A Tempest,

Or, on the flood of interest in:

sentiment analysis,

opinion mining,

and the computational treatment of subjective language

Lillian Lee

Cornell University

http://www.cs.cornell.edu/home/llee

AAAI 2008 invited talk

“Romance should never begin with sentiment. It should begin with science and end with a settlement.” — Oscar Wilde, *An Ideal Husband*

What’s past is prologue

AAAI 1998: Julia Hirschberg, invited talk on the “statistical revolution” in natural language processing (NLP): *the Web will be a new data source*.

At the time:

- Number of pages AltaVista indexed: $\sim$31 million
- Number of weblogs then in existence: $\leq$23 [Source: InformationWeek]

O brave new world

AAAI 1998: Julia Hirschberg, invited talk on the “statistical revolution” in natural language processing (NLP): *the Web will be a new data source*. At the time:

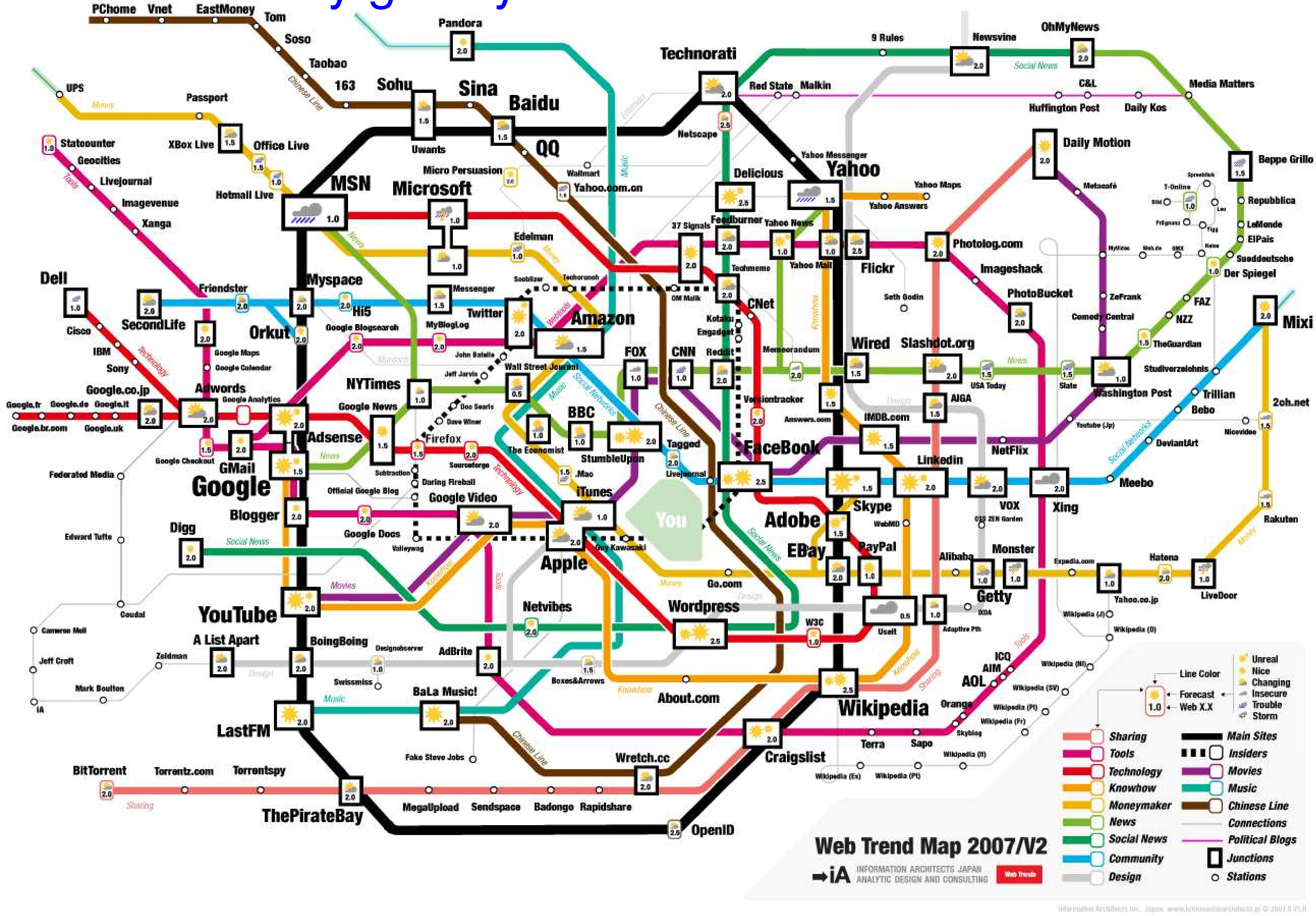
- Number of pages AltaVista indexed: ~31 million
- Number of weblogs then in existence: ≤23 [Source: InformationWeek]

AAAI 2008: Sooooo much more is available online, making for even more opportunities for — and awareness of — NLP.

▷ The Association for Computational Linguistics now has ≥2000 members.

As we all know, we’re not just talking about static documents ...

How many goodly creatures are there here



’Tis new to thee

“What other people think” has *always* been an important piece of information during decision making, even pre-Web.

- asking friends who they plan to vote for
- requesting letters of recommendation from colleagues
- checking *Consumer Reports* regarding dishwasher brands

But now, we can find out the opinions of people who are neither acquaintances nor well-known authorities — that is, people we have never heard of.

And conversely, we can make our opinions known to millions of strangers.

That hath such people in it

People search for, are affected by, and post online ratings and reviews.

- 60% of US residents have done online research on a product at least once, and 15% do so on a typical day.
- 73%-87% of US readers of online reviews of services reported that the reviews had a significant influence on their purchase. (more on economics later)
- 30% of online US residents have posted an online comment or review, as have 18% of online US senior citizens.

But, 58% of US internet users report that online information was missing, impossible to find, confusing, and/or overwhelming.

Creating technologies that find and analyze reviews would answer a tremendous information need.

[Sources: comscore '07, Horrigan Pew survey '08]

Beyond consumption: politics

Of the 31% of US residents — over 60 million people — that gathered information about the 2006 elections online and exchanged views about it by email,

- 28% said that a major reason for their online activities was to get perspectives from *within* their community, and 34% said that a major reason was to get perspectives from *outside* it.
- 28% said that most sites they use *share* their point of view, but 29% said that most sites they use *challenge* their point of view.

[Source: Rainie and Horrigan Pew survey, ’07]

Beyond individual interest in opinions

Corporations care about brand monitoring, keeping tabs on consumer opinions, and other types of business intelligence.

Systems of interest could:

- watch blogs, review aggregation sites, etc. for positive or negative mentions and reviews of a company's (or its competitors') products; or
- automatically process customer feedback

Governmental eRulemaking initiatives (e.g., www.regulations.gov) directly solicit citizen comments on potential new rules

▷ Tools are needed to automatically analyze these comments:
400,000 were received for a single rule on labeling organic food.

Many other applications exist, as well.

A perfect storm

There has been a huge upswell of activity in *sentiment analysis* ...

(a.k.a. opinion mining, subjectivity analysis, review mining, or anything suggestive of the *computational treatment of opinion-oriented, evaluative, and other related types of language*)

... due to the applications we just mentioned, plus improvements in machine learning methods, the new availability of relevant datasets, and the inherent interestingness of the research problems involved (discussed soon).

- 20+ (rough estimate) companies, large and small, are working on sentiment analysis systems (e.g., fellow Cornellian Claire Cardie's startup, Jodange)
- Recent bibliographies list 200-300 references on the subject. [Wiebe '07, Pang&Lee '08]

What's to come

- Challenges: a few examples within a restricted setting showing the complexity of the language phenomena that can be involved
- Selected algorithmic ideas: domain adaptation, modeling label trajectories, and collective classification
- Connections to other areas: politics, economics, etc.

“Easy” case: polarity classification

Consider classifying a subjective text unit as either positive or negative.

- Example: “The most thoroughly joyless and inept film of the year, and one of the worst of the decade.” [Mick LaSalle, describing *Gigli*]
- One application: summarizing reviews (thumbs up or thumbs down)

Can’t we just look for words like “great” or “terrible” ?

Yes, but ...

... learning a sufficient set of such words or phrases is an active challenge. [Hatzivassiloglou&McKeown ’97, Turney ’02, Wiebe et al. ’04, and more than a dozen others, at least]

Can’t we just look for words like “great” or “terrible”?

Yes, but ...

... in a small-scale human experiment [Pang, Lee, & Vaithyanathan ’02]:

	Proposed word lists	Accuracy
Human 1	Positive: dazzling, brilliant, phenomenal, excellent, fantastic Negative: suck, terrible, awful, unwatchable, hideous	58%
Human 2	Positive: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting Negative: bad, cliched, sucks, boring, stupid, slow	64%
Statistics-based	Positive: love, wonderful, best, great, superb, beautiful, **still** Negative: bad, worst, stupid, waste, boring, **?**, **!**	69%

Can’t we just look for words like “great” or “terrible”?

Yes, but ...

- This laptop is a great deal.
- A great deal of media attention surrounded the release of the new laptop.
- This laptop is a great deal ... and I’ve got a nice bridge you might be interested in.

- This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.

Can’t we just look for words like “great” or “terrible”?

Yes, but ...

- She ran the gamut of emotions from A to B. [Dorothy Parker, describing Katharine Hepburn]

- Read the book. [Bob Bland]

State-of-the-art methods using bag-of-words-based feature vectors have proven less effective for sentiment classification than for topic-based classification. [Pang, Lee & Vaithyanathan ’02]

We’ll briefly look at a few particularly interesting algorithmic ideas that have been applied to sentiment analysis.

Underlying theme: relationships (between domains, items, and/or labels)

Domain adaptation

We have already alluded to the fact that certain sentiment-related indicators seem domain-dependent.

- “Read the book.”
- “Unpredictable”: good for movie plots, bad for a car’s steering [Turney ’02]

In general, sentiment classifiers (especially those created via supervised learning) have been shown to often be domain dependent [Turney ’02, Engström ’04, Read 05, Aue & Gamon ’05, Blitzer, Dredze & Pereira ’07].

This holds even though the set of categories (e.g., 1-5 stars) is generally considered constant across domains.

Structural correspondence learning

Blitzer, Dredze, & Pereira ’07 [adapting Blitzer, McDonald, & Pereira ’06, Ando & Zhang ’05]

Book reviews **(source domain: labeled)**	**Kitchen appliance reviews** **(target domain: unlabeled)**
(−) terrible and predictable	terrible leaking thing
(−) absolutely terrible	terrible, always leaking
(+) great	truly terrible
(−) so predictable it’s terrible	OK

Structural correspondence learning

Blitzer, Dredze, & Pereira ’07 [adapting Blitzer, McDonald, & Pereira ’06, Ando & Zhang ’05]

Book reviews
(source domain: labeled)

(–) terrible and predictable

(–) absolutely terrible

(+) great

(–) so predictable it’s terrible

Kitchen appliance reviews
(target domain: unlabeled)

terrible leaking thing

terrible, always leaking

truly terrible

OK

1. Choose *pivot features* that are relatively frequent in *both* domains (need sufficient data on them)and have high mutual information with the source labels (choose words like “terrible”, avoid stopwords like “the”)

Structural correspondence learning

Blitzer, Dredze, & Pereira ’07 [adapting Blitzer, McDonald, & Pereira ’06, Ando & Zhang ’05]

Book reviews
(source domain: labeled)

(−) terrible and predictable

(−) absolutely terrible

(+) great

(−) so predictable it’s terrible

Kitchen appliance reviews
(target domain: unlabeled)

terrible leaking thing

terrible, always leaking

truly terrible

OK

2. Encode correlations between non-pivot features and pivot features in vectors and create a new “feature projection” accordingly (via the SVD).

Idea: learn target-domain features that act like source-domain features.

On the nature of the labels

Another characteristic of many sentiment-analysis problems is that there are interesting relationships between class labels.

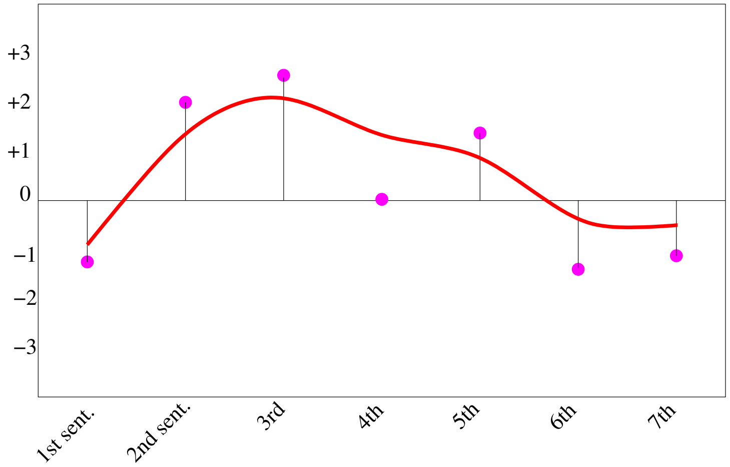
- “1 star” is more like “2 stars” than like “5 stars”
- “high” force of opinion is more like “medium” than “none”
- the “positive”-to-“negative” continuum may not contain the class “objective” ($\neq$ “mixed”)

(cf. degrees of relevance in IR)

Approaches include: (ordinal) regression [Wiebe, Wilson & Hwa ’04], classifier stacks [Koppel&Schler ’05], metric labeling [Pang&Lee ’05], and generalizations of metric labeling [Goldberg&Zhu ’06].

Sentiment flow

Mao and Lebanon ['07]: combine a sequence of local sentiment judgments into a (smoothed) sentiment flow; then, apply nearest-neighbor classification to the flows to make document-level decisions.



- *Isotonic conditional random fields* impose monotonicity constraints that are consistent with sentence labels' ordinal relationships.

Implements intuition of characterizing a document by the "flow" of its discourse.

Collective classification

Relationships between items can be a rich source of information about for performing classification or regression on the items.

- Nearby sentences can share the same subjectivity status, subjective or objective [Pang&Lee ’04]
- Mentions separated by “and” usually have similar sentiment labels; those separated by “but” usually have contrasting labels [Popescu&Etzioni ’05, Snyder&Barzilay ’07]; similar reasoning holds for synonyms and antonyms [Hu&Liu ’04]
- In some domains, references to other speakers generally indicate disagreement [Agrawal et al ’03, Mullen&Malouf ’06, Goldberg, Zhu & Wright ’07] (cf. Adamic&Glance [’05])

Speaking of which ...

A matter of debate

Thomas, Pang, and Lee ['06]:

Given: transcripts of U.S. Congressional floor debates

Goal: classify each *speech segment* (uninterrupted sequence of utterances by a single speaker) as supporting or opposing the proposed legislation

Important characteristics:

1. Discussion context: some speech segments are responses to others
2. Very wide range of topics: flag burning, the U.N., "Recognizing the 30th anniversary of the victory of U.S. winemakers at the 1976 Paris Wine Tasting"
3. Presentation of evidence rather than opinion (see also [Kim&Hovy '06])

   "Our flag is sacred!": is it pro-ban or contra-ban-revocation?

4. Ground-truth labels can be determined automatically (speaker votes)

Using discussion structure

Two sources of information (details suppressed):

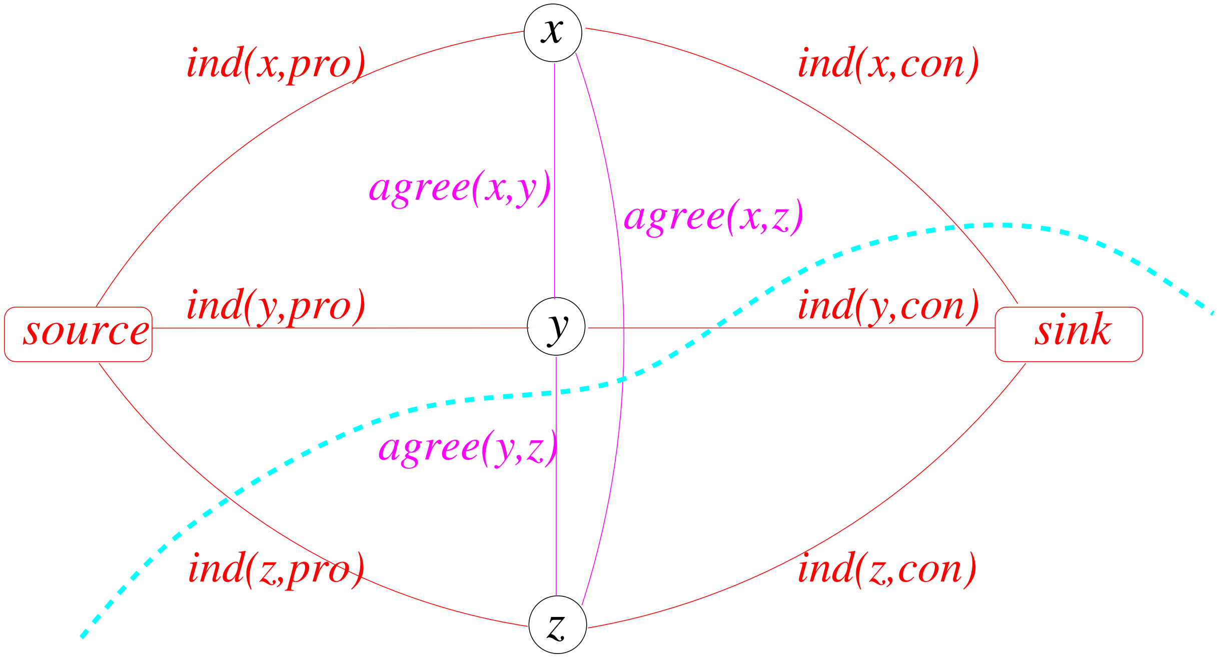
- An individual-document classifier that scores each speech segment x in isolation
- An *agreement classifier* for *pairs* of speech segments, trained to score by-name references (e.g., “I believe Mr. Smith’s argument is persuasive”) as to how much they indicate agreement

Optimization problem: find a classification c that minimizes:

$$\sum_{x} \textit{ind}(x, \overline{c}(x)) + \sum_{x,x':\, c(x) \neq c(x')} \textit{agree}(x, x')$$

(the items’ desire to switch classes due to individual or associational preferences)

A “mitosis” encoding



When the edge weights are non-negative, *network-flow techniques find the min-cost cut efficiently and exactly*.

Negative agreement weights can be handled via pre-processing heuristics [Bansal, Cardie & Lee ’08]

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The “[alteration of] the citizen-government relationship” [Shulman & Schlosberg 2002]

“The transformation of American politics” [*The New York Times*, 2006]

“The End of News?” [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives online.

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The "[alteration of] the citizen-government relationship" [Shulman & Schlosberg 2002]

"The transformation of American politics" [*The New York Times*, 2006]

"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives online.

One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.

Broader implications: politics

The on-line availability of politically-oriented documents, both official (e.g., full text of laws) and non-official (e.g., blogs), means ...

The "[alteration of] the citizen-government relationship" [Shulman & Schlosberg 2002]

"The transformation of American politics" [*The New York Times*, 2006]

"The End of News?" [*The New York Review of Books*, 2005]

More opportunities for sentiment analysis!

Recall: people are searching for political news and perspectives online.

> One ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end.
>
> — George Orwell, "Politics and the English language", 1946

Broader implications: economics

Consumers *report* being willing to pay from 20% to 99% more for a 5-star-rated item than a 4-star-rated item. [Source: comScore]

But, does the polarity and/or volume of reviews have measurable, significant influence on actual consumer purchasing?

▷ Implications for bang-for-the-buck, manipulation, etc.

From a large and lively body of relevant economics literature [Pang & Lee survey, '08], here is a sample quote:

> ...on average, 3.46 percent of sales is attributable to the seller's positive reputation stock. ... the average cost to sellers stemming from neutral or negative reputation scores is $2.28, or 0.93 percent of the final sales price. If these percentages are applied to all of eBay's auctions [$1.6 billion in 2000 4Q], ... sellers' positive reputations added more than $55 million to ... sales, while non-positives reduced sales by about $15 million. [Houser&Wooders '06]

On the quality of reviews

Real/perceived economic impact ⇒ **incentives for fake reviews**

(and suppression of reviews, misattribution of opinions, etc.)

Detection of low-quality reviews [Kim et al. '06, Zhang&Varadarajan '06, Ghose&Ipeirotis '07, Jindal&B. Liu '07, J. Liu et al '07]

Determination of authoritative reviews or sources

Development of algorithms for low-regret product selection from recommendations under adverse conditions, such as only a *constant* fraction of recommenders being honest, and moreover sub-divided into market segments with different tastes [Awerbuch&Kleinberg '05, "collaborative competitive learning"]

Read the book

Even more information about applications, research directions, connections to other fields, and other matters is available ...

Opinion Mining and Sentiment Analysis

Bo Pang and Lillian Lee

www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html

135-150 pp, 330+ references, in print by next week, we’ll also post the pdf when ready

Includes bibliographies, pointers to datasets, more snazzy examples, etc.

Our revels now are ended

We have seen that sentiment analysis...

...has many important applications

...encompasses many interesting research questions

...extends to many areas

You can start working on sentiment analysis right now!

Publicly available datasets include those at http://www.cs.cornell.edu/home/llee/data .

Our revels now are ended

We have seen that sentiment analysis...

...has many important applications

...encompasses many interesting research questions

...extends to many areas

You can start working on sentiment analysis right now!

Publicly available datasets include those at http://www.cs.cornell.edu/home/llee/data .

This is such stuff as dreams are made on!