Detecting Uninteresting Content in Text Streams

Omar Alonso, Chad Carson, David Gerster, Xiang Ji, Shubha U. Nabar
Microsoft Corp.
1065 La Avenida, Mountain View, CA 94043
{omalonso, ccar, dgerster, xiangji, shubhan}@microsoft.com

ABSTRACT

We study the problem of identifying uninteresting content in text streams from micro-blogging services such as Twitter. Our premise is that truly mundane content is not interesting in any context, and thus can be quickly filtered using simple query-independent features. Such a filter could be used for tiering indexes in a micro-blog search engine, with the filtered uninteresting content relegated to the less frequently accessed tiers.

We believe that, due to the nature of textual streams, it should be interesting to leverage the wisdom of the crowds in this particular scenario. We use crowdsourcing to estimate the fraction of the Twitter stream that is categorically not interesting, and derive a single, highly effective feature that separates "uninteresting" from "possibly interesting" tweets.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and software — performance evaluation

General Terms

Experimentation, classification, relevance.

Keywords

Twitter, user study, crowdsourcing.

1. INTRODUCTION

Micro-blogging platforms such as Twitter and Jaiku have recently gained popularity as publishing mechanisms. Millions of users post opinions, observations, ideas and links to articles of interest in the form of status updates. Due to the decentralized and instantaneous nature of publishing on such platforms, these posts contain valuable real-time information. For the same reasons, however, we face the difficult problem of separating the wheat from the chaff. Much of what is published is trivial, of interest to only the publisher and a handful of others. How do we quickly filter out such content so that what remains is of potential interest to a wide audience?

Our motivation for studying this problem arose while building a "real-time" search engine that searches micro-blog updates for real-time information on hot topics. On platforms such as Twitter, users typically generate 50 million updates (tweets) a day on average. The sheer volume of these updates necessitates a tiered index approach wherein potentially interesting updates are indexed in a smaller, more frequently accessed tier, with low-quality, categorically uninteresting updates indexed in the larger tiers. The question that arises then is: how do we identify, on the fly, which tier an update belongs to?

From a content perspective, we would like to explore if the updates appear to be relevant and if there is a class that we can call interesting or appealing based on user judgments.

In this paper, we use crowdsourcing for this exploration. We assign workers on the Amazon Mechanical Turk (AMT)[1] platform the task of labeling posts as "only interesting to author and friends" or "possibly interesting to others", with the premise that no further context is needed for identifying the truly mundane. We chose a crowdsourcing approach because it is cheap and extremely fast for running these types of experiments.

Our studies bring to light certain interesting facts: 57% of the Twitter stream is categorically not interesting, and of these 89% do not contain hyperlinks. Moreover, we find that the simple presence of a link correctly classifies a tweet as "not interesting" or "possibly interesting" more than 80% of the time. This simple rule comes at a price, however, since it incorrectly classifies many tweets as not interesting simply because they do not contain a link.

2. RELATED WORK

Amazon Mechanical Turk has emerged as a viable platform for conducting relevance experiments. Most of the research has been on evaluating relevance assessments and comparing the performance of Mechanical Turk workers versus experts. Examples of this type of research are evaluating a subset of TREC [1] and annotator performance in four different NLP tasks [10].

There have been several recent studies on micro-blogging services. Much of the research has been focused on questions related to the structure and nature of the Twitter community. For example, the geographical and topological properties of the Twitter network are studied in [5] and [6]. In [4] and [11], the authors study motivations for using Twitter and argue that activities on Twitter can be thought of as information seeking or information sharing.
There has also been some work on semantic analysis of the textual content of Twitter updates: The authors in [8] use a partially supervised learning model to map tweets to dimensions that correspond roughly to substance, style, status and social

SIGIR'10, July 19-23, 2010, Geneva, Switzerland.

[1] www.mturk.com

characteristics of the updates. In [3], the authors use Twitter to track consumer sentiment towards certain brands.

The real-time nature of Twitter updates is studied and harnessed in [2] and [9]. Dong et al. [2] uses Twitter signals for url discovery and to improve the ranking of these newly discovered urls in web search results. The authors in [9] used Twitter to build an earthquake reporting system in Japan that outperforms the Japan Meteorological Agency in speed of notification.

The work that comes closest to ours is [7], wherein the authors propose several features for identifying interesting tweets; the features are not, however, experimentally validated.

3. EXPERIMENTS

We performed two experiments using two sets of tweets. For the first experiment, using the Twitter public timeline API[2] we downloaded 100 tweets in the morning and another 100 in the afternoon for five consecutive days (Monday through Friday). After each batch of tweets was downloaded, it was immediately uploaded to AMT, where workers were presented with a set of tweets and asked if the content was "interesting" or "not interesting." Initially, we instructed workers to label tweets "interesting" if the content mentioned "specific information that people might care about" (e.g. "Another earthquake hits Haiti"). We defined "not interesting" content to include "advertisements, opinions, and trivial updates about daily life" (e.g. "Going for lunch with a friend"). Each worker was asked to label multiple tweets, and we collected five distinct judgments for each tweet. No qualification test was used, although we selected only workers having an approval rate (a reputation measure) of at least 97%. The cost of generating labels for each 100-tweet batch was less than $3.

While analyzing the data, we realized that our instructions were unclear. We modified the instructions and defined the labels as "only interesting to author and friends" and "possibly interesting to others." We resubmitted the batches of tweets and found that the quality of the labels improved. Figure 1 shows a large increase in scores of 0/5 or 5/5, which signify unanimous agreement among workers, and a large decrease in scores of 2/5 or 3/5, which signify disagreement.

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)

[2] twitter.com/statuses/public_timeline.xml

For the second experiment, we sampled 1,791 tweets from our internal data system, into which we had loaded a week's worth of status updates from the Twitter "firehose". The agreement among workers in this experiment (Figure 2) showed a similar distribution to the first experiment (Figure 1).

Score	# of Tweets	% of Tweets
Unanimous agreement: 0/5 or 5/5	997	56%
Near-agreement: 1/5 or 4/5	483	27%
Disagreement: 2/5 or 3/5	311	17%
Total	1,791	100%

Figure 2. Agreement among Workers (Experiment 2)

4. DATA ANALYSIS

For each tweet we created a single "interestingness" score, calculated as the number of "possibly interesting to others" AMT labels divided by the total number of labels for that tweet. Each tweet received five labels from five different workers. We observed that 1) 57% of tweets scored 0/5 or 1/5 and 2) within each score band, there was a strong correlation between the fraction of tweets containing a hyperlink and the score (Table 1).

Table 1. Distribution of Interestingness Scores and % of Tweets with Links for each Score (Experiment 1)

Score	# of Tweets	% of Tweets	# of Tweets with Links	% of Tweets with Links
5/5	127	13%	120	94%
4/5	105	11%	82	78%
3/5	79	8%	45	57%
2/5	112	11%	50	45%
1/5	163	17%	41	25%
0/5	394	40%	17	4%
Total	980	100%	355	36%

Read: "78% of tweets having a score of 4/5 contained a link."

Next we defined a class of "uninteresting" tweets having a score of 0/5 or 1/5 (shaded grey in Table 1), with the remainder classified "possibly interesting."

We created multiple textual features, including 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 indicated

by words with a single initial capital letter and 14) percentage of letters in the tweet that do not spell out words.

We attempted to train a decision tree classifier using the above classes and features, but repeatedly found that the "has hyperlink" feature dominated. We then created a simple classifier with a single rule: if a tweet contains a hyperlink, classify it "possibly interesting"; if not, classify it "not interesting." We were surprised to find that this single rule classified tweets with 81% accuracy (Table 2).

Table 2. Confusion Matrix and Accuracy using Single "Has Hyperlink" Rule (Experiment 1)

Confusion Matrix

Classified as →	a	b
a = Not Interesting	499	58
b = Possibly Interesting	126	297

Read: "126 tweets whose actual class was Possibly Interesting were classified as Not Interesting."

Accuracy	#	%
Tweets correctly classified	796	81%
Tweets misclassified	184	19%
Total	980	100%

As the confusion matrix shows, most classification errors (126 out of 184) were due to "possibly interesting" tweets being labeled "not interesting" simply because they did not contain a link. This raises the question of what features might be useful to correctly classify such tweets. Visual inspection of these misclassified tweets shows that many contain named entities ("State of the Union"; "China") and quantities ("$499"; "100K").

We performed the same analyses on the second set of 1,791 tweets labeled using AMT. We were pleased to find that the distribution of interestingness scores was similar to the first experiment, demonstrating that the quality of judgments by AMT workers is high enough to create reproducible results. We noted that the accuracy of the single "has hyperlink" rule increased to 85%.

As with the first experiment, most misclassifications (149 out of 269) were due to tweets with no link being misclassified as not interesting.

To reduce misclassified tweets, we began experimenting with new textual features including the presence of named entities. We saw two ways to generate such features: 1) algorithmic entity extraction and 2) submitting tweets to AMT with instructions on the entities we seek to identify (e.g. "Does this tweet contain the name of a person, organization, or product?").

Table 3. Distribution of Interestingness Scores and % of Tweets with Links for each Score (Experiment 2)

Score	# of Tweets	% of Tweets	# of Tweets with Links	% of Tweets with Links
5/5	103	6%	99	96%
4/5	140	8%	122	87%
3/5	126	7%	87	69%
2/5	185	10%	97	53%
1/5	343	19%	86	25%
0/5	894	50%	34	4%
Total	980	100%	525	29%

Read: "87% of tweets having a score of 4/5 contained a link."

Table 4. Confusion Matrix and Accuracy using Single "Has Hyperlink" Rule (Experiment 2)

Confusion Matrix

Classified as →	a	b
a = Not Interesting	1117	120
b = Possibly Interesting	149	405

Read: "149 tweets whose actual class was Possibly Interesting were classified as Not Interesting."

Accuracy	#	%
Tweets correctly classified	1522	85%
Tweets misclassified	269	15%
Total	1791	100%

Focusing on the second approach, we submitted the 126 misclassified tweets from the first experiment back to AMT and asked workers to judge what types of named entities each tweet contained. Workers were presented with one tweet and asked to judge named entities using the following categories:

- People (John Doe, Mary Smith, joedoe, etc.)
- Places (San Francisco, Germany, UK, etc.)
- Brands or products (Windows 7, Python, iPhone, etc.)
- Organizations (US Congress, Microsoft, etc.)
- Other (State of the Union, US Patent #123456, etc.)
- No. I don't see name(s).

To improve the quality of judgments, we intentionally included the "No" category to avoid workers feeling compelled to find a named entity even when one was not present. Each worker was asked to recognize entities from a single tweet, and we collected five distinct judgments for each tweet. No qualification test was used and the approval rate was 97%. We paid two cents per task for this experiment.

Table 5 shows the distribution of named entity types across these 126 tweets. (For simplicity, only the dominant entity type is

counted for each tweet; in reality, some tweets contained multiple entity types.) . 76% of the tweets had a named entity, highlighting the potential of the named entity feature.

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%

Read: "For 40 tweets, 'Person' was the entity type that received the most judgments."

5. CONCLUSIONS AND FUTURE WORK

Using labels gathered from AMT, we learned that the presence of a hyperlink in a tweet strongly correlates to that tweet's "interestingness" score. This single "has hyperlink" feature classifies tweets with more than 80% accuracy, with most errors due to tweets without hyperlinks being misclassified as "not interesting."

The results are promising, especially given the low cost of the labels. At $3 per 100 tweets, our 980-tweet sample from the first experiment cost less than $30 to label, but still yielded enough information to classify tweets with high accuracy. Because the "has hyperlink" feature is so dominant, however, results may not be representative.

In addition to providing consistent high-quality labels, AMT also shows promise for creating named-entity features that are challenging to compute algorithmically. Such crowdsourced "faux features" could be useful for supervised learning experiments with a small number of labels and therefore a small number of instances. This approach could also be used to evaluate features that are not computationally feasible today, with the goal of quantifying the value of such features if they did become available in the future.

6. REFERENCES

[1] Alonso, O., and Mizzaro, S. 2009. Can we get rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment. In *Proceedings of SIGIR Workshop on the Future of IR Evaluation*.

[2] Dong, A. et al. 2010. "Time is of the Essence: Improving Recency Ranking Using Twitter Data". In *Proceedings of WWW*.

[3] Jansen, B. J. et al. 2009. Twitter Power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*.

[4] Java, A. et al. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of SNA-KDD Workshop*.

[5] Krishnamurthy, B., Gill, P., and Arlitt, M. 2008. A few chirps about Twitter. In *Proceedings of WOSP*.

[6] Kwak, H. et al. 2010. What is Twitter, a Social Network or a News Media? In *Proceedings of WWW*.

[7] Lauw, H., Ntoulas, A., and Kenthapadi, K. 2010. Estimating the Quality of Postings in the Real-time Web. In *Proceedings of SSM.*

[8] Ramage, D., Dumais, S., and Liebling, D. 2010. Characterizing Microblogs with Topic Models. In *Proceedings of ICWSM.*

[9] Sakaki, T. et al. 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of WWW*.

[10] Snow, R. et al. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of EMNLP*.

[11] Zhao, D., and Rosson, M. 2009. How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work. In *Proceedings of GROUP*.