A Holistic Lexicon-Based Approach to Opinion Mining

Xiaowen Ding, Bing Liu and Philip Yu
Department of Computer Science
University of Illinois at Chicago

Introduction – facts and opinions

- Two types of textual information in the world
  - Facts and Opinions
- Current information processing and search focus on facts:
  - I.e., search and read the top-ranked document(s)
    - One fact = multiple facts
- Finding and processing opinions is harder
  - Opinions are hard to express with a few keywords
  - Summarization is needed because
    - One opinion ≠ multiple opinions
    - People do not want to read everything

Introduction – user generated content

- Word-of-mouth on the Web
  - One can express personal experiences and opinions on almost anything, at review sites, forums, discussion groups, blogs ... (called the user generated content.)
  - They contain valuable information
  - Web/global scale: No longer – one's circle of friends
- Mine opinions expressed in the user-generated content is
  - an intellectually challenging problem (it is NLP!)
  - Practically useful
    - Individual consumers and companies.

Opinion mining – the **abstraction**

- We use consumer reviews of products to develop the ideas. Other opinionated contexts are similar.

- Basic components of an opinion
  - Opinion holder: The person or organization that holds a specific opinion on a particular object.
  - Object: on which an opinion is expressed
  - Opinion: a view, attitude, or appraisal on an object from an opinion holder, and more …

Object/entity

- **Definition** (**object**): An object *O* is an entity which can be 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**.
  - *O* is the root node (which also has a set of attributes)
- An opinion can be expressed on any node or attribute of the node.

- To simplify our discussion, we use "**features**" to represent both components and attributes.
- Note: the object *O* itself is also a feature.

Model of a review

- An object O is represented with a finite set of features, $F = \{f_1, f_2, \ldots, f_n\}$.
  - Each feature f_i in F can be expressed with a finite set of words or phrases W_i, which are **synonyms**.
- **Model of a review**: An opinion holder j comments on a subset of the features $S_j \subseteq F$ of object O.
  - For each feature $f_k \in S_j$ that j comments on, he/she
    - chooses a word or phrase from W_k to describe the feature, and
    - expresses a positive, negative or neutral opinion on f_k.

Opinion mining tasks

- At the document (or review) level: opinion on object

  Task: sentiment classification of reviews (Turney 02, Pang et al 02)
  - Classes: positive, negative, and neutral
  - Assumption: each document (or review) focuses on a single object and contains opinion from a single opinion holder.
- At the sentence level (e.g., Rilloff and Wiebe 03)

  Task 1: identifying subjective/opinionated sentences
  - Classes: objective and subjective (opinionated)

  Task 2: sentiment classification of sentences
  - Classes: positive, negative and neutral.
  - Assumption: a sentence contains only one opinion (not true)
  - Then we can also consider clauses or phrases.
- But, still don't know what people liked or disliked

Opinion mining tasks (contd)

- At the feature level (Hu and Liu 2004):

  Task 1: Identify and extract object features *F* that have been commented on by an opinion holder (e.g., a reviewer).

  Task 2: Determine whether the opinions on the features *F* are positive, negative or neutral.

  Task 3: Group feature synonyms.

  - Produce a feature-based opinion summary of multiple reviews.
    - Note: Object itself is also a feature (root of the tree)
- Our focus in this work: Task 2
  - We assume that features have been discovered
  - About Task 1 (see Hu and Liu 2004; Popescu and Etzioni 2005)

Feature-based opinion summary

(Hu and Liu 2004)

GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

….

Feature Based Summary:

Feature1: **picture**

Positive: 12

- The pictures coming out of this camera are amazing.
- Overall this is a good camera with a really good picture clarity.

…

Negative: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, pictures produced by this camera were blurry and in a shade of orange.

Feature2: **battery life**

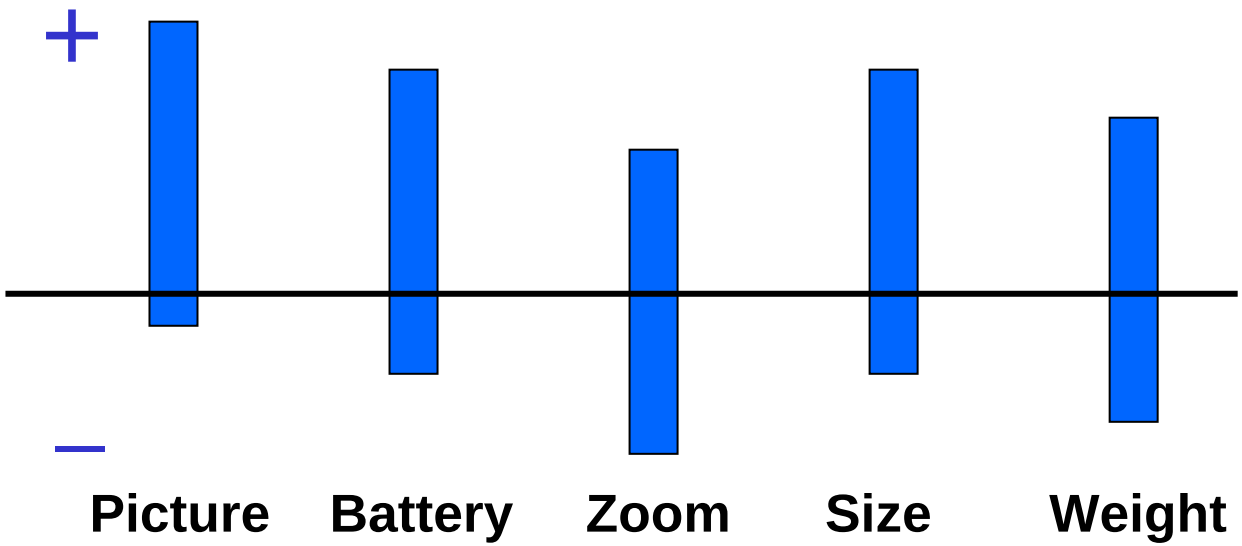
…

Visual summarization & comparison

(Liu et al 2005)

- Summary of reviews of

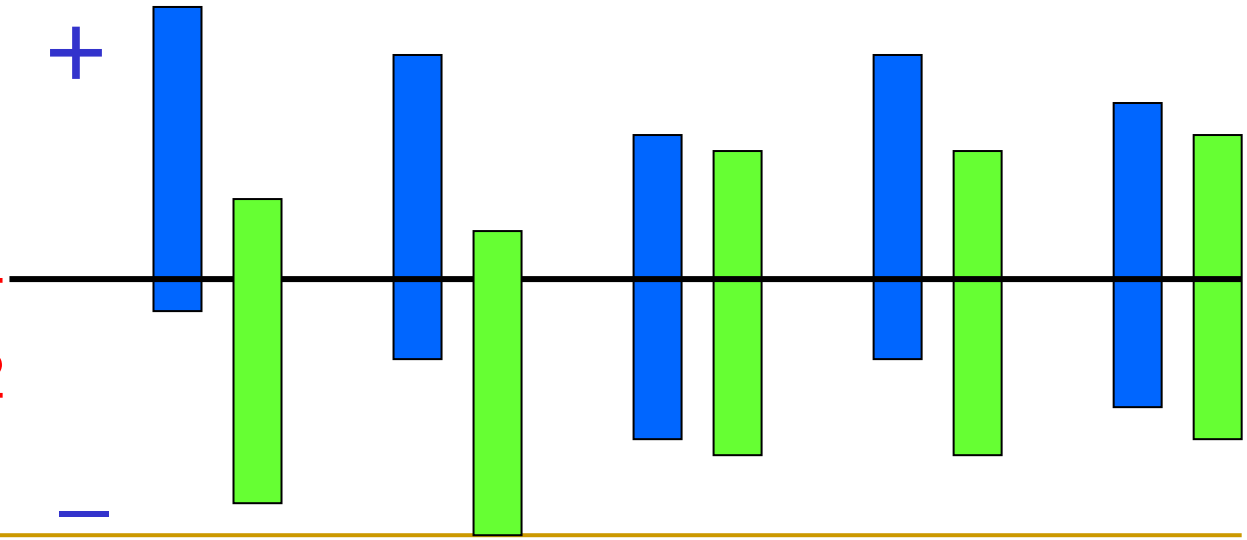
Digital camera 1



- Comparison of reviews of

Digital camera 1

Digital camera 2



Feature-based opinion summary in action (**Microsoft Live Search**)

Canon PowerShot A40 - digital camera

User reviews (42) | **Product details** | Compare prices

Is this useful?

User reviews (42)

The PowerShot A40 is packed full of features and offers excellent value for money. This camera features a high quality 35-105mm (35mm equivalent) 3x optical zoom lens, with a bright maximum aperture of f2.8. Its 2.0M pixel CCD... More...

User reviews at a glance What's this?

All user reviews

General Comments (32 comments) 88% positive

Ease Of Use (31 comments) 97% positive

Features (26 comments) 88% positive

Photo Quality (26 comments) 85% positive

Price (22 comments) 100% positive

All user reviews

Most recent | **Highest rating** | **Lowest rating**

Canon PowerShot A40

Reliable digital camera. Even when accidently soaked with water ,within 24 hours the camera functioned fully. Ive had this unit for 4 years without any problems. More...

www.ciao.co.uk 12/4/2007

Great little Camera

I recently dropped my canon PowerShot A40 2.0. Normally I would have made a trip to Best Buy or Circuit City to buy a newer model with more Pixels, etc, but I had all the lenses for the A40 and the underwaterhousing. The cameral... More...

search.reviews.ebay.com 4/3/2007

Awesome!!! This is one of the best I have ever

Lexicon-based approach (Hu and Liu 2004)

- Our work is based on features in sentences,
  - A sentence may contain multiple features.
  - Different features may have different opinions.
  - E.g., The battery life and picture quality are *great* (+), but the view founder is *small* (-).

- One effective approach is to use *opinion lexicon*, opinion words.
  - Identify all opinion words in a sentence
  - Aggregate these words to give the final opinion to each feature.

Opinion words

- Positive: beautiful, wonderful, good, amazing,
- Negative: bad, poor, terrible, cost someone an arm and a leg (idiom).
- They are instrumental for opinion mining (obviously)
- Two main ways to compile such a list:
  - Dictionary-based approaches
  - Corpus-based approaches
- Important :
  - Some opinion words are context independent (e.g., good).
  - Some are context dependent (e.g., long).

Dictionary-based approaches

- Start from a set of seed opinion words
- Use WordNet's synsets and hierarchies to acquire opinion words
  - Use the seeds to search for synonyms and antonyms in WordNet (Hu and Liu, 2004).
  - Use additional information (e.g., glosses) and learning from WordNet (Andreevskaia and Bergler, 2006) (Esuti and Sebastiani, 2005).
- Advantage: Good to find a lot of such words
- Weakness: Do not find context dependent opinion words, e.g., small, long, fast.

Corpus-based approaches

- Rely on syntactic rules and co-occurrence patterns to extract from large corpora
  - Use a list of seed words
  - A large domain corpus
  - Machine learning
- This approach can find domain (corpus) dependent opinions.

Corpus-based approaches (contd)

- Conjunctions: conjoined adjectives usually have the same orientation (Hazivassiloglou and McKeown 1997).
  - E.g., "This car is *beautiful* **and** *spacious*."(conjunction)

  Since we know "*beautiful*" (seed) is positive, we know that *spacious* is also positive

  - AND, OR, BUT, EITHER-OR, and NEITHER-NOR.
  - Machine learning

- Similar ideas are used or studied in (Popescu and Etzioni 2005; Kanayama and Nasukawa, 2006).

Our approach

- This work also exploits connectives, but with a few differences
  - Context is important
    - One word may indicate different opinions in the same domain.
      "The battery life is *long*" (+)
      "It takes a *long* time to focus" (-).
    - Find domain opinion words is insufficient.
  - Extend it to pseudo and inter-sentence rules.
  - Rules can be applied as the system goes along, no need for a large corpus. Opinions of context dependent words are cumulated with time.

Context dependent opinions

- Intra-sentence conjunction rule
  - Opinion on both sides of "and" should be the same
  - E.g., "This camera takes *great* pictures and has a *long* battery life".
- Not likely to say:
  - "This camera takes *great* pictures and has a *short* battery life."

Pseudo intra-sentence conj. rule

- Sometimes, one may not use an explicit conjunction “and”.
  - Same opinion in same sentence, unless there is a “but”-like clause

- **E.g.,** “The camera has a long battery life, which is *great*”

Inter-sentence conjunction rule

- People usually express the same opinion across sentences
  - unless there is an indication of opinion change using words such as “but” and “however”
- E.g., “The picture quality is amazing. The battery life is long”
- Not so natural to say:
  - “The picture quality is amazing. The battery life is short”

Growing contextual opinion words

- Growing
  - by applying various conjunctive rules
- Verifying the results as the system goes along (see more reviews)
  - Again by those conjunctive rules in additional reviews and sentences
- Only keep those opinions which the system is confident about, controlled by a confidence limit.

Handling of many constructs

- Opinion lexicon is far from sufficient.
- Special handling: Negation, but, etc.
- Not an opinion phrases, but contains an opinion word
  - "a good deal of"
- Not a negation, but contains a negation word, e.g., "not"
  - "not only … but also"
- Not contrary, but has a "but"
  - "not only …but also"

Aggregation of opinion words/phrases

- Input: a pair (*f*, *s*), where *f* is a product feature and *s* is a sentence that contains *f*.
- Output: whether the opinion on *f* in *s* is pos, neg, or neut.
- Two steps:
  - Step 1: split the sentence if needed based on BUT words (but, except that, etc).
  - Step 2: work on the segment s_f containing *f*. Let the set of opinion words in s_f be w_1, .., w_n. Sum up their orientations (1, -1, 0), and assign the orientation to (*f*, *s*) accordingly.

$$\sum_{i=1}^{n} \frac{w_i.o}{d(w_i, f)}$$

Experimental Results

Product name	Opinion Observer			Opinion Observer – No context dependency handling			Opinion Observer – Without using Equation (1)			FBS		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Digital camera 1	0.93	0.92	**0.93**	0.95	0.88	**0.91**	0.91	0.88	**0.89**	0.93	0.80	**0.86**
Digital camera 2	0.96	0.96	**0.96**	0.97	0.92	**0.95**	0.97	0.92	**0.95**	0.98	0.87	**0.92**
Cellular phone 1	0.93	0.90	**0.91**	0.93	0.79	**0.86**	0.91	0.86	**0.88**	0.94	0.70	**0.80**
MP3 player	0.87	0.86	**0.87**	0.89	0.79	**0.83**	0.86	0.82	**0.84**	0.91	0.69	**0.78**
DVD player	0.89	0.88	**0.89**	0.92	0.83	**0.87**	0.9	0.85	**0.87**	0.91	0.72	**0.80**
Cellular phone 2	0.95	0.95	**0.95**	0.95	0.89	**0.92**	0.95	0.92	**0.93**	0.95	0.82	**0.88**
Router	0.84	0.82	**0.83**	0.84	0.78	**0.81**	0.82	0.78	**0.80**	0.83	0.67	**0.74**
Antivirus software	0.90	0.87	**0.88**	0.93	0.72	**0.81**	0.89	0.81	**0.85**	0.94	0.64	**0.76**
Average	0.91	0.90	**0.90**	0.92	0.83	**0.87**	0.90	0.85	**0.87**	0.92	0.74	**0.82**

More results in the paper.

Conclusion

- Lexicon-based approach seems to work.
- But a holistic approach is needed to consider all aspects.
  - A new opinion aggregation function is also given.
  - A new way of looking at context dependent opinion words.
  - Many other important linguistic patterns
- Experiments show the effectiveness.