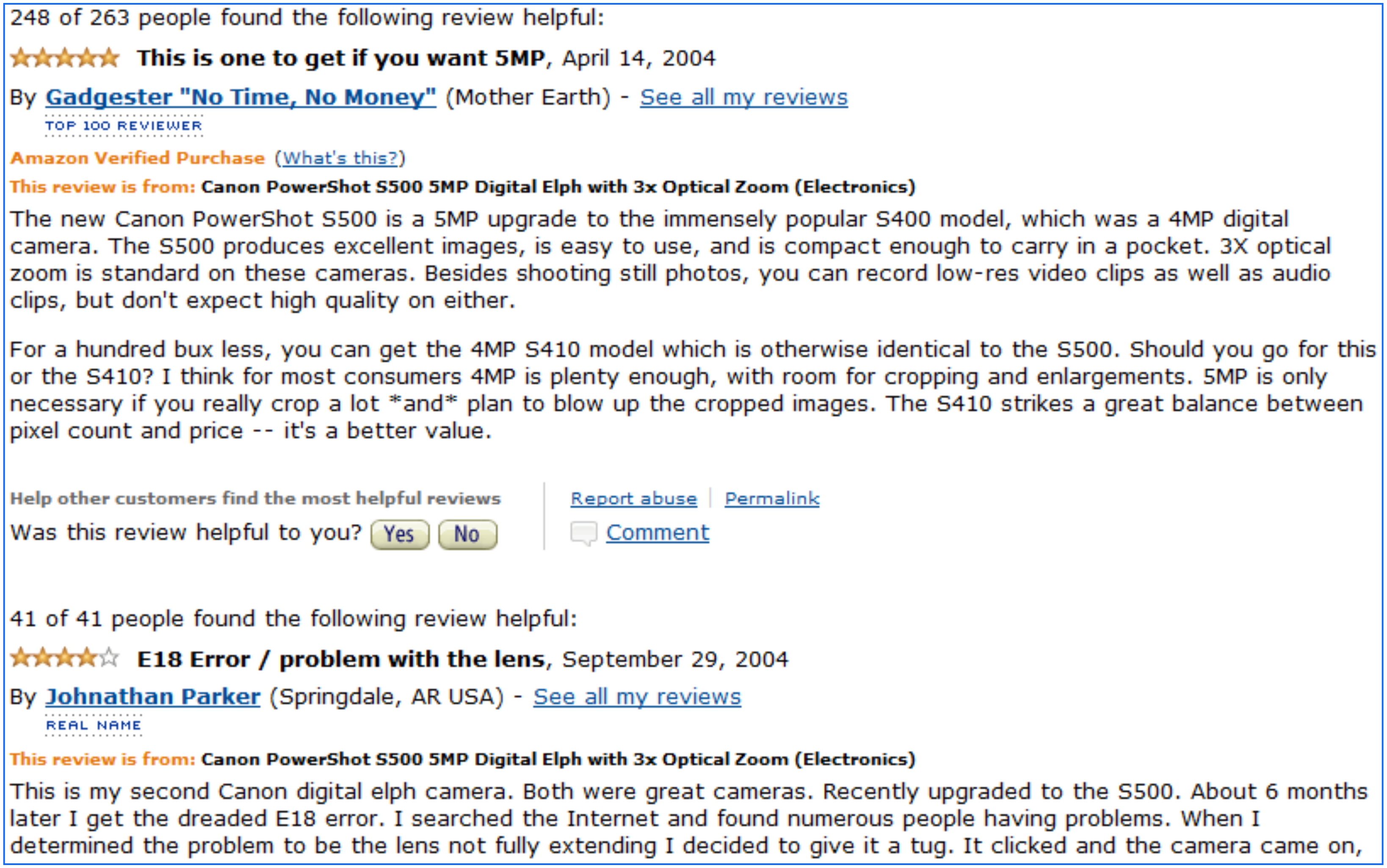
Roadmap

- Problem Definition
- Simple Sentiment analysis
  - Document sentiment classification
  - Sentiment lexica and their generation
- Aspect-based sentiment analysis
- Mining comparative opinions
- Some complications

Sentiment classification

- **Classify a whole opinion document (e.g., a review) based on the overall sentiment of the opinion holder** **(Pang et al 2002; Turney 2002)**
  - Classes: Positive, negative (possibly neutral)
  - Neutral or no opinion is hard. Most papers ignore it.
- **An example review:**
  - *"I bought an iPhone a few days ago. It is such a nice phone, although a little large. The touch screen is cool. The voice quality is clear too. I simply love it!"*
  - Classification: positive or negative?
- **Perhaps the most widely studied problem.**

Some Amazon reviews



248 of 263 people found the following review helpful:

★★★★★ **This is one to get if you want 5MP**, April 14, 2004

By **Gadgester "No Time, No Money"** (Mother Earth) - See all my reviews

TOP 100 REVIEWER

Amazon Verified Purchase (What's this?)

This review is from: **Canon PowerShot S500 5MP Digital Elph with 3x Optical Zoom (Electronics)**

The new Canon PowerShot S500 is a 5MP upgrade to the immensely popular S400 model, which was a 4MP digital camera. The S500 produces excellent images, is easy to use, and is compact enough to carry in a pocket. 3X optical zoom is standard on these cameras. Besides shooting still photos, you can record low-res video clips as well as audio clips, but don't expect high quality on either.

For a hundred bux less, you can get the 4MP S410 model which is otherwise identical to the S500. Should you go for this or the S410? I think for most consumers 4MP is plenty enough, with room for cropping and enlargements. 5MP is only necessary if you really crop a lot *and* plan to blow up the cropped images. The S410 strikes a great balance between pixel count and price -- it's a better value.

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

41 of 41 people found the following review helpful:

★★★★☆ **E18 Error / problem with the lens**, September 29, 2004

By **Johnathan Parker** (Springdale, AR USA) - See all my reviews

REAL NAME

This review is from: **Canon PowerShot S500 5MP Digital Elph with 3x Optical Zoom (Electronics)**

This is my second Canon digital elph camera. Both were great cameras. Recently upgraded to the S500. About 6 months later I get the dreaded E18 error. I searched the Internet and found numerous people having problems. When I determined the problem to be the lens not fully extending I decided to give it a tug. It clicked and the camera came on,

Simplest approach: use a lexicon

- **Count ratio of positive to negative words in the document**
  - More positive words = more positive sentiment
  - Works poorly for individual reviews but works well on average
- **Or use machine learning to discover weights on the sentiment words**
  - Or on all words

Sentiment (or opinion) lexicon

- **Sentiment lexicon: lists of words and expressions used to express people's subjective feelings and sentiments/ opinions.**
  - Not just individual words, but also phrases and idioms, e.g., "cost an arm and a leg"
- **Many sentiment lexica can be found on the web**
  - They often have thousands of terms, and are quite useful

Sentiment lexicon

- **Sentiment words or phrases (also called polar words, opinion bearing words, etc). E.g.,**
  - Positive: beautiful, wonderful, good, amazing,
  - Negative: bad, poor, terrible, cost an arm and a leg.
- **Many of them are context dependent, not just application domain dependent.**
- **Three main ways to compile such lists:**
  - Manual approach: not a bad idea for a one-time effort
  - Corpus-based approach
  - Dictionary-based approach

Sentiment Lexicons

- Bing Liu's Opinion Lexicon
- MPQA Subjectivity Lexicon
- SentiWordNet
- Harvard General Inquirer
- LIWC (Linguistic Inquiry and Word Count)
- NRC Hashtag Sentiment Lexicon (for tweets)
- …

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

Dictionary- vs. Corpus-based

- **Dictionary-based methods**
  - Typically use WordNet’s synsets and hierarchies to acquire opinion words
  - Usually do not give domain or context dependent meanings
- **Corpus-based approaches**
  - Often use a double propagation between opinion words and the items they modify
  - Require a large corpus to get good coverage

Simple sentiment prediction can be done by supervised learning

- **Basically a text classification problem**
  - Features = words in a document
  - Label = the sentiment
- **But different from topic-based text classification.**
  - In topic-based text classification (e.g., computer, sport, science), topic words are important.
  - In sentiment classification, opinion/sentiment words are more important, e.g., great, excellent, horrible, bad, worst, etc.

Simple sentiment assumption and goal

- **Assumption: The doc is written by a single person and express opinion/sentiment on a single entity.**
- **Goal: discover (_, _, *so*, _, _),**

  where e, a, h, and t are ignored
- **Reviews usually satisfy the assumption.**
  - Almost all papers use reviews
  - Positive: 4 or 5 stars, negative: 1 or 2 stars
- **Many forum postings and blogs do not**
  - They can mention and compare multiple entities
  - Many such postings express no sentiments

Supervised learning example

- **Training and test data**
  - Movie reviews with star ratings
    - 4-5 stars as positive
    - 1-2 stars as negative
- **Neutral is ignored**
- **SVM gives the best classification accuracy based on balanced training data**
  - Features: unigrams (bag of individual words)
    - But POS tags, phrases, parsing, and negation help
  - Typical result: 80-90% accuracy

Sentence level analysis

- **Document-level sentiment classification is too coarse for most applications.**
- **So do sentence level analysis**
  - Assumes a single sentiment per sentence
  - Not always true; one can classify clauses instead

Sentence sentiment analysis

- **Usually consists of two steps**
  - Subjectivity classification
    - To identify subjective sentences
  - Sentiment classification of subjective sentences
    - As positive or negative
- **But bear in mind**
  - Many objective sentences can imply sentiments
  - Many subjective sentences do not express positive or negative sentiments/opinions
    - E.g., ”I believe he went home yesterday.”

Subjectivity classification using patterns (Rilloff and Wiebe, 2003)

- **A bootstrapping approach.**
  - A high precision classifier is first used to automatically identify some subjective and objective sentences.
    - Two high precision (but low recall) classifiers are used,
      - A high precision subjective classifier
      - A high precision objective classifier
      - Based on manually collected lexical items, single words and n-grams, which are good subjective clues.
  - A set of patterns are then learned from these identified subjective and objective sentences.
    - Syntactic templates are provided to restrict the kinds of patterns to be discovered, e.g., <subj> passive-verb.
  - The learned patterns are then used to extract more subjective and objective sentences (the process can be repeated).

Roadmap

- Sentiment Analysis Problem
- Simple sentiment classification
- Aspect-based sentiment analysis
  - Based on Information Extraction
- Mining comparative opinions
- Some complications

We need to go further

- **Sentiment classification at both the document and sentence (or clause) levels are useful, but**
  - They do not find what people liked and disliked.
- **They do not identify the targets of opinions, i.e.,**
  - Entities and their aspects
  - Without knowing targets, opinions are of limited use.
- **We need to go to the entity and aspect level.**

This looks a lot like IE!

Recall an opinion is a quintuple

- An *opinion* is a quintuple

  $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

  where

  - e_j is a target entity.
  - a_{jk} is an aspect/feature of the entity e_j.
  - so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Aspect-based sentiment analysis

- Much of the research is based on online reviews
- For reviews, aspect-based sentiment analysis is easier because the entity (i.e., product name) is usually known
  - Reviewers simply express positive and negative opinions on different aspects of the entity.
- For blogs, forum discussions, etc., it is harder:
  - both entity and aspects of entity are unknown,
  - there may also be many comparisons, and
  - there is also a lot of irrelevant information.

Find entities (entity set expansion)

- Uses named entity recognition (NER) and resolution
- E.g., one wants to study opinions on phones
  - given Motorola and Nokia, find all phone brands and models in a corpus, e.g., Samsung, Moto,…
- Entity resolution: do these two names refer to the same brand?

Feature/Aspect extraction

- **Extraction may use:**
  - **Frequent nouns and noun phrases**
    - Sometimes limited to a set known to be related to the entity of interest or using **part discriminators**
      - e.g., for a scanner entity: “of scanner”, “scanner has”,
  - **Opinion and target relations**
    - Proximity or syntactic dependency
  - **Standard IE methods**
    - Rule-based or supervised learning
      - Often HMMs or CRFs (like standard IE)

Standard IE – a quick summary

- IE can be done using templates

    offices in <*loc*>

    operates in <*loc*>

    facilities in <*loc*>

    owned by <*company*>

    <*company*> has positions

    offices of <*company*>

- But in increasingly done using sophisticated NLP

IE – Typical Pipeline

- Extract raw text (html, pdf, ps, gif)
- Tokenize
- Detect term boundaries
  - We extracted *alpha 1 type XIII collagen* from …
  - Their house council recommended …
- Detect sentence boundaries
- Tag parts of speech (POS)
  - *John*/noun *saw*/verb *Mary*/noun.
- Tag named entities
  - Person, place, organization, gene, chemical
- Parse
- Determine co-reference
- Extract knowledge

Parsing is hard

So most people use dependency parsers, chunking, or simple patterns

```
                     S
                 /       \
             /              \
          NP                 VP
  /   /     \     \     \       |  \       \
DT  NNP    NNP   NNP    NNP   VBD  NP       SBAR
 |   |      |     |      |     |   NNP
The Fulton County Grand Jury  said Friday    …
```

```
         S
   NP                  PP   NP
DT   NN             IN   NNP    POS  JJ     JJ      NN
an   investigation  of   Atlanta's   recent primary election …
```

```
( (S
          (NP (DT The) (NNP Fulton) (NNP County) (NNP Grand) (NNP Jury) )
          (VP (VBD said)
            (NP (NNP Friday) )
            (SBAR (-NONE- 0)
              (S
                (NP (DT an) (NN investigation)
                  (PP (IN of)
                        (NP
                          (NP (NNP Atlanta) )
                          (POS 's) (JJ recent) (JJ primary) (NN election) )))
                (VP (VBD produced)
                  (NP (OQUOTE OQUOTE) (DT no) (NN evidence) (CQUOTE CQUOTE)
                        (SBAR (IN that)
                          (S
                            (NP (DT any) (NNS irregularities) )
                            (VP (VBD took)
                              (NP (NN place) ))))))))))
```

Named Entity Recognition

- **Find “all” entities in a document**
  - Label them with entity type
    - Person, place, organization
  - http://demos.basistech.com/rex/
- **Methods**
  - Look up in dictionary
  - Use regression (or SVM or CRF)

    P(x ε person) = f(

    in-name-list(x),

    word-before(x)= “Mr.”,

    single-letter(word-after(x)),

    in-name-list(word-after(x)),

    capitalized(x),

    …)

Named Entity Resolution

- **Brand names** (companies) are relatively easy
  - Need to deal with abbreviations and spelling mistakes
  - But standard lists are incomplete
- **Product models are** more complex
  - Variations in writing styles
    - Honda Civic could be written as “Honda Civic”; “Civic”; “Honda Civic LS”; “Honda Civic LE”; “LE”; “H. Civic”; “Hondah Sivik”
    - Model numbers can be written as: 5, V, Five

      “Asics Speedstar (both I and II), I love the I and II's and can't wait for the III's”
    - Model can be referred to as numbers but numbers do not always refer to models (e.g., “1010 for New Balance 1010”, but $1010)
- **Cities and names have high ambiguity**
  - Cambridge, Rochester, San Jose

Co-reference Resolution

- **This is a huge field, but relatively simple methods such as Lee, Peirsman et al. 2011 work reasonably well.**
  - Locate all noun phrases
  - Identify their properties
    - male/female, singular/plural, …
  - Cluster them in starting with the highest-confidence rules and moving to lower-confidence ones
    - Check first for pronominal/generic-nominal references
    - The do closest first

Co-reference resolution example

- Microsoft announced it plans to acquire Visio. The company said it will finalize its plans within a week.

  Only pronomial and generic nominal references are here. The first "it“ straightforwardly matches "Microsoft". "The company" also matches "Microsoft" first, because it's a pronominal reference, and "Microsoft" is in the subject position. The second "it" matches "the company", and so is also resolved to "Microsoft".

  Mark said that he used Symlin and it caused him to get a rash. He said that it bothered him.

  First "he" resolves to "Mark". "It" does not match "Mark", and so is resolved to "Symlin". Other "Him" and "he" are also resolved to "Mark". The second "it" is resolved to "a rash", because it is closer and neither of the matching candidates "Symlin" and "a rash" - is in the subject position.

How well does IE work?

Level of Information	**Accuracy**
Entities	90-**95%**
Attributes	80-90%
Facts	70-80%

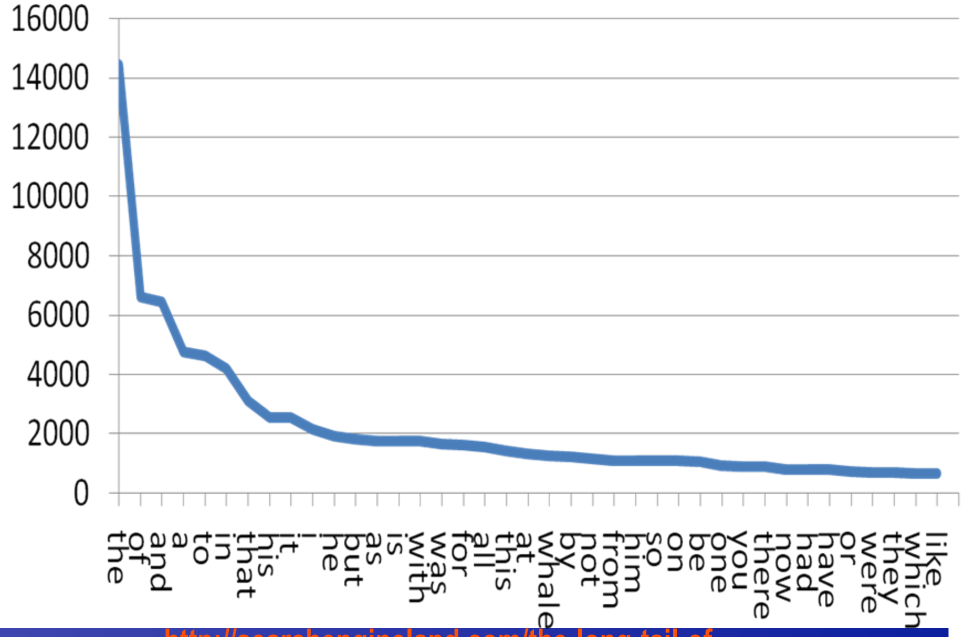
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Why IE is hard

- **Language is complex**
  - Synonyms and Orthonyms
    - *Bush, HEK*
  - Anaphora (and Sortal anaphoric noun phrases)
    - *It, they, the protein, both enzymes*
  - User-generated text is rarely grammatical
  - Complex structure
    - The first time I bought your product, I tried it on my dog, who became very unhappy and almost ate my cat, who my daughter dearly loves, and then when I tried it on her, she turned blue!

Why IE is hard

- **Hand-built systems give poor coverage**
  - Can’t manually list all patterns
  - Zipf's law ensures that most words are rare
- **Statistical methods need training data**
  - Expensive to manually label data



http://searchengineland.com/the-long-tail-of-search-12198

Why IE is easy

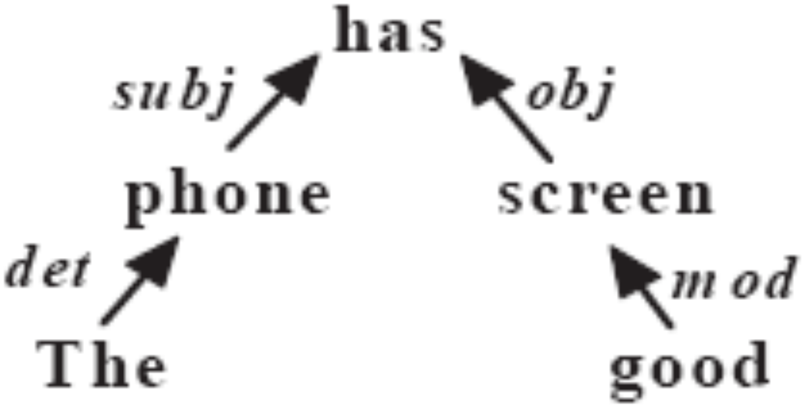
- **Lots of redundant data**
  - Don’t need to get every entity or sentiment right
  - One can generate good summaries even if one is wrong almost half the time.
- **Incomplete, inaccurate answers often useful**
  - EDA
    - Suggest trends or linkages

Extract aspects using Double Propagation (DP) (Qiu et al. 2009; 2011)

- Use *double propagation* (DP)
  - Like co-training
- an opinion should have a target, entity or aspect.
- DP extracts both aspects and opinion words.
  - Knowing one helps find the other.
  - E.g., “*The rooms are spacious*”

The DP method

- **DP is a bootstrapping method**
  - Input: a set of seed opinion words,
  - no aspect seeds needed
- **Based on dependency grammar (Tesniere 1959).**
  - “This phone has good screen”



Roadmap

- Sentiment Analysis Problem
- Simple sentiment classification
- Aspect-based sentiment analysis
- ➡ Mining comparative opinions
- Some complications

Comparative Opinions
(Jindal and Liu, 2006)

- ***Non-Equal Gradable*: Relations of the type *greater* or *less than***
  - *Ex: “optics of camera A is better than that of camera B”*
- ***Equative*: Relations of the type *equal to***
  - Ex: “*camera A and camera B both come in 7MP*”
- ***Superlative*: Relations of the type *greater* or *less than all others***
  - Ex: “*camera A is the cheapest in market*”

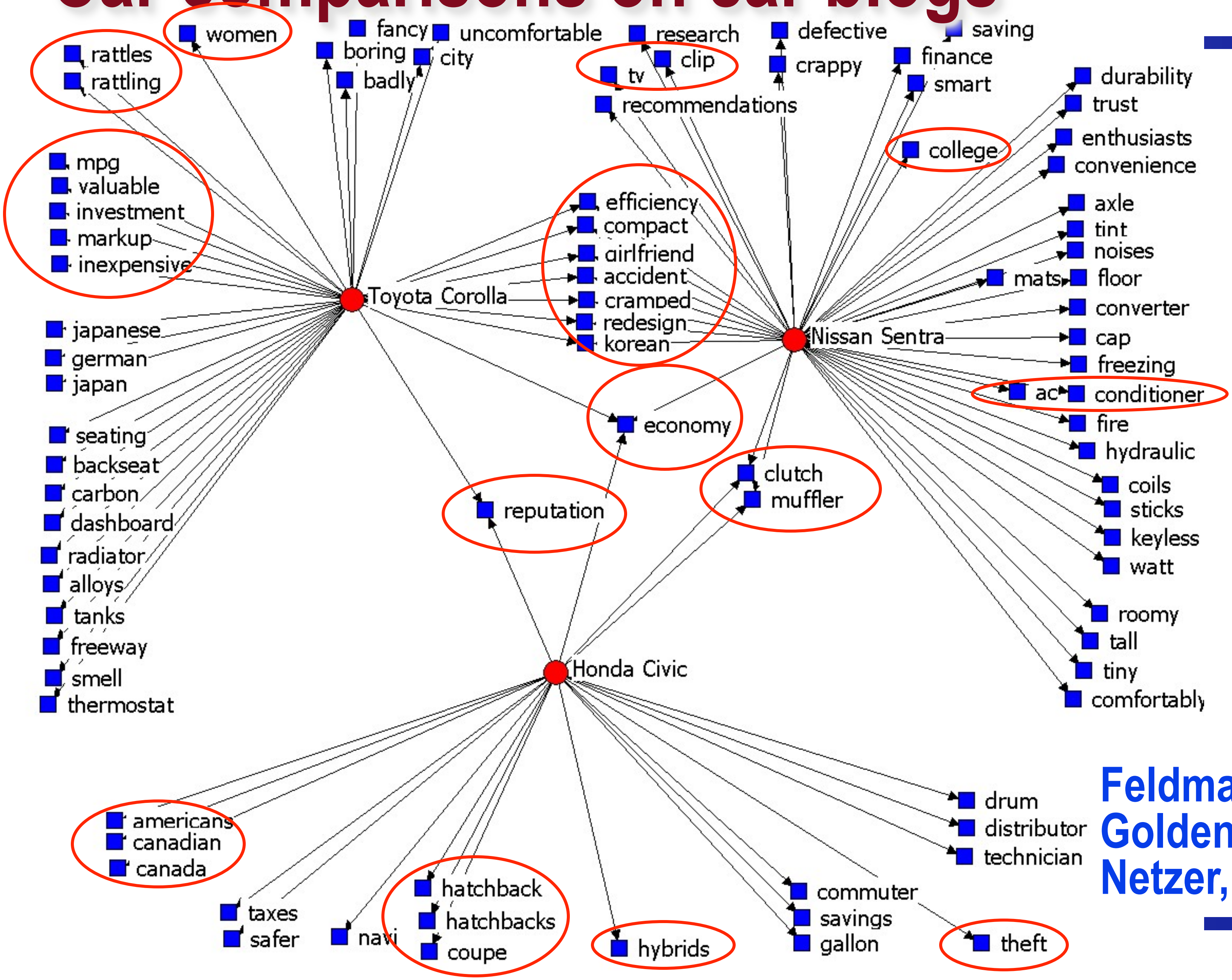
An example

- Consider the comparative sentence
  - “*Canon’s optics is better than those of Sony and Nikon*.”
  - Written by John in 2010.
- The extracted comparative opinion/relation:
  - ({Canon}, {Sony, Nikon}, {optics}, *preferred*:{Canon}, John, 2010)

Common comparatives

- **In English, comparatives are usually formed by adding -*er* and superlatives are formed by adding *-est* to their base adjectives and adverbs**
- **Adjectives and adverbs with two syllables or more and not ending in *y* do not form comparatives or superlatives by adding -*er* or -*est*.**
  - Instead, *more*, *most*, *less*, and *least* are used before such words, e.g., *more beautiful*.
- **Irregular comparatives and superlatives, i.e., *more most*, *less*, *least*, *better*, *best*, *worse*, *worst*, etc**

Car comparisons on car blogs



Feldman, Goldenberg, Netzer, Ungar

Roadmap

- Sentiment Analysis Problem
- Simple sentiment classification
- Aspect-based sentiment analysis
- Mining comparative opinions
- Some complications
  - Sentiment shifters
  - Implicit sentiment

Sentiment shifters (e.g., Polanyi and Zaenen 2004)

- **Sentiment/opinion shifters (or valence shifters)**
  - words and phrases that can shift or change opinion orientations.
- **Negation words like *not*, *never*, *cannot*, etc., are the most common type.**
- **Many other words and phrases can also alter opinion orientations.**
  - E.g., modal auxiliary verbs (e.g., *would*, *should*, *could, etc*)
  - "The brake could be improved."

Sentiment shifters (contd)

- **Some presuppositional items also change opinions, e.g., *barely* and *hardly***
  - “It hardly works.” (comparing to “it works”)
  - It presupposes that better was expected.
- **Words like *fail*, *omit*, *neglect* behave similarly,**
  - “This camera fails to impress me.”
- **Sarcasm changes orientation too**
  - “What a great car, it did not start the first day.”

Explicit and implicit aspects

(Hu and Liu 2004)

- Explicit aspects: Aspects explicitly mentioned as nouns or noun phrases in a sentence
  - The picture quality of this phone is great.
- Implicit aspects: Aspects not explicitly mentioned in a sentence but are implied
  - “This car is so expensive.”
  - “This phone will not easily fit in a pocket.
  - “Included 16MB is stingy”
- Not much work has been done on mining or mapping implicit aspects.

Implicit aspect mapping

- **There are many types of implicit aspect expressions. Adjectives and adverbs are perhaps the most common type.**
  - Most adjectives modify or describe specific attributes of entities.
  - “expensive” ⇒ aspect “price,” “beautiful” ⇒ aspect “appearance”, “heavy” ⇒ aspect “weight”
- **Although manual mapping is possible, in different contexts, the meaning can be different.**
  - E.g., “The computation is expensive”.

Reader’s standing point

- **See this sentence**
  - “I am so happy that Google price shot up today.”
- **Although the sentence gives an explicit sentiment, different readers may feel very differently.**
  - If a reader sold his Google shares yesterday, he will not be that happy.
  - If a reader bought a lot of Google shares yesterday, he will be very happy.
- **Current research mostly ignores the issue.**

Explicit and implicit sentiment

- **Explicit sentiment: “good”, “bad”, “terrible”**
  - The picture quality of this phone is great.
- **Implicit sentiment: “helpful”, “sunny”,**
- “The staff were friendly.”
- “The battery has an exceptionally long life”
- **Context-specific implicit sentiment: “blue”, “small”, “long”**
  - “This car is small.”
  - “I’m feeling blue.
  - “The product is in the black.”

Clause-based sentiment

- **“I changed to Audi because BMW is so expensive.”**
- **“Trying out Google chrome because Firefox keeps crashing.”**
  - The opinion about Firefox is clearly negative, but for Google chrome, there is no opinion.
  - We need to segment the sentence into clauses to decide that “crashing” only applies to Firefox.
  - “Trying out” also indicates no opinion.

Conditionals and Questions

- **Conditional sentences are hard to deal with (Narayanan et al. 2009)**
  - “If I can find a good camera, I will buy it.”
  - But conditional sentences can have opinions
    - “If you are looking for a good phone, buy Nokia”
- **Questions may or may not have opinions**
  - No sentiment
    - “Are there any great perks for employees?”
  - With sentiment
    - “Any idea how to repair this lousy Sony camera?”

Sarcasm

- **Sarcastic sentences**
  - “What a great car, it stopped working in the second day.”
- **Sarcastic sentences are very common in political blogs, comments and discussions.**
  - They make political blogs difficult to handle
  - Many political aspects can also be quite complex and hard to extract because they cannot be described using one or two words.
  - Some initial work by (Tsur, Davidov, Rappoport 2010)

SA may require ‘understanding’

- Two sentences in a medical domain:
  - “I went to see my doctor because of severe pain in my stomach”
  - “After taking the drug, I got severe pain in my stomach”
- If we are interested in opinions on a drug, the first sentence has no opinion, but the second implies negative opinion on the drug.
  - Some understanding seems to be needed?

More understanding?

- **The following two sentences are from reviews in the paint domain.**
  - “For paint_X, one coat can cover the wood color.”
  - “For paint_Y, we need three coats to cover the wood color.
- **We know that paint_X is good and Paint_Y is not, but how by a system.**
  - Do we need commonsense knowledge and understanding of the text?

More interesting sentences

- “My goal is to have a high quality tv with decent sound”
- “The top of the picture was much brighter than the bottom.”
- “Google steals ideas from Bing, Bing steals market share from Google.”
- “When I first got the airbed a couple of weeks ago it was wonderful as all new things are, however as the weeks progressed I liked it less and less.”

Yet more complications

- **Spam sentiment**
  - Common on user product reviews
- **All of NLP**
  - Coreference, entity resolution
  - Parsing
  - Scope of negation
  - …

Beyond positive/negative sentiment

- **Why are you buying/reading/tweeting this?**
  - What leads to word-of-mouth?
  - Why did you cite that paper
- **What emotion does it evoke in you?**
- **Sentiment-analysis style analysis can also be used for**
  - Happiness
  - Personality
  - Style
  - …

Sentiment Method Summary

- **Sentiment analysis has many sub-problems, none of which is fully solved**
  - Despite the challenges, applications are flourishing!
- **Building a reasonably accurate domain specific system is possible.**
  - Building an accurate generic system is hard.
  - But using IE tools and machine learning help.