Natural Language Processing

Lecture 7
Sentiment analysis

The problem of sentiment analysis

Sentiment analysis or **opinion mining** is the computational study of opinions, sentiments and emotions expressed in text.

Example: “(1) *I bought an iPhone a few days ago.* (2) *It was such a nice phone.* (3) *The touch screen was really cool.* (4) *The voice quality was clear too.* (5) *Although the battery life was not long, that is ok for me.* (6) *However, my mother was mad with me as I did not tell her before I bought it.* (7) *She also thought the phone was too expensive, and wanted me to return it to the shop.* ...”

The problem of sentiment analysis

An **object** *o* is an entity which can be a product, person, event, organization, or topic. It is associated with a pair *o*:(*T*, *A*), where *T* is a hierarchy of components (or parts), sub-components etc., and *A* is a set of attributes of *o*.

Example: A particular brand of cellular phone is an **object**. It has a set of **components**, e.g. battery, screen, and also a set of **attributes**, e.g. voice quality, size, weight. The battery component also has its set of attributes, e.g. battery life, battery size.

The problem of sentiment analysis

Let an **opinionated document** be *d*, which can be a product review, a forum post or a blog that evaluates a set of objects. In the most general case, *d* consists of a sequence of sentences $d = <s_1, s_2, \ldots, s_m>$.

An **opinion passage** on a feature *f* of an object *O* evaluated in *d* is a group of consecutive sentences in *d* that express a positive or negative opinion on *f*.

If a feature *f* appears in a sentence *s*, *f* is called an **explicit feature** in *s*. If neither *f* nor any of its synonyms appear in *s* but *f* is implied, then *f* is called an **implicit feature** in *s*.

- Explicit feature: "*The battery life of this phone is too short*";
- Implicit feature: "*This phone is too large*".

The problem of sentiment analysis

The **holder of an opinion** is the person or organization that expresses the opinion.

An **opinion** on a feature *f* is a positive or negative view, attitude, emotion or appraisal on *f* from an opinion holder.

The **orientation of an opinion** on a feature *f* indicates whether the opinion is positive, negative or neutral.

Model of an object: An object *o* is represented with a finite set of features, $F = \{f_1, \ldots, f_n\}$, which includes the object itself as a special feature. Each feature can be expressed with any one of a finite set of words or phrases $W_i = \{w_{i1}, \ldots, w_{im}\}$, which are synonyms of the feature, or indicated by any one of a finite set of feature indicators $I_i = \{i_{i1}, \ldots, i_{iq}\}$ of the feature.

The problem of sentiment analysis

Model of an opinionated document: A general opinionated document *d* contains opinions on a set of objects $\{o_1, \ldots, o_q\}$ from a set of opinion holders $\{h_1, \ldots, h_p\}$. The opinions on each object o_j are expressed on a subset F_j of features of o_j. An opinion can be any one of the following two types:

1. A **direct opinion** is a quintuple $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, where o_j is an object, f_{jk} is a feature of the object o_j, oo_{ijkl} is the orientation or polarity of the opinion on feature f_{jk} of object o_j, h_i is the opinion holder and t_l is the time when the opinion is expressed by h_i.
2. A **comparative opinion** expresses a relation of similarities or differences between two or more objects, and/or object preferences of the opinion holder based on some of the shared features of the objects.

The problem of sentiment analysis

Direct opinions:

1. opinions are directly expresses on an object or features of the object, e.g., "*The voice quality of this phone is great.*";
2. opinions on an object are expressed based on its effect on some other objects, e.g., "*After taking this drug, my left knee felt great*".

Objective of mining direct opinions: Given an opinionated document *d*,

1. discover all opinion quintuples (o_j, f_{jk}, oo_{ijkl}, h_i, t_l) in *d*, and
2. identify all the synonyms (W_{jk}) and feature indicators I_{jk} of each feature f_{jk} in *d*.

The problem of sentiment analysis

Example: “(1) *This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone.* (2) *We called each other when we got home.* (3) *The voice on my phone was not so clear, worse than my previous phone.* (4) *The camera was good.* (5) *My girlfriend was quite happy with her phone.* (6) *I wanted a phone with good voice quality.* (7) *So my purchase was a real disappointment.* (8) *I returned the phone yesterday.*”

An **objective sentence** expresses some factual information about the world (№ 1, 2, 8), while a **subjective sentence** expresses some personal feelings or beliefs.

The problem of sentiment analysis

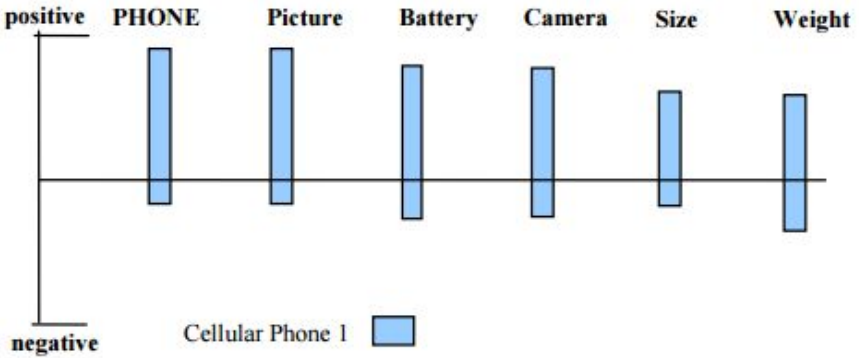
An **explicit opinion** on feature *f* is an opinion explicitly expressed on *f* in a subjective sentence. An **implicit opinion** on feature *f* is an opinion on *f* implied in an objective sentence.

Explicit positive opinion: "*The voice quality of this phone is amazing.*"

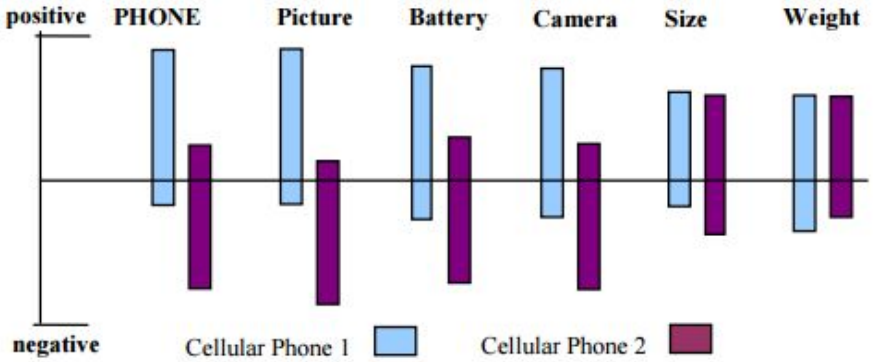
Implicit negative opinion: "*The earphone broke in two days.*"

An **opinionated sentence** is a sentence that expresses explicit or implicit positive or negative opinions. It can be a subjective or objective sentence.

The problem of sentiment analysis



(A) Visualization of feature-based summary of opinions on a cellular phone



(B) Visual opinion comparison of two cellular phones

The problem of sentiment analysis

A simple way to use the results is to produce a **feature-based summary** of opinions on an object or multiple competing objects.

Examples of summary types:

- **Feature buzz summary**: shows the relative frequency of feature mentions. It can tell a company what their customers really care about;
- **Object buzz summary**: shows the frequency of mentions of different competing products. This is useful because it tells the popularity of different products or brands in the market place;
- **Trend tracking**: If the time dimension is added to the above summaries, we get their trend reports. These reports can be extremely helpful in practice because the user always wants to know how things change over time.

Document-Level Sentiment Classification

Task: Given an opinionated document *d* which comments on an object *o*, determine the orientation *oo* of the opinion expressed on *o*, i.e., discover the opinion orientation *oo* on feature *f* in the quintuple (*o*, *f*, *so*, *h*, *t*), where *f* = *o* and *h*, *t*, *o* are assumed to be known or irrelevant.

Assumption: The opinionated document *d* (e.g., a product review) expresses opinions on a single object *o* and the opinions are from a single opinion holder *h*.

Classification Based on Supervised Learning

Features:

- Terms and their frequency;
- Part of speech tags;
- Opinion words and phrases;
- Syntactic dependenies;
- Negation.

Classification Based on Unsupervised Learning

Step 1: extracting phrases containing adjectives or adverbs.

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

In the sentence, "*This camera produces beautiful pictures*", "*beautiful pictures*" will be extracted as it satisfies the first pattern.

Classification Based on Unsupervised Learning

Step 2: estimation of the orientation of the extracted phrases using the **pointwise mutual information** (PMI) measure:

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right).$$

The opinion orientation (*oo*) of a phrase is computed based on its association with the positive reference word “excellent” and its association with the negative reference word “poor”:

$$oo(phrase) = PMI(phrase, \text{“}excellent\text{”}) - PMI(phrase, \text{“}poor\text{”}).$$

$$oo(phrase) = \log_2\left(\frac{hits(\text{phrase } NEAR \text{ "excellent"})hits(\text{"poor"})}{hits(\text{phrase } NEAR \text{ "poor"})hits(\text{"excellent"})}\right).$$

Hits is the number of relevant documents to the search query.

Classification Based on Unsupervised Learning

Step 3: Given a review, the algorithm computes the average *oo* of all phrases in the review, and classifies the review as recommended if the average *oo* is positive, not recommended otherwise.

Sentence-Level Subjectivity

Task: Given a sentence *s*, two sub-tasks are performed:

1. Subjectivity classification: Determine whether *s* is a subjective sentence or an objective sentence;

2. Sentence-level sentiment classification: If *s* is subjective, determine whether it expresses a positive or negative opinion.

Assumption of sentence-level sentiment classification: The sentence expresses a single opinion from a single opinion holder.

Opinion Lexicon Generation

Positive opinion words (*beautiful*, *wonderful*, *good*, and *amazing*) are used to express desired states while **negative opinion words** (*bad*, *poor*, and *terrible*) are used to express undesired states. Apart from individual words, there are also **opinion phrases and idioms**, e.g., *cost someone an arm and a leg*. Collectively, they are called the **opinion lexicon**.

Opinion words types:

- the base type;
- the comparative type, e.g. *better*, *worse*, *best*, *worst*, etc.

Opinion Lexicon Generation

Dictionary-based approach: first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the dictionary (e.g. WordNet) for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration starts. The iterative process stops when no more new words are found.

Corpus-based approach and sentiment consistency: The methods in the corpus-based approach rely on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus.

Feature-Based Sentiment Analysis

1. Identify object features that have been commented on. For instance, in the sentence, "*The picture quality of this camera is amazing*," the object feature is "*picture quality*";
2. Determine whether the opinions on the features are positive, negative or neutral. In the above sentence, the opinion on the feature "*picture quality*" is positive.

Feature Extraction

Format 1 − Pros, cons and the detailed review: The reviewer is asked to describe Pros and Cons separately and also write a detailed/full review.

Format 2 − Free format: The reviewer can write freely, i.e., no separation of Pros and Cons.

My SLR is on the shelf

by camerafun4. Aug 09 '04

Pros: Great photos, easy to use, very small
Cons: Battery usage; included memory is stingy.

I had never used a digital camera prior to purchasing this Canon A70. I have always used a SLR ... **Read the full review**

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 Extraction from Pros and Cons of Format 1

Pros can be separated into three segments:

- great photos ⟨photo⟩
- easy to use ⟨use⟩
- very small ⟨small⟩ ⇒ ⟨size⟩

Cons can be separated into two segments:

- battery usage ⟨battery⟩
- included memory is stingy ⟨memory⟩

Feature Extraction from Pros and Cons of Format 1

The rules are called label sequential rules (LSR), which are generated from sequential patterns in data mining. A **label sequential rule** (LSR) is of the following form, $X \rightarrow Y$, where Y is a sequence and X is a sequence produced from Y by replacing some of its items with wildcards. A wildcard, denoted by a '*', can match any item.

Feature Extraction from Pros and Cons of Format 1

For example, the sentence segment, "*Included memory is stingy*", is turned into the sequence ⟨{included, VB}{memory, NN}{is, VB}{stingy, JJ}⟩.

After labeling, it becomes: ⟨{included, VB}{$feature, NN}{is, VB}{stingy, JJ}⟩.

All the resulting sequences are then used to mine LSRs.

An example rule is: ⟨{easy, JJ }{to}{*, VB}⟩ → ⟨{easy, JJ }{to}{$feature, VB}⟩ confidence = 90%,

where the confidence is the conditional probability, Pr(Y | X), which measures the accuracy of the rule.

Feature Extraction from Reviews of Format 2

- Finding frequent nouns and noun phrases;
- Finding infrequent features by making use of opinion words.

Example: "*picture*" is found to be a frequent feature, and we have the sentence, "*The pictures are absolutely amazing.*" If we know that "*amazing*" is a positive opinion word, then "*software*" can also be extracted as a feature from the following sentence, "*The software is amazing.*" because the two sentences follow the same pattern and "*software*" in the sentence is also a noun.

Opinion Orientation Identification

The lexicon-based approach basically uses opinion words and phrases in a sentence to determine the orientation of the opinion. Apart from the opinion lexicon, negations and but-clauses in a sentence are also crucial and need to be handled. The approach works as follows:

1. Identifying opinion words and phrases: "*The picture quality of this camera is not great***[+1]**, *but the battery life is long***[0]**" ;
2. Handling negations: "*The picture quality of this camera is not great***[-1]**, *but the battery life is long***[0]**";
3. But-clauses: "*The picture quality of this camera is not great***[-1]**, *but the battery life is long***[+1]**";
4. Aggregating opinions:

$$score(f_i, s) = \sum_{op_j \in s} \frac{op_j.so}{d(op_j, f_i)},$$

Sentiment Analysis of Comparative Sentences

Types of comparative relations:

1. Non-equal gradable comparisons: Relations of the type greater or less than that express an ordering of some objects with regard to some of their features, e.g., "*The Intel chip is faster than that of AMD*";
2. Equative comparisons: Relations of the type equal to that state two objects are equal with respect to some of their features, e.g., "*The picture quality of camera X is as good as that of camera Y*";
3. Superlative comparisons: Relations of the type greater or less than all others that rank one object over all others, e.g., "*The Intel chip is the fastest*";
4. Non-gradable comparisons: Relations that compare features of two or more objects, but do not grade them.

Sentiment Analysis of Comparative Sentences

Non-gradable comparisons:

- Object *A* is similar to or different from object *B* with regard to some features, e.g., "*Coke tastes differently from Pepsi*";
- Object *A* has feature f_1, and object *B* has feature f_2 (f_1 and f_2 are usually substitutable), e.g., "*desktop PCs use external speakers but laptops use internal speakers*";
- Object *A* has feature *f*, but object *B* does not have, e.g., "*Cell phone X has an earphone, but cell phone Yoes not have*".

Sentiment Analysis of Comparative Sentences

Given an opinionated document *d*, comparison mining consists of two tasks:

1. Identify comparative sentences in *d*, and classify the identified comparative sentences into different types or classes;
2. Extract comparative opinions from the identified sentences. A comparative opinion in a comparative sentence is expressed with: (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 (objects in O_1 appear before objects in O_2 in the sentence), po is the preferred object set of the opinion holder h, and t is the time when the comparative opinion is expressed.

Sentiment Analysis of Comparative Sentences

Example: Consider the comparative sentence "*Canon's optics is better than those of Sony and Nikon.*" written by John on May 1, 2009.

The extracted comparative opinion is: ({*Canon*}, {*Sony*, *Nikon*}, {*optics*}, preferred:{*Canon*}, *John*, *May-1-2009*).

The object set O_1 is {*Canon*}, the object set O_2 is {*Sony*, *Nikon*}, their shared feature set *F* being compared is {*optics*}, the preferred object set is {*Canon*}, the opinion holder *h* is *John* and the time *t* when this comparative opinion was written is *May-1-2009.*

Text String Element	Automated	Manual/Human
Ability to detect sarcasm?	✗	✓
Emoticon analysis?	✗	✓
Slang and abbreviation analysis?	✗	✓
Time Efficiency	High	Low to Medium
Accuracy	Low to Medium	High
Contextual Analysis	✗	✓

Semantic thesaurus

SentiWordNet: http://sentiwordnet.isti.cnr.it/ ;

SenticNet: http://sentic.net/ ;

WordNet-Affect: http://wndomains.fbk.eu/wnaffect.html .

Thank you for your attention!