Tutorial on

Sentiment Analysis

Presented by

Dr. Pushpak Bhattacharyya

Balamurali A R

Aditya Joshi

IIT Bombay

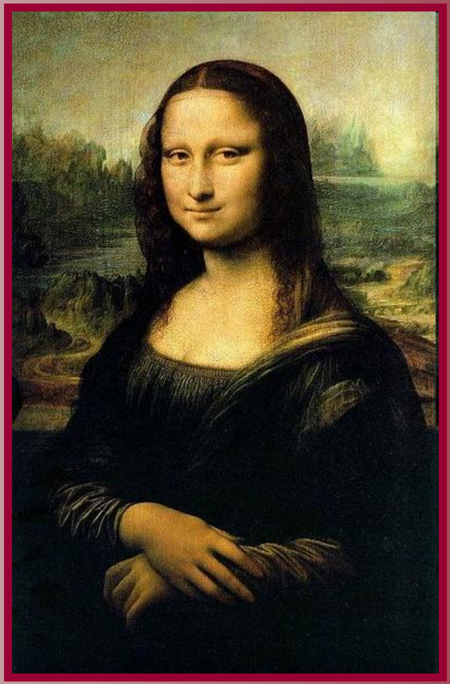
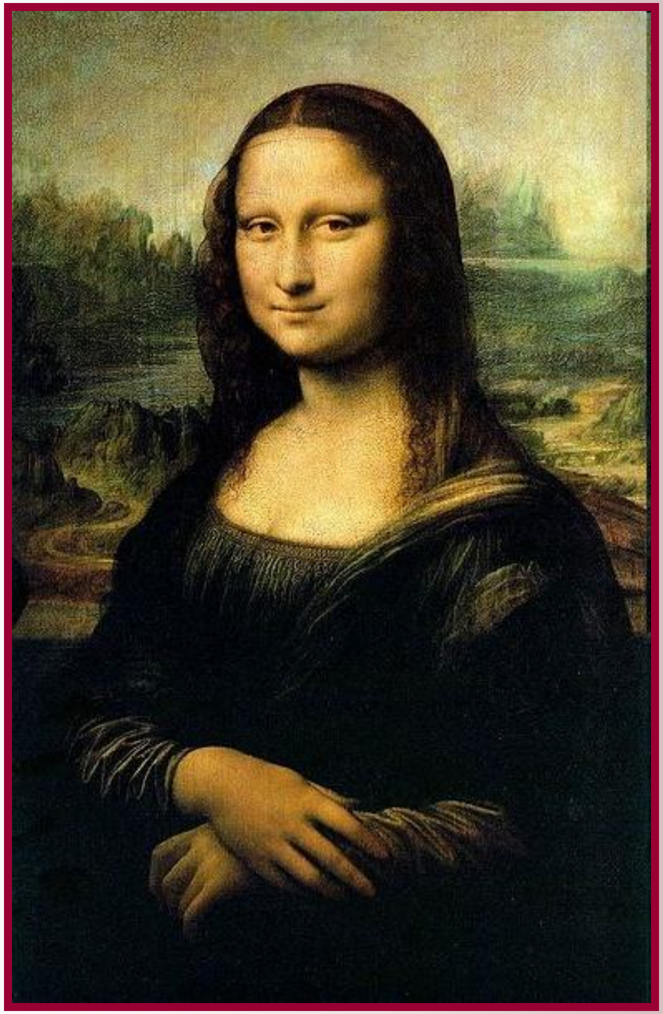


Image from wikimedia commons
Source: Wikipedia



Mona Lisa
16^{th} century
Artist: Leonardo da Vinci

The smile of Mona Lisa

Is she smiling at all?

Is she happy?

What is she smiling about?

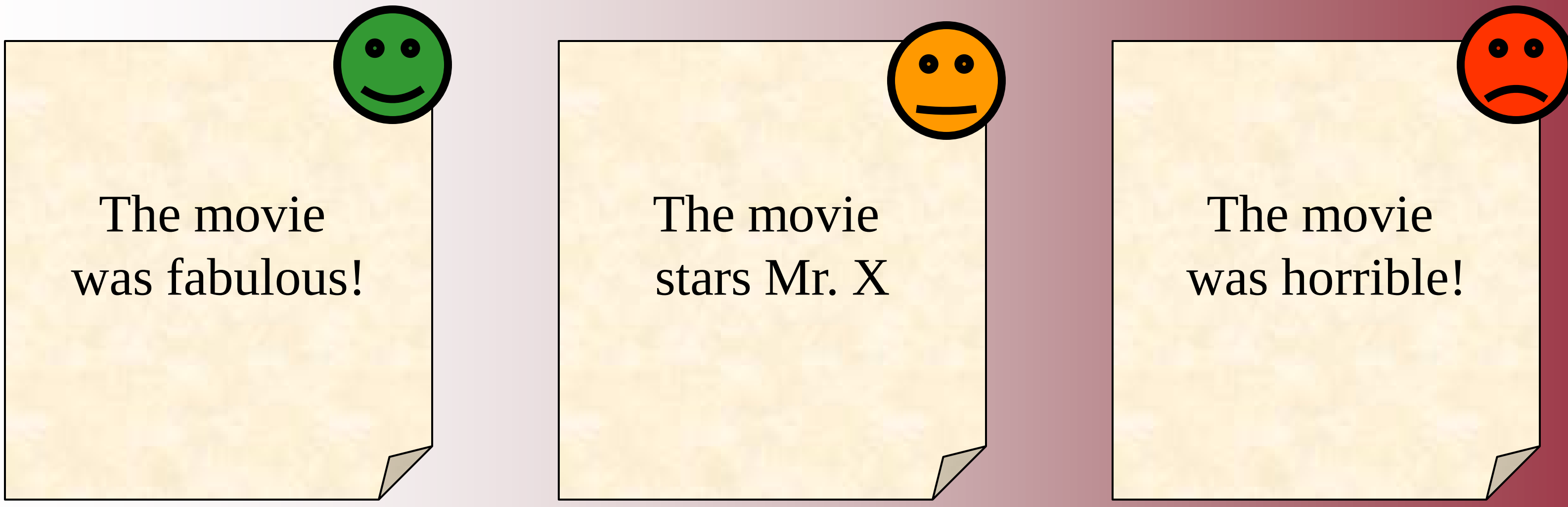
What is she happy about?

What is Sentiment analysis (SA)?

- Given a textual portion,
  - Is the writer expressing sentiment with respect to a topic?
  - What is that sentiment?

What is Sentiment analysis (SA)?

- Identify the orientation of opinion in a piece of text



- Can be generalized to a wider set of emotions

Motivation

- Knowing sentiment is a very natural ability of a human being.

  Can a machine be trained to do it?

- Aims to predict sentiment of a document / phrase / sentence.

  Trivial?

I like this book because it is good.

Reference : [Pang-Lee et al,2008]

Challenges

- Contrasts with standard text-based categorization
- Domain dependent
- Sarcasm
- Thwarted expressions

the sentences/words that contradict the overall sentiment of the set are in majority

Example: *The actors are good, the music is brilliant and appealing. Yet, the movie fails to strike a chord.*

Road map

Motivation & Introduction	Special sentences
Background	Advanced topics
Preliminaries	
Product-related SA	

Road map

Motivation & Introduction

- Perspectivizing SA
- Opinion on the web

Background

Preliminaries

Product-related SA

Special sentences

Advanced topics

‘Perspectiv’izing Sentiment Analysis

Reference : [Riloff et al,2005]

SA & Information extraction

- Goal? To extract facts related to a particular topic from a domain

Topic : ‘Explosion’ in news reports

- Can sentiment nature be used for better IE?

The Minister was outraged by the explosion near the market.

Reference : [Riloff et al,2005]

SA & Information extraction

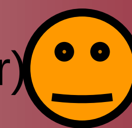
- Extract ‘indicator patterns’ – definitely non-sentiment.
- Retain them for IE

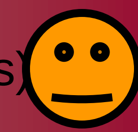
- Improvement by 3% in a terrorism-related data set

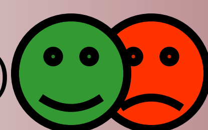
Reference : [Wiebe et al,2006]

SA & Word Sense Disambiguation

Sentiment can be associated with word senses

boil (come to the boiling point and change from a liquid to vapor) 

boil (immerse or be immersed in a boiling liquid, often for cooking purposes) 

boil (be in an agitated emotional state) 

Reference : [Wiebe et al,2006]

SA & Word Sense Disambiguation

- Sentiment-bearing senses more likely in sentiment-bearing sentences

- Sentence sentiment helpful to disambiguate words with sentiment as well as non-sentiment senses

The water is boiling, take it off the stove.

Web has emotions!

- Does web really contain sentiment-related information?
- Where?
- How much?
- What?
  - “Rise of the Web 2.0”
  - a. k. a. “User-generated content on the web”
  - a. k. a. “ Web has emotions”

User-generated content

- Web 2.0 empowers the user of the internet
- They are most likely to express their opinion there
- Temporal nature of UGC: ‘Live Web’
- **Can SA tap it?**

Where?

- Blogs
- Review websites
- Social networks
- User conversations

Conversations between users on one of the above

Reference : www.technorati.com/state-of-the-blogosphere/

How much?

- Size of blogosphere
  - Through the ‘eyes’ of the blog trackers

- Technorati : 112.8 million blogs (excluding 72.82 million blogs in Chinese as counted by a corresponding Chinese Center)
- A blog crawler could extract 88 million blog URLs from blogger.com alone
- 12,000 new weblogs daily

How much opinion?

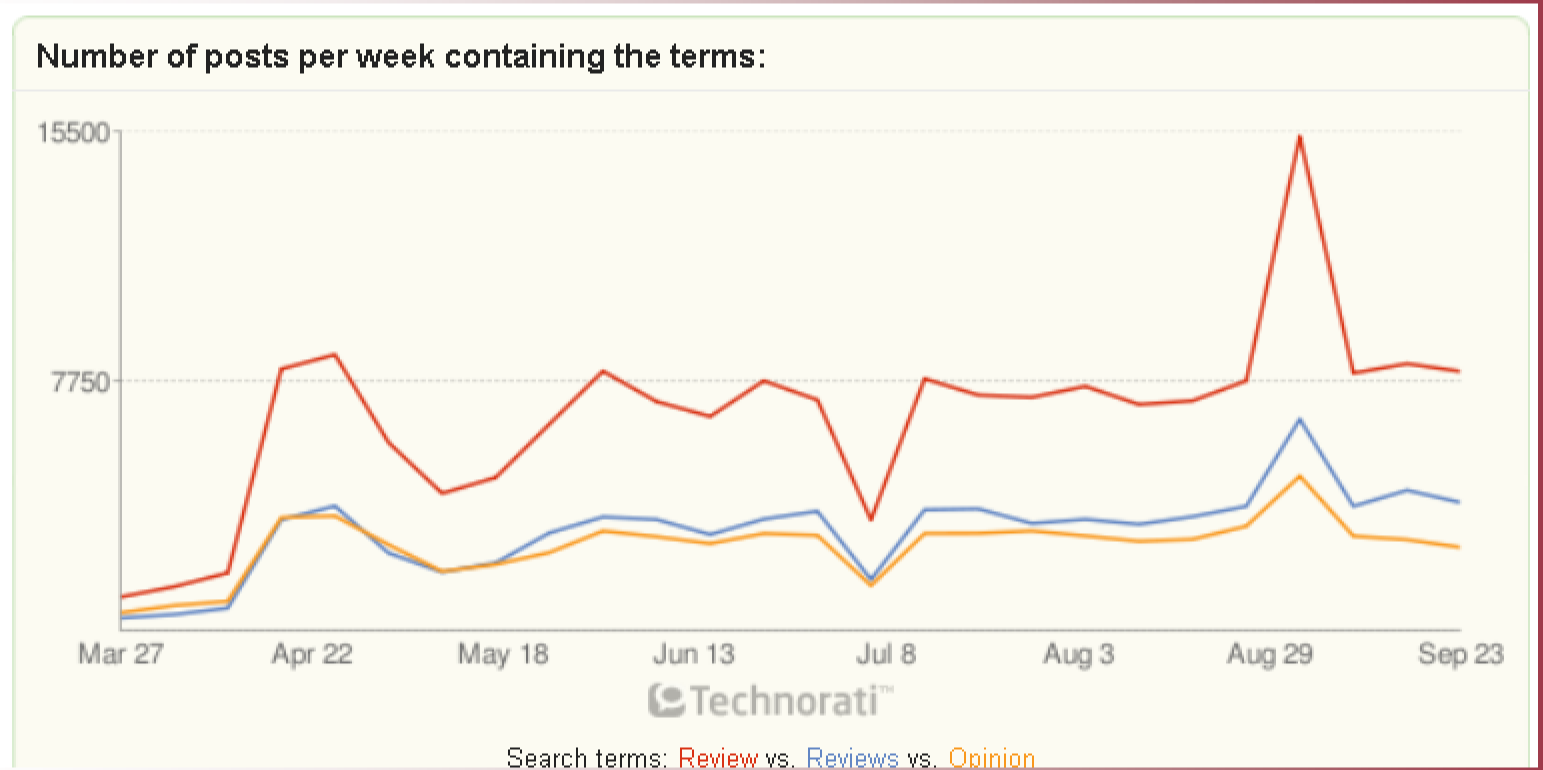


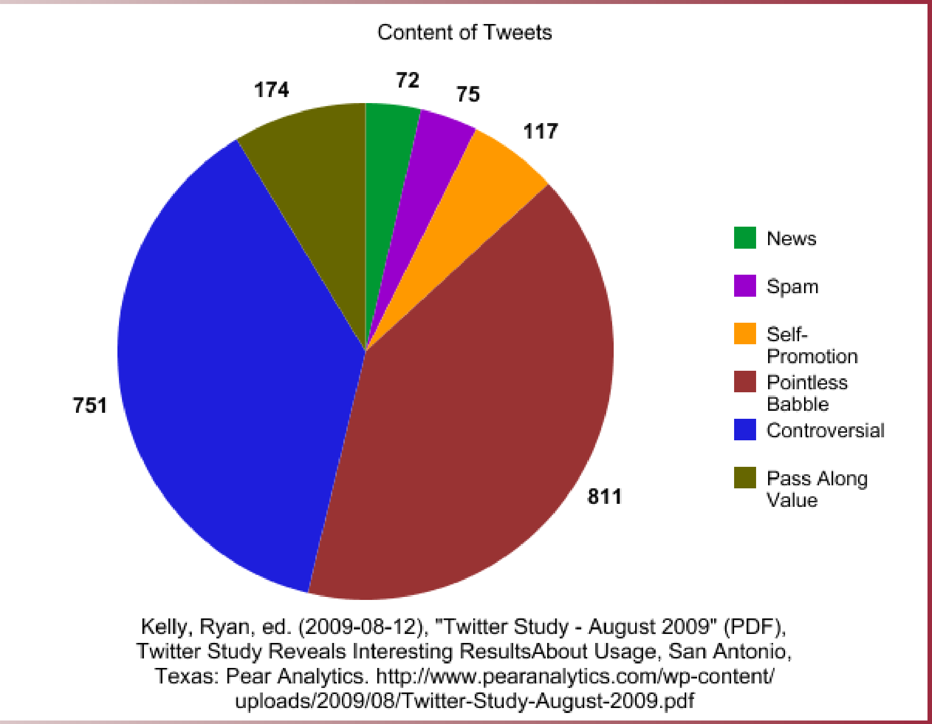
Chart created using : www.technorati.com/chart/

Reference : http://www.ebizmba.com/articles/social-networking-websites

How much?

- 12,22,20,617 unique visitors to facebook in December 2009
- Twitter:

2,35,79,044



What? Reviews

- www.burrrp.com
- www.mouthshut.com
- www.justdial.com
- www.yelp.com
- www.zagat.com
- www.bollywoodhungama.com
- www.indya.com

Restaurant reviews (now, for a variety of 'lifestyle' products/services)

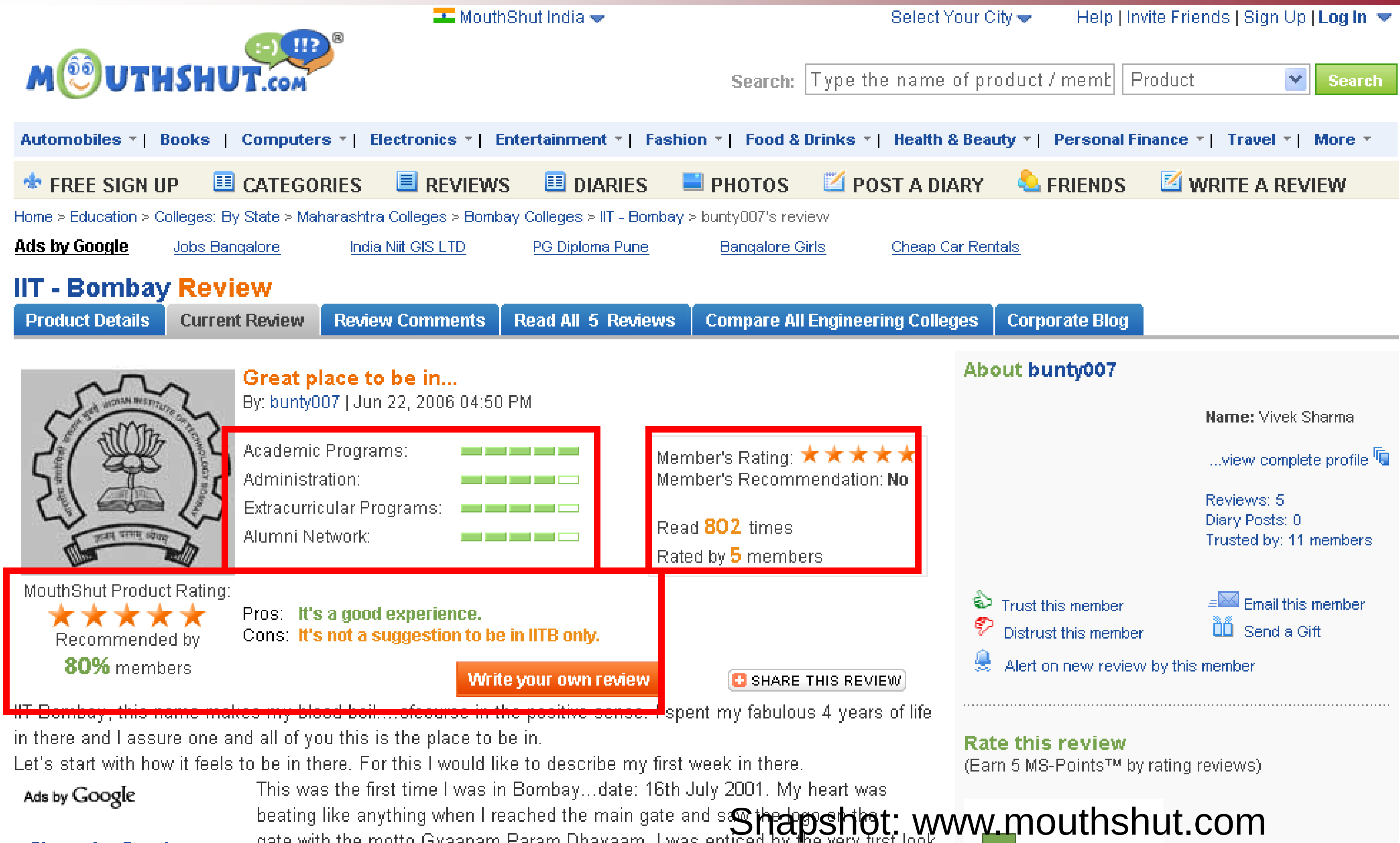
A wide variety of reviews

Professionals: Well-formed

User: More mistakes

Movie reviews by professional critics, users. Links to external reviews also present

A typical Review website



Snapshot: www.mouthshut.com

Sample Review 1

(This, that and this)

FLY E300 is a good mobile which i p . Since this Brand is not familiar in Market a und that E300 was cheap with almost all the and with the same set of features would com one is only 9k.

Touch Screen, good resolution, good talk time, 3.2M and so on...

BUT BEWARE THAT THE CAMERA IS NOT THAT 3.2 MEGA PIXEL, ITS NOT AS GOOD AS MY PREVIOUS MOBILE SONY ERICSSION K750i which is just 2Mega Pixel.

Sony ericsson was excellent with the feature of ca Camera, please excuse. This model of FLY is not regard..

Audio is not bad, infact better than Sony Ericsson

FLY is not user friendly probably since we have just started to use this Brand.

'Touch screen' today signifies a positive feature.
Will it be the same in the future?

Comparing old products

The confused conclusion

From: www.mouthshut.com

Sample Review 2

Hi,

I have Haier phone.. It was good when i was buing this phone.. But I invented A lot of bad features by this phone those are It's cost is low but Software is not good and Battery is very bad..,,Ther are no signals at out side of the city..,, People can't understand this type of software..,, There aren't fe gn is better not good..,, Sound a this side.They are giving heare re giving more talktime and validity these are also good.They are giving colour screen at display time it is also good because other phones aren't this type of feature.It is also low wait.

Lack of punctuation marks, Grammatical errors

Wait.. err.. Come again

From: www.mouthshut.com

Sample Review 3

(Subject-centric or not?)

I have this personal experience of using this cell phone. I bought it one and half years back. It had modern features that a normal cell phone has, and the look is excellent. I was very impressed by the design. I bought it for Rs. 8000. It was a gift for someone. It worked fine for first one month, and then started the series of multiple faults it has. First the speaker didnt work. I took it to the service centre (which is like a govt. office with no work). It took 15 days to repair the handset, moreover they charged me Rs. 500. Then after 15 days again the mike didnt work, then again same set of time was consumed for the repairs and it continued. Later the camera didnt work, the speakes were rubbish, it used to hang. It started restarting automatically. And the govt. office had staff which I doubt have any knoledge of cell phones??

These multiple faults continued for as long as one year, when the warranty period ended. In this period of time I spent a considerable amount on the petrol, a lot of time (as the service centre is a govt. office). And at last the phone is still working, but now it works as a paper weight. The company who produces such items must be sacked. I understand that it might be fault with one prticular handset, but the company itself never bothered for replacement and I have never seen such miserable cust service. For a comman man like me, Rs. 8000 is a big amount. And I spent almost the same amount to get it work, if any has a good suggestion and can gude me how to sue such companies, please guide.

For this the quality team is faulty, the cust service is really miserable and the worst condition of any organisation I have ever seen is with the service centre for Fly and Sony Erricson, (it’s near Sancheti hospital, Pune). I dont have any thing else to say.

From: www.mouthshut.com

Sample Review 4

(Good old sarcasm)

"I've seen movies where there was practically no plot besides explosion, explosion, catchphrase, explosion. I've even seen a movie where nothing happens. But *White on Rice* was new on me: a collection of really wonderful and appealing characters doing completely baffling and uncharacteristic things."

Review from: www.pajiba.com

What? Social networks

- Expressing opinion an important element

1. Comments (on photographs, status msgs.)
2. Status messages / tweets

   ‘Pritesh Patel loved the pasta he had at Pizza hut today’

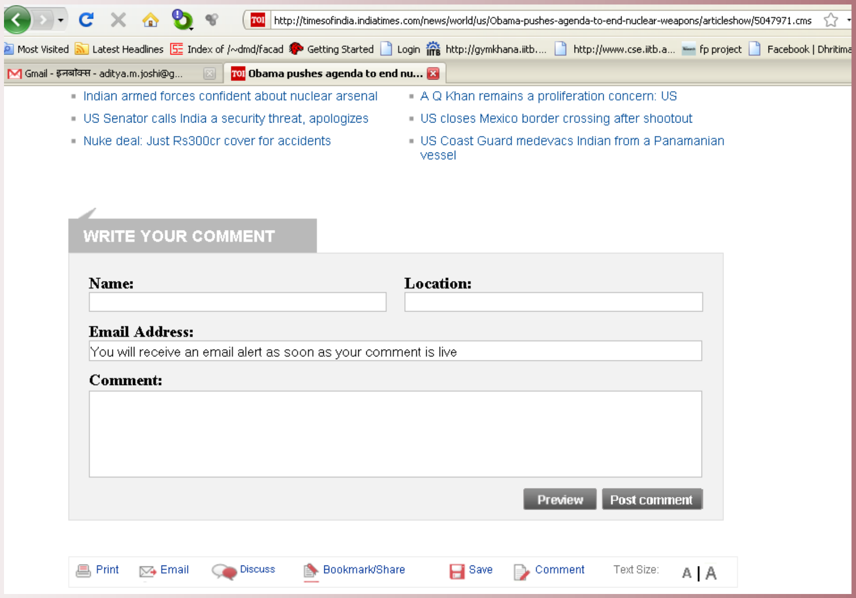
3. ‘Become a fan’ on facebook

   ‘Nokia E51. Become a fan’.

   ‘4 of your friends are a fan of Ganpati. Become a fan’.

What? Comments

- In what form does opinion exist on the web?
- Comments everywhere



From: www.timesofindia.com

What? Comments

- Two types of comments:
  - Comments about the article/ blogpost:
    - *Very well-written indeed…*
  - Comments about the topic of the article:
    - *I agree with you.. I used to love **'s movies at a point of time but these days all he comes out with is trash. <Often leads to a conversation>*

(- Comments about the blogger:

- *If you think Shahid Kapoor is ugly, go buy glasses. While you are at it, buy yourself a brain too*

)

Road map

Motivation & Introduction	Special sentences
• Perspectivizing SA • Opinion on the web	
Background	**Advanced topics**
• Terminology • Classifiers	
Preliminaries	
Product-related SA	

Terminology

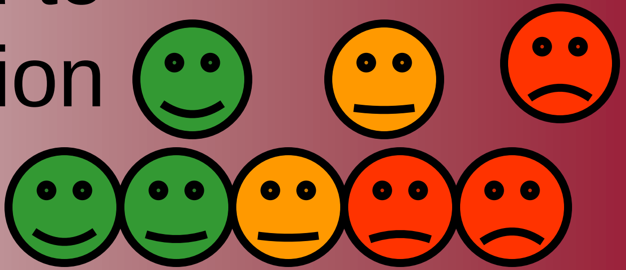
- The road till now…
  - What is SA?
  - How is it related to other fields?
  - Do we have enough data to work on?

- Delving into the details of SA
  - Starting with the basics…

Reference : http://www.colour-journal.org/2007/1/2/07102article.htm

Sentiment Analysis, Emotion Analysis

- Sentiment Analysis: Limited to positive/negative classification



- Emotion Analysis: Works with a wider range of emotions.
  - 6 basic emotions: anger, surprise, disgust, sadness, happiness and fear

Subjectivity

- Subjectivity: Bearing opinion content

Positive / negative/neutral/both

- Objectivity: Without opinion content

"I feel both happy and sad about it.
Happy because….. Sad because…. "

Reference : http://www.cs.pitt.edu/mpqa/databaserelease/Database.2.0.README

Annotating a sentiment corpus

- Simple:
  - Sentiment value to a word
  - Sentiment value to a sentence / document

- Nested: (used in MPQA corpus)
  - Representation using a private state

boil *(reach boiling point) : Objective*

Reference : http://www.cs.pitt.edu/mpqa/databaserelease/Database.2.0.README

Private state

- “A state that is not open to objective observation”
  - Opinion, observation
  - Speculations, beliefs
- Also have an intuitive intensity

“The US fears a spill-over”, said Xirao-Nima.

Reference : http://www.cs.pitt.edu/mpqa/databaserelease/Database.2.0.README

Description

- Source:
  - Who expressed?
  - Source could be nested. **Xirao-Nima -> US**
- Span
  - Span of text that represents the private state
- Intensity

“The US fears a spill-over”, said Xirao-Nima.

Classifiers for SA

Classification task

- Input: Document, sentence, phrase, word
- Categorical output among: Positive, negative, neutral

.. granularity may be different in some cases

Naïve Bayes classifiers

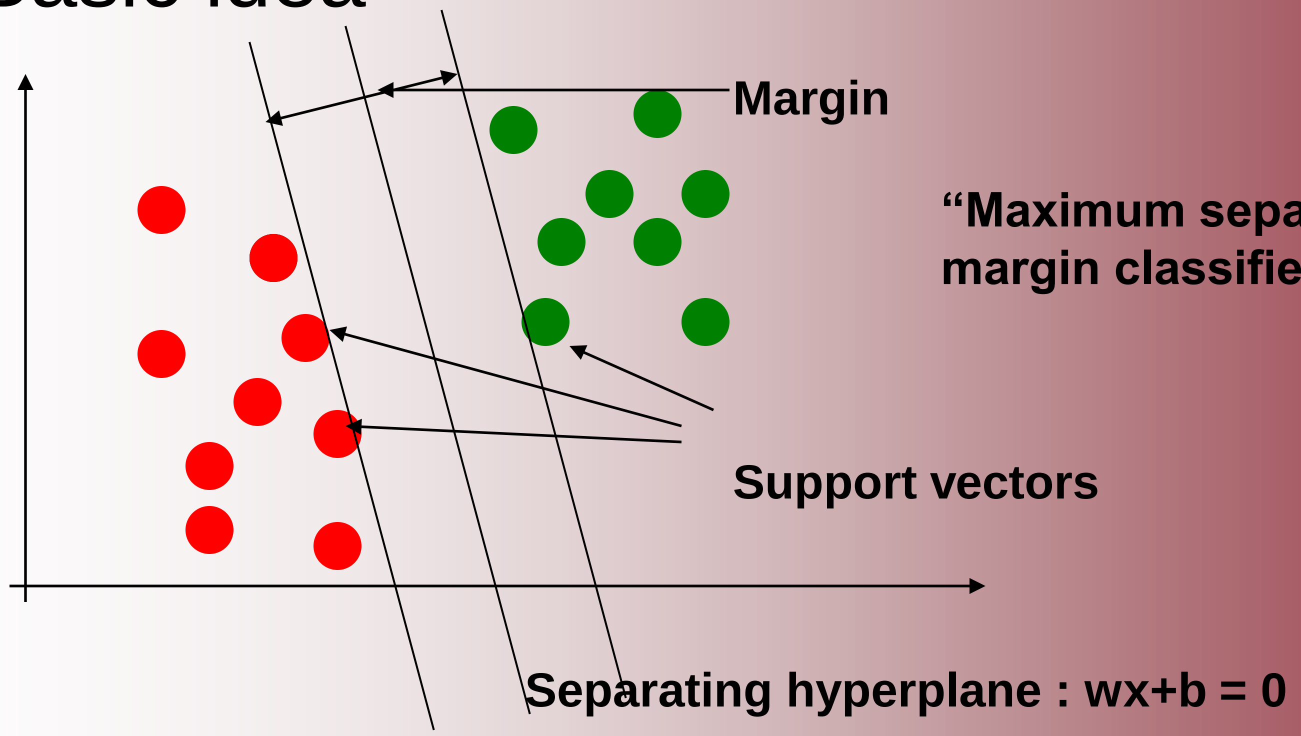
- Based on Bayes rule
- Naïve Bayes : Conditional independence assumption

$$P(C_i | X) = \frac{P(X | C_i) \cdot P(C_i)}{P(X)}$$

$$P(X | C_i) = \prod_{k=1}^{d} P(x_k | C_i)$$

Support vector machines

- Basic idea



“Maximum separating-margin classifier”

Multi-class SVM

- Multiple SVMs are trained:
  - True/false classifiers for each of the class labels
  - Pair-wise classifiers for the class labels

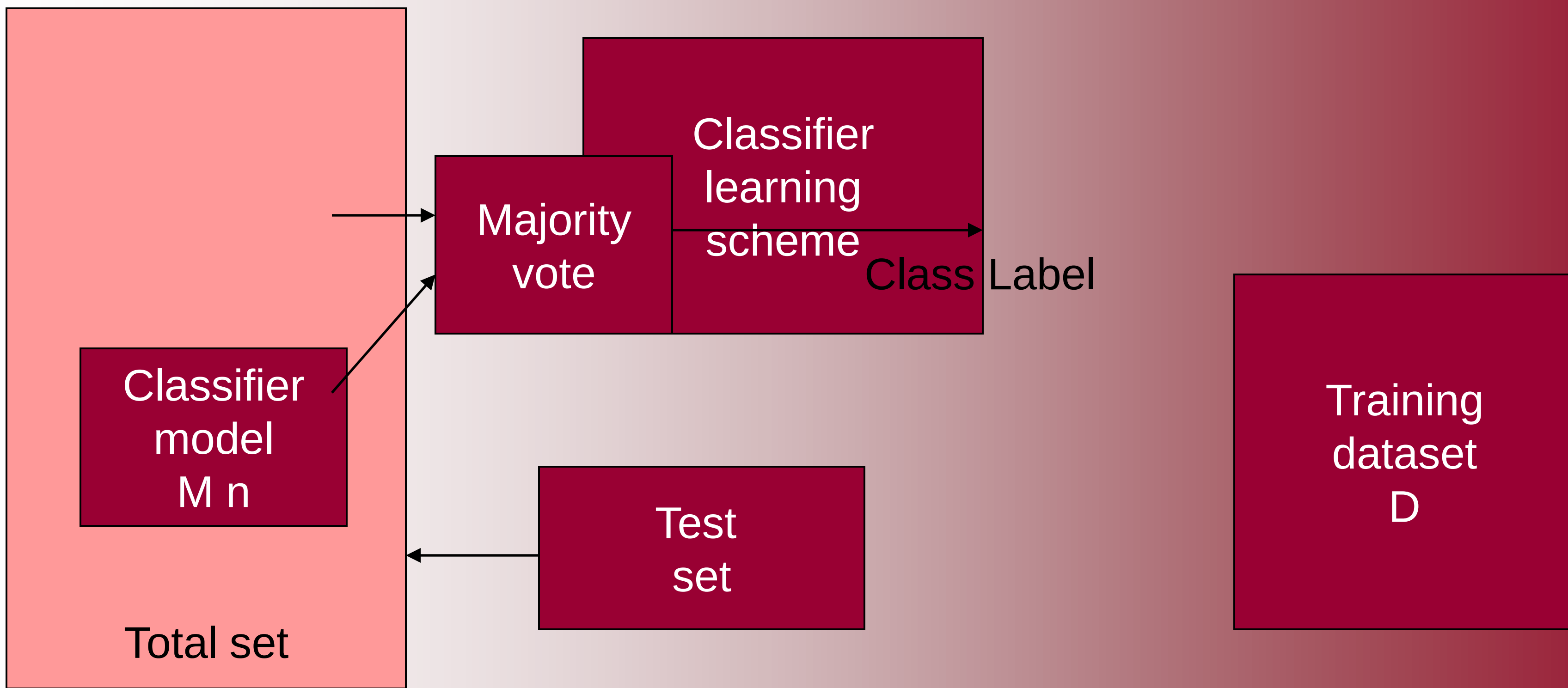
Reference : Scribe by Rahul Gupta, IIT Bombay

Combining Classifiers

- ‘Ensemble’ learning
- Use a combination of models for prediction
  - Bagging : Majority votes
  - Boosting : Attention to the ‘weak’ instances
- Goal : An improved combined model

Reference : Scribe by Rahul Gupta, IIT Bombay

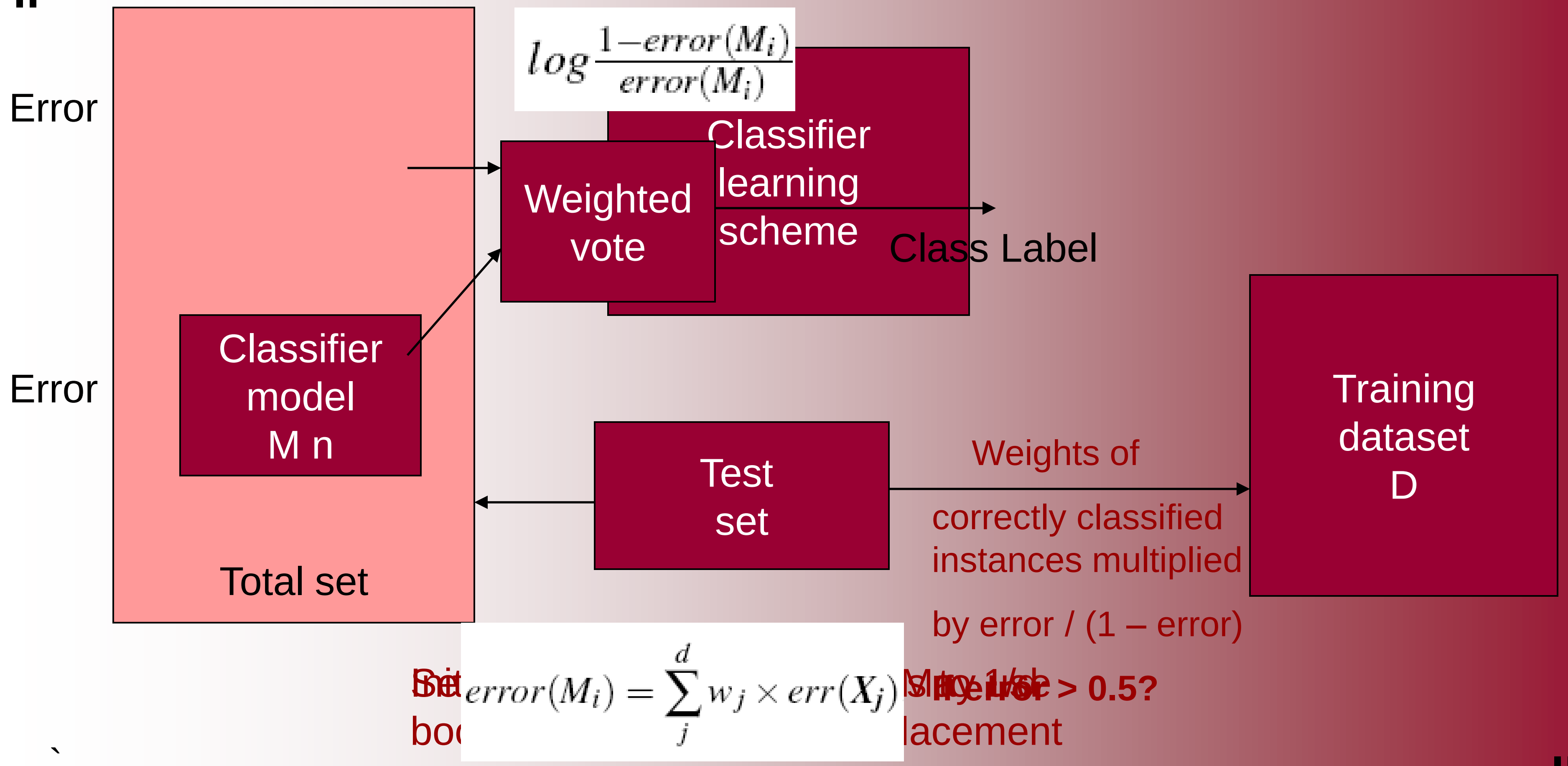
Bagging



Selected at random. May use bootstrap sampling with replacement

Reference : Scribe by Rahul Gupta, IIT Bombay

Boosting (AdaBoost)



Road map

Motivation & Introduction	Special sentences
• Perspectivizing SA • Opinion on the web	

Background

- Terminology
- Classifiers

Preliminaries

- Lexical resources
- Contextual polarity
- Subjectivity detection

Product-related SA

Advanced topics

Opinion lexical resources

I love my country

Need for resources

- Document level?
- Sentence level?
- What resources?

General Inquirer (GI)

(Stone, et al., 1966),

WordnetAffect

(Valitutti,et al., 2004),

SentiWordNet

(Esuli & Sebastiani, 2006)

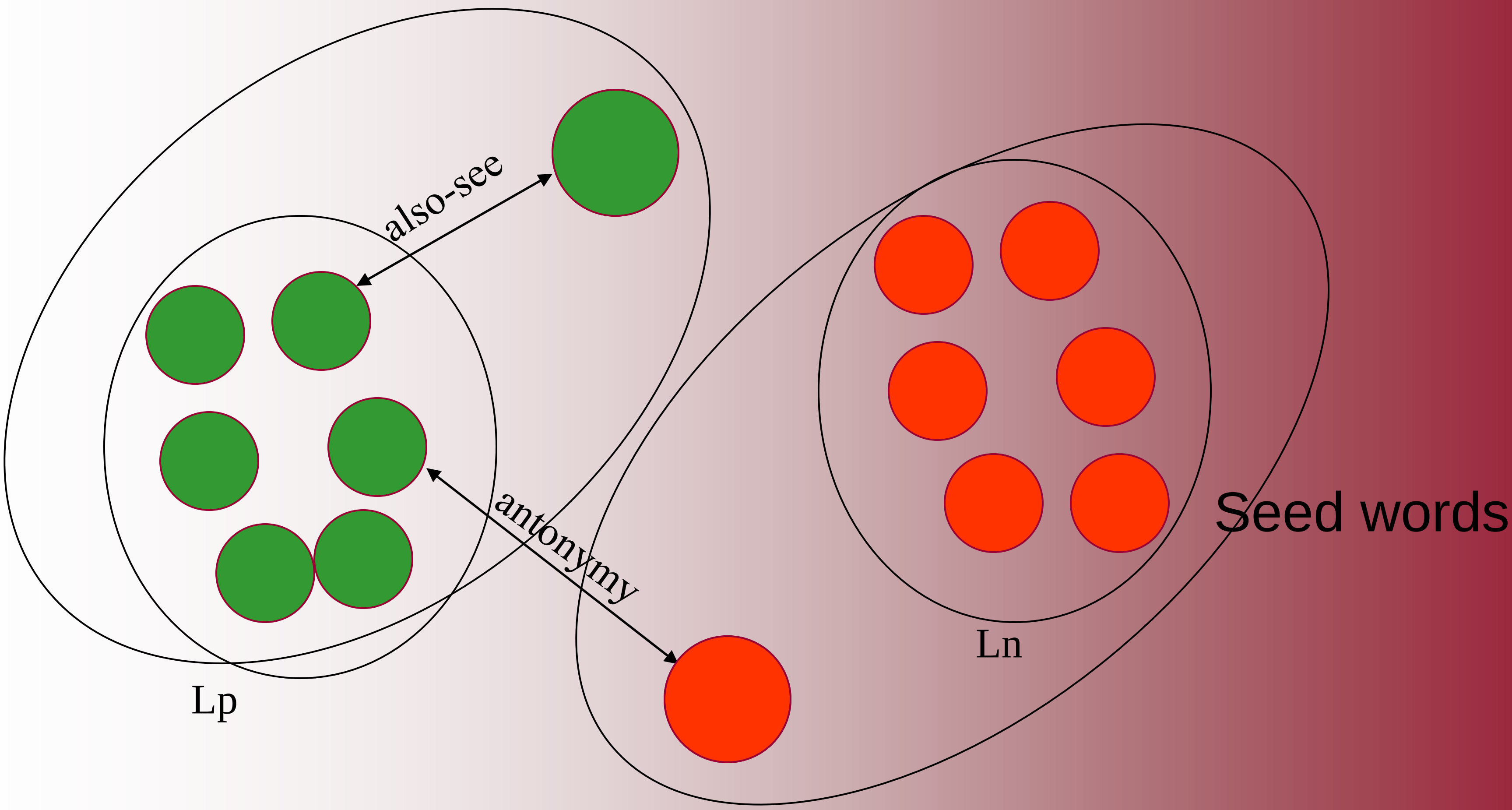
Reference : [Esuli et al,2006]

SentiWordnet

- WorldNet 2.0 marked with polarity based on gloss definition
- Three scores
- Interpreting scores

- Intensity of each category with resp. to sense
- Percentage usage in each category
- Uncertainty of annotator in labeling them

Seed-set expansion in SWN



The sets at the end of kth step are called Tr(k,p) and Tr(k,n)

Tr(k,o) is the set that is not present in Tr(k,p) and Tr(k,n)

Building SentiWordnet

- Classifier combination used: **Rocchio (BowPackage) & SVM(LibSVM)**
  - Different training data based on expansion
  - POS –NOPOS and NEG-NONEG classification
- Total eight classifiers
- Score Normalization

Scoring SentiWordnet

- Maximum of triple score (for labeling)
- Difference of polarity score (for semantic orientation)

pestering
P = 0,
N = 0.625,
O = 0.375

Diff(P,N) = - 0.625
Negative

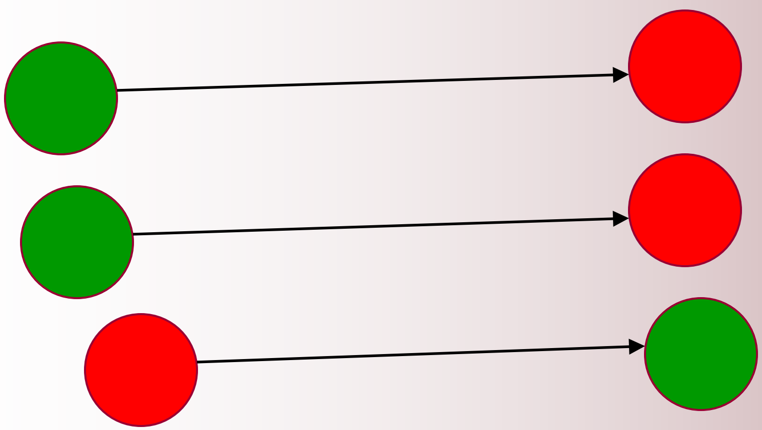
Reference : [Saif et al,2009]

Another lexicon-MSOL

- A highly scalable resource –
  - Process applicable to all existing lexical resources
  - Not just to WordNet alone
- Can include multiword expressions
- No manual annotation needed

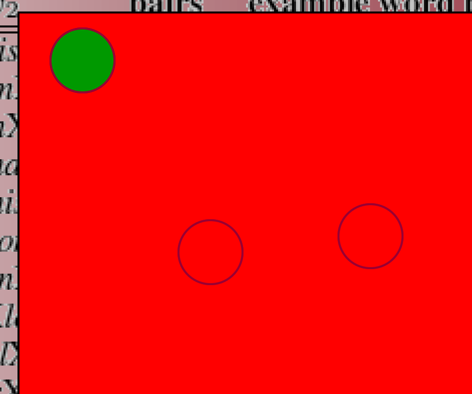
“A bit of all right”

Building MSOL



Words with polarity
(seed words)

Affix pattern w_1	w_2	# word pairs	example word pair
X	dis		
X	im		ble
X	inX		sistent
X	ma		
X	mi		ne
X	no		
X	un		
X	Xl		
lX	illX		
rX	irX	48	responsible–irresponsible
Xless	Xful	51	harmless–harmful



Affix patterns from Macquarie dictionary
Paragraphs in Roget Thesaurus

Reference : [Saif et al,2009]

A snapshot

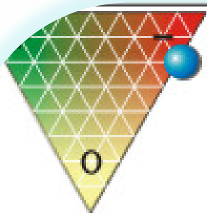
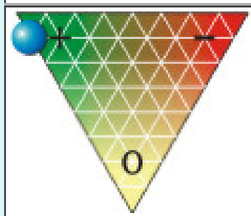
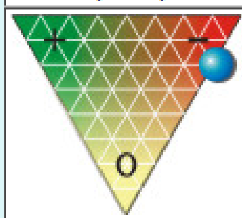
- MSOL (scaled with words from GI)
  - – Total words -76,400
  - – #Positives -30,458
  - – #Negatives 45,942

Snapshot of multiwords in MSOL

```
a_big_yawn negative
a_bit_hot positive
a_bit_much negative
a_bit_of_all_right positive
a_bit_of_fluff positive
a_bit_on_the_nose negative
a_bit_on_the_side negative
a_bit_rough negative
```

Reference : [Esuli et al,2006], [Saif et al, 2009], [Denecke et al,2009]

SA lexicon : What is missing

 P = 0, N = 0.75, O = 0.25	happy(2) pleased(3) *experiencing pleasure or joy; "happy you are here"; "pleased with the good news"*
 P = 0.875, N = 0, O = 0.125	happy(3) felicitous(2) *marked by good fortune; "a felicitous life"; "a happy outcome"*
 P = 0, N = 0.75, O = 0.25	happy(4) *satisfied; enjoying well-being and contentment; "felt content with her lot"; "quite happy to let things go on as they ar*
 …5, N = 0.5, O = 0.375	happy(5) euphoric(1) *exaggerated feeling of well-being or elation*

“Mi

Twist for ‘irreverent’?

Recognizing Contextual Polarity

“Millions follow Gandhi’s **irreverent** quest for truth.”

Contextual Polarity

- May be different from word’s prior polarity
- Many things to be considered in assessing CP.
- For example,
  - Local negation
  - negation of the proposition
  - negation of the subject

no one *thinks that it’s good*

Training data creation

- MPQA - Subjective expressions marked with contextual polarity (Weibi et al ,2005)
  - Positive tag
  - Negative tag
  - Both tags
  - Neutral tag

- Prior-Polarity Subjectivity Lexicon created
  - Expanded using GI word list
  - Tagged with prior polarity

Jerome says the hospital feels no different than a hospital in the states.

Algorithm

Complete text

Neutral –Polarity Classifier

BoosTexter AdaBoost.HM

T=5000

Reference : [Wilson et al,2005]

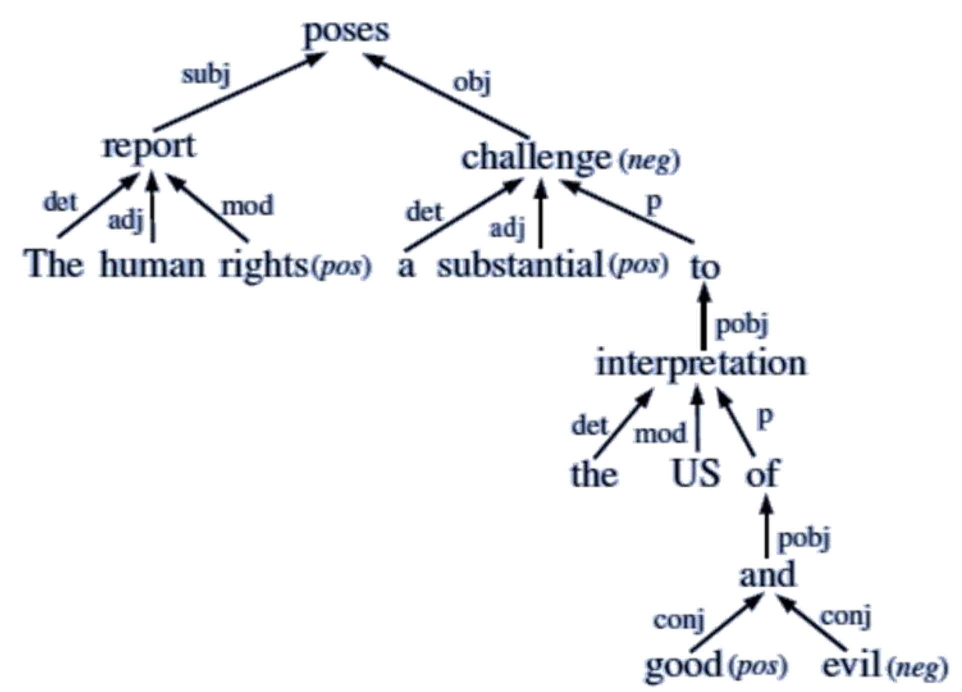
Features-NP classifier

Word Features	Sentence Features	Structure Features
word token	strongsubj clues in current sentence: count	in subject: binary
word part-of-speech	strongsubj clues in previous sentence: count	in copular: binary
word context	strongsubj clues in next sentence: count	in passive: binary
prior polarity: positive, negative, both, neutral	weaksubj clues in current sentence: count	
reliability class: strongsubj or weaksubj	weaksubj clues in previous sentence: count	
Modification Features	weaksubj clues in next sentence: count	**Document Feature**
preceeded by adjective: binary	adjectives in sentence: count	document topic
preceeded by adverb (other than not): binary	adverbs in sentence (other than not): count	
preceeded by intensifier: binary	cardinal number in sentence: binary	
is intensifier: binary	pronoun in sentence: binary	
modifies strongsubj: binary	modal in sentence (other than will): binary	
modifies weaksubj: binary		
modified by strongsubj: binary		
modified by weaksubj: binary		

Reference : [Wilson et al,2005]

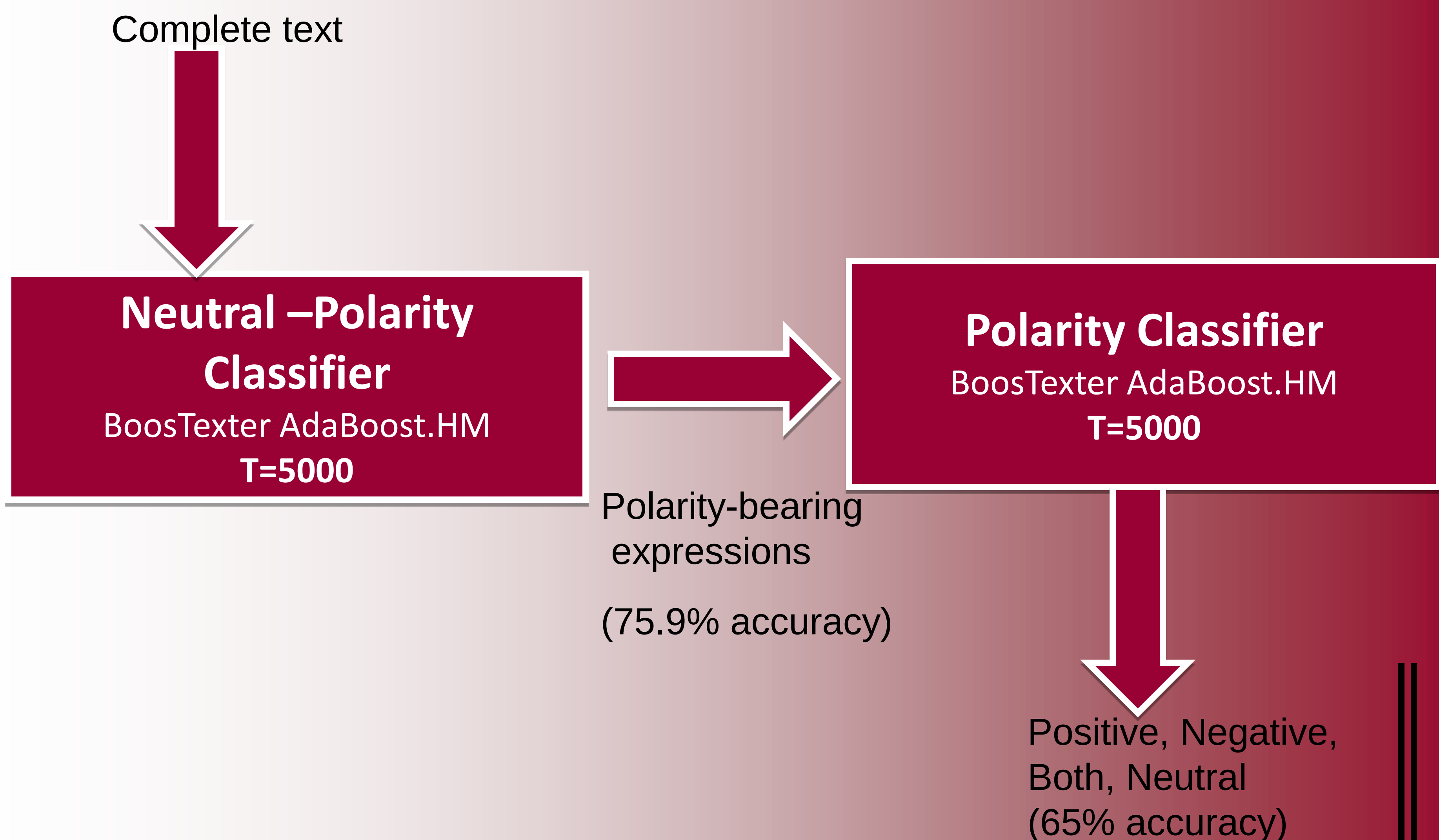
Parsing

.



The human rights report poses a substantial challenge to the US interpretation of good and evil.

Algorithm


Complete text
Neutral –Polarity Classifier
BoosTexter AdaBoost.HM
T=5000
Polarity-bearing expressions
(75.9% accuracy)
Polarity Classifier
BoosTexter AdaBoost.HM
T=5000
Positive, Negative, Both, Neutral
(65% accuracy)

Reference : [Wilson et al,2005]

Features : Polarity classifier

Word Features
word token
word prior polarity: positive, negative, both, neutral
Polarity Features
negated: binary
negated subject: binary
modifies polarity: positive, negative, neutral, both, notmod
modified by polarity: positive, negative, neutral, both, notmod
conj polarity: positive, negative, neutral, both, notmod
general polarity shifter: binary
negative polarity shifter: binary
positive polarity shifter: binary

Reference : [Wilson et al,2005]

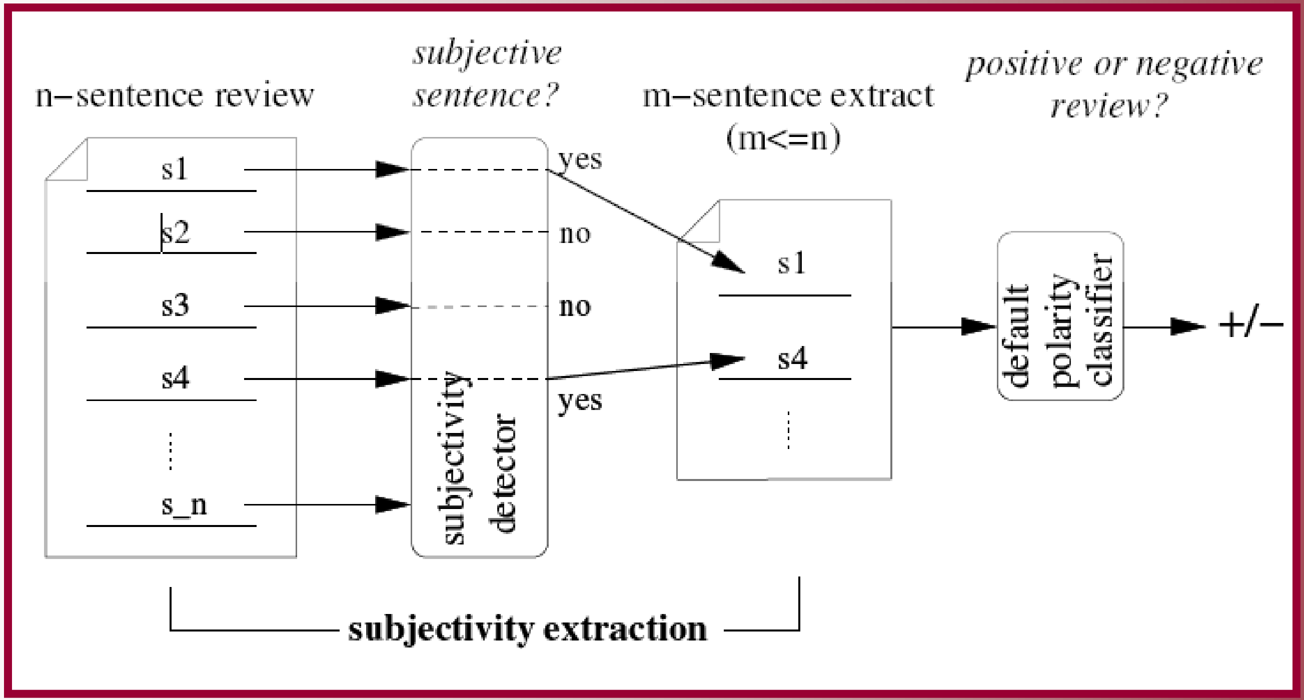
Issues

- Accuracy is low (65%)
- Computationally daunting
  - Creation of parse tree required
- Need of domain-specific training data
  - MPQA corpus is domain-specific
- *“Research to be continued”*

Subjectivity detection

Subjectivity detection

- **Aim**: To extract subjective portions of text
- **Algorithm used**: Minimum cut algorithm



Reference : [Pang-Lee,2004]

Constructing the graph

- Why graphs?
- Nodes and edges?
- Individual Scores
- Association scores

Prediction whether two sentences should have the same subjec $Pr^{NB}_{sub}(s_i)$ l

$$assoc(s_i, s_j) \stackrel{def}{=} \begin{cases} f(j-i) \cdot c & \text{if } (j-i) \leq T; \\ 0 & \text{otherwise.} \end{cases}$$

T : **Threshold** – maximum distance upto which sentences may be considered proximal

f: The **decaying** function

i, j : **Position** numbers

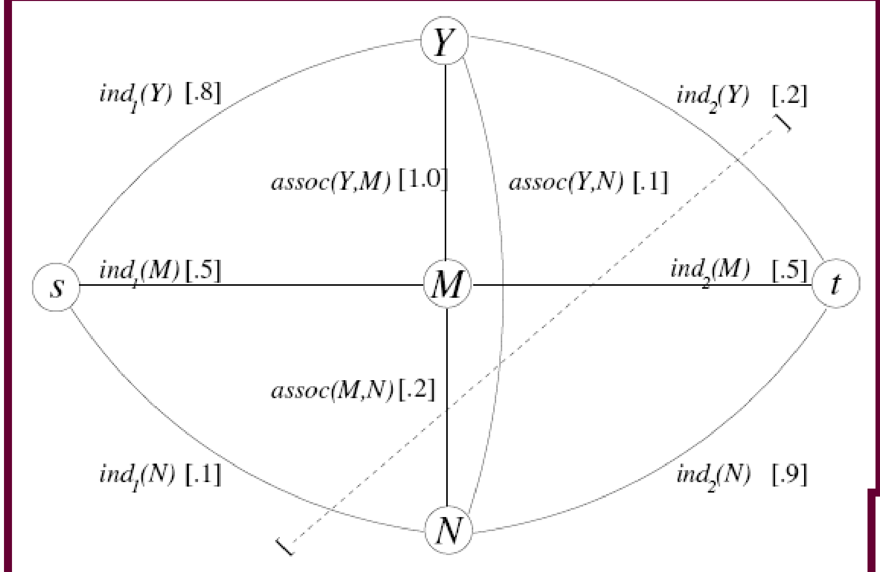
Reference : [Pang-Lee,2004]

Constructing the graph

- Build an undirected graph G with vertices $\{v1, v2\ldots,s, t\}$ (sentences and s, t)
- Add edges (s, v_i) each with weight $ind_1(x_i)$
- Add edges (t, v_i) each with weight $ind_2(x_i)$
- Add edges (v_i, v_k) with weight $assoc\ (v_i, v_k)$
- Partition $$\sum_{x \in C_1} ind_2(x) + \sum_{x \in C_2} ind_1(x) + \sum_{\substack{x_i \in C_1, \\ x_k \in C_2}} assoc(x_i, x_k).$$

Reference : [Pang-Lee,2004]

Example

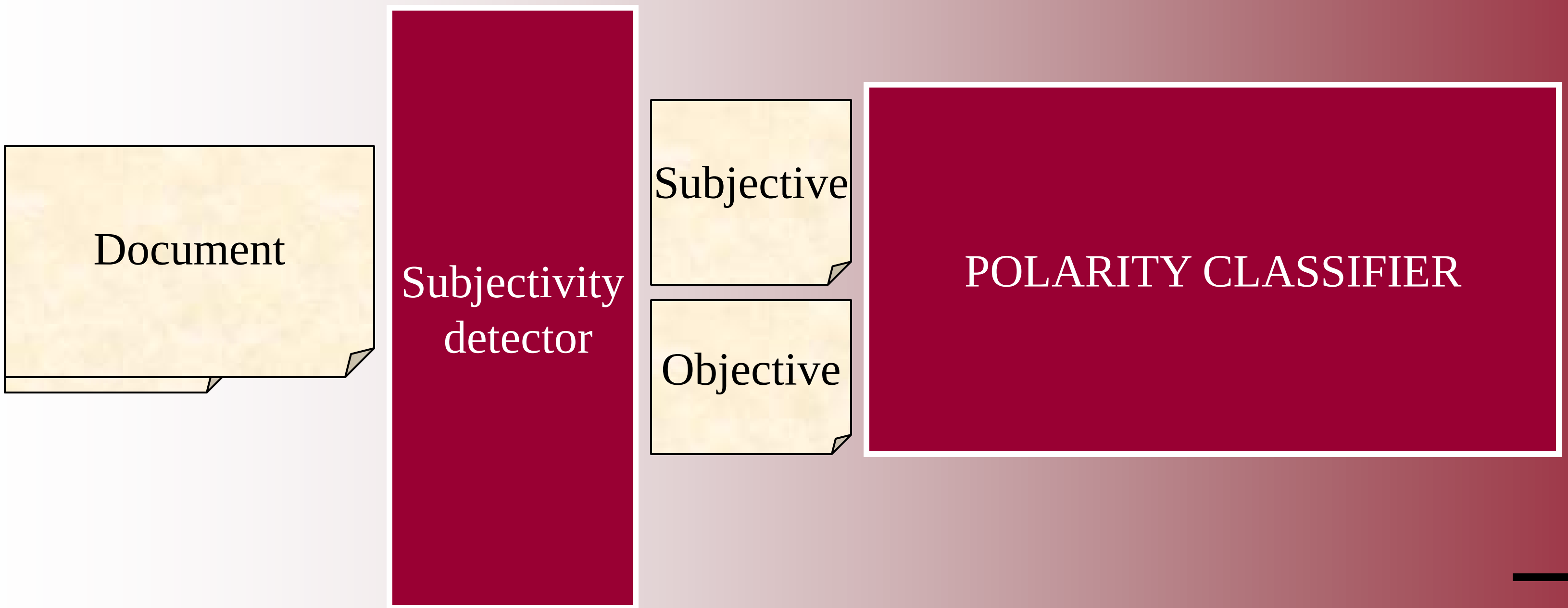


Sample cuts:

C_1	Individual penalties	Association penalties	Cost
{Y,M}	.2 + .5 + .1	.1 + .2	1.1
(none)	.8 + .5 + .1	0	1.4
{Y,M,N}	.2 + .5 + .9	0	1.6
{Y}	.2 + .5 + .1	1.0 + .1	1.9
{N}	.8 + .5 + .9	.1 + .2	2.5
{M}	.8 + .5 + .1	1.0 + .2	2.6
{Y,N}	.2 + .5 + .9	1.0 + .2	2.8
{M,N}	.8 + .5 + .9	1.0 + .1	3.3

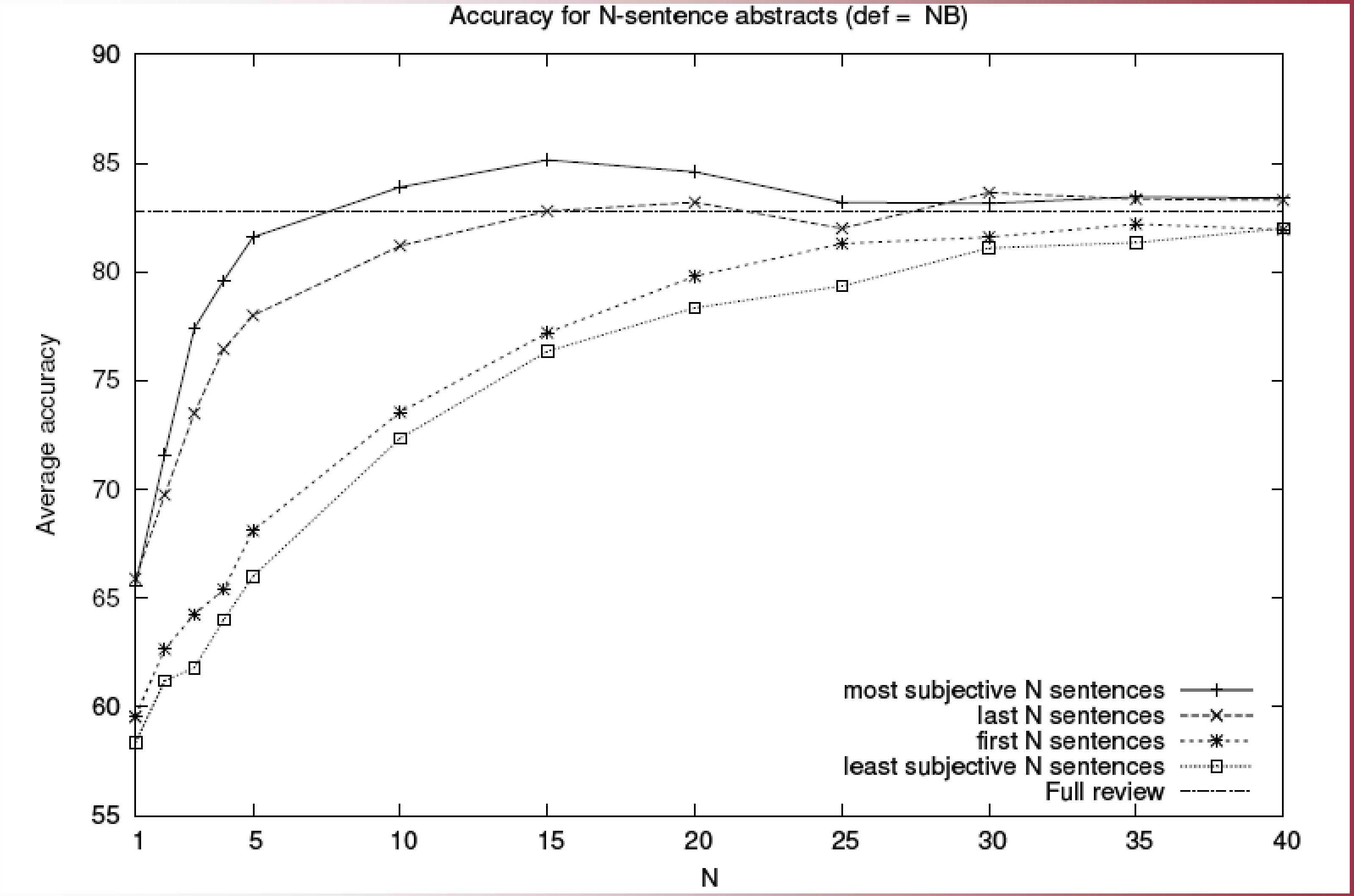
Results (1/2)

- Naïve Bayes, no extraction : 82.8%
- Naïve Bayes, subjective extraction : 86.4%
- Naïve Bayes, ‘flipped experiment’ : 71 %



Reference : [Pang-Lee,2004]

Results (2/2)



Road map

Motivation & Introduction

- Perspectivizing SA
- Opinion on the web

Background

- Terminology
- Classifiers

Preliminaries

- Lexical resources
- Contextual polarity
- Subjectivity detection

Product-related SA

- Product review domain
- Document-level SA
- Feature engineering
- Product feature-based SA

Special sentences

Advanced topics

Product review domain for SA

Reference : [Liu et al,2009]

Analyze this

I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought it. She also thought the phone was too expensive, and wanted me to return it to the shop.

Reference : [Liu et al,2009]

Terminology (1/3)

Object (*O*) : *(T, A)*

Entity

(person / event / product)

Components Attributes

O: Cell phone *T: Battery, Screen*

A: Size, weight, etc.

Reference : [Liu et al,2009]

Terminology (2/3)

- Explicit features:
  - Feature *f* or any synonym appearing
- Implicit features
  - Neither *f* nor any of its synonyms are explicitly mentioned
  - But *f* is implied

The camera is blurry.

Reference : [Liu et al,2009]

Terminology (3/3)

- Opinion
- Opinion Holder
- Opinion orientation

Orientation of an opinion on a feature f

<Microsoft> stated they were happy about the presales of windows 7.

Reference : [Liu et al,2009]

Product Domain Model

- Model of an object :

Object : $F = \{f_1, f_2 \ldots f_n\}$

Words $= \{w_{i1}, w_{i2} \ldots w_{in}\}$

Feature indicators $= \{i_{i1}, i_{i2} \ldots i_{in}\}$

- Model of an opinionated document
  – Document d with a set of objects $\{o_1, o_2, \ldots\}$
  – A set of opinion holders $\{h_1, h_2, .. h_p\}$
  – Opinion on each object O_j is expressed on a subset F_j of features of O_j

Reference : [Liu et al,2009]

Different Types of Opinion

- Direct Opinion
- Comparative Opinion

- A relation of similarities or differences
- Expressed through a comparative or superlative form of an adjective/adverb

Canon EXS rebel is better than Nikon DX0

And the objective is….

- Identify all synonyms and feature indictors
- Find orientation
- Create summary

Document-level sentiment analysis

What documents?

Includes but not limited to…

- Web pages: Blogs
- Transcripts of parliamentary proceedings
- Reviews of a variety of domains

Document-level SA

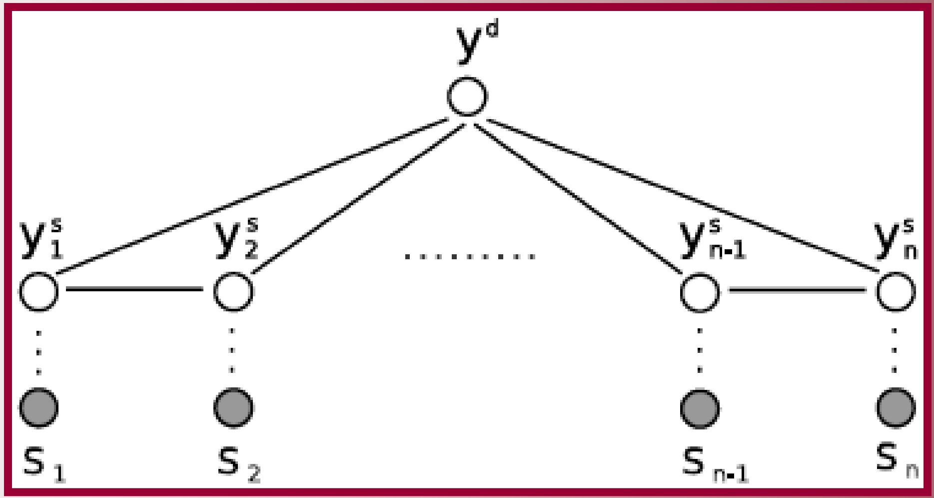
- Calculating overall sentiment of a document based on its contents (sentences)

- Can be useful in calculating an overall trend across documents

Reference : [McDonald **et al**, 2007]

Sentence-document model

- $S_1 \ldots S_n$: sentences
- Y_s .. : Sentiment labels of sentences
- Y_d : Document sentiment



Sentiment of a document

- Equal weightage to all sentences to contribute to the sentiment of the document

- Using position of a sentence to study its sentiment contribution

Reference : [Agarwal et al,2005]

Sentiment of many documents

- Using similarity between documents to find their sentiment value
- Use similarity between feature vectors to calculate Mutual similarity co-efficients

$$MSC(d_i, d_j) = \frac{\sum_k (F_i(f_k) * F_j(f_k)) - s_{min}}{s_{max} - s_{min}}$$

- $F_i(f_k)$: 1 if k^{th} feature is present in i^{th} doc.
- s_{max}, s_{min}: largest and smallest value of common features between documents

Reference : [Agarwal et al,2005]

Sentiment of many documents

- Min-cut algorithm for graph representation
- Source and sink : Positive and negative sentences

Reference : [Pang-Lee, 2002]

Traditional classifiers for document analysis

- Naïve Bayes

$$P_{\mathrm{NB}}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)}$$

- Max Entropy

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right)$$

  – $\lambda_{i,c}$: feature weight parameters

So the big question is..

- What are features?
- Where do they come from?

- What are good features?
  – Features that increase the accuracy of sentiment prediction at document level
- So, how to get them?

Feature Engineering

Feature engineering

Reference : [Pang-Lee,2008]

Feature Engineering

- Designing features to aid sentiment analysis
  - Term presence v/s frequency
  - Unigrams v/s bigrams
  - POS tagging
  - Syntax
  - Negation
  - Topic-oriented features

Reference : [Pang-Lee,2008]

Some common features (1/2)

- Term presence v/s frequency?
  - Presence: Binary valued
  - Hapax legomena
- Unigrams v/s bigrams?
  - Subsumption hierarchy
  - Contrastive distances
- POS tagging

Term presence: ‘useful’ : 1 / 0
Frequency: ‘useful’ : 236 (count)

Reference : [Pang-Lee,2008]

Some common features (2/2)

- Syntax
  - Dependency-based features
  - Valence shifters
- Negation
- Topic-oriented features
  - Checks whether a phrase follows a reference in a given topic

THIS_WORK is better than most other OTHER_WORKS by the author.

Product feature Based SA

Camera :

{Lens, Weight, Size, Strap}

Reference : [Hu et al,2005] ,[B,Liu et al, 2005]

Reviews

- Three types of Review Formats:-
  1. Pros & Cons –. E.g. *cnet.com*
  2. Pros, cons & detailed review – E.g. *eopinions.com*
  3. Free Format - E.g. *amazon.com*

No batteries available, September 22, 2009

By **A. Broadwell** (California) - See all my reviews

REAL NAME™

This is a great camera--takes good pictures, easy to use, very small. The huge, huge, drawback is that there are no replacement batteries available anywhere in the world. Panasonic has locked the camera so that it can only use Panasonic batteries, but there are none to be had. This is only a drawback if you are going on a long trip in areas without electricity, so that there is nowhere to recharge a battery, as I am. I would not have bought this camera had I known about the lack of spare batteries.

Help other customers find the most helpful reviews | Report this | Permalink

Was this review helpful to you? Yes No | Comment

Pros & Cons tend to be brief

Use full sentences

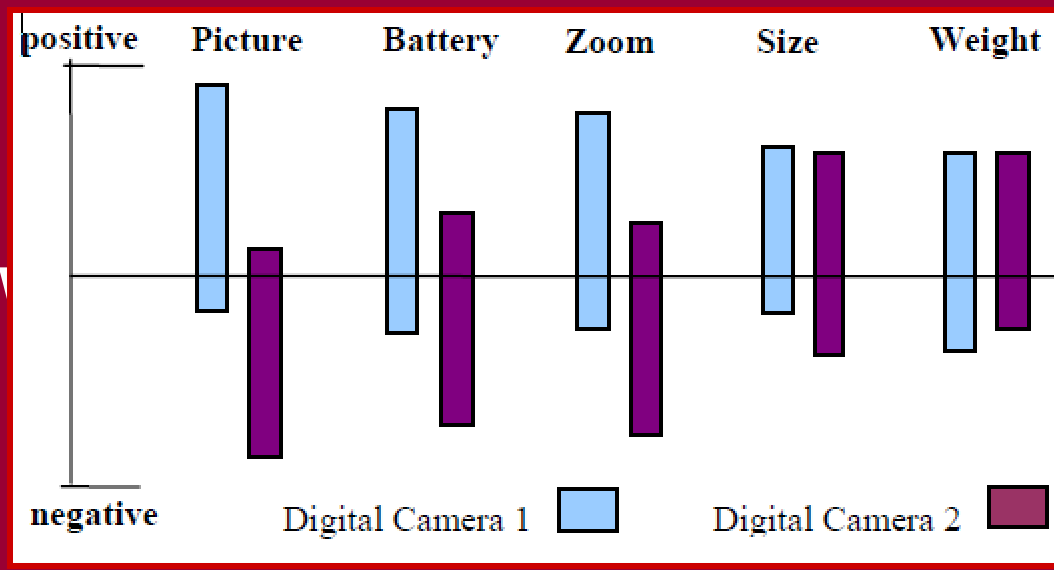
Opinion orientation of features are separated

Reference : [Jindal **et al**, 2006]

Part 1 : Handling type 2 reviews

Goals:

- **Extract product features**
  - **Why review type 2?**
- **Compare products**

Revie  **omplete**

Heavy, bad picture quality, battery life too short

Steps of processing

- Find and Download reviews
- Extracting Product features
- Identifying Opinion orientation
- Visual representation

Find & Download reviews

- **Find and Download reviews**
- **Extracting Product features**
- **Identifying Opinion orientation**
- **Visual representation**

- **Finding reviews**:
  - If the system is not at a dedicated review site

Extraction rules to identify reviews on the website pages

- Learnt from the user annotation of review pages on a website

Extracting product features

- Find and Download reviews
- **Extracting Product features**
- Identifying Opinion orientation
- Visual representation

- Preprocessing
- Rule generation
- Post-processing
- Feature refinement

Extracting product features

- Find and Download reviews
- Extracting Product features
- Identifying Opinion orientation
- Visual representation

- Preprocessing
- Rule generation
- Post-processing
- Feature refinement

<V>Included <N>[feature] <V>is

<N>[feature] <V>is <Adj>stingy

… etc.

To find general language patterns,

- Perform POS tagging and remove digits
- Replace actual feature words with [feature]
- Produce trigrams to act as itemsets

Extracting product features

Find and Download reviews

Extracting Product features

Identifying Opinion orientation

Visual representation

- Preprocessing
- Rule generation
- Post-processing
- Feature refinement

Rule 1: <V>Included <N>[feature] -> [feature]

Rule 2: <N1>, <N2> -> [feature]

Rule 3: <N1>, [feature] -> <N2> … etc.

Association mining (with 1% support) to generate rules

Extracting product features

- **Find and Download reviews**
- **Extracting Product features**
- **Identifying Opinion orientation**
- **Visual representation**

- Preprocessing
- Rule generation
- Post-processing
- Feature refinement

Association rule mining does not consider the sequence nature of data

- Sequence is crucial in NLP
- Validate against training data to maintain the sequence

Extracting product features

- Find and Download reviews
- **Extracting Product features**
- Identifying Opinion orientation
- Visual representation

- Preprocessing
- Rule generation
- Post-processing
- Feature refinement

Why refine?

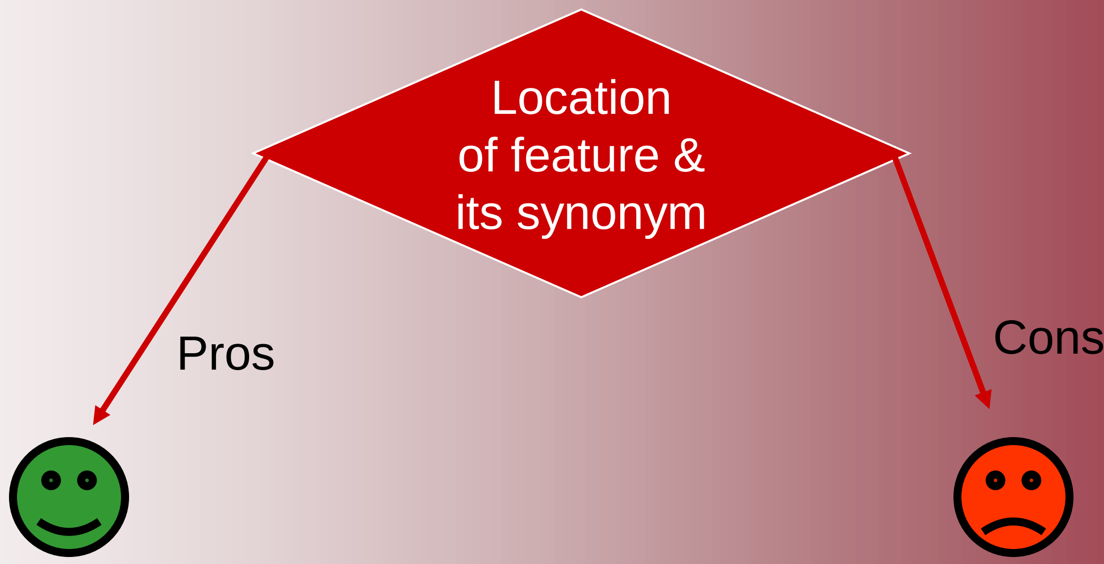
- Feature conflict : Two candidate features in one sentence segment
- Selecting 'more' suitable features

How? In case of conflict, use the feature with

"…slight hum for subwoofer when not in use"

Identifying opinion orientation

- Find and Download reviews
- Extracting Product features
- Identifying Opinion orientation
- Visual representation



Visual representation

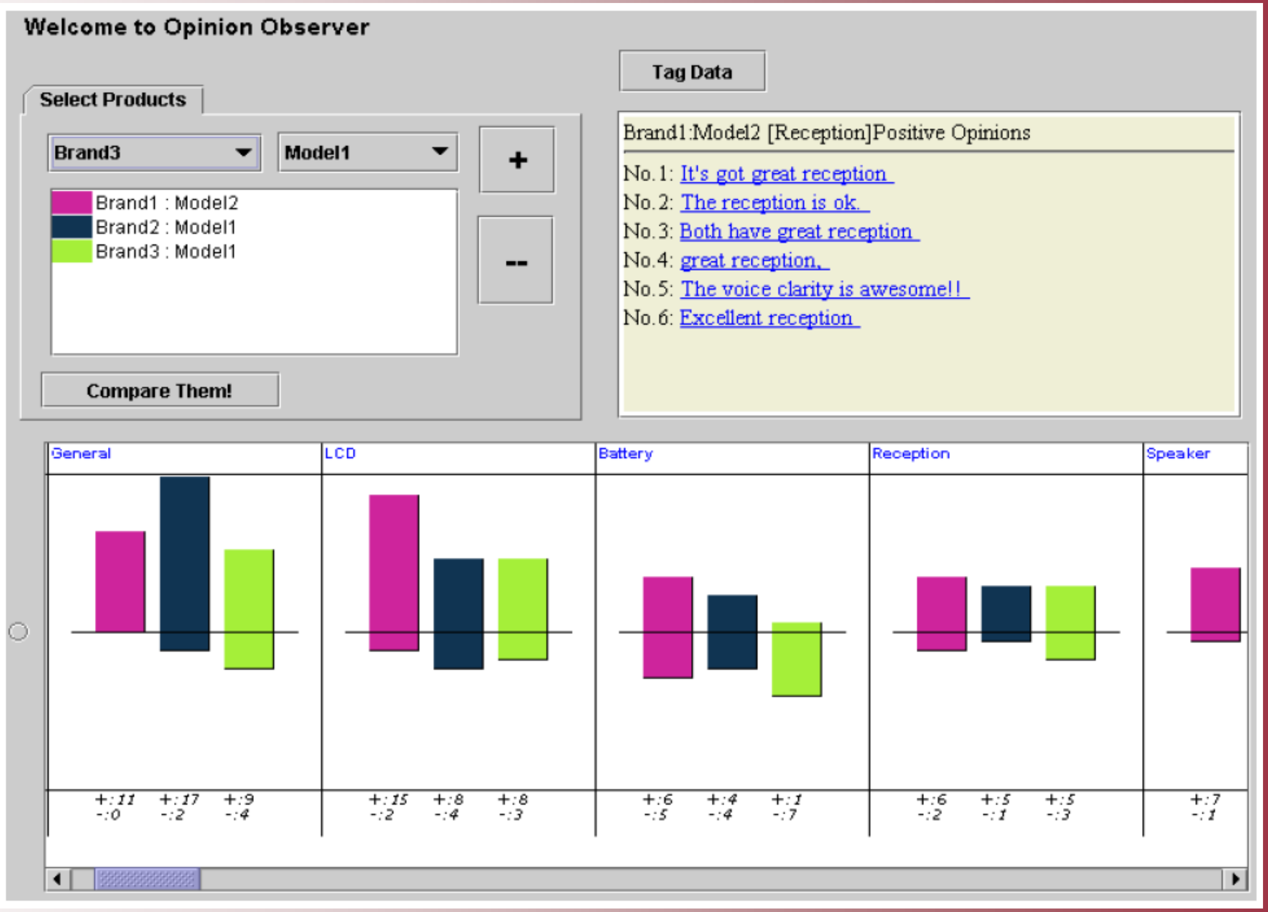
Find and Download reviews

Extracting Product features

Identifying Opinion orientation

Visual representation

Snapshot:



Part 2: Handling type 1 & 3 reviews

"The PSPgo Is great! i would recommend to anyone"

by PoofyPoofman on October 9, 2009

Pros: The sliding action is very nice the analog stick looks like its in an awkward position but when you hold it and play a game its actually very nice

Cons: The M2 memory card they should of kept the old one

Summary: Great product sony should sell a nice amount ... **read more**

Reply to this review

Type 1 Example: Cnet Review

No batteries available, September 22, 2009

By **A. Broadwell** (California) - See all my reviews

REAL NAME™

This is a great camera--takes good pictures, easy to use, very small. The huge, huge, drawback is that there are no replacement batteries available anywhere in the world. Panasonic has locked the camera so that it can only use Panasonic batteries, but there are none to be had. This is only a drawback if you are going on a long trip in areas without electricity, so that there is nowhere to recharge a battery, as I am. I would not have bought this camera had I known about the lack of spare batteries.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report this | Permalink

Comment

Type 3 Example: Amazon Review

- **Find and Download reviews**
- Frequent Feature identification
- Opinion Word extraction
- Word-level Opinion Orientation
- Infrequent Feature identification
- Sentence-level Opinion Orientation
- Summary generation

Find & Download reviews

- **Same as for type 1**

- **Finding reviews**:
  - If the system is not at a dedicated review site

Extraction rules to identify reviews on the website pages

- Learnt from the user annotation of review pages on a website

Frequent feature identification

- Same as association mining in type 1
- **Rule generation**

Rule 1: <V>Included <N>[feature] -> [feature]
Rule 2: <N1>, <N2> -> [feature]
Rule 3: <N1>, [feature] -> <N2> … etc.

Association mining (with 1% support) to generate rules

Frequent feature identification

- Same as association mining in type 1

- Rule generation
- **Feature pruning**

Why?

Not all candidate features are genuine features

Example:

The digital image CCD does not work.

I had searched fro a digital camera for three months

This is the best digital camera on the market

How?

Compact pruning

Redundancy pruning

Find and Download reviews

Frequent Feature identification

Opinion Word extraction

Word-level Opinion Orientation

Infrequent Feature identification

Sentence-level Opinion Orientation

Summary generation

Frequent feature identification

- Same as association mining in type 1

- **Rule generation**
- **Feature pruning**
  - **Compact pruning**

- A feature F is compact in sentence S if…

  any two-word sequence in F is not more than three in distance

Example: ~~Digital image CCD is not good.~~

This digital camera is so awesome.

I bought a new digital camera.

Prune features that do not satisfy above definition

Find and Download reviews

Frequent Feature identification

Opinion Word extraction

Word-level Opinion Orientation

Infrequent Feature identification

Sentence-level Opinion

Opinion word extraction

- Select sentences having features
- Find adjectives in these sentences (Presence of adjectives is useful for predicting *opinion)*

The strap is horrible and gets in the way of parts of the camera you need access to.

Find and Download reviews

Frequent Feature identification

Opinion Word extraction

Word-level Opinion Orientation

Infrequent Feature identification

Sentence-level Opinion Orientation

Summary generation

Word-level opinion orientation

- Seed set containing polarity-affixed adjectives
- Expanded using synonymy in WordNet
- Match adjectives extracted in previous step
- Assign the corresponding polarity

Infrequent feature identification

- Extract nearest noun and noun group of opinion word

The pictures are absolutely amazing.
The software that comes with it is amazing.

Find and Download reviews

Frequent Feature identification

Opinion Word extraction

Word-level Opinion Orientation

Infrequent Feature identification

Sentence-level Opinion Orientation

Summary generation

Sentence-level opinion orientation

Majority opinion of the words

↓

Orientation of the sentence

Summary generation

Example output:

Feature: picture

No. of positive occurences: 12

- Overall this is a good camera with a really good picture clarity.
- The pictures are absolutely amazing - the camera captures the minutest of details

….. etc.

No. of negative occurences: 2

- The pictures come out hazy if your hands shake even for a moment during the entire process of taking a picture.

….. etc.

Road map

Motivation & Introduction
- Perspectivizing SA
- Opinion on the web

Background
- Terminology
- Classifiers

Preliminaries
- Lexical resources
- Contextual polarity
- Subjectivity detection

Product-related SA
- Product review domain
- Document-level SA
- Feature engineering
- Product feature-based SA

Special sentences
- Comparative sentences
- Conditional sentences
- Implicit sentiment

Advanced topics

Part I : Comparative Sentences

- “This movie is good but the other movie was definitely superior.”
- “The food here isn’t half as good as the other restaurant.”

Reference : [Jindal **et al**, 2006]

Part I : Comparative Sentences

- What are they?
- Why for SA?
- **Challenges?**

A common way to evaluate is to compare

I cannot agree with you more.

India has a growth rate of x % while China has a growth rate of y %

Reference : [Jindal **et al**, 2006]

Part I : Comparative Sentences

- Comparative tags
- Tasks

JJ : Adjectives
erb
omparative
uperlative
mparative
uperlative

Extract comparative sentences

Extract sentiment in these sentences

The car has higher mileage than others in its class

Reference : [Jindal **et al**, 2006]

Extracting comparative sentences

- Comparative relations
- Type
- How?

Class-sequential rules
Pattern → Label

<{NN} {VBZ} {RB} {more JJR} {NN} {NN} {NN}> → Comparative

Steps

Reference : [Murthy **et al**, 2008]

Opinion in comparatives

- Types
- Opinionated

increasing comparative + word of sentiment X → sentiment X

decreasing comparative + word of sentiment X → sentiment Y

For 'more' or 'less', use specific rules

This has more energy than that.

Context-based comparatives

One-sided association (OSA) :

$$OSA(F,C) = log\frac{Pr(F,C)Pr(C|F)}{Pr(F)Pr(C)}$$

If C & F (and synonym of C & F) co-occur in pros, count as 1.

If antonym of C & F co-occur in cons, count as 1

Words and synonyms in pros, count as 1

Antonyms of words in cons, count as 1

OSA $_{pros}$ (F, C) > OSA $_{cons}$ (F, C) : Prefer, else No

Pros: High mileage

Cons: Low steering flexibility

Reference : [Murthy **et al**, 2008]

Results

Pointwise Mutual Information :

$$PMI(w1, w2) = \frac{Hits(w1ANDw2)}{Hits(w1)Hits(w2)}$$

	EntityS1 Preferred			EntityS2 Preferred		
	Prec.	*Rec.*	*F*	*Prec.*	*Rec.*	*F*
PCS (OSA)	0.967	0.966	0.966	0.822	0.828	0.825
PCS: No Pros & Cons	0.925	0.980	0.952	0.848	0.582	0.690
PCS (PMI)	0.967	0.961	0.964	0.804	0.828	0.816

Part II : Conditional sentences

- *"If your Nokia phone is not good, buy this great Samsung phone."*

Reference : [Jindal **et al**, 2006], **Narayanan et al 2009]**

Part II: Conditional Sentences

- What?
- Connectives
- Components

Two clauses:

1) Condition clause
2) Consequent clause

"If (…), then (…)"

Reference : [Narayanan et al, 2009]

And about opinion expressed…

Even if opinion words are present –

- Sentences may express no opinion
- May express opinion
- Both the condition and consequent together may determine the opinion

If someone makes a beautiful and reliable car,

I will buy it.

Type of conditionals (1/2)

- **Zero Conditional**
- **First Conditional**
- **Second Conditional**
- **Third conditional**

If I had bought the a767, I would have hated it

Type of conditionals (2/2)

- How to identify?

1. Tense patterns
2. Semantic meaning

- Advantage taking former style

"….different types can be detected easily because they depend on tense which can be produced by a part-of-speech tagger"

Reference : [Narayanan et al, 2009]

Identifying patterns

Type	Linguistic Rule	Conditional POS tags	Consequent POS tags
0	If + simple present → simple present	VB/VBP/VBZ	VB/VBP/ VBZ
1	If + simple present → will + bare infinitive	VB/VBP/VBZ /VBG	MD + VB
2	If + past tense → would + infinitive	VBD	MD+ VB
3	If + past perfect → present perfect	VBD+VBN	MD + VBD

Feature Engineering

- Sentiment words/phrases and their locations
- POS tags of sentiment words
- Words indicating no opinion
- Tense patterns
- Special characters
- Conditional connectives
- Negation words

Classification

- Classifier used: SVM
- Two classifiers used for sentence classification:

1. One of these:
   a. Condition Classifier
   b. Consequent Classifier

2. A topic classifier for identifying topic

Based on the presence of topic detected in conditional clause or consequent clause

Whole-sentence-based classification

- Used multiple instances of the same sentence if more than one topic found as test vector

- Two extra features added
  - *Topic location*
  - *Opinion weight*

Reference : [Narayanan et al, 2009]

Observations

- Highest F-score reported for whole-sentence based classification

- Consequent usually plays the key role in determining the sentiment of the sentence

Reference : [Stephen et al, 2009]

Part III : Detecting Implicit Sentiment

- On November 25, A soldier veered his jeep into a crowded market and killed three civilians.

- On November 25, A soldier’s jeep veered into a crowded market, causing three civilian deaths.

Reference : [Stephen et al, 2009]

Implicit sentiment

- Verbal descriptions of an event carries an underlying attitude

- The speaker twists to promote
  - A particular problem definition
  - Causal interpretation
  - Moral evaluation
  - Treatment recommendation

Reference : [Stephen et al, 2009]

How could it be done?

- Lexical choice

- Syntactic choices

"Mistakes were made" - Ronald Reagan [Iran Contra scandal].

Reference : [Stephen et al, 2009]

Implicit sentiment – A linguist's view

- Syntactic diathesis alternations
- Core idea
- Feature engineering

A set of
thirteen semantic
properties
were selected
as features

X murders Y entails that X started event

Reference : [Stephen et al, 2009]

Phenomena

- Transitive form of the verb held more implicit sentiment than its nominal counterpart
- Ergative class of same verb does not convey much sentiment

Suffocation kills 24-year-old woman
Man suffocates 24-year old woman

Reference : [Stephen et al, 2009]

Feature Engineering

- Find domain terms
- Include term-related syntactic dependency features
- Two construction-specific features added
  - TRANS:v – represents v in a canonical, syntactically transitive usage
  - NOOBJ:v – represents v used without a direct object

Reference : [Stephen et al, 2009]

Classification

- Dataset used – pro & anti-death penalty websites
  - Domain term used – “killed”
  - Also mined frequent terms
- Along with bigram features ,above were added to get a better classification using SVMs

Road map

Motivation & Introduction

- Perspectivizing SA
- Opinion on the web

Background

- Terminology
- Classifiers

Preliminaries

- Lexical resources
- Contextual polarity
- Subjectivity detection

Product-related SA

- Product review domain
- Document-level SA
- Feature engineering
- Product feature-based SA

Special sentences

- Comparative sentences
- Conditional sentences
- Implicit sentiment

Advanced topics

- Opinion Spam
- Opinion Flame
- Opinion Search
- Temporal SA
- Wishlist analysis
- Cross-lingual/Cross-domain SA

Opinion spam

Opinion spam: A side-effect of UGC

- Reviews contain rich user opinions on products and services
- Anyone can write anything on the Web
  – No quality control
- Result
- Incentives

Positive opinion ->
Financial gain for
organization

Types of spam reviews

- **Type 1 (untruthful opinions)**
- **Type 2 (reviews on brands only)**
- **Type 3 (non-reviews)**

Advertisements
Other irrelevant reviews containing no opinions

It’s from nikon, what more you want..

Status of opinion spam-handling

- Review's Review done manually mostly
- Some customer review sites do have sophisticated algorithms to tackle them
- But not all
- And definitely not all types

Opinion flame

Opinion Flame

- Flame
- Risky discussion
- Where?

E-mails, discussions, chat conversations, etc.

Reference : [Pazienza **et al**]

Linguistics of flame recognition

- Characterized by:
  - Offensive language
  - Off-the-topic
  - Repetitive cites from other posts
  - Repetitive address to a specific reader
  - Ironic expressions / unusual politeness

Reference : [Spertus, 1997]

Smokey

- Mailbox filter for flame detection
- Uses rule classes and C4.5 decision trees
- Noun appositions
- Imperative sentence
- Bad/negative words
- Scare quotes
- Profanity rules

“%$#%@%..”

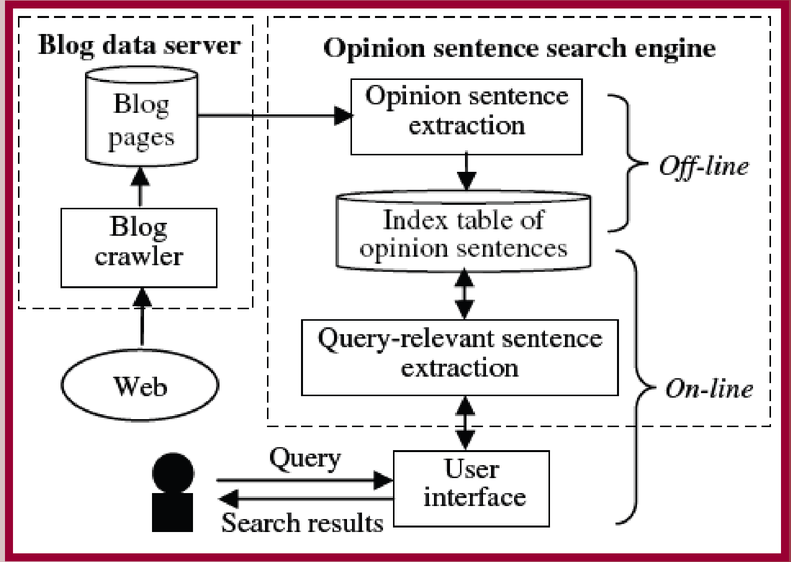
Opinion search

Reference : [Furuse, 2007]

Opinion Search

- Goal: Search engine that extracts opinion sentences relevant to blog pages

- Two components:
  - Opinion content
  - Query Relevance



Components of Opinion Search

- Opinion Identification

1. Clue expressions
2. Semantic categories
3. Parts of speech

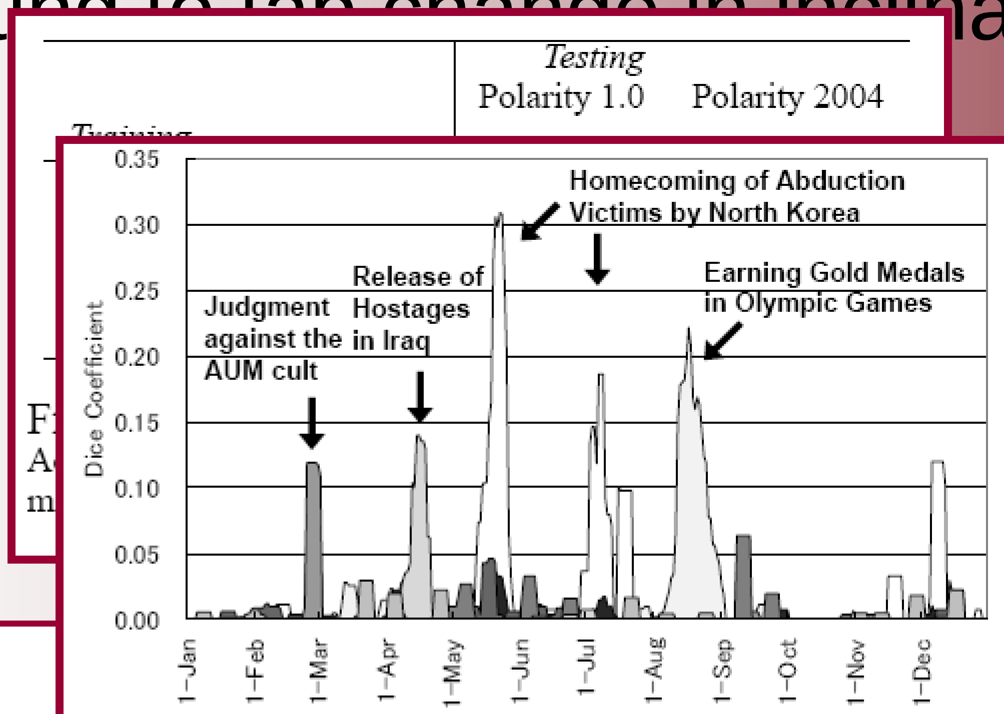
- Query relevance

a) Query phrase in sentence or the one before it
b) Query phrase in sentence or its ‘chunk’

Temporal SA

Temporal Sentiment Analysis

- ‘Time’ factor in trends
- Interesting to tap change in inclination / moods



Now: This cell phone has a touch screen.

Wish-list analysis

Wish-list analysis

- Wish : Desire or hope for something to happen
- Highly domain-specific

- Can we track what user's wishes are?

I wish for world peace.

Cross-Lingual SA

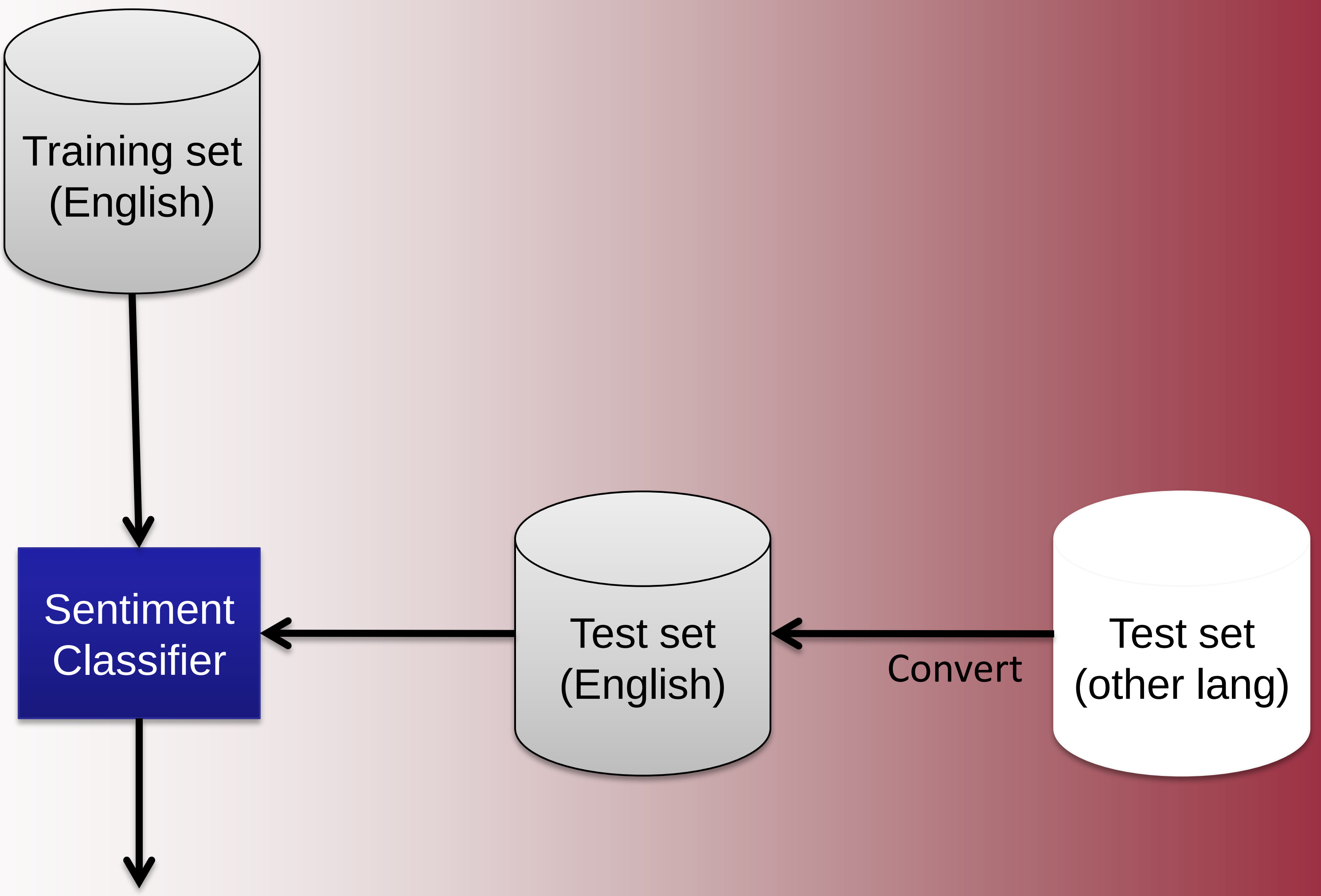
Cross-lingual SA

- **Why?**
  – Majority focus on English Sentiment Classification
  – Unavailability of annotated corpora

- How to leverage existing corpora for sentiment classification of other languages

Reference : [Wan et al, 2008]

Naïve idea (1/2)



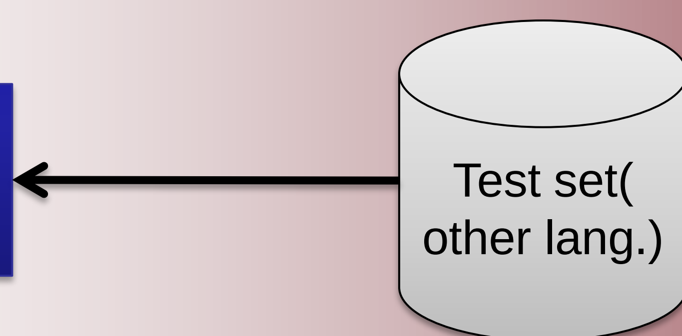
Reference : [Wan et al, 2008]

aïve idea (2/2)

For the naïve ideas,

Result not promising

Performance depends upon underlying distribution of words between original language and translated language



Reference :[Whitehead et al, 2008]

Cross-Domain Sentiment Analysis

Why?

- To create a general Classifier for all domains

or

- Labeled Data needed for training
- Gathering training data
  - If numeric rating present : easy
  - Manual & expensive

  e.g. Political opinions, Blogs

Some observations

- Domain differences are substantial
  - One domain classifier cannot beat even baseline of other domain
- Within a domain a specific low level feature worked better
  - In target domain another or combination of low level feature worked better

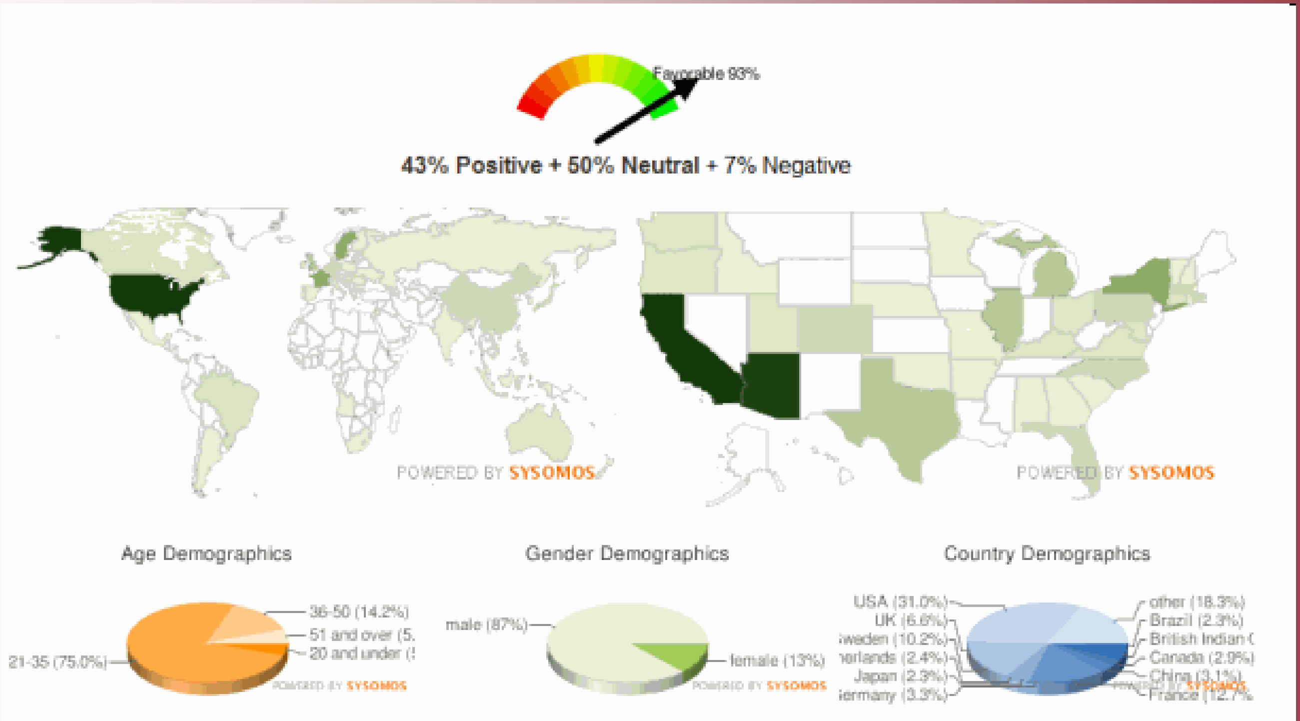
Reference : http://www.readwriteweb.com/archives/sentiment_analysis_is_ramping_up_in_2009.php

Sentiment Analysis in 2009

Actual sentiment analysis applications

Social media monitoring

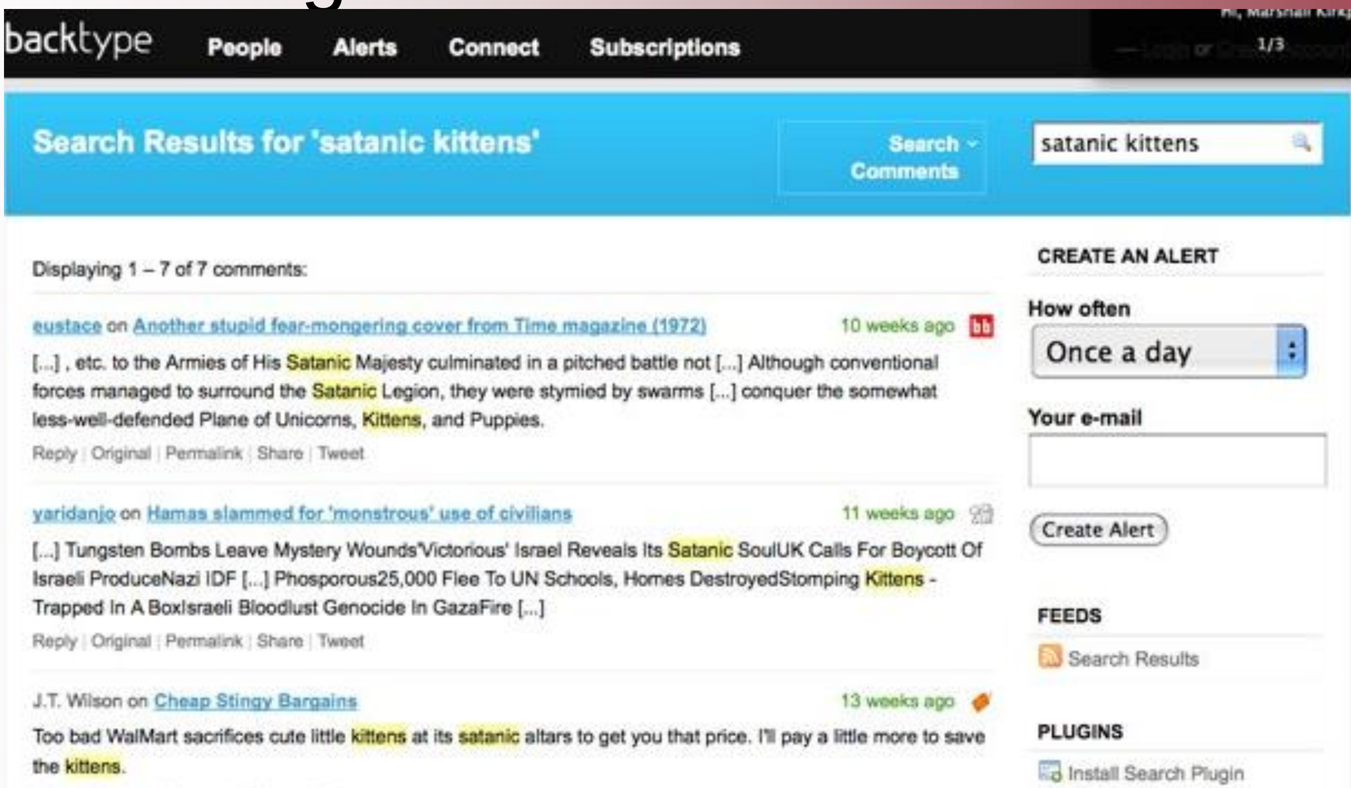
- Spans blogs, social media, news media reports



Snapshot: Sysomos

Conversation analysis

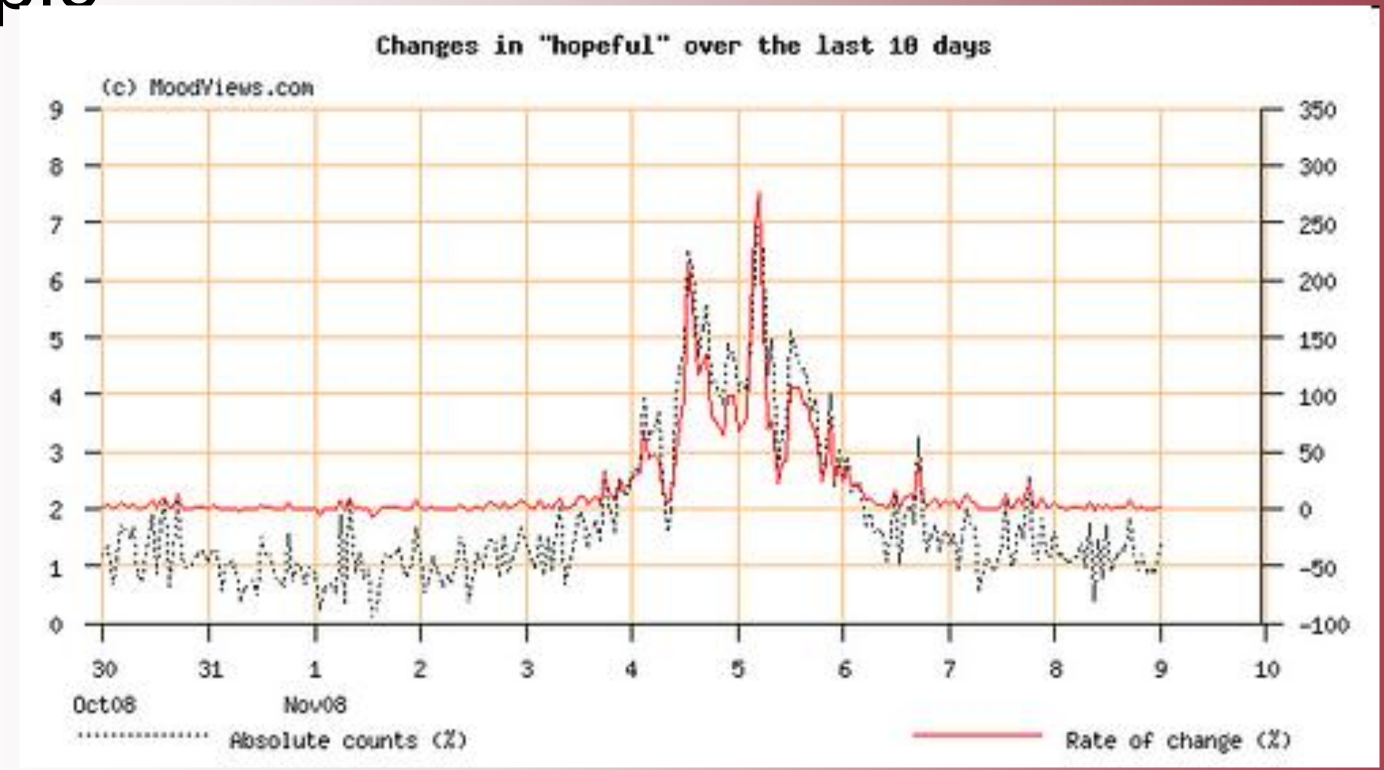
- Tracking conversation on social networking sites



Snapshots: Backtype

Mood analysis

- Real-time updation of moods w. r. t. a topic



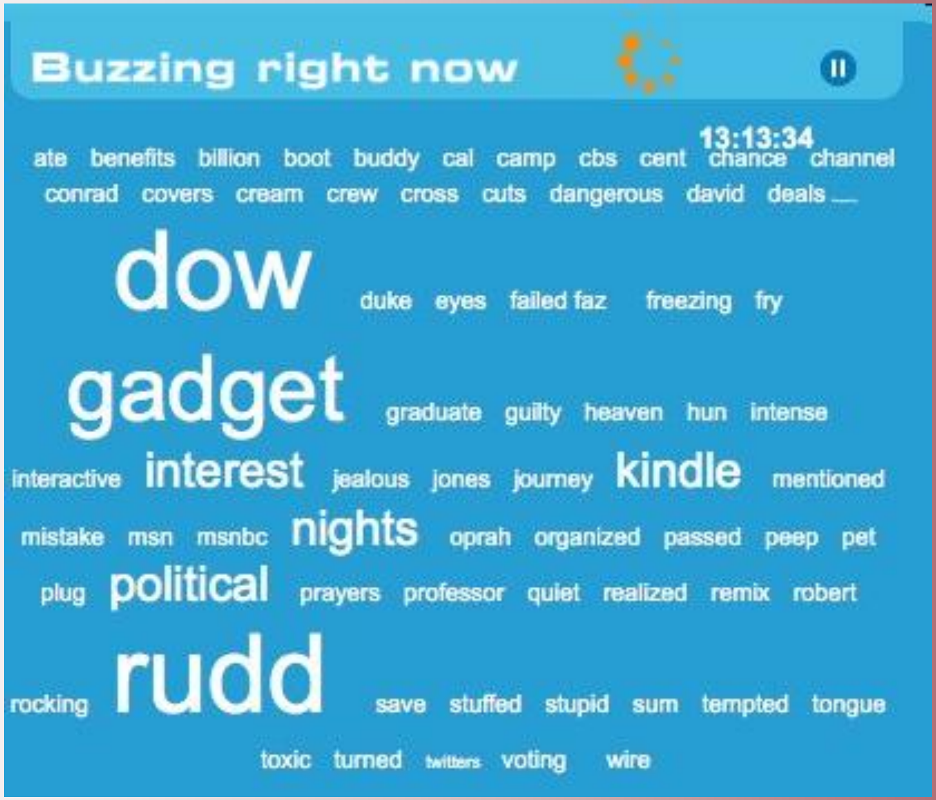
Snapshot: MoodViews

Semantic search

- Sentiment search API by Evri
- Claims to allow deeper answers like “who”, “why”

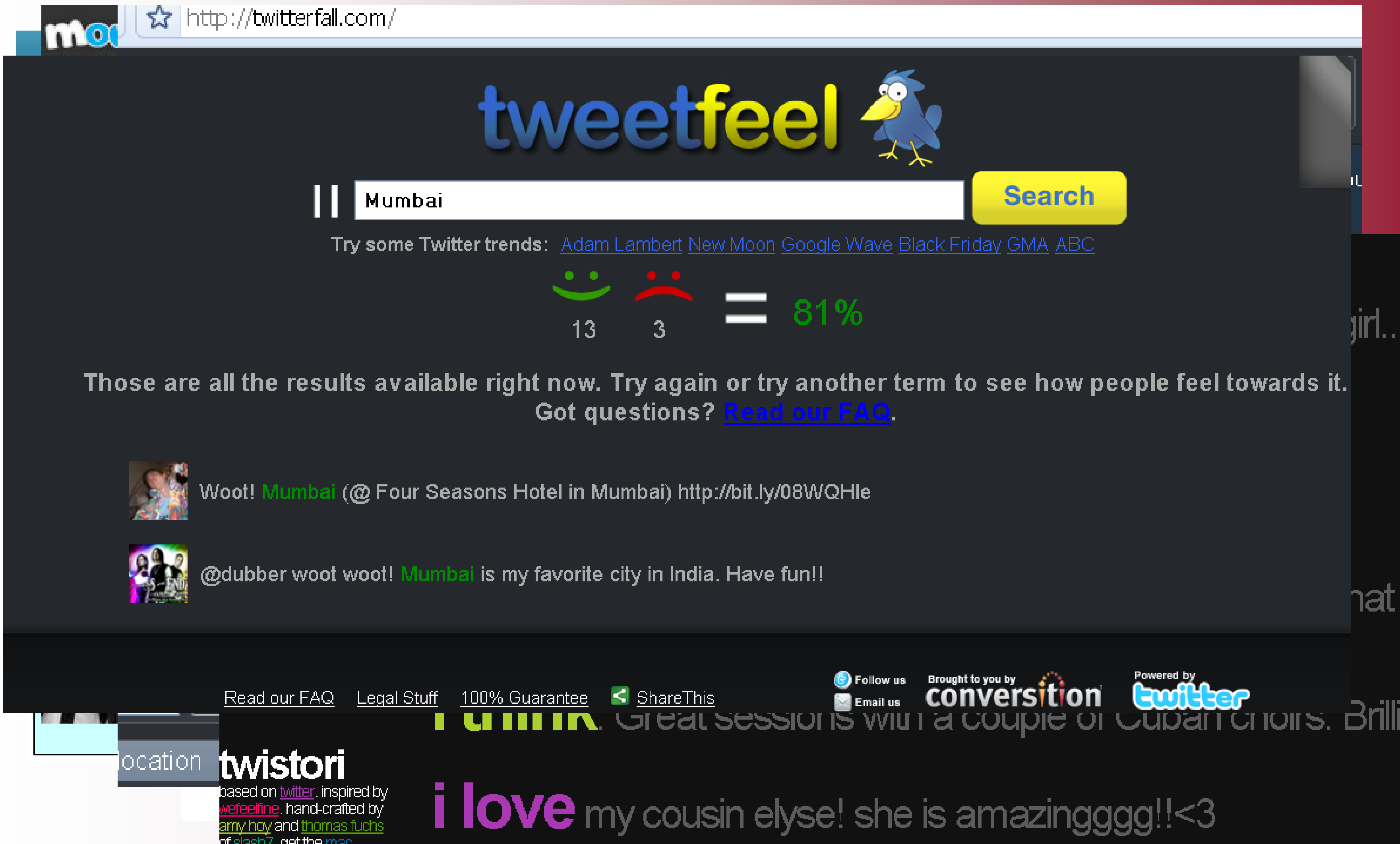
A zeitgeist

- Understanding the ‘climate’



Snapshot: Twitscoop

… and many more



Open questions for a researcher

- Opinion Spam/ Opinion Flame/ Opinion Search/ Temporal Sentiment analysis/ Wishlist analysis/ Cross-domain SA/ Cross-lingual SA

- Alternative approaches for subjectivity extraction
- Alternative approaches for document-level sentiment analysis
- Domain-specific lexical resource for SA
- Handling sarcastic statements for SA
- Handling thwarted expressions for SA
- Detecting sentiment for implicit product features
- SA applied to other NLP tasks

Standard datasets for SA

– Congressional floor-debate transcripts

http://www.cs.cornell.edu/home/llee/data/convote.html

– Cornell movie-review datasets

http://www.cs.cornell.edu/people/pabo/movie-review-data/

– Customer review datasets

http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip

– Economining

http://economining.stern.nyu.edu/datasets.html

– MPQA Corpus

http://www.cs.pitt.edu/mpqa/databaserelease

– Multiple-aspect restaurant reviews

http://people.csail.mit.edu/bsnyder/naacl07

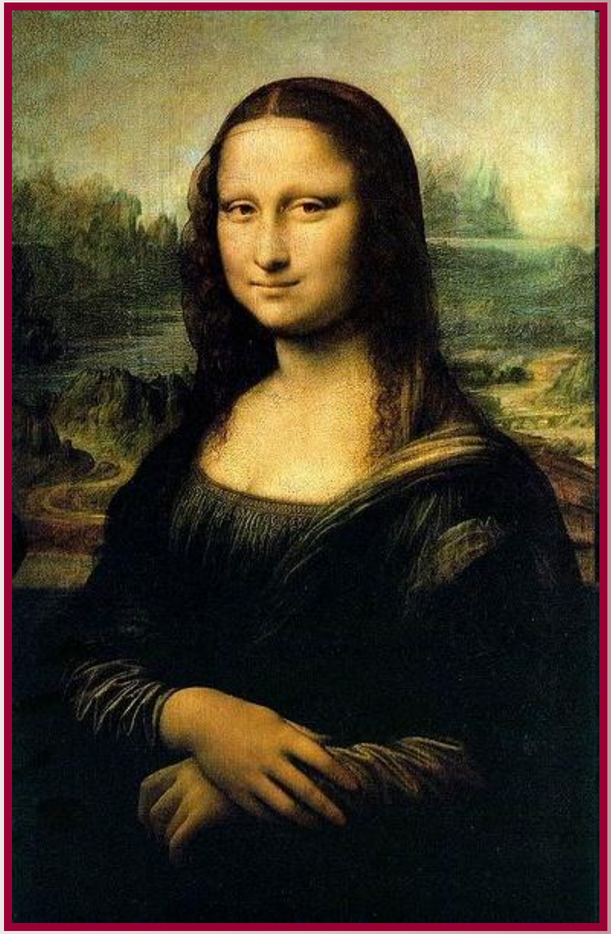
– Review-search results sets

http://www.cs.cornell.edu/home/llee/data/search-subj.html

References

- J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proceedings of the ACL students research workshop, Association for Computational linguistics, 2005.
- N. Jindal, B. Liu.Mining. Comparative sentences and relations. American association for artificial intelligence. 2006.
- N. Jindal and B. Liu, “Identifying comparative sentences in text documents,” Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), 2006.
- N. Jindal and B. Liu, “Opinion spam and analysis,” Proceedings of the Conference on Web Search and Web Data Mining (WSDM), pp. 219–230, 2008.
- Jindal,Nitin and Liu,Bing Mining Comparative Sentences and Relations, American Association for Artificial Intelligence,2006.
- Klenner, M and A Fahrni. "Old wine and warm beer: Targetspecific." AISB. Aberdeen,scotland, 2008.
- Liu,Bin Hu,M and Cheng,J “Opinion observer: Analyzing and comparing opinions on the web,” Proceedings of WWW, 2005.
- Liu,Bing,Sentiment Analysis and Subjectivity,Handbook of Natural Language Processing,CRC Press,2009
- G. Murthy, B. Liu. Mining opinions in comparative sentences. Proceedings of the 22nd International conference on computational linguistics. 2008.
- R. McDonald, K. Hannan. MStructured models for fine-to-coarse sentiment analysis. Association for Computational Linguistics. 2007.
- .B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” Proceedings of the Association for Computational Linguistics (ACL), pp. 271–278, 2004.
- B. Pang and L. Lee, “Opinion mining and sentiment analysis.” Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.
- M. Pazienza, A. Stellato. Frames, Risky discussions, no flames recognition in forums.
- A.-M. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.
- Ramanathan Narayanan, Bing Liu and Alok Choudhary. "Sentiment Analysis of Conditional Sentences." Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-09). August 6-7, 2009.
- E. Riloff and J. Wiebe. Exploiting Subjectivity classification to improve information extraction. Proceedings of the 20th National conference on artificial intelligence, 2005.
- Saif Mohammad; Cody Dunne; Bonnie Dorr, Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus, EMNLP,ACL PP-599-608,August 2009.
- E. Spertus. Smokey: Automatic recognition of hostile messages. American association for artificial intelligence. 1997.
- Strapparava, C. and Valitutti, A. WordNet-Affect: an affective extension of WordNet, Proceedings of LREC, 2004, pp-1083-1086.

Stone, Philip J. and Dunphy, Dexter C. and Smith, Marshall S. and Ogilvie, Daniel M., The General Inquirer: A Computer Approach to Content Analysis, MIT Press,1966.



Questions?