# Sentiment Analysis:
## A discovery challenge

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

# Introduction

- Opinion mining or sentiment analysis
  - computational study of opinions, sentiments, appraisal, and emotions expressed in text.
    - Reviews, Twitter, blogs, discussions, comments, etc
- Why is it 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.
    - True for individuals and organizations

# A Fascinating Problem!

- Intellectually challenging
  - A popular research topic in NLP, text mining, and even management sciences!
  - Although there has been so much research,
    - the progress has not been fast!
- Wide spread applications in every domain
  - More than 60 companies in USA alone
    - Many have died and many new ones are still coming
  - One CEO said "Our sentiment analysis is as bad as everyone else's"

# Abstraction (1): what is an opinion?
## - Structure the unstructured

- **Id: Abc123 on 5-1-2008** "*I bought an iPhone* today*. It is such a nice phone. The touch screen is 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, …*"
- We see: Each opinion has a
  - target
  - Sentiment: positive and negative
  - opinion holder: person who holds the opinions
  - time when the opinion was given

# What is an opinion?
(Hu and Liu, 2004; Liu. in NLP handbook)

- 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 a aspect of the entity $e_j$.
  - $so_{ijkl}$ is the sentiment value of the opinion. $so_{ijkl}$ 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.
- Note the simplification: *target* = $(e_j, a_{jk})$

# Structure the unstructured

- Objective: Given an opinionated document,
  - Discover all quintuples $(e_j, a_k, so_{ijkl}, h_i, t_l)$,
  - 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.

# Abstraction (2): Opinion Summary
(Hu & Liu, 2004)

We need quantitative summary

""*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, …*"

**Aspect-Based Summary:**

Opinion summary on 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

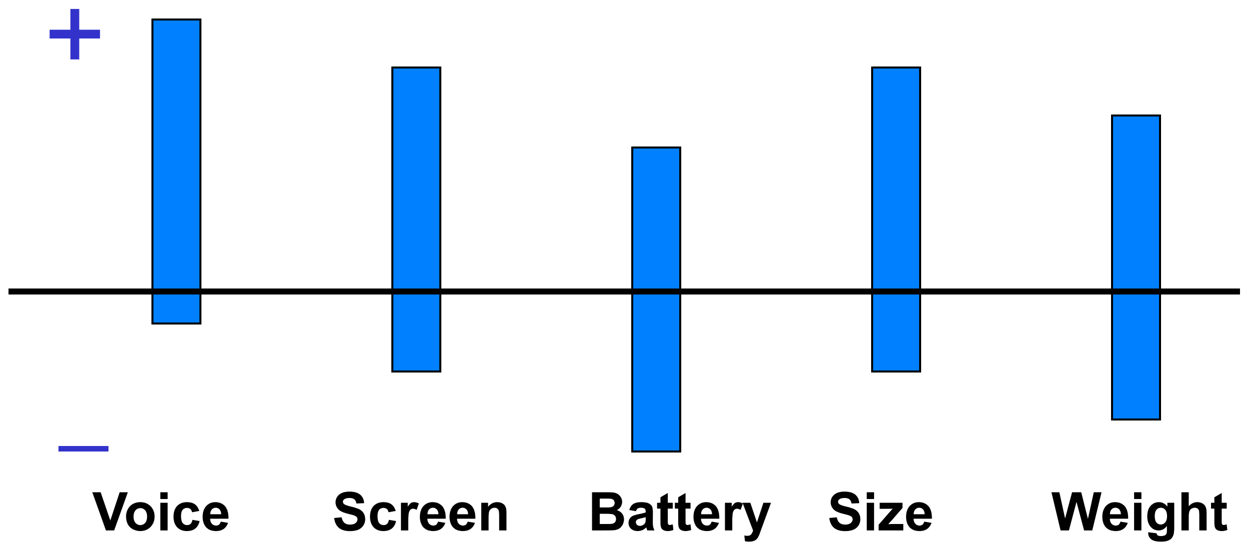
- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…
**Feature2**: **voice quality**
…

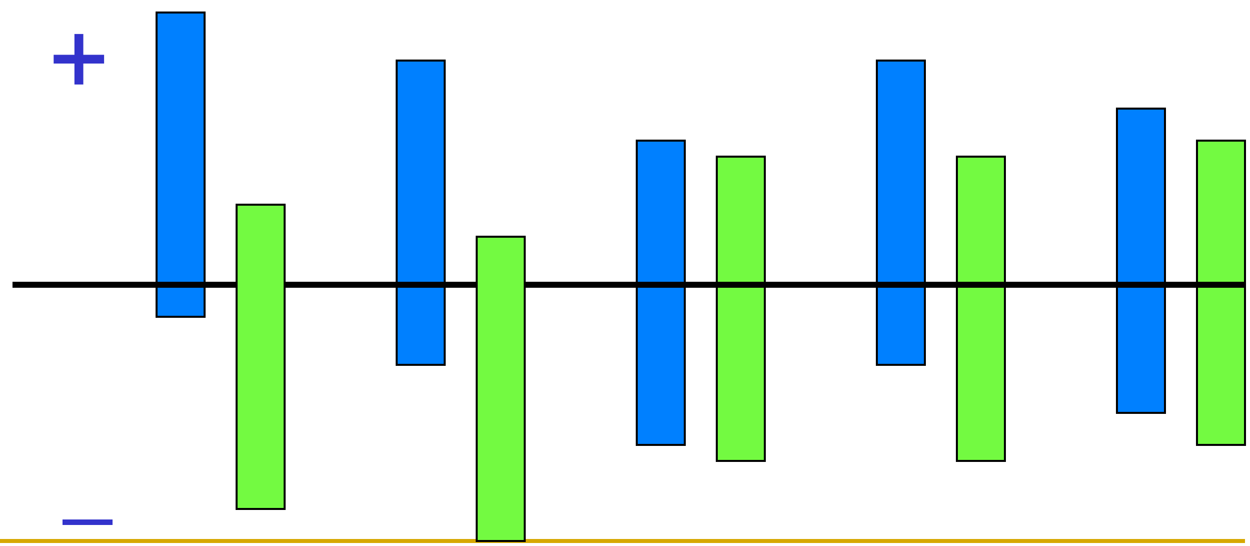
*Note: We omit opinion holders*

# Opinion observer - visualization (Liu et al. 05)

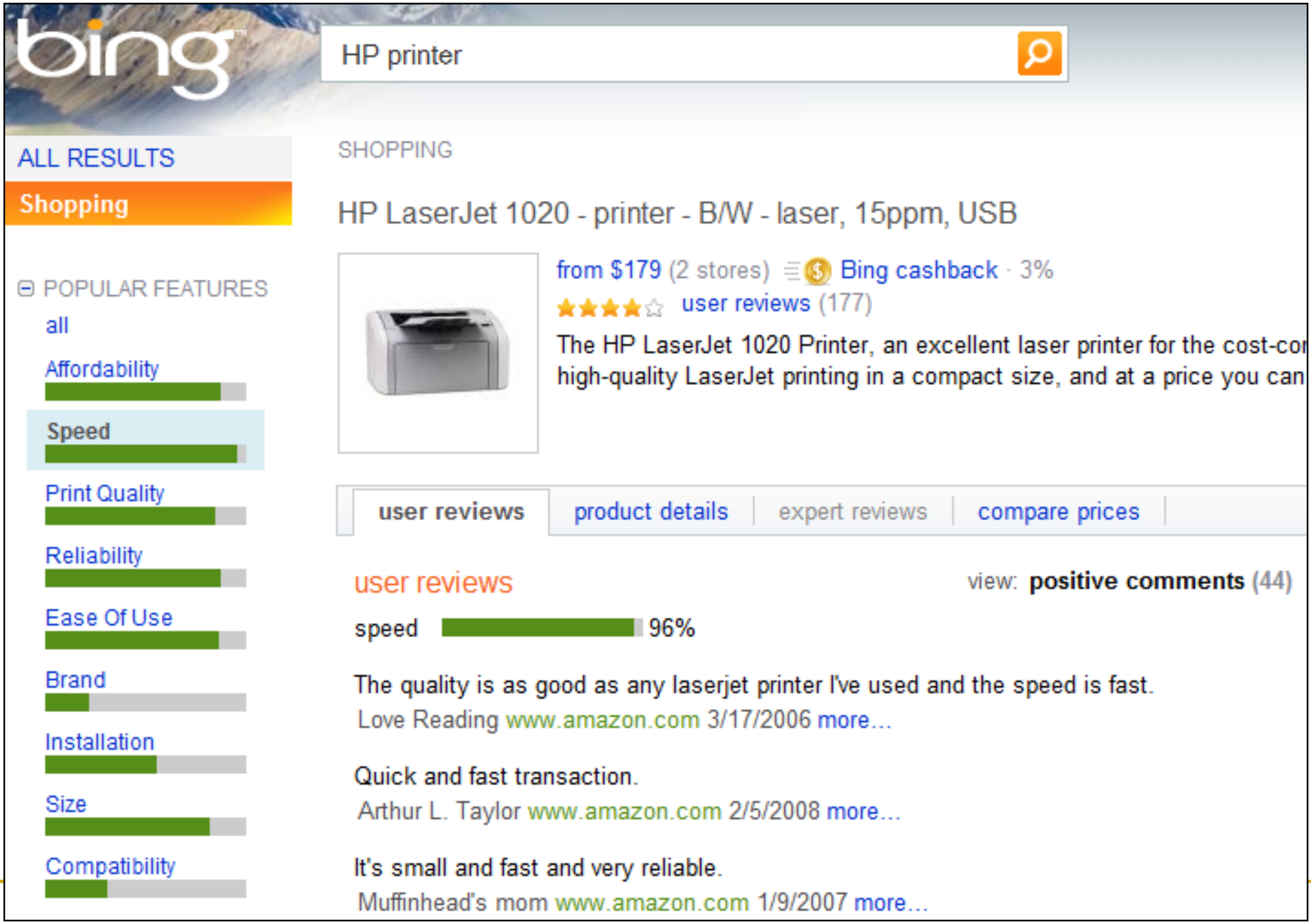
- Summary of reviews of
  Cell Phone 1




- Comparison of reviews of
  Cell Phone 1
  Cell Phone 2



# Feature/aspect-based opinion summary



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

# Not just one problem

- ($e_j$, $f_{jk}$, $so_{ijkl}$, $h_i$, $t_l$),
  - $e_j$ - a target entity: Named Entity Extraction (more)
  - $f_{jk}$ - a feature/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
- Natural language processing issues
  - Coreference resolution
  - Synonym match (voice = sound quality)
  - …

# Highly researched sub-problems

- Document-level
  - Classify reviews as positive or negative
- Sentence-level
  - Subjectivity and sentiment classification, but note
    - both subjective & objective sentences can have opinion.
    - Many subjective sentences have no +ve or –ve opinion
- Aspect-level sentiment analysis
  - Aspect extraction
  - Aspect sentiment classification
- A key challenge is about discovery

# Entity discovery/extraction

- Given BMW and Ford, find all car brands and models and different ways of writing them in a text collection
  - Although similar, it is different from the traditional named entity recognition (NER).
- **Formulation:** Given a set *Q* of seed entities of a particular class *C*, and a set *D* of candidate entities, we wish to determine which of the entities in *D* belong to *C*.
- A classification problem. It needs a binary decision for each entity in *D* (belonging to *C* or not)
  - But it's normally solved as a ranking problem

# Some methods (Li et al 2010, Zhang and Liu, 2011)

- Distributional similarity: This is the traditional method used in NLP, which compare the surrounding text of candidates.
  - It performs poorly.
- PU learning: learning from positive and unlabeled examples.
  - S-EM algorithm (Liu et al. 2002)
- Bayesian Sets: We extended the method given in (Ghahramani and Heller, NIPS-05).

# Determine sentiment is hard!

- Most algorithms use sentiment terms and/or classification to determine sentiments.
  - Sentiment terms do not go very far.
- There is a long tail of cases that sentiment terms cannot handle
  - There seem to be a unlimited number of ways that one can use to express opinions
    - Every domain has some peculiar cases, which make the general opinion mining very hard in practice.
- We need a lot of knowledge discovery

# Some Example Sentences

- I am so happy because my new iPhone is nothing like my old ugly Nokia phone.
- After my wife and I slept on the mattress for a week, I found a hill in the middle.
- Since I had a lot of pain on my back, so my doctor put me on the drug, and only two days after, I have no more pain.
- After taking the drug, my blood pressure went to 400.
- Trying out Google chrome because Firefox keeps crashing
- Anyone know a good Sony camera?
- Anyone know how to fix this lousy washer?
- If I can find a good Sony camera, I will buy it.
- If you are in for a good camera, go for Canon S500.
- What a great car, it stopped working in the second day.

# Basic rules of opinions (Liu, 2010)

- Opinions/sentiments are governed by many rules, e.g.,
  - *Opinion word or phrase, ex: "This is a good car"*

    P ::= a positive opinion word or phrase

    N ::= an negative opinion word or phrase
  - *Desirable or undesirable facts, ex:* "After my wife and I slept on it for two weeks, I noticed a mountain in the middle of the mattress"

    P ::= desirable fact

    N ::= undesirable fact

# Basic rules of opinions

- *High, low, increased and decreased quantity of a positive or negative potential item*, ex: "The battery life is long."

  PO ::= no, low, less or decreased quantity of NPI
  | large, larger, or increased quantity of PPI

  NE ::= no, low, less, or decreased quantity of PPI
  | large, larger, or increased quantity of NPI

  NPI ::= a negative potential item

  PPI ::= a positive potential item

# Basic rules of opinions

- *Decreased and increased quantity of an opinionated item, ex:* "This drug reduced my pain significantly."

  PO ::= less or decreased N
  | more or increased P
  NE ::= less or decreased P
  | more or increased N

- *Deviation from the desired value range*: "This drug increased my blood pressure to 200."

  PO ::= within the desired value range
  NE ::= above or below the desired value range

# Basic rules of opinions

- *Producing and consuming resources and wastes, ex:* "This washer uses a lot of water"

  PO ::= produce a large quantity of or more resource
  | produce no, little or less waste
  | consume no, little or less resource
  | consume a large quantity of or more waste

  NE ::= produce no, little or less resource
  | produce some or more waste
  | consume a large quantity of or more resource
  | consume no, little or less waste

# Desirable or undesirable facts
(Zhang and Liu, 2011)

- "After sleeping on the mattress for one month, a valley has formed in the middle."
- In most sentiment analysis task, we need opinion words, e.g., good, bad, hate, crap, junk, etc
- But objective nouns indicating desirable and undesirable facts can imply opinions too.
  - E.g., How to discover such nouns from a domain corpus?

# The technique

- Sentiment analysis to determine whether the context is +ve or –ve.
  - E.g., "I saw a valley in two days, which is terrible."
  - This is a negative context.
- Statistical test to find +ve and –ve candidates.

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Pruning to move those unlikely ones though *sentiment homogeneity*.

# Pruning

- For an aspect with an implied opinion, it has a fixed opinion, either +ve or –ve, but not both.
- We find two direct modification relations using a dependency parser.
  - Type 1: *O → O-Dep → A*
    - e.g. " *This TV has good picture quality*."
  - Type 2: *O → O-Dep → H ← A-Dep ← A*
    - e.g. " *The springs of the mattress are bad*. "
- If an aspect has mixed opinions based on the two dependency relations, prune it.

# Opinions implied by resource usage
(Zhang and Liu, 2011)

- Resource usage descriptions often imply opinions (as mentioned in rules of opinions)
  - E.g., "This washer uses a lot of water."
- Two key roles played by resources usage:
  - An important aspect of an entity, e.g., water usage.
  - Imply a positive or negative opinion
- Resource usages that imply opinions can often be described by a triple.

  (verb, quantifier, noun_term),

  - Verb: uses, quantifier: "a lot of ", noun_term: water

# The proposed technique

- The proposed method is graph-based.
  - Stage 1: Identifying Some Global Resource Verbs
    - Identify and score common resource usage verbs used in almost any domain, e.g., "use" and "consume"
  - Stage 2: Discovering Resource Terms in each Domain Corpus
    - Use a graph-based method considering occurrence probabilities.
    - With resource verbs identified from stage 1 as the seeds.
    - Score domain specific resource usage verbs and resource terms.

# The algorithm

**Algorithm:** MRE ($Q$, $G$)

**Input:** A global resource verb set $Q$ with their hub scores computed from HITS in stage 1, and $G$ is the bipartite graph

**Output:** a ranked list of candidate resource terms

1. $u^0(i) \leftarrow H(i)$ of verb $i$, if $verb\ i \in Q$
2. $u^0(i) \leftarrow \arg\min_{r \in Q}\{H(r)\}$, if $verb\ i \notin Q$
3. **Repeat** till convergence
4. $\quad r^{n+1}(j) = \sum_{(i,j)\in L} p_{ij} u^n(i)$
5. $\quad u^{n+1}(i) = \sum_{(i,j)\in L} p_{ji} r^n(j)$
6. $\quad$ normalize $r(j)$ and $u(i)$
7. Output the ranked candidate resource terms based on their $r(j)$ score values.

# Coreference resolution: semantic level?

- Coreference resolution (Ding and Liu, 2010)
  - "I bought the Sharp tv a month ago. The picture quality is so bad. Our other Sony tv is much better than this Sharp. ***It*** is also so expensive".
    - "it" means "Sharp"
  - "I bought the Sharp tv a month ago. The picture quality is so bad. Our other Sony tv is much better than this Sharp. ***It*** is also very reliable."
    - "it" means "Sony
- Sentiment consistency.

# Coreference resolution (contd)

- "The picture quality of this Canon camera is very good. ***It*** is not expensive either."
  - Does "it" mean "Canon camera" or "Picture Quality"?
    - Clearly it is Canon camera because picture quality cannot be expensive.
    - Commonsense knowledge, but can be discovered.
- For coreference resolution, we actually need to
  - do sentiment analysis first, and
  - mine adjective-noun associations using dependency
- Finally, use supervised learning

# 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 in market*"

# Analyzing Comparative Opinions

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

  $(E_1, E_2, F, po, h, t)$,

  where $E_1$ and $E_2$ are the entity sets being compared based on their shared features/aspects $F$, $po$ 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.

# Deal with comparative opinions

- Gradable comparative sentences can be dealt with *almost* as normal opinion sentences.
  - E.g., "*optics of camera A is better than that of camera B*"
  - Positive: "*optics of camera A*"
  - Negative: "*optics of camera B*"
- Difficulty: recognize non-standard comparatives
  - E.g., "I am so happy because my new iPhone is nothing like my old slow ugly Droid."
  - ?

# Some techniques (Jindal and Liu, 2006, Ding et al, 2009)

- Identify comparative sentences
  - Using class sequential rules as attributes in the data, and then
  - Supervised learning
- Extraction of different items
  - Label sequential rules
  - conditional random fields
- Determine opinion orientations
  - Parsing and opinion lexicon
    - We have not used supervised learning

# Group aspects synonyms (Zhai et al. 2011a, b)

- Once aspects expressions are discovered, group them into /aspect categories.
  - Power usage and battery life are the same.
- A variety of information is used in clustering
  - Lexical similarity based on WordNet
  - Distributional information
  - Syntactical information/constraints
- Two Methods:
  - Clustering: EM-based method.

# The EM-based method

- WordNet similarity

$$Jcn(w_1, w_2) = \frac{1}{IC(w_1) + IC(w_2) - 2 \times Res(w_1, w_2)}$$

- EM-based probabilistic clustering

$$\mathrm{P}(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{ti}\mathrm{P}(c_j|d_i)}{|V| + \sum_{m=1}^{|V|}\sum_{i=1}^{|D|} N_{mi}\mathrm{P}(c_j|d_i)}$$

$$\mathrm{P}(c_j) = \frac{1 + \sum_{i=1}^{|D|} \mathrm{P}(c_j|d_i)}{|C| + |D|}$$

$$\mathrm{P}(c_j|d_i) = \frac{\mathrm{P}(c_j)\prod_{\mathrm{k}=1}^{|d_i|} \mathrm{P}(w_{d_{i,k}}|c_j)}{\sum_{\mathrm{r}=1}^{|C|} \mathrm{P}(c_r)\prod_{\mathrm{k}=1}^{|d_i|} \mathrm{P}(w_{d_{i,k}}|c_r)}$$

# Constrained Topic Modeling

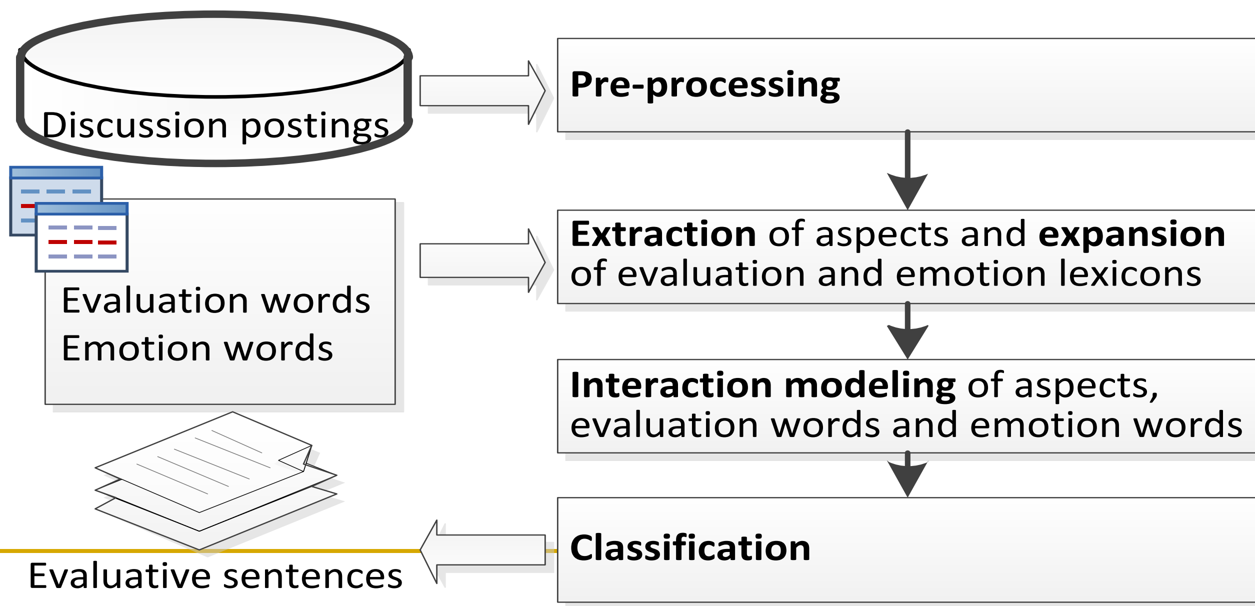
- Constrained topic model: Constrained-LDA
- In topic modeling, we add probabilistic constraints
  - Must-links
  - Cannot link
- In Gibbs sampling, we consider constraints to guide its topic assignments of aspect terms.

# Find evaluative opinions in discussions (Zhai et al. 2011)

- Existing research focuses on product reviews
  - reviews are opinion-rich and
  - contain little irrelevant information.
- Not true about online discussions.
  - Many of the postings do not express opinions about the discussion topic.
  - Evaluative opinions, "*The German defense is strong.*"
  - Non-evaluative opinions, "*I feel so sad for Argentina.*" "*you know nothing about defense*"
- Goal: discover evaluative opinion sentences.

# 3. The Proposed Technique

- Intuitions: (1) An **evaluative** opinion should comment on a topic/ entity or some aspects of it. (2) **Evaluation words** and **emotion words** are indications of evaluative and emotional sentences, respectively.
- Overview: Given the raw discussion postings, the algorithm works in 4 steps to identify ***evaluative*** sentences.



## **3.1** Extraction of Aspects and Expansion of Evaluation and Emotion Lexicons

**Input**: Text corpus *R* ; Evaluation word seeds *vas*;
Emotion word seeds *mos*. // Not sufficient

**Output**: All evaluation words *VA*; All emotion words *MO*;
All aspects: *A*

**Task 1**. Extract **aspects** using evaluation/emotion words;

**Task 2**. Extract **aspects** using extracted aspects;

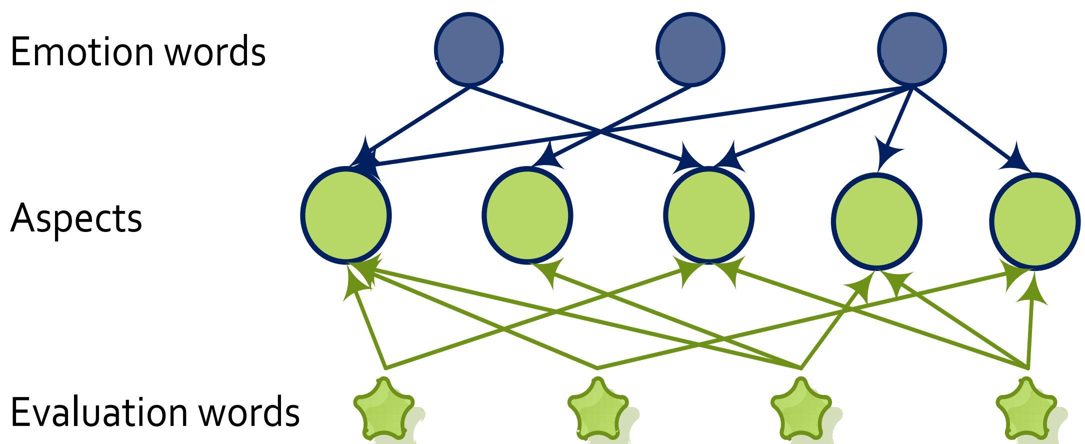
**Task 3**. Extract **evaluation words** and **emotion words** using the given or extracted evaluation words and emotion words respectively.

# Double-Propagation (DP)

- We use the Double Propagation method in (Qiu et al 2009; 2011).
- The idea is that an opinion has a target.
  - Ex: This Sony camera is great.
- This technique needs a dependency parser.
- In this work, we are interested in Chinese microblog (weibo) discussions
  - But Chinese dependency parsers are not accurate.
- We approximate the DP method using POS tags

# 3.2 Aspects, Evaluation Words and Emotion Words Interaction

- An extracted aspect that is associated with many *evaluation words* is more likely to indicate an evaluative sentence. Then, we want to give a high score to the aspect.
- An extracted aspect that is associated with many *emotion words* is not a good indicator of an evaluative sentence. It should be assigned a low score.



# 3.2 Aspects, Evaluation Words and Emotion Words Interaction

- An evaluation word that does not modify *good* (high scored) aspects are likely to be a wrong evaluation word, and should be weighted down.
- The more evaluative the aspects are, the less emotional their associated emotion words should be.

$$asp(a_i) = \lambda \times \sum_{(i,j)\epsilon E_{va-a}} eva(va_j) - (1-\lambda) \times \sum_{(i,k)\epsilon E_{mo-a}} emo(mo_k) \quad (1)$$

$$eva(va_j) = \sum_{(i,j)\epsilon E_{va-a}} asp(a_i) \quad (2)$$

$$tmp(mo_k) = \sum_{(i,k)\epsilon E_{mo-a}} asp(a_i) \quad (3)$$

$$emo(mo_k) \propto -tmp(mo_k) \quad (4)$$

$$emo(mo_k) = -tmp(mo_k) + max = max - tmp(mo_k) \quad (5)$$

$$max = \max\{tmp(mo_1), tmp(mo_2), \dots, tmp(mo_{|V_{mo}|})\} \quad (6)$$

# Summary

- Opinion mining or sentiment analysis is a fascinating NLP or text mining problem.
- It is also restricted NLP problem
  - Because we only need to understand one aspect of the semantic meaning.
- General NLP is probably hopeless.
- But can we solve this restricted problem?
  - Although many challenges, there are already numerous applications.
  - I am optimistic.

# References

- See my page and the book:
  - http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
  - B. Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. *Second Edition*, Springer, July, 2011.