

Aspect Summarization from Blogosphere for Social Study

Chia-Hui Chang (張嘉惠)
National Central University, Taiwan
12/31/2008
(網路社群服務計算暨探勘技術研討會)



Part I



A Brief Survey on Sentiment and Opinion Summarization

Introduction

- Opinionated documents
 - News article vs. editorial reviews
- Opinionated documents on Web
 - Online reviews: Amazon, IMDB, netflix, epinion.com, etc.
 - bboards, newsgroups, forums
 - Blogs

Motivations and Applications

- **Product survey:** What features of the ThinkPad T43 do customers like and which do they dislike?
- **Political opinion polls:** Should abortion be banned?
- **Sentiment classification for advertisement analysis**
- **Prediction (election outcomes, market trends):**
Will Clinton or Obama win?
- **Opinion question answering:** What does [X] feel about [Y]?

Research Topics

- Research topics
 - Subjectivity filtering (Blog opinion retrieval)
 - Sentiment classification (Polarity orientation)
 - Product feature (opinion target) extractions
 - Opinion holder (source) extraction
 - Aspect analysis (Latent topic analysis)
 - Visualization of sentiments
 - ...

Sentiment Classification

- Definition:
 - Differentiate between Positive, Negative and Neutral

- Supervised:
 - Manually created adjective words with positive and negative labels,
 - Semantic features based on substitutions and proximity,
 - Sentiment wordlist and WordNet

- Unsupervised:
 - Turney (ACL2002) used internet-based method to determine semantic orientation of reviews.

Turney ACL 2002



Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

Peter D. Turney

Institute for Information Technology
National Research Council of Canada
Ottawa, Ontario, Canada, K1A 0R6
peter.turney@nrc.ca

Turney ACL 2002

Thumbs up or thumbs down?

- Goal: classify reviews (whole document) as “positive” or “negative”.
- Method:
 - Find (possibly) meaningful phrases from review (e.g., “bright display”, “inspiring lecture”, ...)
 - Estimate “semantic orientation” of each candidate phrase (based on POS patterns, like ADJ NOUN)
 - Assign overall orientation of review by averaging orientation of the phrases in the review

Semantic orientation (SO) of phrases

$$PMI(word_1, word_2) = \log_2 \left[\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)} \right]$$

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

$$= \log_2 \left[\frac{\text{hits}(phrase \text{ NEAR } \text{“excellent”}) \text{ hits}(\text{“poor”})}{\text{hits}(phrase \text{ NEAR } \text{“poor”}) \text{ hits}(\text{“excellent”})} \right]$$

Table 2. An example of the processing of a review that the author has classified as *recommended*.⁶

| Extracted Phrase | Part-of-Speech Tags | Semantic Orientation |
|------------------------------|---------------------|----------------------|
| online experience | JJ NN | 2.253 |
| low fees | JJ NNS | 0.333 |
| local branch | JJ NN | 0.421 |
| small part | JJ NN | 0.053 |
| online service | JJ NN | 2.780 |
| printable version | JJ NN | -0.705 |
| direct deposit | JJ NN | 1.288 |
| well other | RB JJ | 0.237 |
| inconveniently located | RB VBN | -1.541 |
| other bank | JJ NN | -0.850 |
| true service | JJ NN | -0.732 |
| Average Semantic Orientation | | 0.322 |

Table 3. An example of the processing of a review that the author has classified as *not recommended*.

| Extracted Phrase | Part-of-Speech Tags | Semantic Orientation |
|------------------------------|---------------------|----------------------|
| little difference | JJ NN | -1.615 |
| clever tricks | JJ NNS | -0.040 |
| programs such | NNS JJ | 0.117 |
| possible moment | JJ NN | -0.668 |
| unethical practices | JJ NNS | -8.484 |
| low funds | JJ NNS | -6.843 |
| old man | JJ NN | -2.566 |
| other problems | JJ NNS | -2.748 |
| probably wondering | RB VBG | -1.830 |
| virtual monopoly | JJ NN | -2.050 |
| other bank | JJ NN | -0.850 |
| extra day | JJ NN | -0.286 |
| direct deposits | JJ NNS | 5.771 |
| online web | JJ NN | 1.936 |
| cool thing | JJ NN | 0.395 |
| very handy | RB JJ | 1.349 |
| lesser evil | RBR JJ | -2.288 |
| Average Semantic Orientation | | -1.218 |

Table 5. The accuracy of the classification and the correlation of the semantic orientation with the star rating.

| Domain of Review | Accuracy | Correlation |
|---------------------|----------|-------------|
| Automobiles | 84.00 % | 0.4618 |
| Honda Accord | 83.78 % | 0.2721 |
| Volkswagen Jetta | 84.21 % | 0.6299 |
| Banks | 80.00 % | 0.6167 |
| Bank of America | 78.33 % | 0.6423 |
| Washington Mutual | 81.67 % | 0.5896 |
| Movies | 65.83 % | 0.3608 |
| The Matrix | 66.67 % | 0.3811 |
| Pearl Harbor | 65.00 % | 0.2907 |
| Travel Destinations | 70.53 % | 0.4155 |
| Cancun | 64.41 % | 0.4194 |
| Puerto Vallarta | 80.56 % | 0.1447 |
| All | 74.39 % | 0.5174 |

Key ideas in Turney's paper

- Simplification:
 - classify an entire document, not a piece of it. (Many reviews are mixed.)
- Focus on what seems important:
 - Extract semantically oriented words/phrases from the document. (Phrases are less ambiguous than words).
- Bootstrapping/semi-supervised learning:
 - To assess orientation of phrases, use some kind of contextual similarity of phrases

Subjectivity Filtering

- Definition:
 - Differentiate “facts” from “opinions”
- Blog opinion retrieval track @ TREC 2006
 - Attract the most participants in TREC 2007
- Supervised
 - Yu and Hatzivassiloglou (EMNLP2003)
 - use news as facts, editorials and letters as opinions to train a model
 - Riloff and Wiebe (EMNLP2003)
 - Learning Extraction Patterns for Subjective Expressions
- Unsupervised

Yu and Hatzivassiloglou EMNLP 2003

Towards Answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences

- Finding opinion sentences
 - Similarity approach
 - Naïve bayes classifier
 - use news as facts, editorials and letters as opinions to train a model
 - Multiple naïve bayes classifiers
- Identifying the polarity of opinion sentences
 - Score each word W_i with POS_j
 - Turney's approach with k seeds
 - Averaging via over per word log likelihood score
 - Positive $> t_p >$ neutral $> t_n$

Pang, Lee, Vaithyanathan EMNLP 2002

Thumbs up? Sentiment Classification using Machine Learning Techniques



- Movie review (IMDB) classification
 - 752 negative reviews and 1301 positive reviews
- Baseline
 - Count human-provided polar words (50% to 69%)
- Naïve Bayes, Maximum Entropy, SVM
 - 3-fold cross validation
 - Naïve Bayes (bag of words): 78.7
 - SVM-lite “set of words”: 82.9 was best result
 - Results do not reach levels achieved in **topic categorization**
- Various feature combinations
 - Unigram (**presence** works best), bigram, POS, text position

Dave, Lawrence, Pennock WWW 2003

Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews

- Product-level review-classification
- Train **Naïve Bayes classifier** using a corpus of self-tagged reviews available from major web sites (C|net, amazon)
- **Refine the classifier using the same corpus before evaluating it on sentences mined from broad web searches**

Some of the slides are pilfered from [W. Cohen's slides](#) for the Course "Analysis of Social Media."

Product Feature Extractions

- Definition

- Find all explicit and implicit product features
- Produce the positive and negative opinion sets for each feature

- Supervised rule mining

- Hu & Liu KDD 2004, WWW 2005

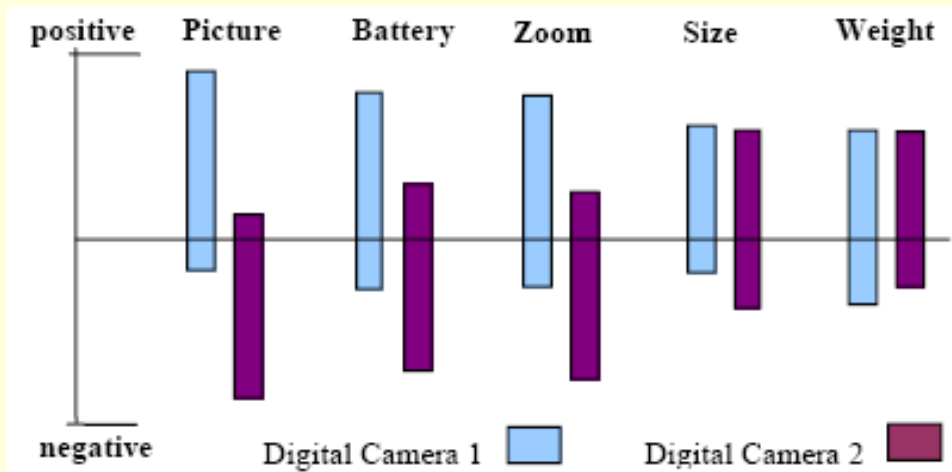
Mining Opinion Features in Customer Reviews

Opinion observer: Analyzing and comparing opinions on the Web

Hu & Liu KDD 2004, WWW 2005

Mining Opinion Features in Customer Reviews

- Supervised rule mining
 - Perform POS tagging
 - Replace actual feature words with [feature]
 - Use n-gram (n=3) to produce shorter segments from long ones
- Extract association rules (item-sets) based on support (>1%)
 - Generate rules with [feature] on the RHS
 - Impose ordering on LHS



Popescu & Etzioni 2005

- Report on a product review mining system that extracts and labels opinion expressions their attributes
- They use the **relaxation-labeling technique** from computer vision to perform unsupervised classification satisfying local constraints (which they call **neighborhood features**)
- The system tries to solve several classification problems (e.g. opinion and target finding) at the same time rather than separately.

Visualization of Sentiment

1) [opinmind.com](http://www.opinmind.com)

searches for positive/negative sentiments about search term

sample queries:

google vs. microsoft

2) opine (Popescu, Nguyen, Etzioni)

sentiment-feature labeling of hotel reviews

<http://www.cs.washington.edu/research/knowitall/opine/>

new york, attribute:bed

3) OASYS

sentiment analysis of news sources, sliced by country and source

<http://oasys.umiacs.umd.edu/oasysnew/oasys.php>

2006 Horizon Awards Winner: University of Maryland's Oasys — Opinion Thermometer

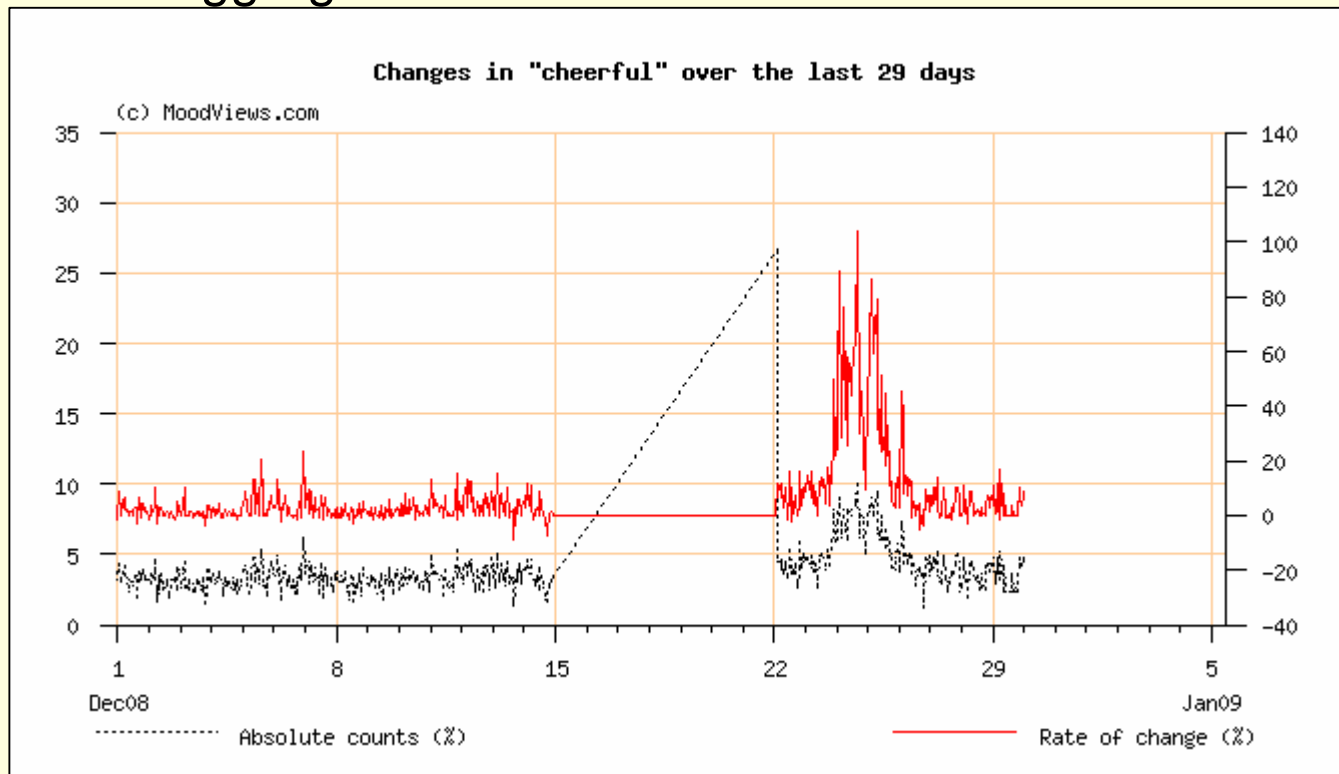
Visualization of Sentiment (Cont.)

4) Moodviews

<http://ilps.science.uva.nl/MoodViews/>

Moodteller - predict aggregate mood from text

Moodspotter - explain discrepancies between predicted and actual aggregate mood



Visualization of Sentiment (Cont.)

5) TextMap

search engines for entity & sentiment over news & blogs

news: <http://www.textmap.com>

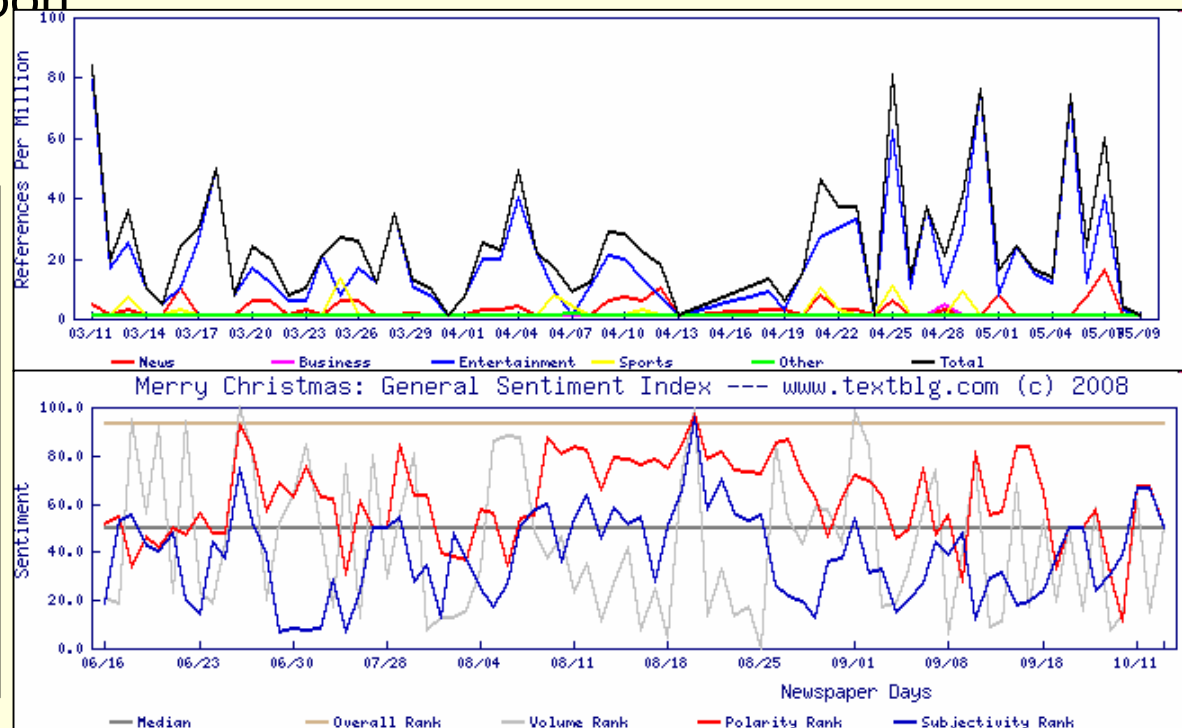
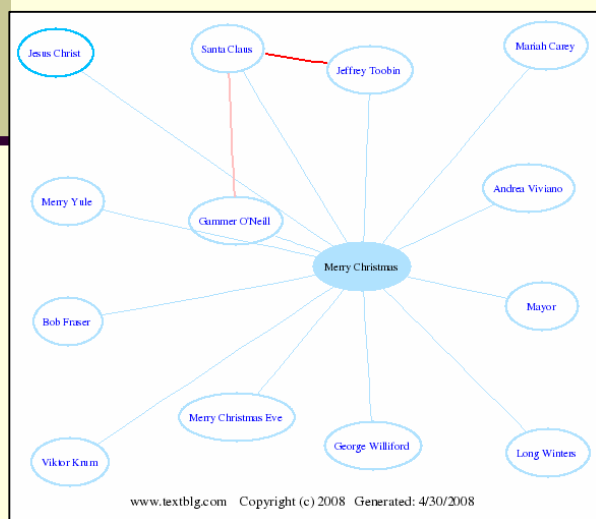
blog: <http://www.textblg.com>

<http://www.icwsm.org/papers/3--Godbole-Srinivasaiah-Skienna.pdf>

Daily sentiment report

Top entities

Heatmaps



Extraction of Opinion Holders

- Bethard et al. 2004
Automatic Extraction of Opinion Propositions and their Holders
- Choi et al. 2005
Identifying sources of opinions with conditional random fields and extraction patterns
- Kim & Hovy 2006
Extracting opinions, opinion holders, and topics expressed in online news media text
- Choi, Breck, Cardie 2006
Joint extraction of entities and relations for opinion recognition

Aspect Analysis

- Hofmann, Machine Learning 2001
 - Unsupervised learning by probabilistic latent semantic analysis.
- Blei, et al. Machine Learning, 2003.
 - Latent dirichlet allocation.
- Mei, et al. WWW 2007
 - Topic sentiment mixture: modeling facets and opinions in weblogs.
- Titov and McDonald. WWW 2008
 - Modeling online reviews with multi-grain topic models.

Trends & Buzz

- Koppel & Shtrimberg(2004)
 - Correlate positive/negative news stories about publicly traded companies and the stock price changes

- Glance et al. (2005)
 - Market Intelligence from message boards, forums, blogs.

- Mishne and Glance (2006)
 - Analyze Blog sentiments about movies and correlate it with its sales

Part II

Aspect Summarization from Blogosphere for
Social Study

Chia-Hui Chang & Kun-Chang Tsai
ICDM Workshops 2007

Outline

- Motivation
- Related Work
- Tasks for Aspect Summarization
- Experiments
- Conclusions

Motivation

- The popularity of weblogs opened up a lot of new opportunities in **social study**.
 - In the past, researchers in social study had to spent huge efforts in collecting data.
 - Now, we can collect public opinions non-intrusively from the blogosphere by querying articles and comments on a specific topic during a specific time period.

Introduction

- Previous studies on the opinions extraction focus on polarity identification.
- In the context of social study,
 - knowing what people approve or disapprove is not enough
 - the main rationale of why they agree or disagree on are more important
- Balanced politics
<http://www.balancedpolitics.org/>

Related Work by Domain

- **Opinion Extraction from Different Domain:**
 - Product reviews (Turney, 2002; Dave et al., 2003; Hu and Liu, 2004),
 - Movie reviews (Zhuang et al., 2006),
 - Editorial reviews (Kim and Eduard, 2004).
 - Blog articles (Ku, et al., 2006)

Related Work by Task

1. Sentiment orientation: either positive or negative
 - From product review to blogs
2. Opinion or fact, subjective or objective:
 - Especially in blogs
3. Feature extraction
 - Product features (Hu and Liu, 2005)
 - Rationale/Reasons behind the opinion

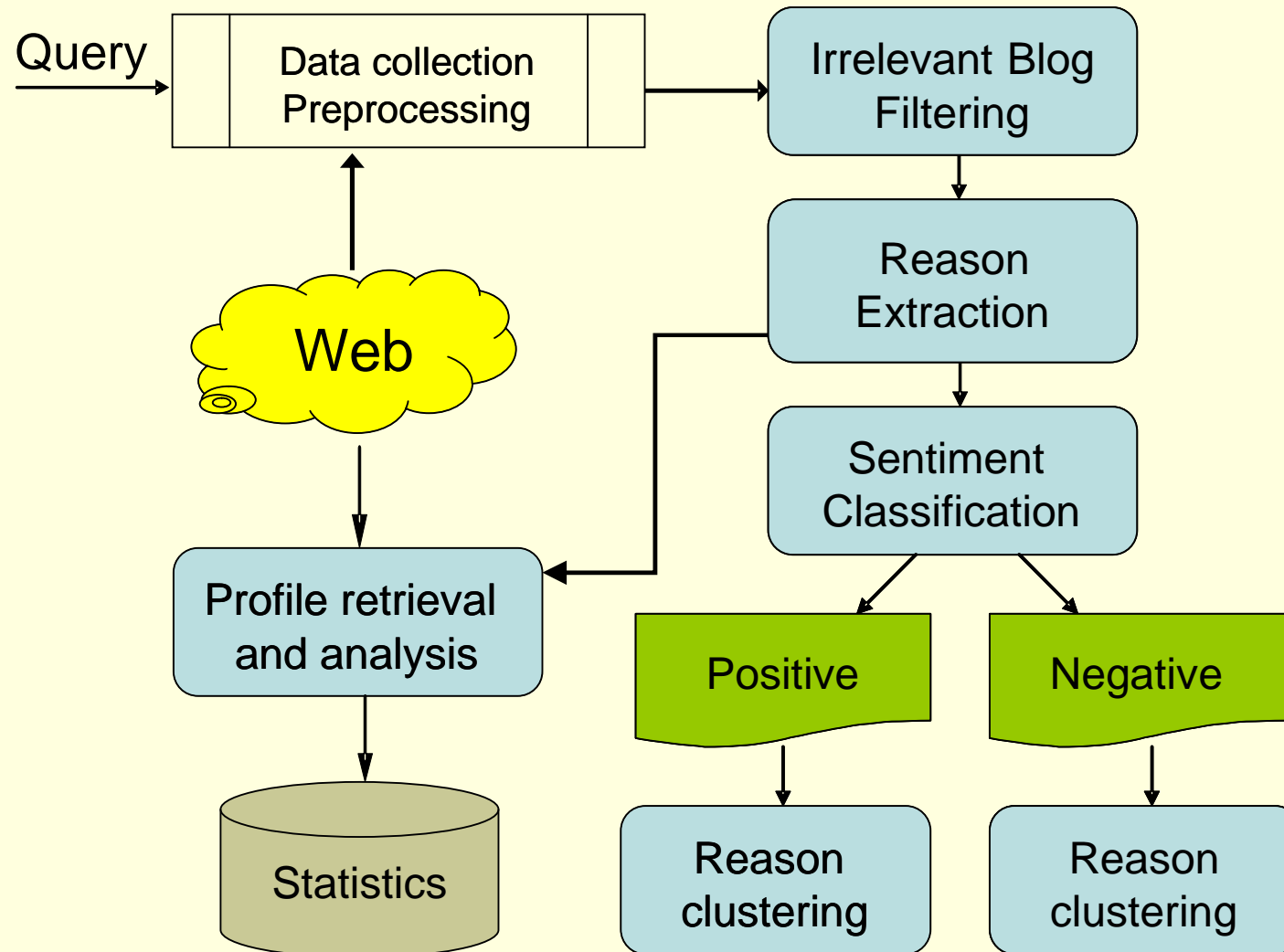
Related Work by Approach

- On sentiment classification
 - Supervised:
 - Manually created adjective words with positive and negative labels,
 - Semantic features based on substitutions and proximity,
 - Sentiment wordlist and WordNet
 - Unsupervised:
 - Turney (ACL2002) used internet-based method to determine semantic orientation of reviews.

Related Work by Approach (Cont.)

- Classification of opinion or fact:
 - Riloff and Wiebe (EMNLP2003) separate subjective and objective sentences using supervised machine learning approach.
 - Yu and Hatzivassiloglou (EMNLP2003) use news as facts, editorials and letters as opinions to train a model.

Aspect Summarization Flowchart



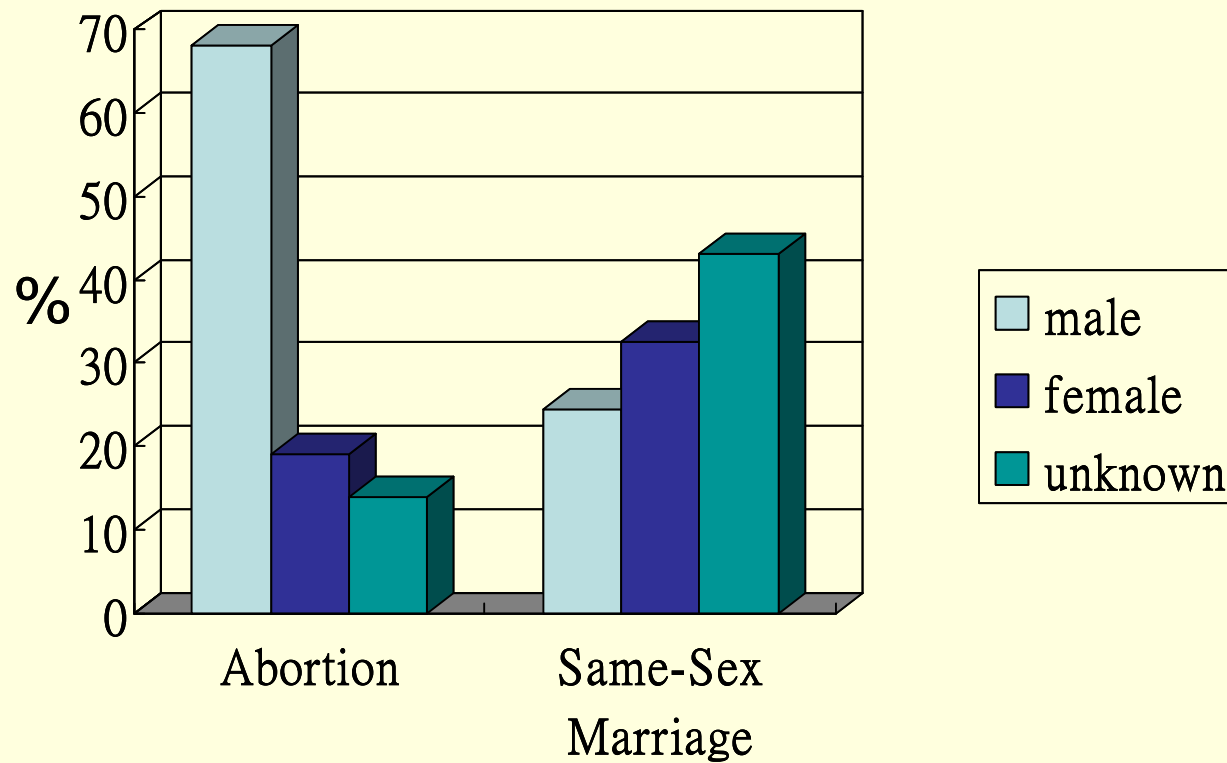
On Same Sex Marriage

| Positive reasons | Negative Reasons |
|---|---|
| <ol style="list-style-type: none"><li data-bbox="322 544 1070 715">1. The first positive reason (with 10 supports) is that same-sex couple wants to be treated equally.<li data-bbox="322 727 1070 898">2. The second (with 4 supports) describe that same-sex marriage doesn't threat heterosexual ones.<li data-bbox="322 911 1070 1201">3. The third main reason simply states that government has no business in people's private, sexual lives, they can't ban same-sex marriage.<li data-bbox="322 1214 1070 1441">4. The fourth one explains that as long as two people love each other, they can marry no matter what their sexes are. | <ol style="list-style-type: none"><li data-bbox="1111 544 1993 898">1. The first negative reason (with 5 supports) describes that same-sex marriage shouldn't be legal because they often adopt children, and the children in this kind of family will have huge emotional problems.<li data-bbox="1111 911 1993 1082">2. The second reason (with 4 supports) states that church shouldn't be forced to hold same-sex marriage.<li data-bbox="1111 1094 1993 1201">3. The third reason illustrates that same-sex marriage is unnatural. (3 supports)<li data-bbox="1111 1214 1993 1441">4. The fourth one explains that we should not break the traditional definition of marriage as a union of one man and one woman. (3 supports) |

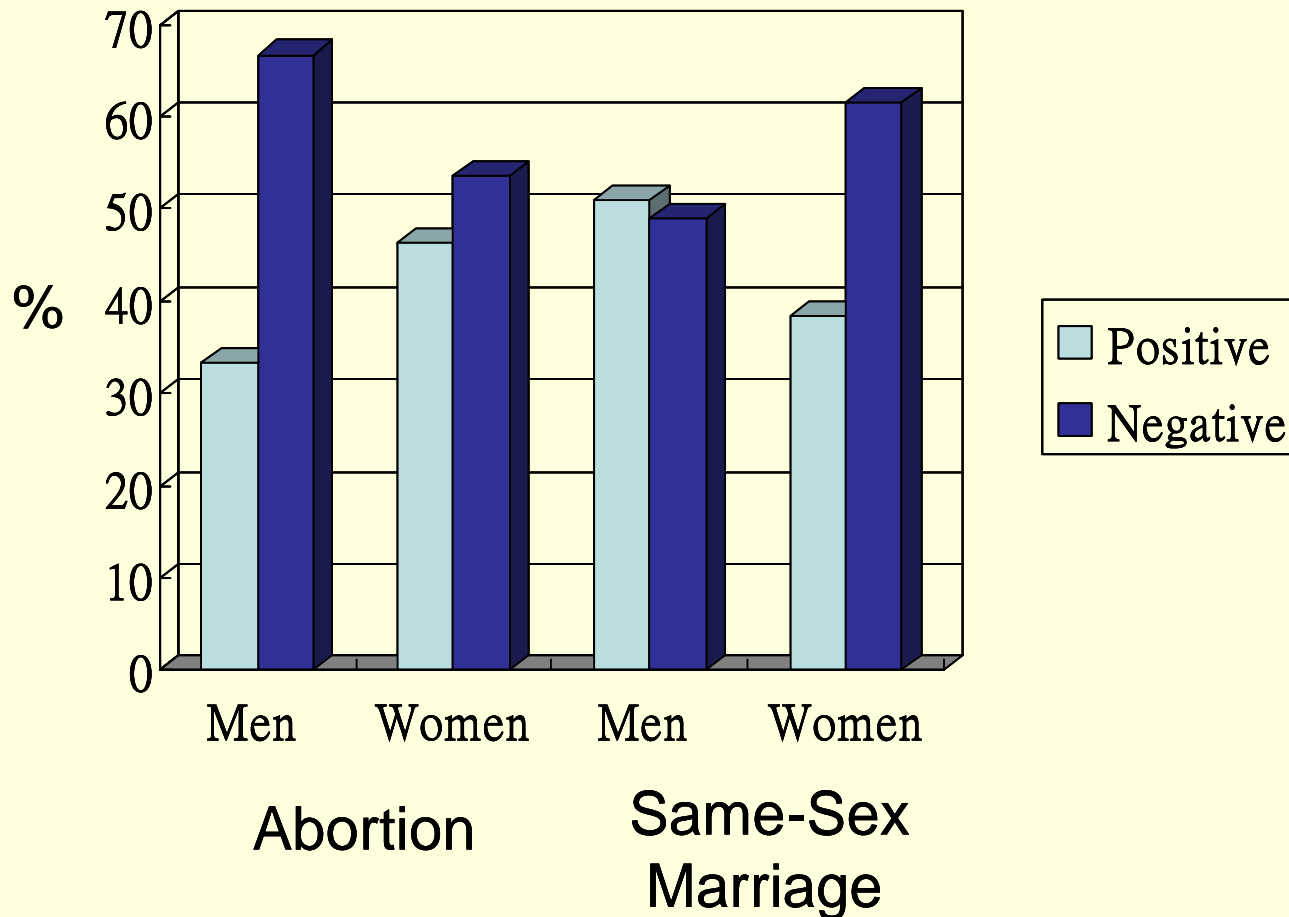
On Abortion

| Positive reasons | Negative Reasons |
|--|--|
| <ol style="list-style-type: none"><li data-bbox="286 544 1099 735">1. The first reason (with 6 supports) describes that human body is private property and women can decide by herself.<li data-bbox="286 759 1099 903">2. The second one (with 6 supports) comes from <i>eugenics</i> which concerns babies with serious defect.<li data-bbox="286 927 1099 1070">3. The third one (with 6 supports) says there might be no support to mothers or they wouldn't abort.<li data-bbox="286 1094 1099 1238">4. The fourth one explains that some women accidentally got pregnant and abortion should be one of the choices.<li data-bbox="286 1262 1099 1445">5. The fifth main reason claims that some couples believe their family is unbalanced without a son so they want to use abortion. | <ol style="list-style-type: none"><li data-bbox="1131 544 2033 735">1. The first negative reason (with 7 supports) toward "abortion" describes that abortion is not necessary because women can choose adoption instead.<li data-bbox="1131 759 2033 903">2. The second one (with 4 supports) talks about how cruel the procedure of abortion surgery is.<li data-bbox="1131 927 2033 1015">3. The third one simply indicates that abortion is murder.<li data-bbox="1131 1038 2033 1182">4. The fourth one explains that there are always alternative solutions to overcome the difficulty.<li data-bbox="1131 1206 2033 1350">5. The fifth main reason claims that an unborn baby should be treated as a human with the right been protected |

Support Counts by Gender



Sentiment Orientation by Gender



Preprocessing

- Sentence break
 - CCG sentence segmentation tool
- Stemming
 - Porter stemmer
- Part-of-Speech tagging
 - Genia Tagger

Task 1: Irrelevant Blogs Filtering

- Not all of the search result blogs contain reasons.
- Topic word density in sentence levels
 - If the density of a blog below a threshold, we consider it doesn't contain any reason.

$$\frac{\text{\# of sentences which contain at least one topic words}}{\text{\# of sentences in the blog}}$$

Paragraph segmentation

- An aspect usually contains more than one sentence to explain the rationale. Therefore, we use paragraph as a unit.
- Paragraphs can be easily divided by html tag.
 - The boundary tag of paragraphs are <div>, <h>, <p> and consecutive
.
- Usually a paragraph contains less than 10 sentences.

Task 2: Reason Extraction

- Question:
 - Is there unsupervised approach to extract paragraphs with the reasons?
- Observation:
 - The words that people use to rationale for a specific query Q have some relationship with Q .
- Define topic related words (noun, verb, adj) by
 - Pointwise mutual information (PMI)
 - Logarithm of the odds ratio (LODR)
- The higher of the scores, the more relevant the words are to the topic

Find topic related words

■ **PMI score** $PMI(t, Q) = \log\left(\frac{P(t, Q)}{P(t)P(Q)}\right) = \log\left(\frac{p}{p(t)}\right)$

- $P(t, Q)$: probability of the word t co-occur with any topic words in the same paragraph.
- $P(t)$: probability of the word t occurs in a paragraph
- $P(Q)$: probability of any topic words occur in a paragraph

■ **LODR score** $LODR(t) = \log\frac{p/(1-p)}{q/(1-q)} = \log\frac{p(1-q)}{q(1-p)}$

- $p = P(t|Q)$: probability of the word t occurs when any word of Q appears
- $q = P(t|\sim Q)$: probability of the word t occurs when none of the words in Q appears

Identify Paragraphs Containing Reasons

- Using either one of the measures as the weight, we calculate the strength of a paragraph to the topic by:

$$TS(P) = \sum_{t \in P} weight(t) * f(t | P)$$

- $f(t|P)$: frequency of the word t in paragraph P
- If the weight is greater than a threshold, we consider the paragraph contains reason.

Task 3: Sentiment classification

- Using Turney's internet-based method
 - Pre-compute for each word in sentiment dictionary, General Inquirer (GI)
- For each extracted reason, we sum up scores of the sentiment words in the reason:
 - If it is greater than θ , the reason is positive, otherwise, it's negative.

Reason Clustering

- Clustering tool:
 - Frequent itemset hierarchical clustering tool, FIHC
- Each paragraph is represented by bag of words.
 - 1. all words
 - 2. $\frac{3}{4}$ LODR
 - 3. $\frac{3}{4}$ PMI

Experiments

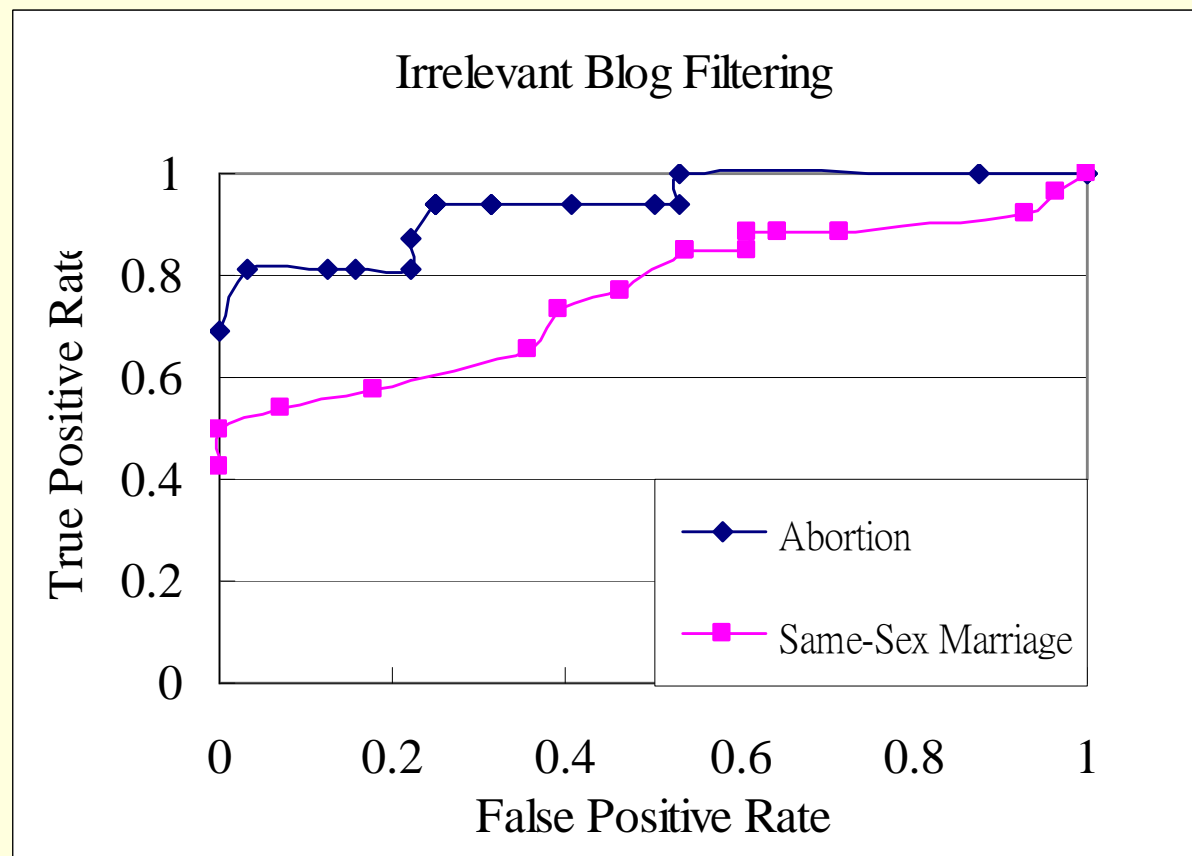
- Two topics: “Abortion” and “Same-sex Marriage”
 - Use only the first 100 relevant blog articles and manually label the standard answers for each of the four tasks
- Binary classification evaluation
 - Irrelevant blog filtering: ROC curve
 - Reason extraction: Precision-Recall curve
 - Sentiment classification: ROC curve
- Clustering evaluation
 - Consistency
 - User view

Evaluation method

- Binary classification evaluation
 - Irrelevant blog filtering: ROC curve
 - Reason extraction: Precision-Recall curve
 - Sentiment classification: ROC curve
- Clustering evaluation
 - Consistency
 - User view

Experiments on Irrelevant Blog Filtering

- ROC Curve (best F1=0.87 and 0.70)

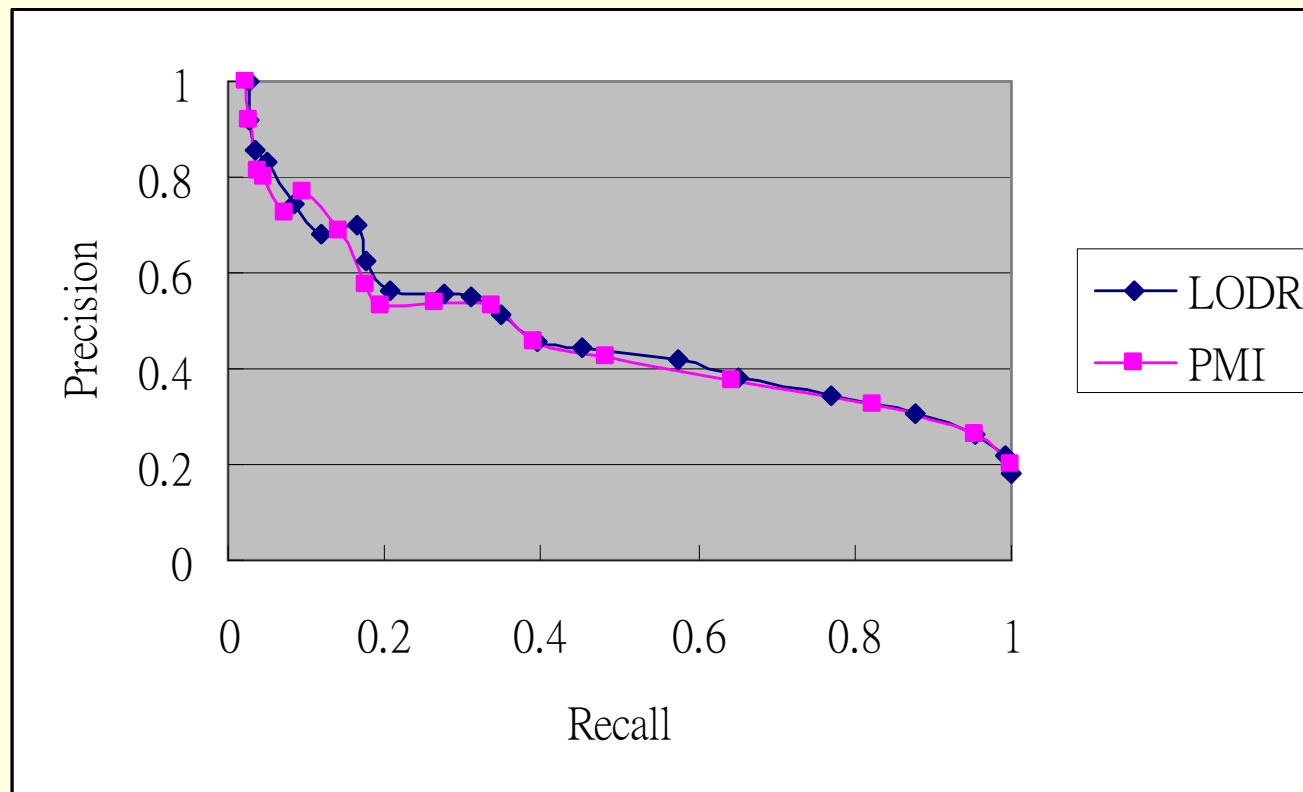


Experiment on Reason Extraction

- We choose the best F1-measure of previous step as input.
 - Abortion: 694 paragraphs and only 88 paragraphs contain reasons.
 - Same-Sex Marriage: 701 paragraphs after filtering and only 86 paragraphs contain reasons.
- A threshold is used to determine whether it is a paragraph containing a reason

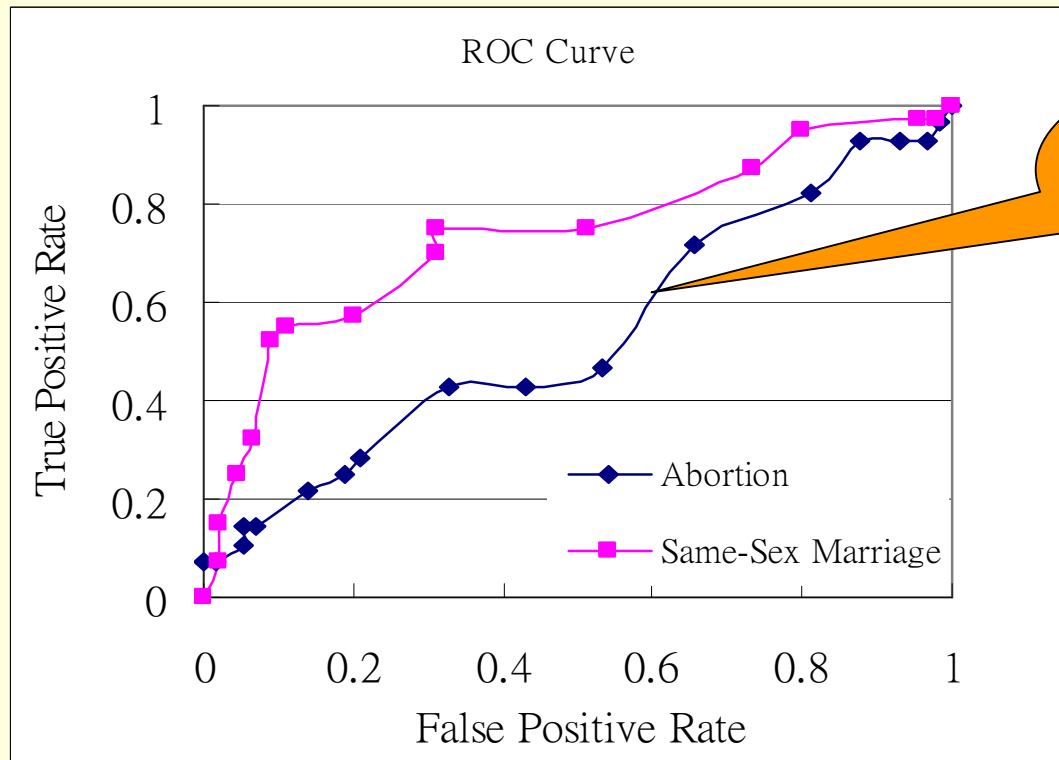
Performance on Reason Extraction

- Precision-Recall Curve (best F1=51.2% and 47.9%)



Performance on Sentiment Classification

- Test only on paragraphs containing reasons



More negative words used

Average Consistency

| Conformity | | All words | Top 3/4 words |
|----------------------|----------|-----------|---------------|
| Abortion | Positive | 0.559 | 0.655 |
| | Negative | 0.597 | 0.547 |
| Same-sex Marriage | Positive | 0.721 | 0.661 |
| | Negative | 0.619 | 0.513 |
| Average | | 0.624 | 0.594 |

Conclusion 1

- Aspect summarization can help us understand how people feel about some specific topic, especially on social events.
- We proposed a simple but powerful unsupervised approach to extract aspects from blogs.
- Profile analysis can be helpful in many circumstances.

Conclusion 2

- Research topics
 - Subjectivity filtering (Blog opinion retrieval)
 - Sentiment classification (Polarity orientation)
 - Product feature (opinion target) extractions
 - Opinion holder (source) extraction
 - Aspect analysis (Latent topic analysis)
 - Visualization of sentiments
 - Rationale/Reason analysis
 - Blog, bboards, analysis
 - ...