Information Retrieval

INFO 4300 / CS 4300

- Classification applications in IR
  - Spam detection
  - Sentiment classification
  - Online advertising
  - **Identifying fake online reviews**

Classification

- Classification is the task of automatically applying labels to items
- Useful for many search-related tasks
  - Spam detection
  - Sentiment classification
  - Online advertising
  - **Identifying fake online reviews**

I don't like spam

- Spam
  - Any content generated for malevolent purposes

http://www.youtube.com/watch?v=g8huXkSaL7o

Spam, Spam, Spam

- Classification is widely used to detect various types of spam
- There are many types of spam
  - Email, advertisement, blog, web page
- Develop specialized classifiers for each type

Example Spam Assassin Output

```
To: ...
From: ...
Subject: non profit debt
X-Spam-Checked: This message probably not SPAM
X-Spam-Score: 3.853, Required: 5
X-Spam-Level: *** (3.853)
X-Spam-Tests: BAYES_50,DATE_IN_FUTURE_06_12,URIBL_BLACK
X-Spam-Report-rig: ---- Start SpamAssassin (v2.6xx-cscf) results
        2.0 URIBL_BLACK       Contains an URL listed in the URIBL blacklist
                      [URIs: bad-debtyh.net.cn]
        1.9 DATE_IN_FUTURE_06_12   Date: is 6 to 12 hours after Received: date
        0.0 BAYES_50         BODY: Bayesian spam probability is 40 to 60%
                      [score: 0.4857]

Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards, No-collateral
Bank Loans, Personal Loans,
Medical Bills etc.
http://www.bad-debtyh.net.cn
```

Spam, Spam, Spam

- Web page spam
  - Link spam
    - » Adding links to message boards
    - » Link exchange networks
    - » Link farming
  - Term spam
    - » URL term spam
    - » Dumping
    - » Phrase stitching
    - » Weaving

[Gyongyi and Garcia-Molina, 2005]

Spam Example

Website:

BETTING NFL FOOTBALL PRO FOOTBALL
SPORTSBOOKS NFL FOOTBALL LINE
ONLINE NFL SPORTSBOOKS NFL

Players Super Book

When It Comes To Secure NFL Betting And Finding
The Best Football Lines Players Super Book Is The
Best Option! Sign Up And Ask For 30 % In Bonuses.

MVP Sportsbook

Football Betting Has Never been so easy and secure!
MVP Sportsbook has all the NFL odds you are looking for.
Sign Up Now and ask for up to

30 % in Cash bonuses.

Term spam:

pro football sportsbooks nfl football line online nfl sportsbooks nfl football
gambling odds online pro nfl betting pro nfl gambling online nfl football
spreads offshore football gambling online nfl gamblibg spreads online
football gambling line online nfl betting nfl sportsbook online online nfl
betting spreads betting nfl football online online football wagering online
gambling online gambling football online nfl football betting odds offshore
football sportsbook online nfl football gambling ...

Link spam:

MVP Sportsbook Football Gambling Beverly Hills Football Sportsbook
Players SB Football Wagering   Popular Poker Football Odds
Virtual Bookmaker Football Lines   V Wager Football Spreads
Bogarts Casino Football Point Spreads   Gecko Casino Online Football Betting
Jackpot Hour Online Football Gambling   MVP Casino Online Football Wagering
Toucan Casino NFL Betting   Popular Poker NFL Gambling
All Tracks NFL Wagering   Bet Jockey NFL Odds
Live Horse Betting NFL Lines MVP Racebook NFL Point Spreads
Popular Poker NFL Spreads Bogarts Poker NFL Sportsbook ...

Spam Detection via ML

- Useful features
  - Unigrams
  - Formatting (invisible text, flashing, etc.)
  - Misspellings
  - IP address
- Different features are useful for different spam detection tasks
- Email and web page spam are by far the most widely studied, well understood, and easily detected types of spam

Information Retrieval

INFO 4300 / CS 4300

- Classification applications in IR
  - Spam detection
  → - Sentiment classification
  - Online advertising
 - **Identifying fake online reviews**

Sentiment

- Blogs, online reviews, and forum posts are often opinionated
- Sentiment classification attempts to automatically identify the polarity of the opinion
  - Negative opinion
  - Neutral opinion
  - Positive opinion
- Sometimes the strength of the opinion is also important
  - "Two stars" vs. "four stars"
  - Weakly negative vs. strongly negative

Classifying Sentiment

- Useful features
  - Unigrams
  - Bigrams
  - Part of speech tags
  - Adjectives
- SVMs with unigram features have been shown to be outperform hand built rules

Sentiment Categorization

Document → Is the overall sentiment in the document positive? negative? → ☺
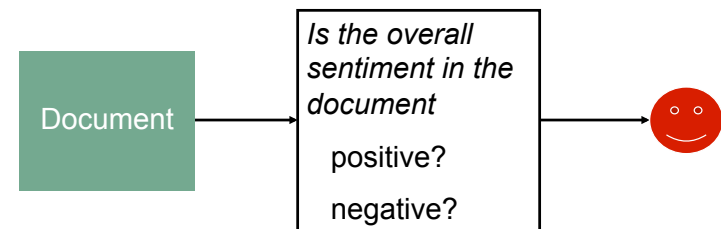
→ Pang et al., *EMNLP 2002*
Turney, *ACL 2002*
Turney & Littman, *TOIS 2003*
…

Sentiment Categorization

- Applied standard text categorization algorithms
  - Features = bag of words
  - Classifier = machine learning algorithms
    - » SVMs, naïve Bayes, maxent
- Data
  - Reviews from IMDb archive
  - 752 negative, 1301 positive

Sentiment categorization appears to be harder than categorizing by topic, e.g. ~82% accuracy for movie reviews.

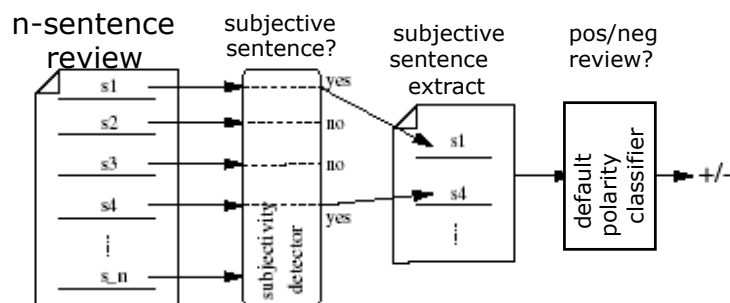[E.g., Pang, Lee, Vaithyanathan, *EMNLP 2002*]

What is the problem?

– This laptop is *a great deal*.

– *A great deal* of media attention surrounded the release of the new laptop model.

– If you think this laptop is *a great deal*, I've got a nice bridge for you to buy.

– …The protagonist tries to protect her *good* name…

[Examples from Lillian Lee]

Sentiment Categorization

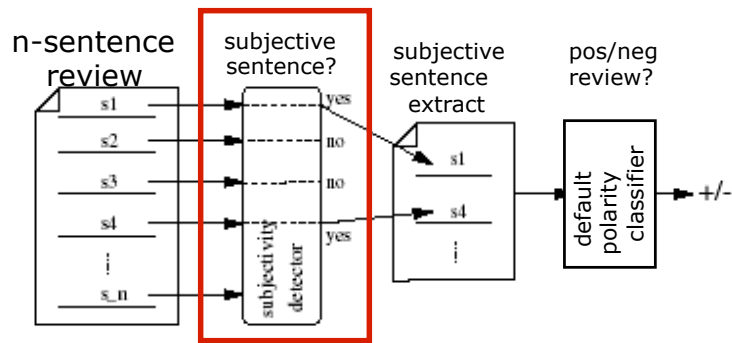**Hypothesis:** sentence-level subjectivity decisions would help.

[Pang and Lee, *ACL 2004*]

Sentence-level Subjectivity Detector

- Text categorization
  - Training data????
    - » Subjective sentences: movie review snippets
      - ◆ www.rottentomatoes.com
    - » Objective sentences: movie plot summaries
      - ◆ IMDb.com
  - Test sentence
    - » Output a **subjectivity score**

[Pang and Lee, *ACL 2004*]

Sentiment Categorization

n-sentence review → subjective sentence? → subjective sentence extract → pos/neg review?

subjectivity detector / default polarity classifier / +/-

Not good enough...

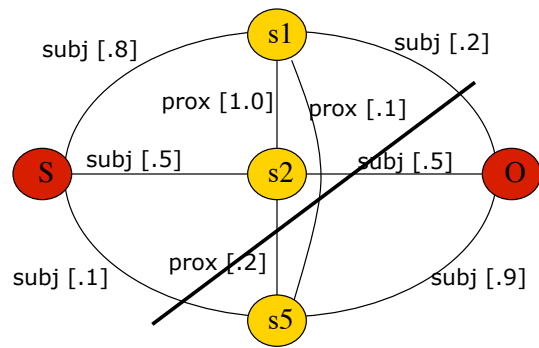[Pang and Lee, *ACL 2004*]

Minimum Cut Classification Algorithm

- Based on Blum & Chawla [2001]; Boykov et al. [1999]
- Subjectivity constraints
  - Believe the sentence-level **subjectivity** classifier score [0..1]
- Coherence constraints
  - Assumption: nearby sentences should share subjectivity status
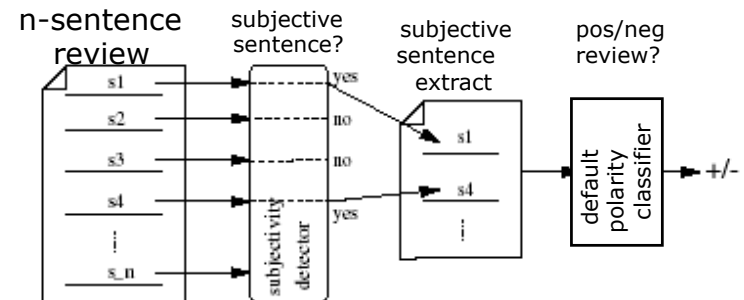 - Sentence-level **proximity** scores

[Pang and Lee, *ACL 2004*]

Sentence-level Subjectivity Detector

- Minimum cuts in graphs

subj [.8]   subj [.2]
prox [1.0] prox [.1]
subj [.5]   subj [.5]
prox [.2]
subj [.1]   subj [.9]

[Pang and Lee, *ACL 2004*]

Sentiment Categorization

n-sentence review → subjective sentence? → subjective sentence extract → pos/neg review?
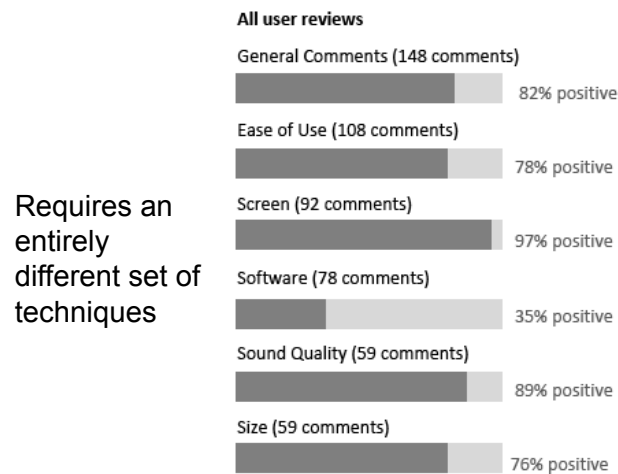
Performance increases: ~82.8% → ~86.4%
  Better than: first N, last N, most subjective N sentences
  Min-cut algorithm better than subjectivity detector alone
  Proximity constraints are important

Aspect-based Sentiment Summarization

Requires an entirely different set of techniques

All user reviews

General Comments (148 comments)
82% positive

Ease of Use (108 comments)
78% positive

Screen (92 comments)
97% positive

Software (78 comments)
35% positive

Sound Quality (59 comments)
89% positive

Size (59 comments)
76% positive

NEW YORK, Dec 3 (Reuters) - U.S. stock index futures fell on Tuesday, indicating the S&P 500 may fall for a third consecutive day on concerns a reduction in stimulus by the Federal Reserve may be on the horizon earlier than some had anticipated.

– The benchmark S&P index fell on Monday as stronger-than-expected data on manufacturing and construction spending prompted speculation the Fed may lean towards scaling back its stimulus of $85 billion in monthly bond purchases.

While virtually all market participants accept the central bank will begin to trim the stimulus at some point, the timing remains in question, with many analysts expecting the announcement in March.

"Very strong data would certainly set the stage for a January move to taper, but I'd be surprised to see them take any significant step in the waning weeks of the year," said Rick Meckler, president of investment firm LibertyView Capital Management in Jersey City, New Jersey.

"But the fact remains that it is coming, and they have given us so much time to prepare for it that it may well be not as significant an event when they finally do and maybe that has been their plan all along."

The Fed has said it would begin to slow the program when certain economic measures meet its targets. The economic calendar is packed this week with data that may provide some insight, culminating with the November payrolls report on Friday.

Sentiment analysis for blogs, forums, etc.

we must banned junkfoods in school because chioldren woul rely on it always as there snack and junkfoods contain some articial preservatives which is not good especially for children.

Login To Reply   43 weeks 3 days ago

It is true that some junk food contains artificial ingredients and/or preservatives, but so do other foods, like jam, dried fruit and meat, and as long as junk food is eaten in controlled amounts than it doesn't pose a serious problem. Kids need a snack in school anyways, and even if the school banned all junk food students would still bring it.

Login To Reply   22 weeks 20 hours ago

Information Retrieval
INFO 4300 / CS 4300

- Classification applications in IR
  - Spam detection
  - Sentiment classification
  - Online advertising
    - **Identifying fake online reviews**