The effect of wording on message propagation: *Topic- and author-controlled* natural experiments on Twitter

Chenhao Tan, Lillian Lee, Bo Pang

Cornell University, Google

How to get messages across more effectively?

What factors determine the success of messages?

Important factors [Milkman and Berger, 2012; Romero et al. 2013; Suh et al. 2010; etc]

- Characteristics of the author, author's social network
- Message topic
- Message timing

How to get messages across more effectively?

- Find a good topic [Guerini et al. 2011]
- Become influential or find influential users to help spread [Kempe et al. 2003]

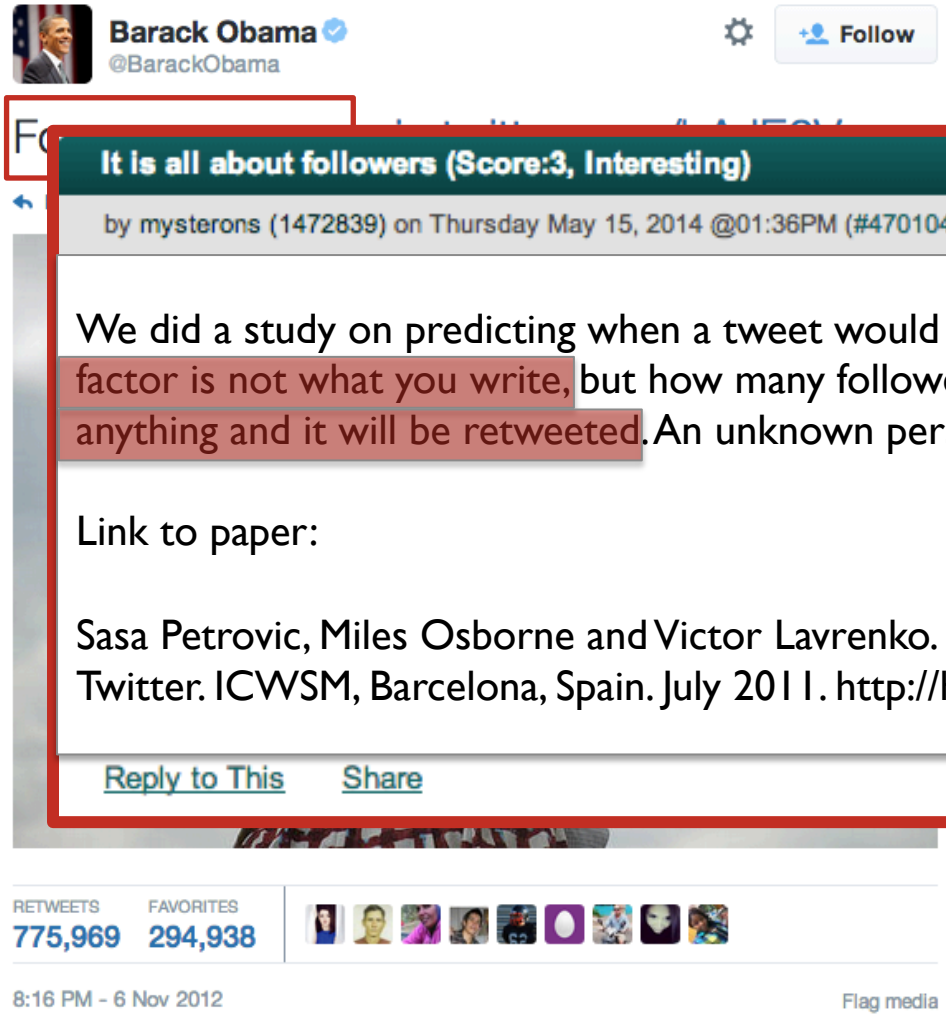
How to get messages across more effectively?

- Find a good topic [Guerini et al. 2011]
- Become influential or find influential users to help spread [Kempe et al. 2003]
- Improve the quality of the content
  - Image [Isola et al. 2011]
  - ***Wording***

    humor, informative, emphasize certain aspects

Revisit the example: Does wording actually matter?

Revisit the example: Does wording actually matter?



How can we focus on the effect of wording?

Add more control to better understand the effect of wording

- Author control
  - Obama *vs.* me
- Topic control
  - Presidential election *vs.* this talk

What if BarackObama had posted about re-election using a different wording?

e.g. "4 more years to prove that we can!"

The same users post multiple tweets on the same topic

Topic- and author-controlled pairs

Topic- and author-controlled pairs are common!

- *2.4 Million* topic- and author-controlled tweet pairs
  - 1.77M differing in more than just spacing
  - 632K whose difference was only spacing

More cleaning up is required for natural experiments!

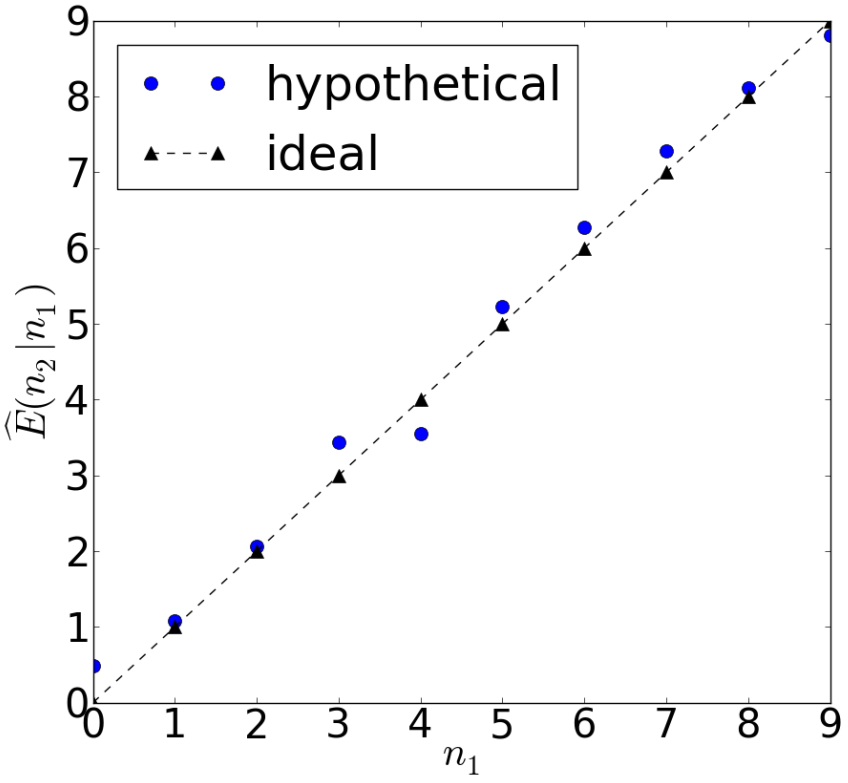
- Timing can matter (thankfully, Twitter doesn't re-rank posts, but presents strictly in chronological order)
  - The first one may enjoy a first-mover advantage
  - The second one may be preferred as the updated one

- Number of followers also has complicated effects

Use *identical pairs* to find an "ideal" setting

- Notation
  - n_1 : number of retweets for the first tweet
  - n_2 : number of retweets for the second tweet
- Difference between n_1 and n_2

$$D = \sum_{0 \le n_1 < 10} |\widehat{E}(n_2|n_1) - n_1|$$

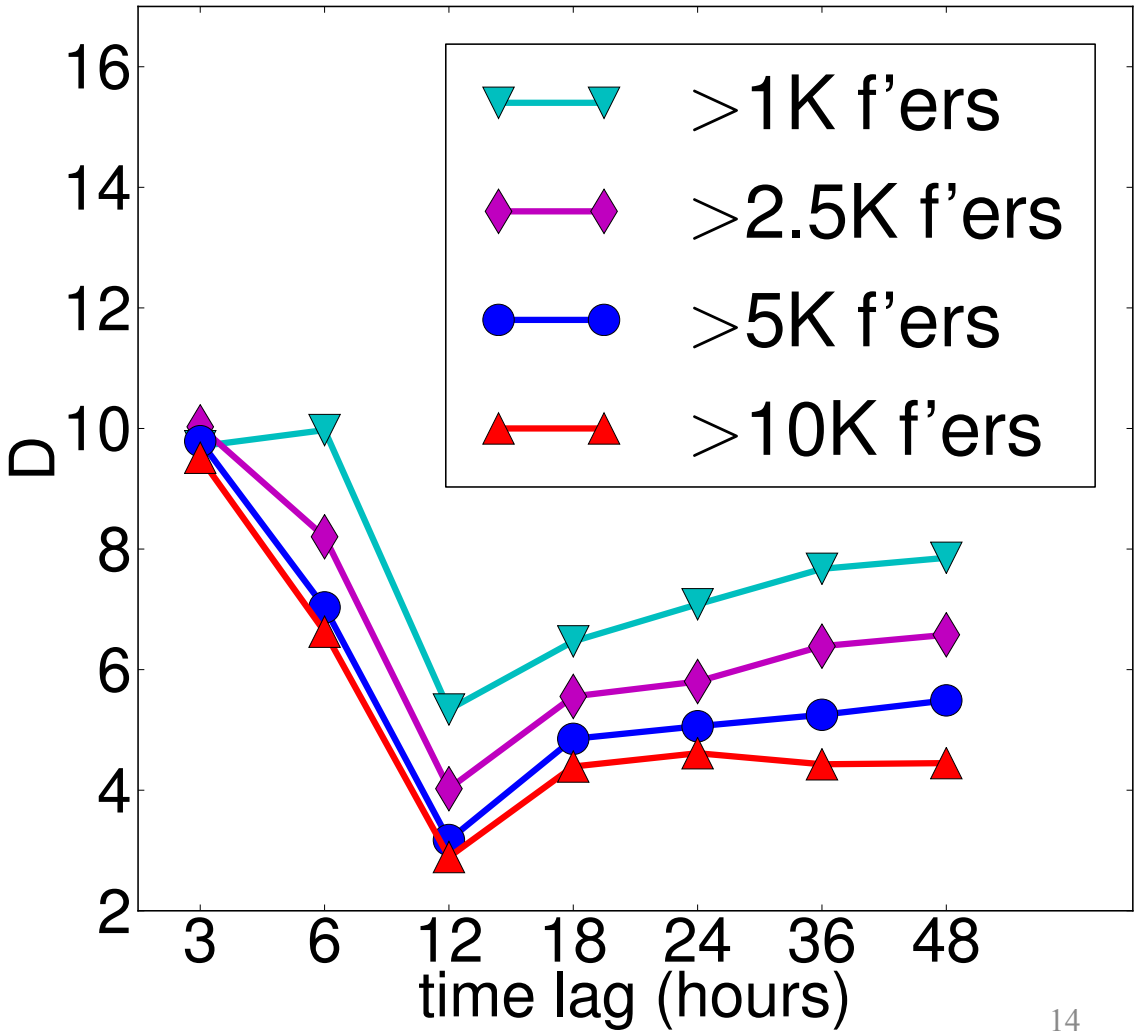


Use *identical pairs* to find an "ideal" setting

$$D = \sum_{0 \le n_1 < 10} |\widehat{E}(n_2|n_1) - n_1|$$

As time lag increases, *D* decreases as we get more data and then increases

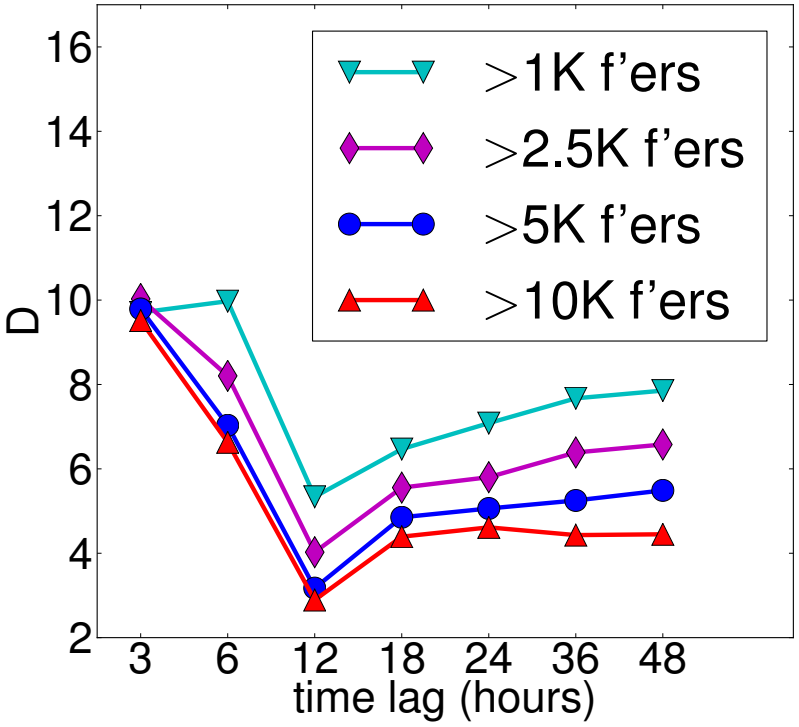
As number of followers increases, *D* decreases

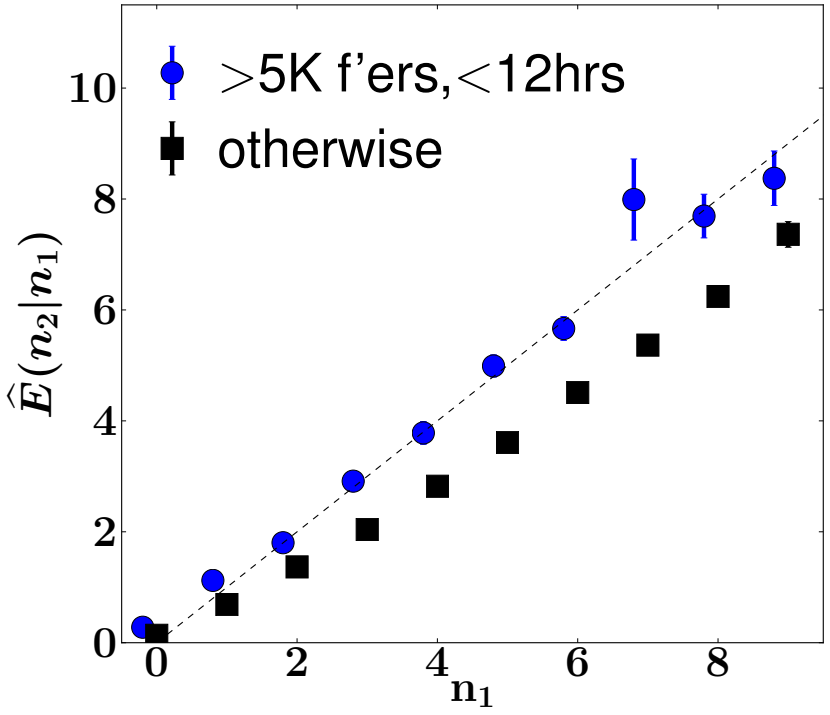


The ideal setting found through *identical* pairs:

users who have more than 5K followers

two tweets are posted within12 hours





More filtering

- Ideal setting: >5K followers, <12 hours
- Non-trivial textual changes
  - Similarity below median to avoid typos, etc
- Significant changes in retweet numbers
  - Take top 5% and bottom 5% in terms of $n_2 - n_1$
- Limit the number of pairs by an author to 50

This brings us 11K topic- and author- controlled pairs for natural experiments!

Does wording matter?

Wording does not matter

Humans should not be able to tell which one in a pair was retweeted more

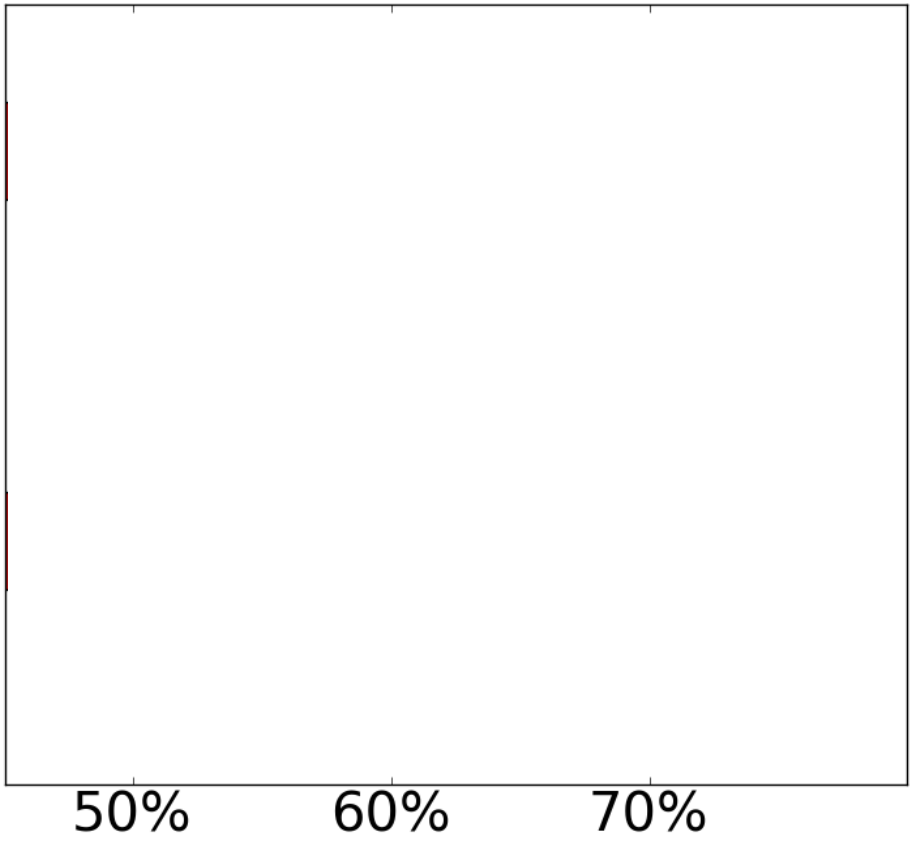
Humans can tell which one in a pair was retweeted more (accuracy > 50%)

Wording matters!

Can humans tell which tweet will be retweeted more?

- Randomly sample 100 pairs
- 20 pairs a task on Amazon Mechanical Turk
- 39 judgments for each pair

Can humans tell which tweet will be retweeted more?



Average accuracy for each labeler: 61.3%

Accuracy of the majority label for each pair: 73%

Predict which tweet will be retweeted more within a pair

- Cross validation experiments: 11K topic- and author-controlled pairs (5-fold cross validation)
- *Heldout* experiments: 1.8K topic- and author-controlled pairs from a different group of users that have never been used

  (Only used once, 6 days before submission!)

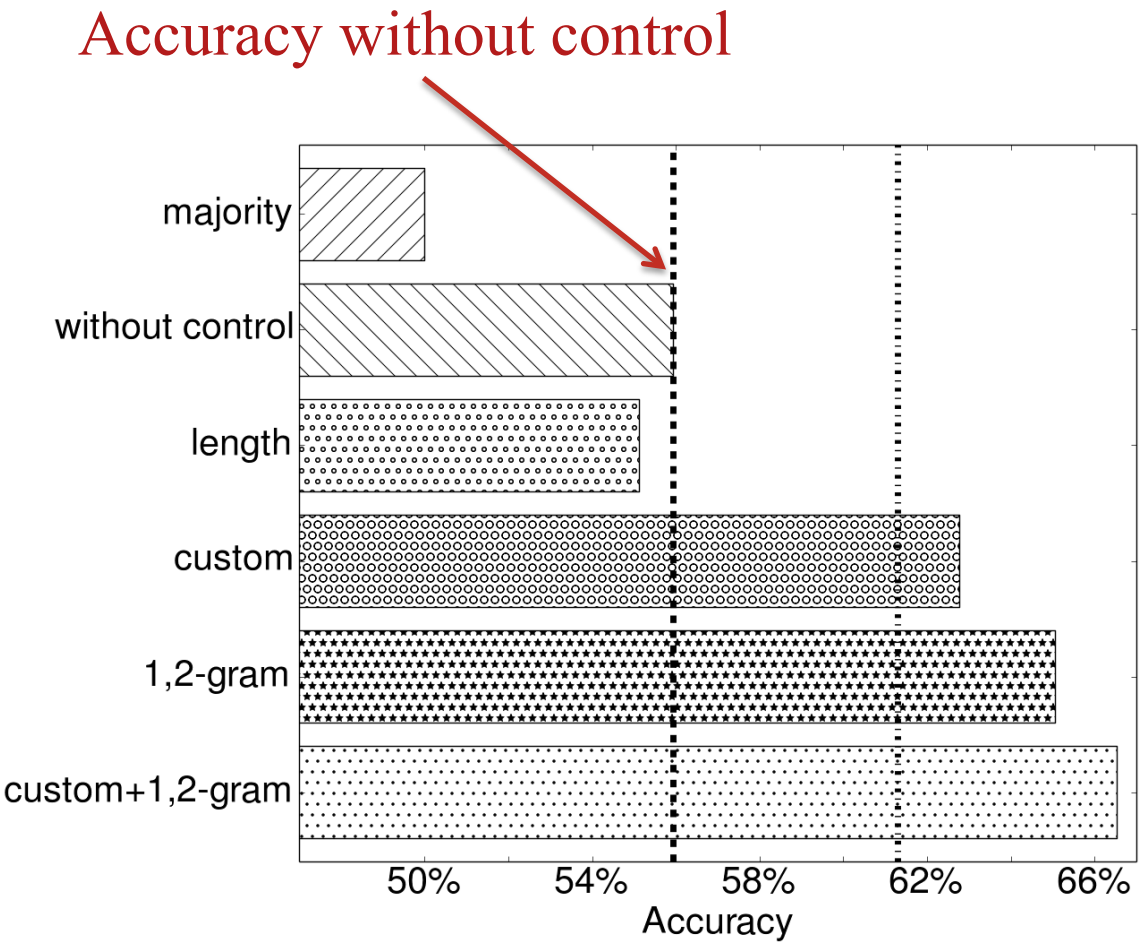
Predict which tweet will be retweeted more within a pair

- Features
  - Custom features that we proposed: lexicons, informativeness, language model features, etc (39 features)
  - Bag of words: unigram+bigram (7K features)
- Approach
  - Take the difference between features for two tweets in a pair after linear normalization
  - Logistic regression

Predict which tweet will be retweeted more within a pair

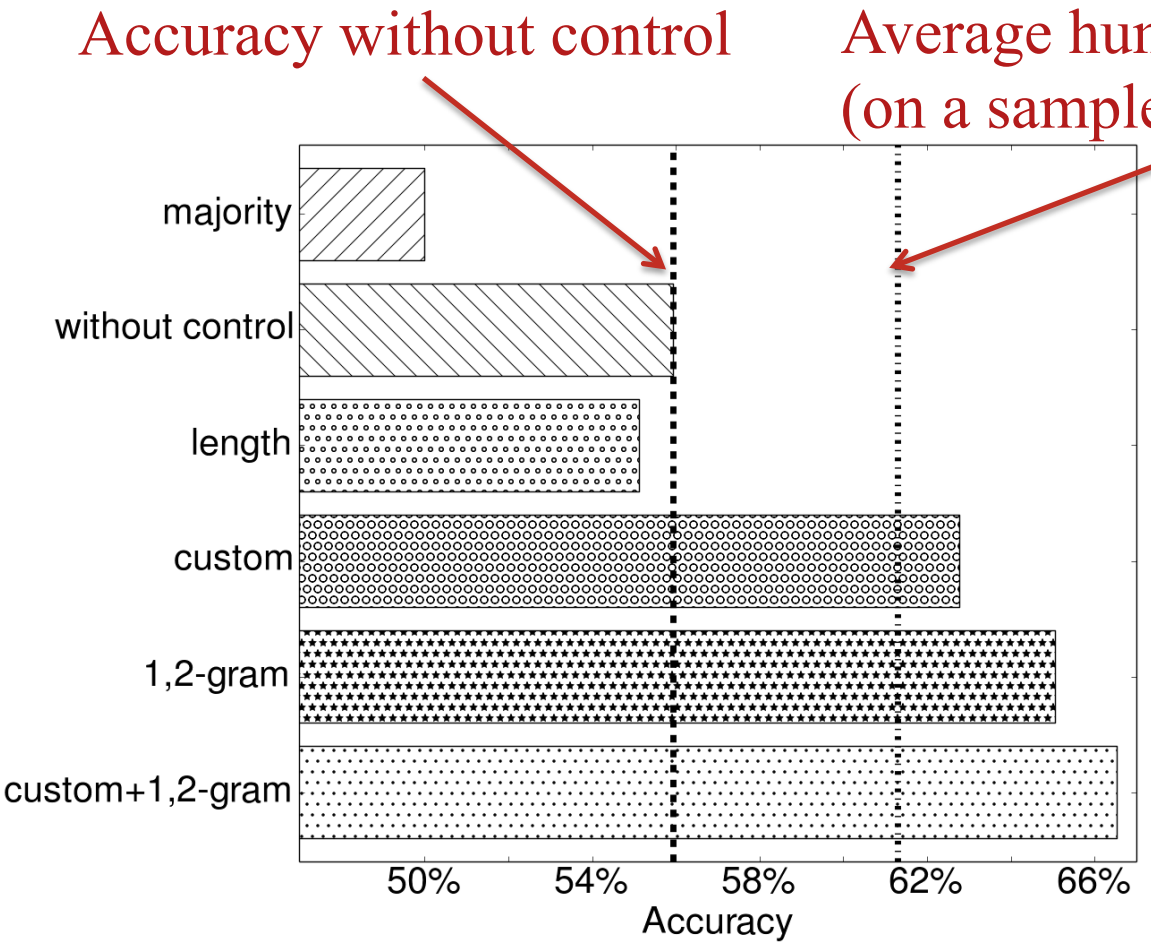
- A strong baseline
  - A classifier to distinguish10K most retweeted unpaired tweets from 10K least retweeted unpaired tweets
  - Use bag-of-words features, [number of followers and timing]
  - Cross validation accuracy 98.8%

Cross-validation performance: is control necessary?



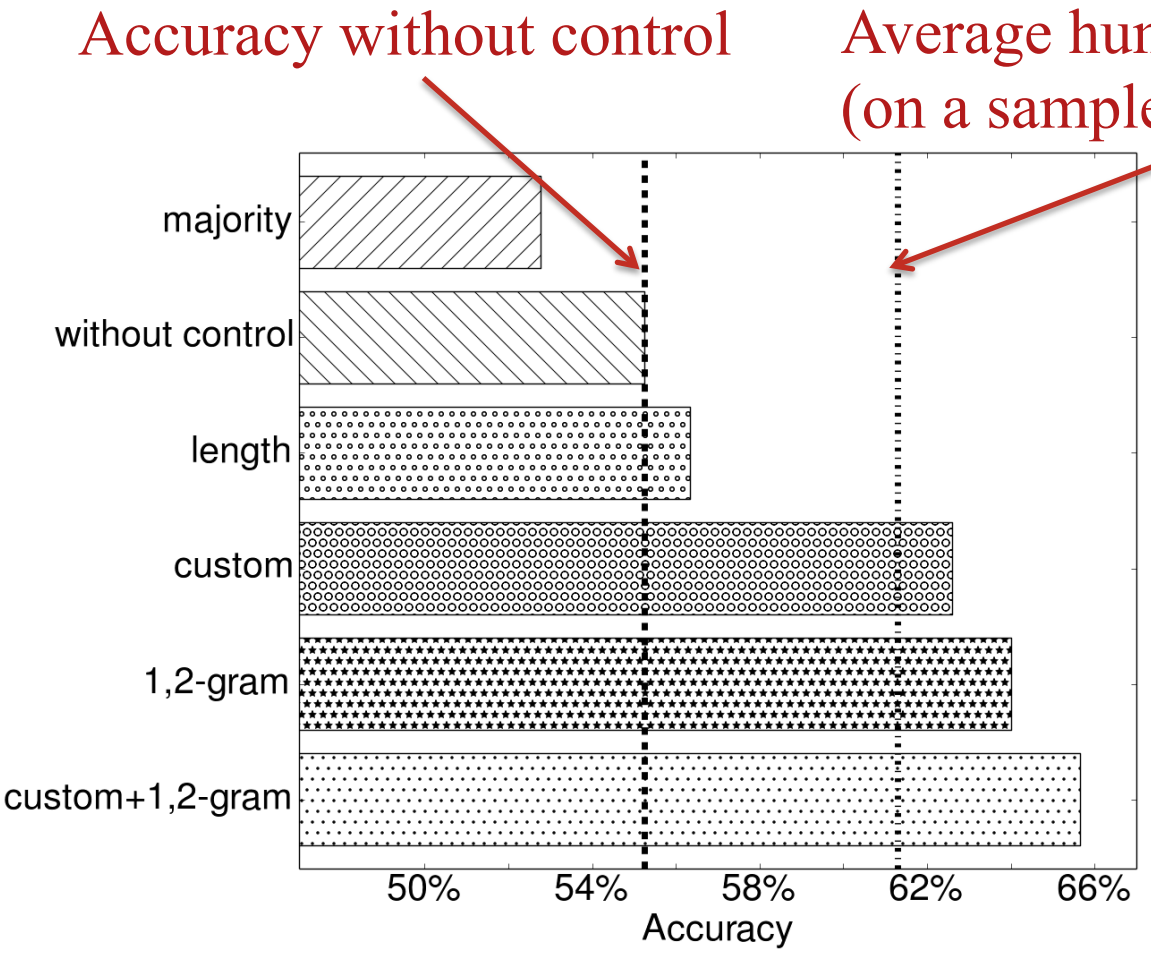
- Best method outperforms the baseline by more than 10%

Cross-validation performance



- Best method outperforms the baseline by more than 10%
- Custom does pretty well by itself, and outperforms average human accuracy
- Adding custom improves bag-of-words

Fortunately, same results hold in heldout data



- Best method outperforms the baseline by more than 10%
- Custom does pretty well by itself, and outperforms average human accuracy
- Adding custom improves bag-of-words

Should we conform to community norm?

- Train language models using non-paired tweets
- Compute unigram, bigram language model score

  higher score = closer to twitter language
- Test whether more retweeted tweets have a larger score

Be like the community (conformity)

- Train language models using non-paired tweets
- Compute unigram, bigram language model score

  higher score = closer to twitter language
- Test whether more retweeted tweets have a larger score

	Effective?
Twitter unigram language model	$p < 0.001$
Twitter bigram language model	$p < 0.001$

Should we maintain personal style?

- Train language models using history of each person
- Compute unigram, bigram language model score

  higher score = closer to personal history
- Test whether more retweeted tweets have a larger score

Be true to yourself

- Train language models using history of each person
- Compute unigram, bigram language model score

  higher score = closer to personal history
- Test whether more retweeted tweets have a larger score

	Effective?
Personal unigram language model	$p < 0.001$
Personal bigram language model	———

Take away

- We used topic- and author-controlled pairs to show that wording matters!
- Average human is not perfect in telling which is better; computers can do better
- Controlling topics and authors can improve predictive performance significantly over an approach without control

Thank you & Questions?

- Data

http://chenhaot.com/pages/wording-for-propagation.html

- Demo

http://chenhaot.com/retweetedmore

- Quiz

http://chenhaot.com/retweetedmore/quiz