Spam Detection in Twitter Trending Topics

Grant Stafford, Professor Louis Yu, advisor

April 11, 2013

Motivation

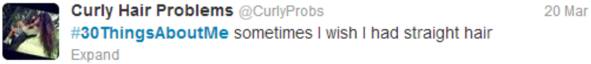
"Spam will be a thing of the past in two years' time."
-Bill Gates, 2004

"IHOP #Answer4Everything http://t.co/OLS7RpwL COOL VIDEO TELLS A METHOD TO EARN $700+ DAILY!"
-Twitter Spam, 2013

Motivation

Why investigate spam in the Twitter trending topics?

1. Help users see only relevant information, such as



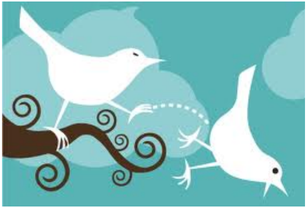
2. Verify integrity of trends in the social network:

   Close to 32% of messages in Sina Weibo come from spammers (Yu 2012).

Spam Criterion

What is "spam"?

1. Contains a URL to a website completely unrelated to the topic or hashtag on a tweet.
2. Retweets in which legitimate links are changed to illegitimate ones, obfuscated by URL shorteners.



Related Work

General Approach

1. Select textual and/or structural attributes
2. Develop classifier via machine learning techniques
3. Determine effectiveness of classifier and apply it

Detecting Spam Bots in Online Social Networking Sites (Wang, 2010)

1. Naive Bayes with structural and textual attributes is best.
2. Roughly 3% of tweets are spam.

Detecting Spammers on Twitter (Benevenuto et. al., 2010)

1. SVM trained on 1.8 billion tweets.
2. Evaluated which attributes are most effective.

Gathering Data

Obtaining Tweets for Labelling

- Used tweepy python library to connect to Twitter streaming API.
- Filtered hourly on the trending topics worldwide for the 'en' language code.
- Program ran from 2/1/13 to 2/7/13 on a computing cluster in the CS lab, gathering data on over 9 million tweets across 801 distinct trending topics.

Construcing a Labelled Collection

- Hand-labelled nearly 1500 tweets randomly sampled from the data
- Ensured examples from each of the 170 hours and over 40 spam examples.

Labelled Collection Overview

Non-Spam Instances	Spam Instances
1453	42

Attributes

Attributes Identified by Previous Research (Beneventuo et al. 2010)

- URLs per word
- Total number of words
- Number of numeric characters
- Number of total characters
- Number of URLs
- Number of hashtags
- Number of mentions
- Number of retweets
- Whether the tweet was a reply
- *Rank of topic (added by us)

Evaluation

χ^2 attribute selection, then compare mean values of attributes between classes.

Classifier

Naive Bayes Classifier

Applies Bayes theorem from probability with assumption that the attributes are all independent, allowing us to compute:

$$P(Spam)\prod_{i=1}^{d}(X_i|Spam) \quad \text{and} \quad P(NonSpam)\prod_{i=1}^{d}(X_i|NonSpam)$$

and assign the tweet to the class with the higher value.

Classifier Evaluation

- Standard information retrieval metrics: precision, recall, and F-measure (Macro and Micro F1) obtained by 10-fold cross validation.
- Compared against baseline classifier that classifies all as non-spam.

Spam Impact Evaluation

Spam Percentage in Trending Topics Overall

Simply find the percentage of spam across our entire dataset.

Variance in Spam Percentage Among Trending Topics

Use Pearson's χ^2 goodness of fit test to establish whether observed distribution of frequencies differs from an expected distribution with equal percentages for all topics.

Effect of Spam on Topic Rankings

Count the number of topics which change rank after filter is applied.

Attribute Evaluation

Attributes Ranked by Signficance

Attribute	χ^2 Statistic
URLs per word	116
URLs	111
Number of hashtags	71
Numeric characters	17
Rank of topic	12
Whether tweet was a reply	3

Attributes by Class

Attribute	Non-Spam Mean	Spam Mean
URLs per word	0.0077	0.0476
URLs	0.0847	0.5714
Number of hashtags	0.8671	1.0238
Numeric characters	1.3896	3.2177
Rank of topic	4.2638	6.1429

Classifier Evaluation

Confusion Matrix

		Predicted	
		Spam	Non-Spam
True	Spam	1327	125
	Non-Spam	19	23

Information Retrieval Metrics

		Metric		
		Precision	Recall	F1
Class	Non-Spam	0.986	0.914	0.949
	Spam	0.155	0.548	0.242

Comparison to Baseline

The Micro-F1 measure was 0.929 and the Macro-F1 measure was 0.596. Micro-F1 is 3% worse than baseline but Macro-F1 is 24% better.

Spam Impact

Overall Impact

- Previous research (Wang, 2010) estimated 3% spam messages overall.
- Our hand-labelled collection contained about 2.8% spam messages.
- Classifier marked 9.9% of the training dataset as spam.
- Classifier found average of 9.0% spam on test data.

Discussion

Suggests trending topics do not have significantly more or less spam than Twitter overall.

Spam Impact

Variation of Spam Percentage Across Topics

- Every one of the 170 tests was significant at the 5% level.
- The average value of the chi-squared statistics was 7008.

Discussion

Strongly suggests that the percentage of spam is not anywhere close to uniform across the trending topics.

Spam Impact

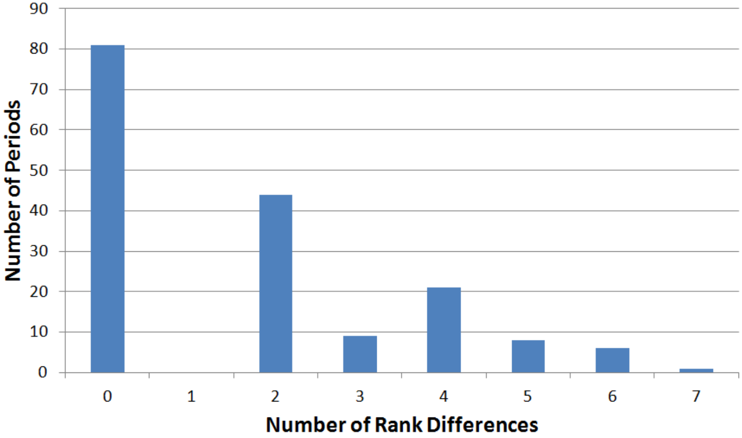
Spam Impact on Rankings of the Trending Topics

- 47% (81 of the 170) time periods saw no change.
- On average, 1.66 topics differed from previous ranking.

Discussion

Any change in rankings requires 2 topics out of position, suggesting that rankings were not greatly affected by the presence of spam.

Frequency of Rank Changes Across Periods

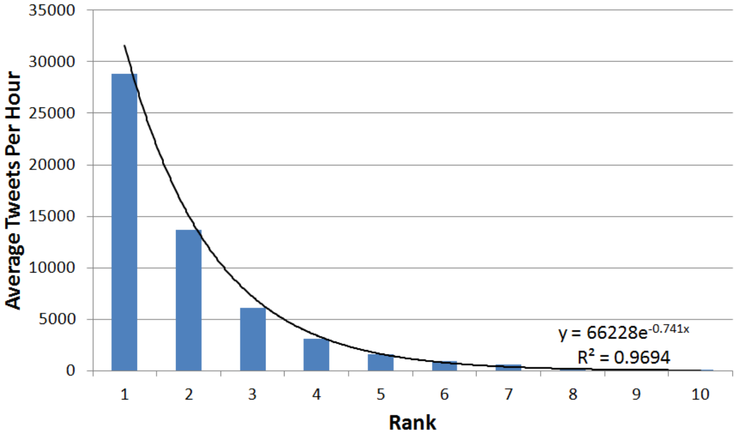


Spam Impact

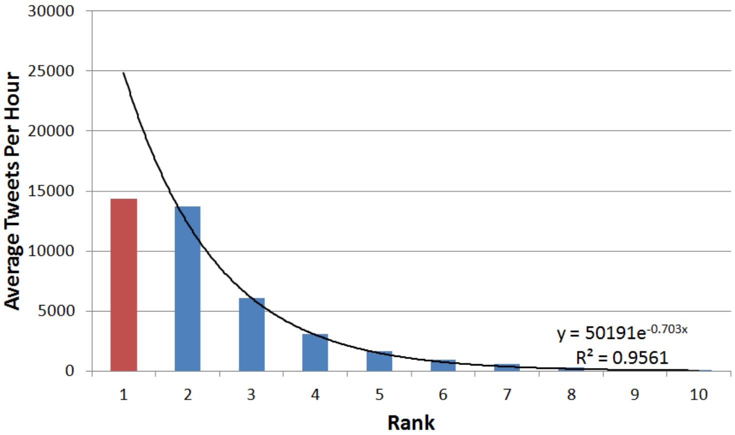
Explaining the Findings

How can spam have so little impact on the rankings despite being far from proportionally distributed across topics?

Average Tweets Per Hour by Rank



Average Tweets Per Hour by Rank



Conclusions

Observations

- Some topics, such as news stories, often contain URLs. Spammers may take advantage of this fact when targeting topics.
- Personal topics, such as #30FactsAboutMe, rarely contain links and tend to elicit short responses from users. Increased user participation or decreased spammer targeting may be factors for these.

Takeaways

- Spammers don't drive trending topic popularity; they piggyback on topics they find to be most effective for spreading their messages.
- Due to spam prevention techniques or other factors, Twitter trends represent users fairly truthfully.

Questions?

Any questions?

Feel free to use more than 140 characters!

Future Work

Deeper

Given limited time to gather, process, and classify data and to analyze results on this complex topic, we obtained interesting findings, but could go deeper:

- Quantify topic vulnerability: how do news and personal topics compare?
- Including structural attributes: enhanced support for these conclusions.