Sentiment Analysis

Bing Liu
University Of Illinois at Chicago
liub@cs.uic.edu

Introduction

- Two main types of textual information.
  - Facts and Opinions
- Most current text information processing methods (e.g., web search, text mining) work with factual information.
- Sentiment analysis or opinion mining
  - computational study of opinions, sentiments and emotions expressed in text.
- Why opinion mining now? Mainly because of the Web; huge volumes of opinionated text.

Introduction – user-generated media

- Importance of opinions:
  - Opinions are so important that whenever we need to make a decision, we want to hear others' opinions.
  - In the past,
    - Individuals: opinions from friends and family
    - businesses: surveys, focus groups, consultants …
- Word-of-mouth on the Web
  - User-generated media: One can express opinions on anything in reviews, forums, discussion groups, blogs ...
  - Opinions of global scale: No longer limited to:
    - Individuals: one's circle of friends
    - Businesses: Small scale surveys, tiny focus groups, etc.

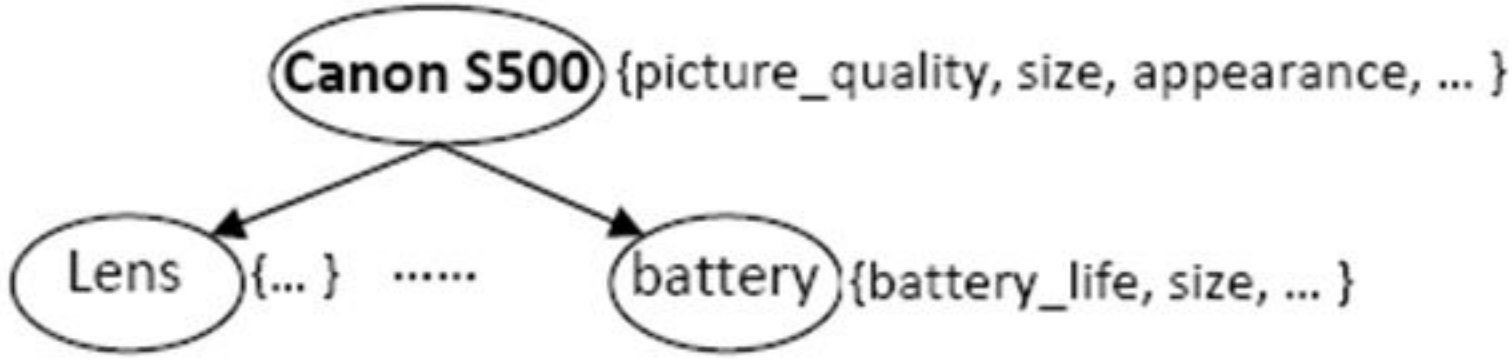
An Example Review

- *"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. 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, and wanted me to return it to the shop. …"*

- What do we see?
  - **Opinions, targets of opinions, and opinion holders**

Target Object (Liu, Web Data Mining book, 2006)

- **Definition** (**object**): An *object o* is a product, person, event, organization, or topic. *o* 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.



- An opinion can be expressed on any node or attribute of the node.
- To simplify our discussion, we use the term ***features*** to represent both components and attributes.

What is an Opinion? (Liu, a Ch. in NLP handbook)

- An *opinion* is a quintuple

  $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$,

  where

  - o_j is a target object.
  - f_{jk} is a feature of the object o_j.
  - so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature f_{jk} of object o_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.

Objective – structure the unstructured

- Objective: Given an opinionated document,
  - Discover all quintuples (o_j, f_{jk}, so_{ijkl}, h_i, t_l),
    - i.e., mine the five corresponding pieces of information in each quintuple, and
  - Or, solve some simpler problems

- With the quintuples,
  - Unstructured Text → Structured Data
    - Traditional data and visualization tools can be used to slice, dice and visualize the results in all kinds of ways
    - Enable qualitative and quantitative analysis.

Sentiment Classification: doc-level

(Pang and Lee, Survey, 2008)

- Classify a document (e.g., a review) based on the overall sentiment expressed by opinion holder
  - Classes: Positive, or negative
- Assumption: each document focuses on a single object and contains opinions from a single op. holder.
- *E.g., thumbs-up or thumbs-down?*
  - *"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. 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, and wanted me to return it to the shop. …"*

Subjectivity Analysis: sent.-level

(Wiebe et al 2004)

- Sentence-level sentiment analysis has two tasks:
  - Subjectivity classification: Subjective or objective.
    - Objective: e.g., *I bought an iPhone a few days ago.*
    - Subjective: e.g., *It is such a nice phone.*
  - Sentiment classification: For subjective sentences or clauses, classify positive or negative.
    - Positive: *It is such a nice phone.*
- But (Liu, a Ch in NLP handbook)
  - subjective sentences ≠ +ve or –ve opinions
    - E.g., *I think he came yesterday.*
  - Objective sentence ≠ no opinion
    - Imply –ve opinion: *The phone broke in two days*

Feature-Based Sentiment Analysis

- Sentiment classification at both document and sentence (or clause) levels are not enough,
  - they do not tell what people like and/or dislike
  - A positive opinion on an object does not mean that the opinion holder likes everything.
  - An negative opinion on an object does not mean …..
- Objective (recall): Discovering all quintuples

  $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$
- With all quintuples, all kinds of analyses become possible.

Feature-Based Opinion Summary

(Hu & Liu, KDD-2004)

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. 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, and wanted me to return it to the shop. …"

Feature Based Summary:

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: **battery life**

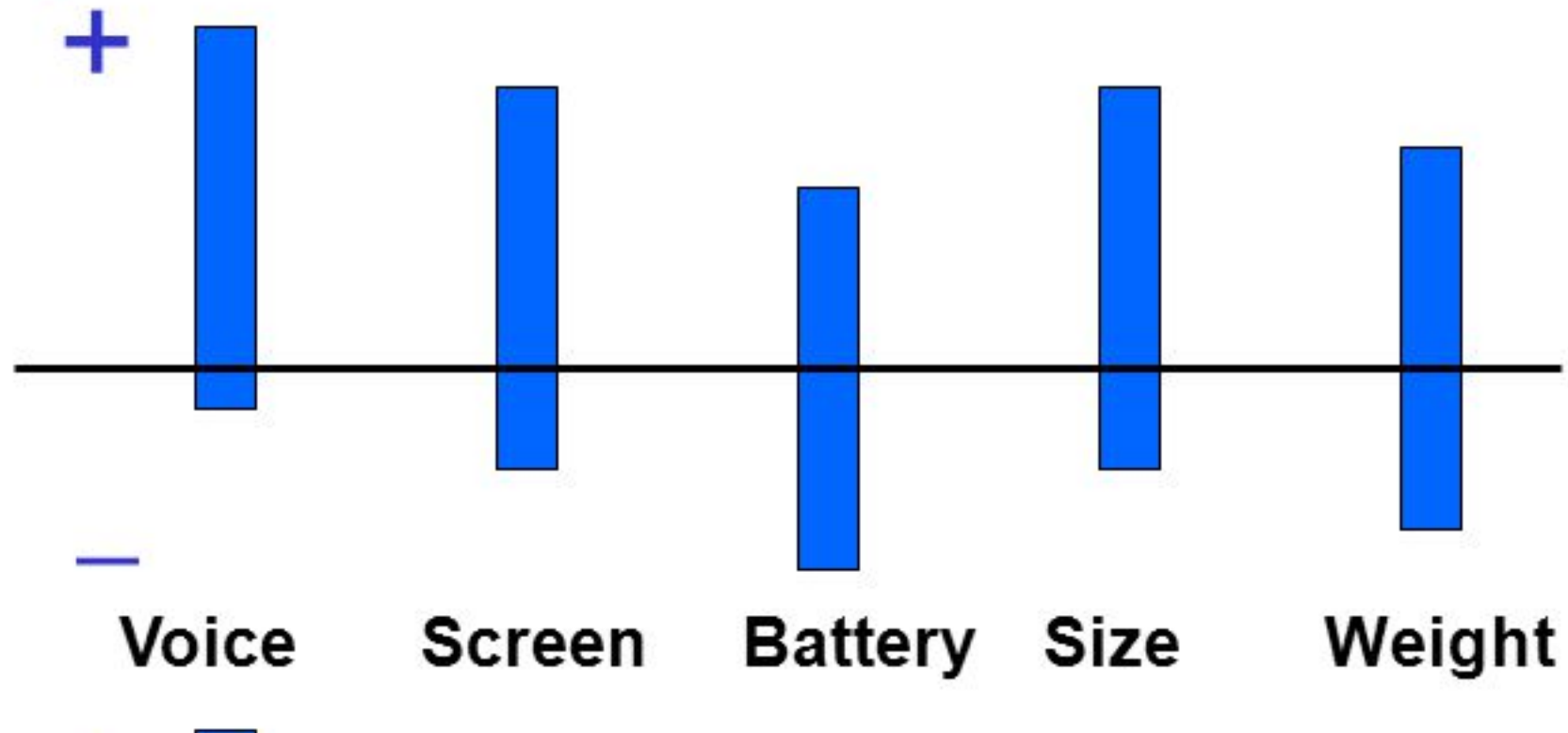
…

Note: We omit opinion holders

Visual Comparison (Liu et al. WWW-2005)

- Summary of reviews of

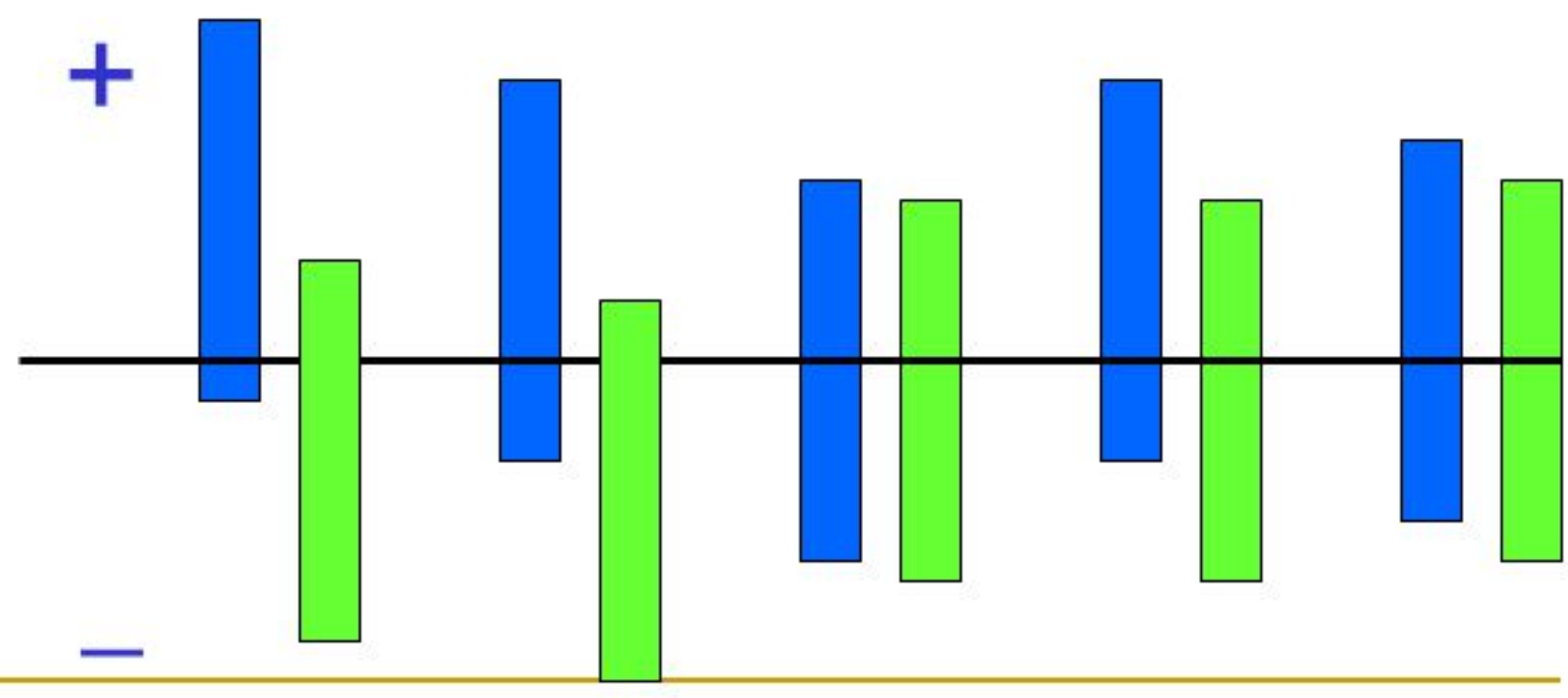
Cell Phone 1



- Comparison of reviews of

Cell Phone 1

Cell Phone 2



Feat.-based opinion summary in Bing

Sentiment Analysis is Hard!

- "*This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth. We called each other when we got home. The voice on my phone was not so clear, worse than my previous phone. The battery life was long. My girlfriend was quite happy with her phone. I wanted a phone with good sound quality. So my purchase was a real disappointment. I returned the phone yesterday.*"

Senti. Analy. is not Just ONE Problem

- $(o_j, f_{jk}, so_{ijkl}, h_i, t_l)$,
  - o_j - a target object: Named Entity Extraction (more)
  - f_{jk} - a feature of o_j: Information Extraction
  - so_{ijkl} is sentiment: Sentiment determination
  - h_i is an opinion holder: Information/Data Extraction
  - t_l is the time: Data Extraction
- Co-reference resolution
- Relation extraction
- Synonym match (voice = sound quality) …
- None of them is a solved problem!

Accuracy is Still an Issue!

- Some commercial solutions give clients several example opinions in their reports.
  - Why not all? Accuracy could be the problem.
- Accuracy: both
  - Precision: how accurate is the discovered opinions?
  - Recall: how much is left undiscovered?
  - Which sentence is better? (cordless phone review)
    - (1) The voice quality is great.
    - (2) I put the base in the kitchen, and I can hear clearly from the handset in the bed room, which is very far.

Easier and Harder Problems

- Reviews are easier.
  - Objects/entities are given (almost), and little noise
- Forum discussions and blogs are harder.
  - Objects are not given, and a large amount of noise

- Determining sentiments seems to be easier.
- Determining objects and their corresponding features is harder.
- Combining them is even harder.

Manual to Automation

- Ideally, we want an automated solution that can scale up.
  - Type an object name and then get +ve and –ve opinions in a summarized form.
  - Unfortunately, that will not happen any time soon.

Manual ----------|-------- Full Automation

  - Some creativity is needed to build a scalable and accurate solution.

I am Optimistic

- Significant researches are going on in several academic communities,
  - NLP, Web, data mining, information retrieval, …
  - New ideas and techniques are coming all the time.
- Industry is also trying different strategies, and solving some useful aspects of the problem.
- I believe a reasonably accurate solution will be out in the next few years.
  - Use a combination of algorithms.

Two Main Types of Opinions

- Direct Opinions: direct sentiment expressions on some target objects, e.g., products, events, topics, persons.
  - E.g., "the picture quality of this camera is great."
- Comparative Opinions: Comparisons expressing similarities or differences of more than one object. Usually stating an ordering or preference.
  - E.g., "car x is cheaper than car y."

Comparative Opinions (Jindal and Liu, 2006)

- *Gradable*
  - *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 camera available in market*"

Mining Comparative Opinions

- Objective: Given an opinionated document *d*,. Extract comparative opinions:

  (O_1, O_2, F, po, h, t),

  where O_1 and O_2 are the object sets being compared based on their shared features *F*, *po* is the preferred object set of the opinion holder *h*, and *t* is the time when the comparative opinion is expressed.

- Note: not positive or negative opinions.

Opinion Spam Detection (Jindal and Liu, 2007)

- Fake/untruthful reviews:
  - Write undeserving positive reviews for some target objects in order to promote them.
  - Write unfair or malicious negative reviews for some target objects to damage their reputations.
- Increasing number of customers wary of fake reviews (biased reviews, paid reviews)

An Example of Practice of Review Spam

Belkin International, Inc

- Top networking and peripherals manufacturer | Sales ~ $500 million in 2008
- Posted an ad for writing fake reviews on amazon.com (65 cents per review)

Timer: 00:00:00 of 60 minutes

Want to work on this HIT? Accept HIT

Want to see other HITs? Skip HIT

Write Product Reviews 25-50 Words

Requester: Mike Bayard

Qualifications Required: HIT approval rate (%) is not less than 95

Jan 2009

Write a Positive 5/5 Review for Product on Website

Positive review writing.

- Use your best possible grammar and write in US English only
- Always give a 100% rating (as high as possible)
- Keep your entry between 25 and 50 words
- Write as if you own the product and are using it
- Tell a story of why you bought it and how you are using it
- Thank the website for making you such a great deal
- Mark any other negative reviews as "not helpful" once you post yours

Instructions:

The link below leads to a product on a website. Read-through the product's features and write a positive review for it using the guidelines above to the best of your ability. I have also provided the part number for this product and you can click on the links below to see it on several alternative websites. In order to post some reviews you will need to create an account on the site. You can use your own email address or open a new free webmail account (gmail, yahoo...) and use it to post with.

Experiments with Amazon Reviews

- June 2006
  - 5.8mil reviews, 1.2mil products and 2.1mil reviewers.
- A review has 8 parts
  - *<Product ID> <Reviewer ID> <Rating> <Date> <Review Title> <Review Body> <Number of Helpful feedbacks> <Number of Feedbacks> <Number of Helpful Feedbacks>*
- Industry manufactured products "*mProducts*"

  e.g. electronics, computers, accessories, etc
  - 228K reviews, 36K products and 165K reviewers.

Some Tentative Results

- Negative outlier reviews tend to be heavily spammed.
- Those reviews that are the only reviews of some products are likely to be spammed
- Top-ranked reviewers are more likely to be spammers
- Spam reviews can get good helpful feedbacks and non-spam reviews can get bad feedbacks

Summary

- We briefly defined and introduced
  - Direct opinions: document, sentence and feature level
  - Comparative opinions: different types of comparisons
  - Opinion spam detection: fake reviews.
- A lot of applications.
- Technical challenges are still huge.
- But I am optimistic. Accurate solutions will be out in the next few years.
  - Maybe it is already out there that I do not know of.

References

- B. Liu, "Sentiment Analysis and Subjectivity." A Chapter in *Handbook of Natural Language Processing*, 2nd Edition, 2009 or 2010 (*email me if you want a softcopy*).
  - (An earlier version) B. Liu, "Opinion Mining", A Chapter in the book: Web Data Mining, Springer, 2006.
- B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2(1-2), 2008.