Identifying Uninteresting Content in Text Streams

Omar Alonso, Chad Carson, David Gerster, Xiang Ji, Shubha Nabar
Microsoft

23 July 2010

SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation

bing

Motivation

- Lots of people use Twitter on a daily basis
- Do other folks care about the content?
- Relevance and text length
  - What can we get from 140 chars?
- Tier index

The question

- How do we identify on the fly if a tweet is interesting or not?
- How do we define “interestingness”?

Our approach

- Crowdsourcing to label if a tweet is interesting or not
  - Only interesting to authors and friends
  - Possibly interesting to others
- 5 workers
- 5 tweets per task
  - Easy to read
- Two data sets
  - 980 tweets
  - 1791 tweets

Experiment design

Please label the following tweets according to the instructions below.

Please read the tweets below and mark them as "only interesting to author and friends" or "possibly interesting to others". "Only interesting to author and friends" means that only the author of the tweet (or someone who knows them) would find the tweet worth reading, while "possibly interesting to others" means that other people might find the tweet worth reading.

Examples of "*only* interesting to **author** and **friends**":

- Going for lunch with my friend.
- I love the Simpsons, but can't stand the musical episodes
- go read a book @m0delchiik
- Today, you hurt me once, shame
- Bout to go to work...

Examples of "*possibly* interesting to **others**":

- Another earthquake hits Haiti.
- No source, but allegedly #TonightShow staffers upset with Conan. http://tinyurl.com/ybts7sm
- Congrats to Kurt Warner on a heck of a career. Now make room for Mr Leinart.
- Leap shares decline after JPMorgan down

Tip: Some tweets will be hard to label. Because of how people use twitter, most tweets will be "only interesting to author and friends". Please try to be consistent.

Paying a little visit to the fam...=)	○ *Only* interesting to **author** and **friends**	○ *Possibly* interesting to **others**
@leevigraham Hi Leevi, any pointers for creating an xml sitemap using LG Better Meta v1.9.0 with a site built using structure?	○ *Only* interesting to **author** and **friends**	○ *Possibly* interesting to **others**
@TheBoobLady It is a eye opener!	○ *Only* interesting to **author** and **friends**	○ *Possibly* interesting to **others**
HORRIBLE DAY! I was suffering all day! I'm sick, sound like a nasal freak, I was in pain all day! =[	○ *Only* interesting to **author** and **friends**	○ *Possibly* interesting to **others**
@jasminebridgerx ah! goodnight monster, Happy latish birthday? xx lol sorry!	○ *Only* interesting to **author** and **friends**	○ *Possibly* interesting to **others**

bing

Careful with instructions

- Instructions v1
  - Interesting: specific information that people might care about
  - Not interesting: advertisements, opinions, and trivial updates about daily life
- Instructions v2
  - Only interesting to authors and friends
  - Possibly interesting to others

Clear instructions are important

Score	Initial Labels	Revised Labels	Change
Unanimous agreement: 0/5 or 5/5	30%	53%	+23%
Near-agreement: 1/5 or 4/5	32%	27%	-5%
Disagreement: 2/5 or 3/5	38%	20%	-19%
Total	100%	100%	

Figure 1. Clearer Instructions Yield More Agreement Among Workers (Experiment 1)

Experiment results

- Result was an “interestingness” score for each tweet (0/5 to 5/5)
- Data set #1
  - 40% of tweets scored a 0/5
  - 17% of tweets scored a 1/5
- Data set #2
  - 50% of tweets scored a 0/5
  - 19% of tweets scored a 1/5

Textual features

1. Presence of a hyperlink
2. Average word length
3. Maximum word length
4. Presence of first person parts of speech
5. Largest number of consecutive words in capital letters
6. Whether the tweet is a retweet
7. Number of topics as indicated by the “#” sign
8. Number of usernames as indicated by the “@” sign
9. Whether the link points to a social media domain (e.g twitpic.com)
10. Presence of emoticons and other sentiment indicators
11. Presence of exclamation points
12. Percentage of words not found in a dictionary
13. Presence of proper names as by words with a single initial capital letter
14. Percentage of letters in the tweet that do not spell out words

Classification

- Created labels of two classes: “not interesting” (0/5 or 1/5) and “possibly interesting” (2/5 or higher)
- Decision tree classifier with 14 features
  - Has hyperlink feature dominates
- Build a simple classifier with single rule:

  If tweets contains a link -> possible interesting

  else -> not interesting
  - Single rule classified tweets with 81% accuracy
- Same test with second data set
  - 85% accuracy

Decision stump

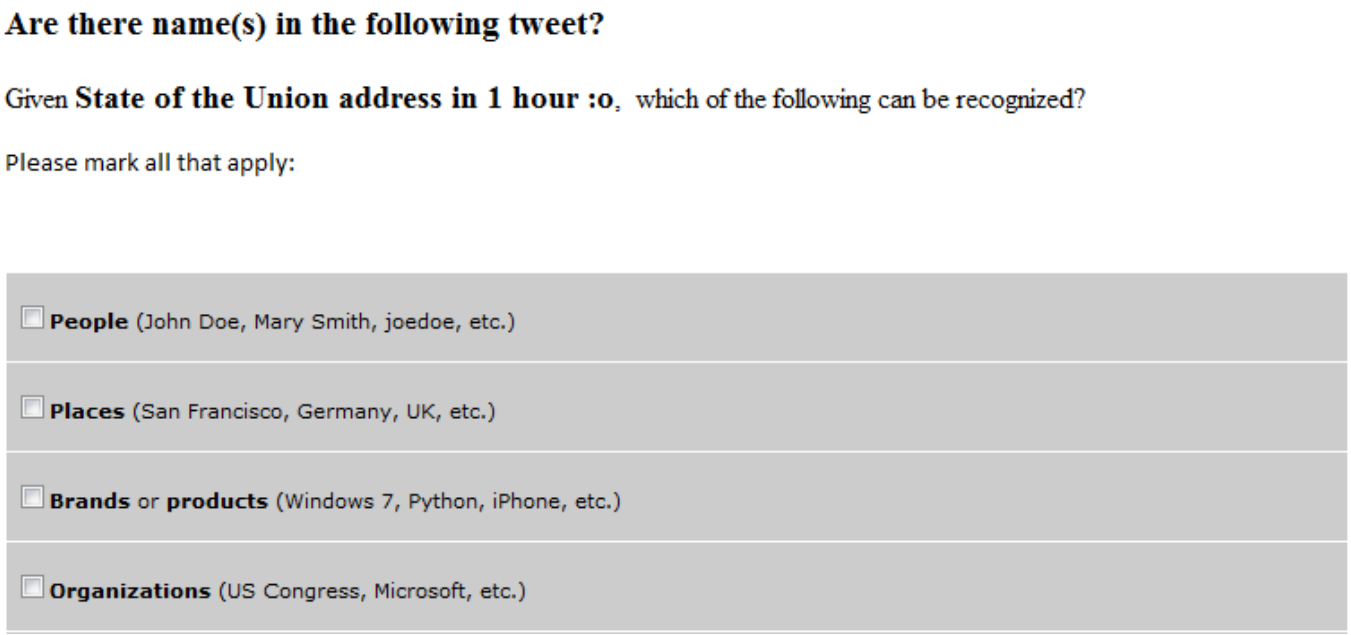
- Cases of miss-classifications are interesting tweets that don't have a link
- Eyeballing these interesting no-link tweets, looks like many contain named entities
- More generally, need better textual features

New textual features

1. Algorithmic named-entity extraction
2. Human computation named-entity extraction

Human named-entity extraction

- Asked workers to identify people, places, products and organizations
- High agreement
- Only tested a small subset



bing

Practice what you preach

- Didn't test the experiment enough
- Worker feedback asking what to do if there is no category
- Re-visit the design and included a "No. I don't see name(s)" category

NER results

Table 5. Named Entity Types for 126 "Interesting" Tweets with no Links (Experiment 1)

Entity Type	# of Tweets	% of Tweets
Person	40	32%
No entity	20	24%
Place	21	17%
Technology	21	17%
Other	10	8%
Organization	4	3%
Total	126	100%

A note on payments

- $3 per 100 tweets for interestingness
- $0.02 per tweet for NER

Conclusions

- Most tweets are not of general interest (57%)
- Crowdsourcing to label “interesting” and “uninteresting” tweets and train a classifier
- Very fast
- Faux features
- Ideally, don’t bother storing / indexing tweets identified as uninteresting
- Other features to explore: temporal, social

Future work

- Currently working on
  - Larger data set #3
  - NER for all data set
  - NER features for classification
- Next
  - Temporal analysis
  - More work on classification & clustering