Sentiment Analysis and Opinion Mining

Introduction

- Opinion mining or sentiment analysis
  - Computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, etc., expressed in text.
    - Reviews, blogs, discussions, news, comments, feedback, or any other documents
- Terminology:
  - Sentiment analysis is more widely used in industry.
  - Both are widely used in academia
- But they can be used interchangeably.

Why are opinions important?

- "Opinions" are key influencers of our behaviors.
- Our beliefs and perceptions of reality are conditioned on how others see the world.
- Whenever we need to make a decision, we often seek out the opinions of others. In the past,
  - Individuals: seek opinions from friends and family
  - Organizations: use surveys, focus groups, opinion polls, consultants.

Introduction – social media + beyond

- 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.
- Global scale: No longer – one's circle of friends
- Organization internal data
  - Customer feedback from emails, call centers, etc.
- News and reports
  - Opinions in news articles and commentaries

Introduction – applications

- Businesses and organizations
  - Benchmark products and services; market intelligence.
    - Businesses spend a huge amount of money to find consumer opinions using consultants, surveys and focus groups, etc
- Individuals
  - Make decisions to buy products or to use services
  - Find public opinions about political candidates and issues
- Ads placements: Place ads in the social media content
  - Place an ad if one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- Opinion retrieval: provide general search for opinions.

A fascinating problem!

- Intellectually challenging & many applications.
  - A popular research topic in NLP, text mining, and Web mining in recent years (Shanahan, Qu, and Wiebe, 2006 (edited book); Surveys - Pang and Lee 2008; Liu, 2006 and 2011; 2010)
  - It has spread from computer science to management science (Hu, Pavlou, Zhang, 2006; Archak, Ghose, Ipeirotis, 2007; Liu Y, et al 2007; Park, Lee, Han, 2007; Dellarocas, Zhang, Awad, 2007; Chen & Xie 2007).
  - 40-60 companies in USA alone
- It touches every aspect of NLP and yet is confined.
  - Little research in NLP/Linguistics in the past.
- Potentially a major technology from NLP.
  - But it is hard.

A large research area

- Many names and tasks with somewhat different objectives and models
  - Sentiment analysis
  - Opinion mining
  - Sentiment mining
  - Subjectivity analysis
  - Affect analysis
  - Emotion detection
  - Opinion spam detection
  - *Etc.*

Roadmap

- **Opinion Mining Problem**
- Document sentiment classification
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Aspect-based opinion summarization
- Opinion lexicon generation
- Mining comparative opinions
- Some other problems
- Opinion spam detection
- Utility or helpfulness of reviews
- Summary

Structure the unstructured (Hu and Liu 2004)

- Structure the unstructured: Natural language text is often regarded as unstructured data.
- The problem definition should provide a structure to the unstructured problem.
  - Key tasks: Identify key tasks and their inter-relationships.
  - Common framework: Provide a common framework to unify different research directions.
  - Understanding: help us understand the problem better.

Problem statement

- It consists of two aspects of abstraction

(1) Opinion definition. What is an opinion?

- Can we provide a structured definition?
  - If we cannot structure a problem, we probably do not understand the problem.

(2) Opinion summarization. why?

- Opinions are subjective. An opinion from a single person (unless a VIP) is often not sufficient for action.
- We need opinions from many people, and thus opinion summarization.

Abstraction (1): 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

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)

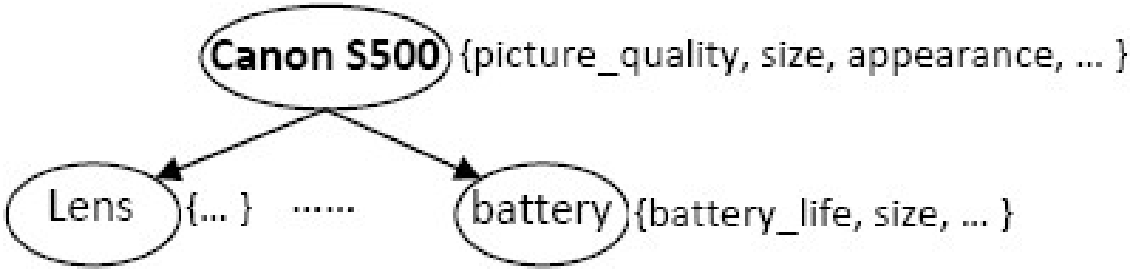
- Regular opinions: Sentiment/opinion expressions on some target entities
  - Direct opinions:
    - "The touch screen is really cool."
  - Indirect opinions:
    - "After taking the drug, my pain has gone."
- Comparative opinions: Comparisons of more than one entity.
  - E.g., "iPhone is better than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

A (regular) opinion

- Opinion (a restricted definition)
  - An opinion (or regular opinion) is simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity (Hu and Liu 2004; Liu 2006) from an opinion holder (Bethard et al 2004; Kim and Hovy 2004; Wiebe et al 2005).
- Sentiment orientation of an opinion
  - Positive, negative, or neutral (no opinion)
    - Also called *opinion orientation*, *semantic orientation*, *sentiment polarity*.

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.



- 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 feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or more granular ratings.
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.

Some remarks about the definition

- Although introduced using a product review, the definition is generic
  - Applicable to other domains,
  - E.g., politics, social events, services, topics, etc.
- (e_j, a_{jk}) is also called the opinion target
  - Opinion without knowing the target **is of limited use.**
- The five components in $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$ must correspond to one another. Very hard to achieve
- The five components are essential. Without any of them, it can be problematic in general.

Some remarks (contd)

- Of course, one can add any number of other components to the tuple for more analysis. E.g.,
  - Gender, age, Web site, post-id, etc.
- The original definition of an entity is a hierarchy of parts, sub-parts, and so on.
  - The simplification can result in information loss.
    - E.g., "The seat of this car is rally ugly."
    - "seat" is a part of the car and "appearance" (implied by ugly) is an aspect of "seat" (not the car).
  - But it is usually sufficient for practical applications.
    - It is too hard without the simplification.

"Confusing" terminologies

- Entity is also called object.
- Aspect is also called feature, attribute, facet, etc
- Opinion holder is also called opinion source
- Some researchers also use topic to mean entity and/or aspect.
  - Separating entity and aspect is preferable
- In specific applications, some specialized terms are also commonly used, e.g.,
  - Product features, political issues

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 either implicitly assumes a standing point, or ignores the issue.

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.

Structure the unstructured

- Goal: Given an opinionated document,
  - Discover all quintuples (e_j, f_{jk}, so_{ijkl}, h_i, t_l),
  - Or, solve some simpler forms of the problem
    - E.g., sentiment classification at the document or sentence level.
- With the quintuples,
  - Unstructured Text → Structured Data
    - Traditional data and visualization tools can be used to slice, dice and visualize the results.
    - Enable qualitative and quantitative analysis.

Two closely related concepts

- Subjectivity and emotion.
- **Sentence subjectivity**: An *objective sentence* presents some factual information, while a *subjective sentence* expresses some personal feelings, views, emotions, or beliefs.
- **Emotion**: Emotions are people's subjective feelings and thoughts.

Subjectivity

- Subjective expressions come in many forms, e.g., opinions, allegations, desires, beliefs, suspicions, speculations (Wiebe 2000; Wiebe et al 2004; Riloff et al 2006).
  - 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"

Emotion

- No agreed set of basic emotions of people among researchers.
- Based on (Parrott, 2001), people have six main emotions,
  - love, joy, surprise, anger, sadness, and fear.
- Strengths of opinions/sentiments are related to certain emotions, e.g., joy, anger.
  - However, the concepts of emotions and opinions are not equivalent.

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”.

Sentiment, subjectivity, and emotion

- Although they are clearly related, these concepts are not the same
  - Sentiment $\neq$ subjective $\neq$ emotion
- Sentiment is not a subset of subjectivity (without implied sentiments by facts, it should be)
  - sentiment $\not\subset$ subjectivity
- The following should hold
  - emotion $\subset$ subjectivity
  - sentiment $\not\subset$ emotion, …

Abstraction (2): opinion summary

- With a lot of opinions, a summary is necessary.
  - A multi-document summarization task
- For factual texts, summarization is to select the most important facts and present them in a sensible order while avoiding repetition
  - 1 fact = any number of the same fact
- But for opinion documents, it is different because opinions have a quantitative side & have targets
  - 1 opinion $\neq$ a number of opinions
  - Aspect-based summary is more suitable
    - Quintuples form the basis for opinion summarization

Aspect-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, …"

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**

…

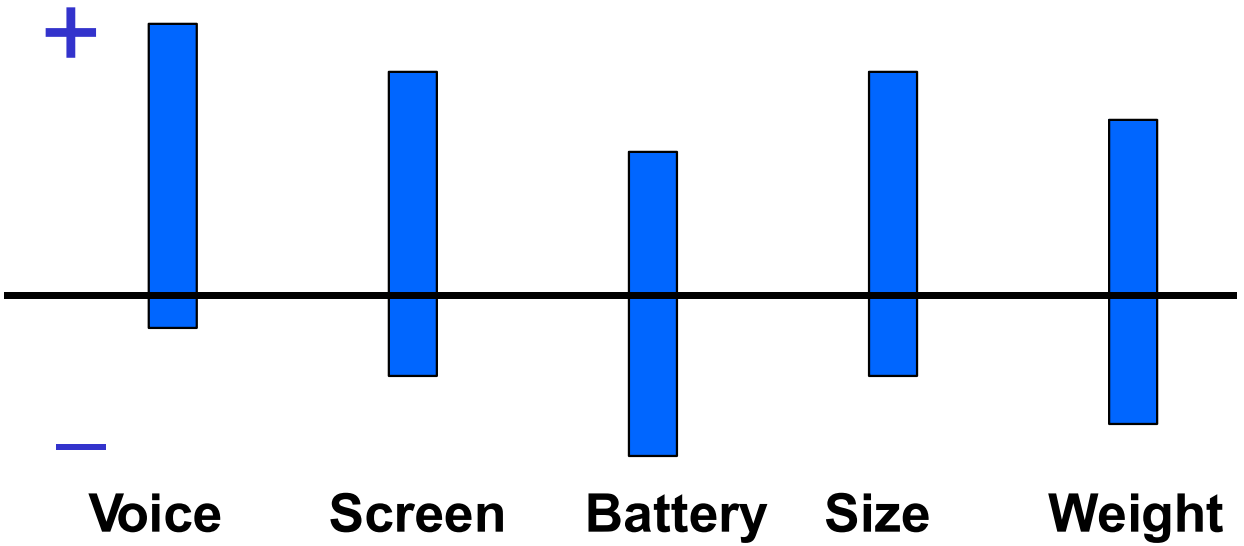
Note: We omit opinion holders

1. Originally called ***feature-based opinion mining and summarization***

Opinion Observer (Liu et al. 2005)

- Summary of reviews of

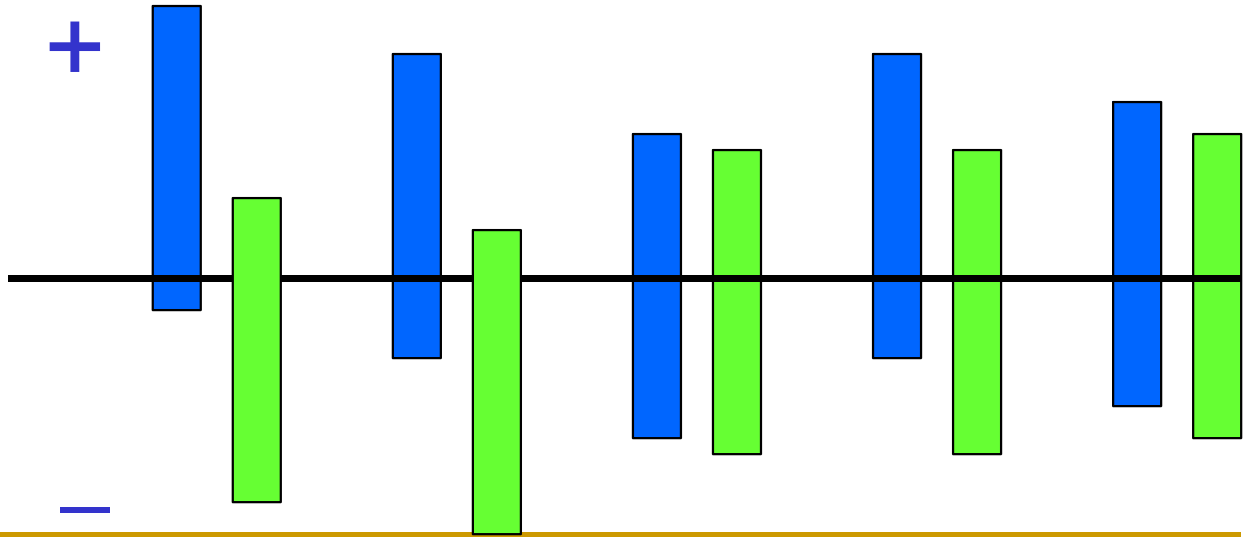
Cell Phone 1



- Comparison of reviews of

Cell Phone 1

Cell Phone 2



Aspect-based opinion summary

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

Google products | sony camera | Search Products

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

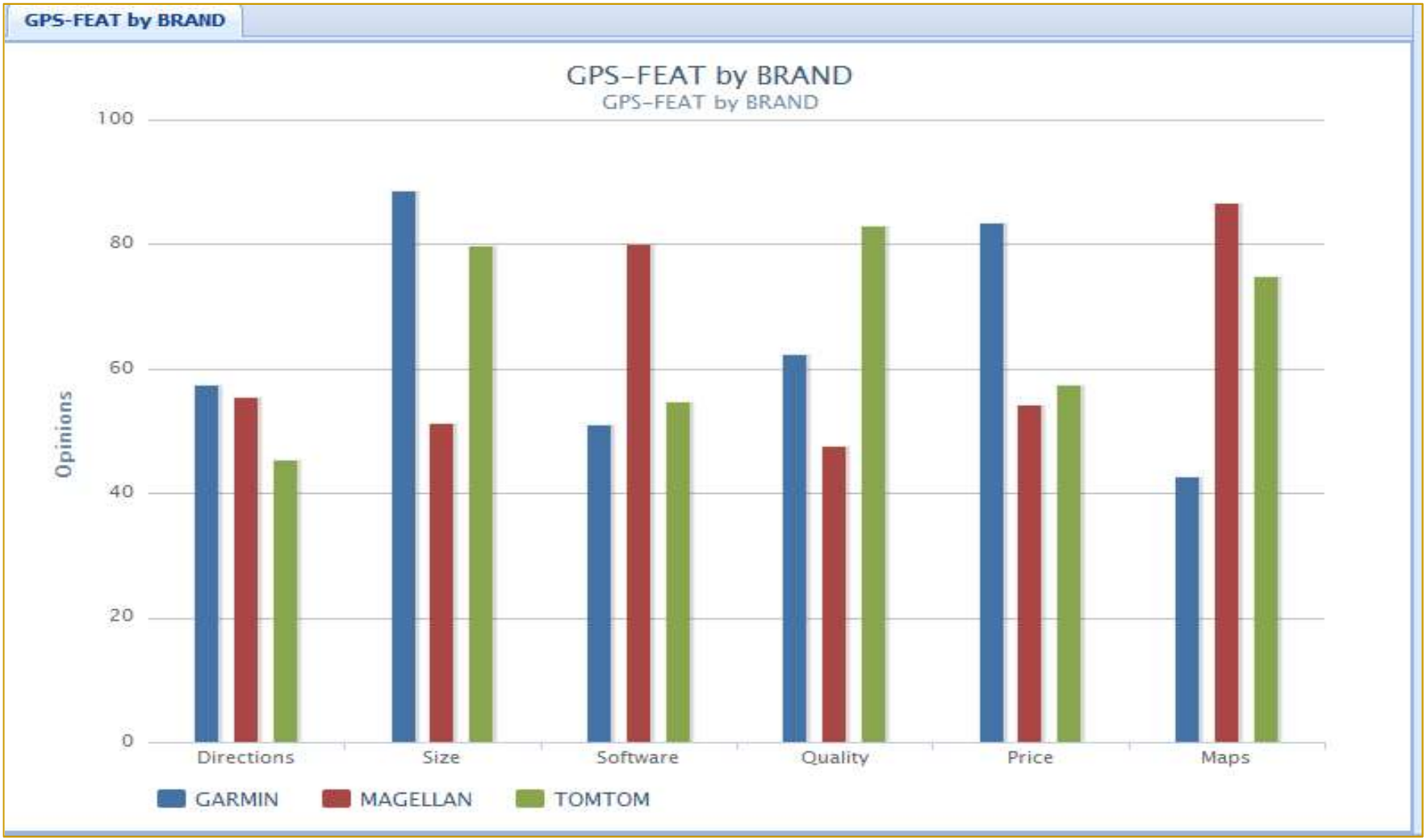
Summary - Based on 159 reviews

1 | 2 | 3 stars | 4 stars | 5 stars

What people are saying

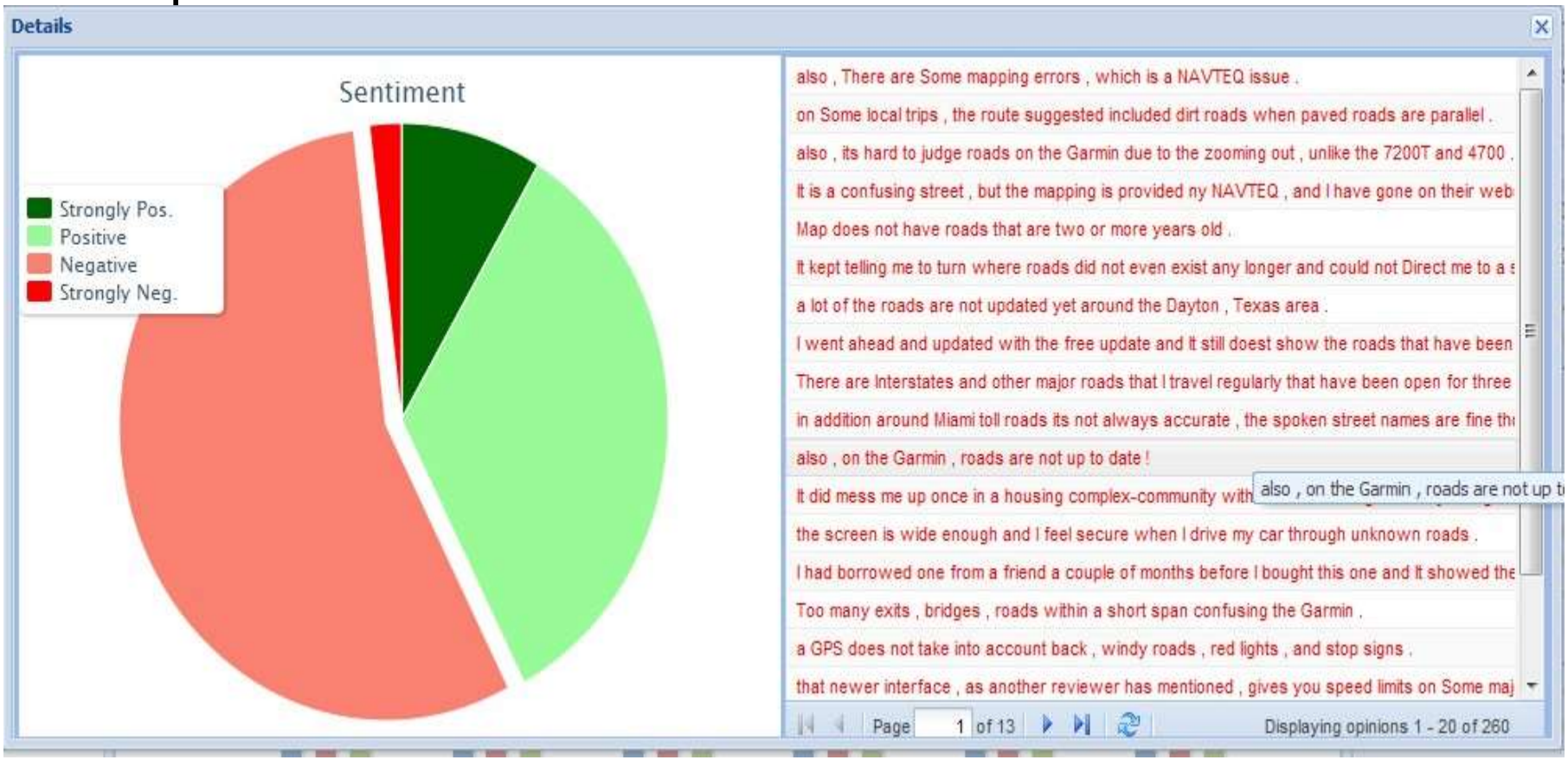
pictures	"We use the product to take quickly photos."
features	"Impressive panoramic feature."
zoom/lens	"It also record better and focus better on sunny days."
design	"It has the slightest grip but it's sufficient."
video	"Video zoom is choppy."
battery life	"Even better, the battery lasts long."
screen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

Some examples from OpinionEQ

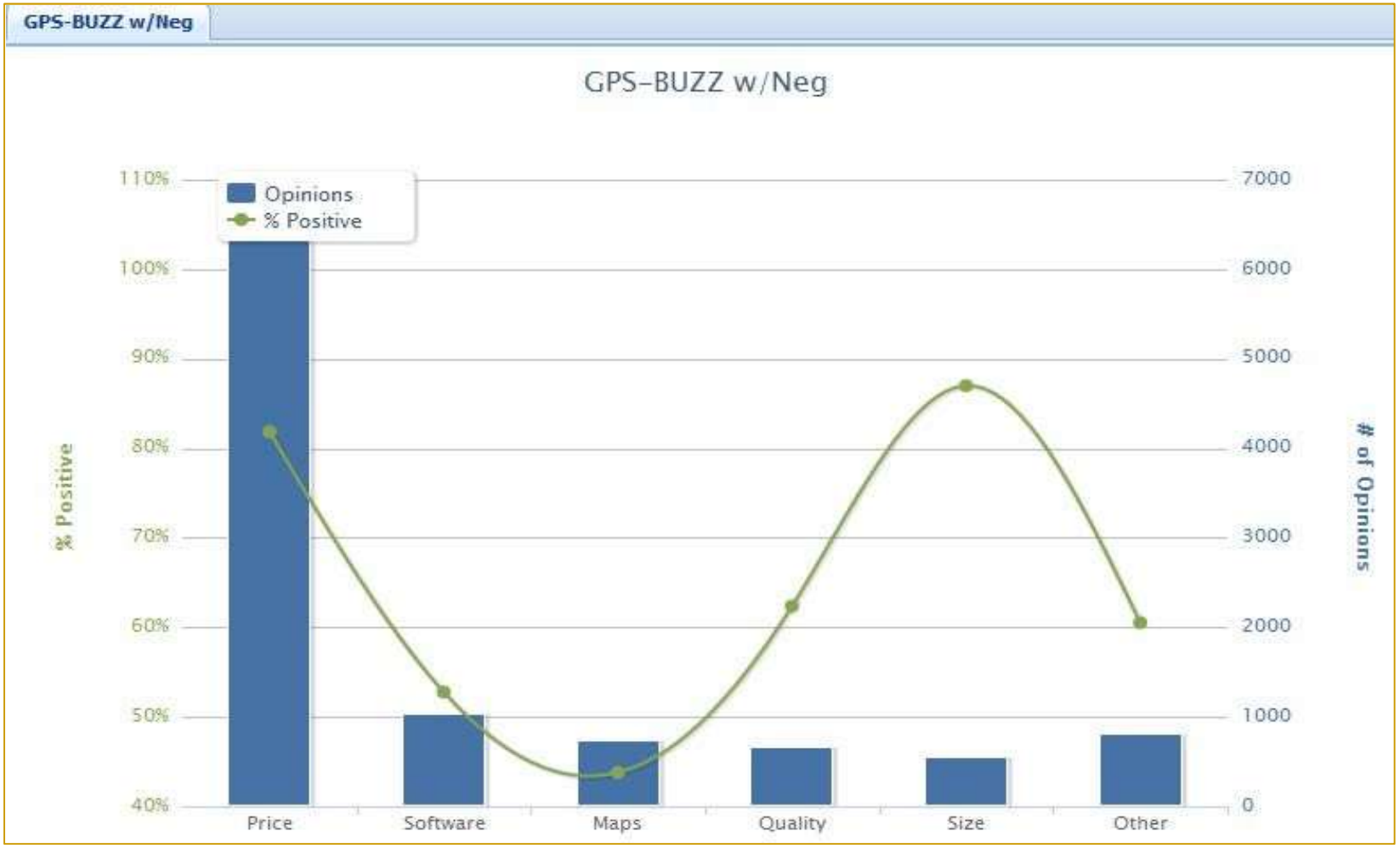


Detail opinion sentences

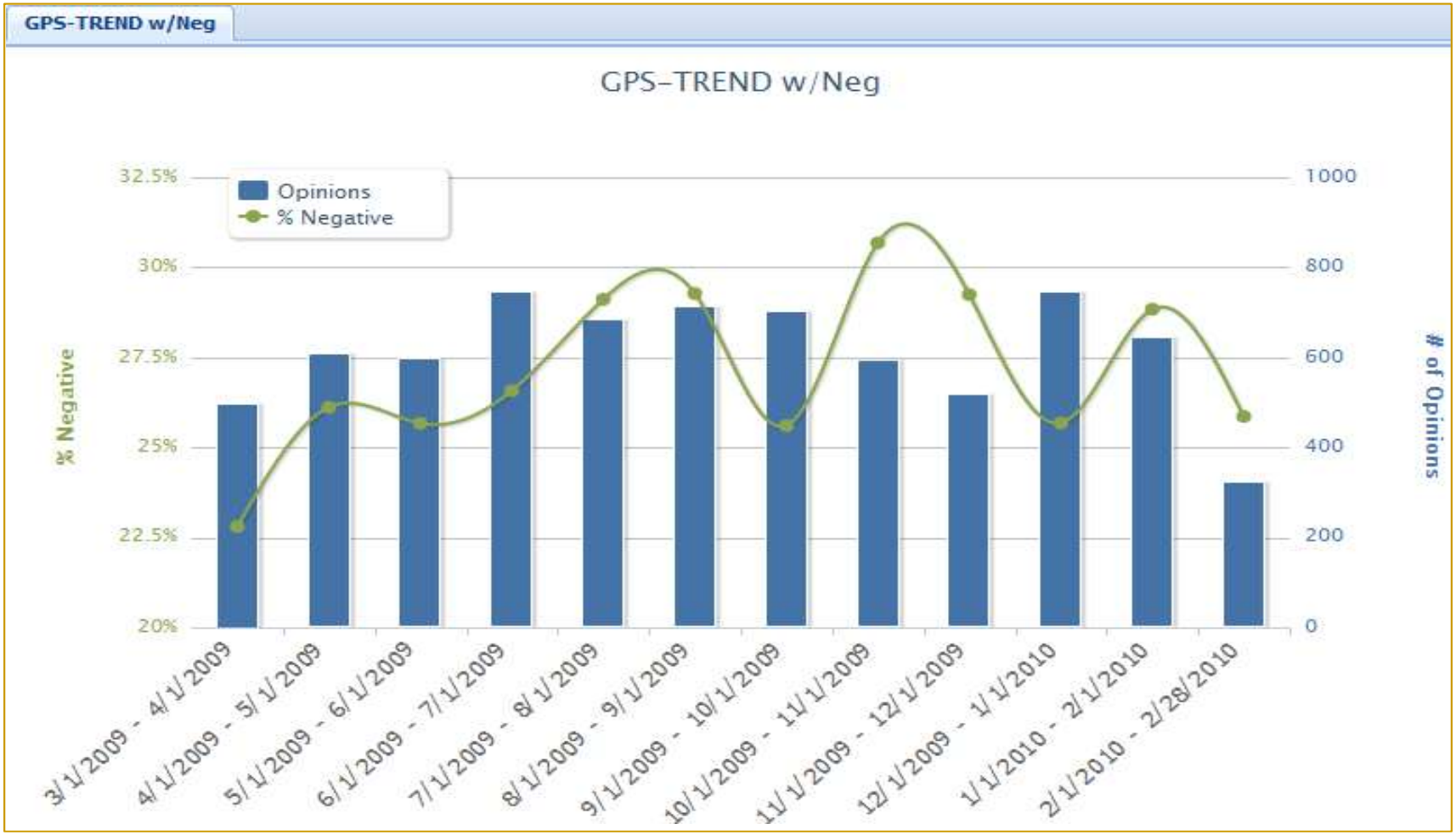
- Click on any bar (previous slide) to see the opinion sentences. Here are negative opinion sentences on the maps feature of Garmin.



% of +ve opinion and # of opinions

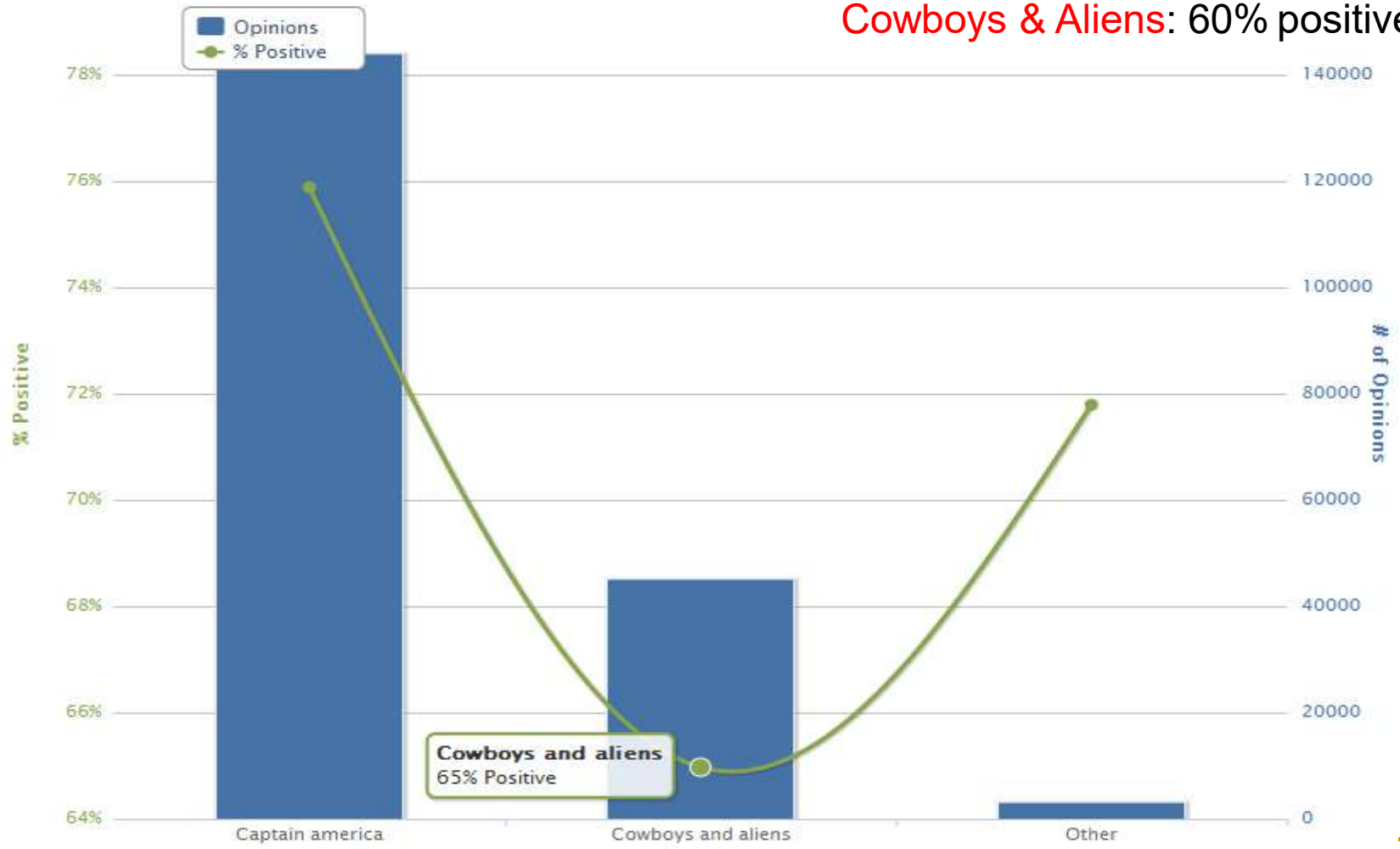


Aggregate opinion trend



Live tracking of two movies (Twitter)

User ratings from Rotten Tomatoes: Captain America: 81% positive
Cowboys & Aliens: 60% positive



July 8, 2011 to Present

Not just ONE problem

- $(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,
  - e_j - a target entity: Named Entity Extraction (more)
  - a_{jk} – an aspect 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
  - 5 pieces of information must match
- Coreference resolution
- Synonym match (voice = sound quality)
- …

Opinion mining 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 Samsung phone. The battery life was short too. 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.*"

Easier and harder problems

- Tweets from Twitter are 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
- Determining sentiments seems to be easier.
- Extracting entities and aspects is harder.
- Combining them is even harder.

Opinion mining in the real world

- Source the data, e.g., reviews, blogs, etc

  (1) Crawl all data, store and search them, or

  (2) Crawl only the target data
- Extract the right entities & aspects
  - Group entity and aspect expressions,
    - Moto = Motorola, photo = picture, etc …
- Aspect-based opinion mining (sentiment analysis)
  - Discover all quintuples

    (Store the quintuples in a database)
- Aspect based opinion summary

Roadmap

- Opinion Mining Problem
- **Document sentiment classification**
- Sentence subjectivity & sentiment classification
- Aspect-based sentiment analysis
- Aspect-based opinion summarization
- Opinion lexicon generation
- Mining comparative opinions
- Some other problems
- Opinion spam detection
- Utility or helpfulness of reviews
- Summary