9/2/14
lec. 3

(make copies) { 2 handouts:
outline (the 'red parts' @ top of each sheet)
some helpfulness cues from other papers
Set up Firefox tabs for figs, talk slides for other figures

fake, joke, sarcasm

Review "quality"/"helpfulness":

reviews considered independently

~~reviews or pot~~ potentially also considering context

(diff. connotations)

[intro] (1) prediction ~~in and of itself~~: ~~useful~~ some features, ~~techniques for biased labels~~. (features helpful for projects)

[focus] (2) ~~predic~~ a lens for studying social influence ~~/interesting techniques~~.

~~(3) for.~~

(naturally, will talk about techniques)

We talked about review helpfulness last time (~~and thought~~ remember we wrote down and evaluated some features of helpfulness).

Helpfulness ratings ~~is~~ are being used as a navigation tool by sites like Amazon, since so many reviews

When these evaluations are produced by users, you have social navigation.

[Is this from the GK reading?]

You could ask, why do this "socially" if it ~~could~~ can be done automatically.

And needless to say, there's been quite a bit of ~~energy~~ research energy devoted here.

W.r.t. considering this to be a prediction problem, there are a # of options:

- the "given" label is "x out of y" found this helpful.

→ regression problem?
binary classification (threshold on ratio x/y)

- what features?

Otterbacher (have handout; also project using desktop 1)

→ First example: Jahna Otterbacher '09 'Helpfulness'... ~~as a measure of online~~, Table 3 (pg 958)

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

numerous better.

I wanted to show this table b/c it divides the description nicely into ~~"~~ ~~"~~ the "~~to~~ general concept" of each feature vs. how it was implemented. ~~Some of these~~ You might want to choose a few for your pilot studies

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.

And, ~~not @ all a~~ she was not the first to introduce many of these ~~...~~ in the context of review mining.

Link to survey of previous work on the course homepg.

4th column (; "Explanation/Justification" is perhaps easiest to go through?)
Observations?

(13, 14, 16, 17) = readability

intrinsic quality: ~~both the review and ... (s) (t), the reviews~~

(1-3): "trick" "commonly used for coverage/objectivity of a review
Note: citation ~~to~~ to [6] is probably wrong.

q's: should one expect 'extreme' to be ↑ or ↓ correlated? (can do regression/classification w/out hypothesis)

just talk

[4], [10, 11]
extreme → different

q: perplexity?

let p_1, p_2 two distributions identified with two samples s_1, s_2. (We'll blur the distinction). $H(p_1, p_2) = -\sum_i p_1(x_i) \log(p_2(x_i))$. p_1 is considered "true", p_2 the "alternate code". ~~Perplexity is 2^H, and converts to bits as unit.~~

(12) ~~similar~~ ~~same~~ commonalities (this is like ~~that~~ Gilbert & Karahalios found some of their interviewees saying: "A completely unique review wouldn't serve any purpose"

→ perhaps best to consider extreme wrt other review ratings

q's about this

[5-9]: ~~prope~~ other properties [some authorial, some may be for 'smoothing from past']
contextual row: " ~~as we saw w/ our example 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.~~

~~note that Amazon didn't use to sort reviews by helpfulness, we think.~~

⟼ Looking @ the 2011 version of Ghose & Ipeirotis '11, table 1. as another example of features that have been tried'.

(* note: some of these features were used to predict sales rank, so things like ~~in the reviewer's~~ helpfulness rating not used in helpfulness prediction)

more reviewer characteristics

more ~~fea~~ readability features ~~(Lst)~~ (5th row)

interest in "SMOG", the "Simple Measure of Gobbledygook". Explained constants as result of fitting to data.

(*) subjectivity measures, using same idea that the product description can be treated as ~~ob~~ definitely objective material (as it happens, to train the subj. detector)

~~comment about devprob: ... confidence for subj. prob, or <sthg short I can't recall>~~

And other studies look @ similar features, as well.

devprob: relate to the UBC slashdot/thread visualizer we talked about before.

~~So, ... 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?~~

Discussion re: fake reviews, sarcasm, etc: noted that machine classificat'n results are surprisingly high.
q: are fake reviews 'product-dependent'?

How well do such features (esp text-only features) do?

J. Liu et al, EMNLP-CoNLL

Re-annotate reviews according to own written standard b/c of biases:
~~Mismatch b/c~~ - bias toward eval'ing as 'helpful' (half of 23k sample has >90% helpful)
- some reviews don't get eval'd (~~early bird bias~~) rich-get-richer)
(temporal: earlier get more.

Are there social factors behind mismatch?

(2) ~~Study~~ Using quality/helpfulness as a lens on social influence
→ focus of this course

[Sipos, Ghosh, Joachims WWW '14]: (mis)-~~ordering of~~ ranking by community topic
(to some degree - also proprietary Amazon factors; tau = .84 for helpfulness votes vs. actual Amazon ranking
revs w/ ≥ 10 votes) ~~polarity~~

- "true" quality by "final" {future} ranking
4 mos in future
Attempts to avoid 'self-fulfilling prophecy' technique for biased labels
- effect on helpfulness vote, and on whether they vote.

stress the daily-snapshot idea

<Sent, Fig 3> → correction of mis-ranking
point out 'high' rank = 0
(color = prob of next eval coming in as "helpful")

q: re: how to measure participation decision:
- paper assumes "constant # of pageviews (for each position) in each period between snapshots".
Doing relative # of pageviews then should be OK. (take it or leave it).
- also a measure of 'participation' to handle smoothing of sparse data.
→ motivated to participate to correct mis-ranking

Note a point of the message (of paper): this is an 'opposite' effect to how clicks on rankings of web-search results accumulate to the already-highly ranked.

- [non-ranking setting]
- overall, ~~seeing other~~ prior ⊕ w/ ~~rates~~ increased prob. of ⊕ eval. ⊖: not much effect.

Muchnik, Aral, Taylor, Science 2013] · manipulation study ("~~fixing~~ sidestep true quality") opt-in

comment rating, so ranking is not a factor (this is not Amazon).

"the Web site performed" [not by the experimenters, therefore]

Artificially seeded the comments' initial vote w/ a ⊕ or ⊖ @ ~~creation~~ eval, 101K comments altogether

$\underbrace{4049 \quad 1942}_{\text{reflect natural (observed) prior...}}$

lots of "no initial eval" as control.

most demonstrative <Fig 1.c> ~~(don't show for time, just make ref. on webpg?)~~

explain bars are confidence interval

~~data analyzed~~

so perhaps not so relevant in terms of students being able to use

- note asymmetric effect: herding for ⊕.
- use the "no artificial 1st eval" as baseline.

[Danescu-Niculescu-Mizil, Kossinets, Kleinberg, Lee WWW '09] : effect of conformance to group opinion
cultural diffs in

: natural experiment: "plagiarized" reviews to sidestep true quality

clipped slides

explain what grey bars are (and expect, since for 5-star avg no one can be +4 stars away)

not just "more (+) = more helpful", but for same degree of deviation, better to be (+)

note that wrt one ~~review~~ version of a plagiarized review being more appropriate, the example plagiarized pair show the latter to be more helpful, and our studies showed no sig diff of helpfulness for @ the pair.

[Cheng, Danescu-Niculescu-Mizil, Leskovec ICWSM '14] = effect of evaluations on the author
: propensity matching; ~~to~~ 'natural' experiment,
but requires measuring text quality to
pair "similar" reviews.

post quality: evals drop significantly after a neg. eval.
rise is not significant after pos. eval.

(other controls in pairing: # words, # posts written, general ~~helpfulness~~ quality evaluations previously).

fig 4 in survey.

* = sig diff.

q: maybe the people who ~~to~~ got ⊖ eval were trolls, and so were motivated
to write worse posts b/c that's what ~~they~~ trolls do?

- interesting that people who received "no" feedback most likely to drop out