Lecture #3 1 handout.
Correlates of review helpfulness
The making of a research project paper. (www 2009)

(a) is continuing on the themes we were discussing last time.
(b) is a project on this topic that I was involved in.

I want to discuss it not just b/c it's on-topic, but also to give some insight into the research process,

"starting w/ a little question, vs. starting w/ a big question"

Our particular case was this:

Sometimes people are resistant to starting w/ little questions.
And then are people who protest how incremental the work seems to be..."these days..."

On the other hand, I think that for younger grad students, too much focus only on "big" questions can be a bit paralyzing.

Also, I don't know of any conference where everything is a breakthrough, the pace of science just usually isn't that fast.

"The joint progress of the research community tends to go at a rate of a little bit @ time." (These years where there were lots of breakthroughs proceed...)

Anyway, the version of the story I'll tell here is a bit different from how I gave talks on this work.

O.K., last time we did @ that exercise where you guys guys write down some factors you personally look for in deciding whether a review is helpful.

Now let's look @ some factors that have been considered in the literature.

---

First example: Getherbank '09. "Helpfulness... the measure of enticing", Table 3 (p. 258)

(use reference to, so people can see pub info (screen width/ resolution permitting))

---

I wanted to show this table b/c it divides the description nicely into:

- the "general concept" of each feature vs. how it was implemented:

  Also, the thing is super-convenient, (so, in terms of presentation, this is nice)

- These are what she tried, not all of which turned out to be significant in the end.

An: not @ all a she was not the first to introduce many of these link to survey of previous work on the course homepage.
4th column ("Explanatory/Justification" is perhaps easiest to go through?)

Observations:

Intrinsic quality: both the review and, in (5)-(9), the reviewer
(1-3) "trick" commonly used for coverage/objectivity of a review
Note: citation to [6] in probably wrong.

(4, 10, 11)

extreme \rightarrow different

(12) [this is like the Gilbert Kamalakies found some of their interviews
commonalities saying: "A completely unique review wouldn't serve any purpose"
p, perhaps best to consider extreme wth other review ratings

contextual row: as on one of our examples last time, sometimes helpfulness
        depends on other reviews, if you
        seems to be a nod to the idea that relationship to other reviews
        can matter.
        But the actual features in this category don't seem that linked to
        that concept.

Looking at the 2011 version of Ghose Tjerjost's IL table 1.

as another example of features that have been tried.

\textit{Note:} some of these features were used to predict sales rank, so things like
in the individual (i.e., helpfulness rating not used in helpfulness prediction)

more reviewer characteristics
more readability features (5th row)
objectivity measures, using same idea that the product description can be treated as

\textit{definitely objective material} (as it happens, to train the subj. detector)

comment about descript. could be to get confidence for subj. prob, or \textit{<smart\ I can't recall>}

And other studies look at similar features, as well.

*Self:* these were the kinds of things that were in the air

Any comments/questions, esp. re: people's potential pilot studies for A1?
recommendations

one q.: finding a very specific feature that few people have talked about, like the fault in
the review example we saw last time?
O.K. now I want to talk about the project that a group of collaborators and I did on this topic, while such ideas were "in the air."

+ Shaw dreamware entry for today, to see names.

- at the time we started this, which was early 2008, there was a few papers looking at predicting review helpfulness.

And, Jim Kleinberg, who's a professor of computer science, was working w/ Georgi Kassides, his postdoc, whose PhD was in sociology w/ Duncan Watts.

Together, they were fooling around with a crawl of Amazon reviews.

Meanwhile Christian was a first-year grad student who'd taken my grad course that fall, and was interested in working with me.

Jon: Georgi asked us if we'd like to join in on testing for interesting things to do w/ the review text (e.g., post-prediction @ Cornell on sentiment analysis).

We thought we'd start w/ the "helpfulness prediction," as a potentially interesting variant of sentiment categorization (when you'd try to predict the # of stars a review gives based on the text.)

"helpfulness prediction," "real data": "x out of y found this helpful."

= regression? or (binary) classification?

we ended up going w/ this, choosing 60% as the gold-label decision threshold.

[ based on dataset stats, although it will turn out this didn't matter b/c we tuned away from classification.]

@ problem: text features not adding to accuracy (vs. meta-features)

we tried a bunch of things, as you can imagine, although we weren't as sophisticated.

didn't try anything as sophisticated as what's on your handout.

so do you stop there and go, oh well?

Sometimes the answer is yes! And maybe not such an interesting prob anymore...

But, we thought at least, let's try to analyze errors a bit more. So, block @ the data, try to figure out what's wrong.

reaction: any of these interesting correlates/bottom text + meta-features?
(b) many duplicate reviews (removed via plagiarism detection techniques)
writing this claim may be useful for your projects.

due to cross-posting by Amazon among editions, we recrawled all the data
This is a difficult decision to make—-to throw out all your data and start over.
But sometimes you gotta do it.

All the same, sometimes it's a bad idea.

(c) went through lots of correlation plots...

one point: someone noticed a correlation between agreement w/ avg. helpfulness.

--- but maybe everyone simply has the same opinions?

idea: look at places where there's polarization (where people don't all have same opinion). (slide 37 of 57)

--- but maybe this just looks like it's not textual, and really there's a correlation in the text.

show plagiarized reviews.

natural experiment: highly similar text pairs.

How do you "sell" an "obvious" result?