Effect of Spam on Hashtag Recommendation for Tweets

Surendra Sedhai Aixin Sun
Nanyang Technological University, Nanyang Avenue, Singapore 639798
surendra001@e.ntu.edu.sg axsun@ntu.edu.sg

ABSTRACT

Presence of spam tweets in a dataset may affect the choices of feature selection, algorithm formulation, and system evaluation for many applications. However, most existing studies have not considered the impact of spam tweets. In this paper, we study the impact of spam tweets on hashtag recommendation for hyperlinked tweets (*i.e.,* tweets containing URLs) in HSpam14 dataset. HSpam14 is a collection of 14 million tweets with annotations of being spam and ham (*i.e.,* non-spam). In our experiments, we observe that it is much easier to recommend "correct" hashtags for spam tweets than ham tweets, because of the near duplicates in spam tweets. Simple approaches like recommending most popular hashtags achieves very good accuracy on spam tweets. On the other hand, features that are highly effective on ham tweets may not be effective on spam tweets. Our findings suggest that without removing spam tweets from the data collection (as in most studies), the results obtained could be misleading for hashtag recommendation tasks.

Keywords

Hashtag recommendation; Microblog; Tweets; Spam

1. INTRODUCTION

With the popularity of the Twitter platform, spam has become a major issue. It is reported that the click-rate of spam links shared in Twitter is two orders of magnitude higher than that in email [1]. Due to the effective propagation and higher click-rate of spam content, spamming activities have become a serious issue in Twitter. Growing number of spam may not only affect user experiences [2] but also the applications using Twitter data, from many perspectives including feature selection, algorithm formulation, and system evaluation. However, most existing studies simply use all tweets collected from the Twitter stream without removing the spam tweets. To the best of our knowledge this is the first study on the effect of spam tweets on hashtag recommendation.

We argue that the presence of spam tweets may result in misleading results because features/methods that are effective to spam tweets may not necessarily be effective to ham tweets. The availability of the HSpam14 dataset makes it feasible to study the impact of spam tweets for applications based on tweets. HSpam14 is a dataset containing 14.07 million tweets in English which are annotated with spam and ham (or non-spam) labels [4]. In this study, we evaluate the 7 methods detailed in [3] for hashtag recommendation for hyperlinked tweets using subsets of spam tweets, ham tweets, and the mixture of the two subsets from HSpam14.

Our results show that: (i) Spammers use few popular hashtags which are relatively easy to recommend; simple approaches achieve high accuracy in recommending hashtags for spam tweets. (ii) Features/methods which are effective to ham tweets may not be effective to spam tweets; and (iii) Performance on spam dataset is substantially better than ham dataset across all evaluation measures. The same hashtag recommendation method may achieve much better accuracy on the dataset containing spam tweets than a dataset containing only ham tweets. Without removing spam in a data collection, the results obtained for a hashtag recommendation method may be (misleadingly) better than its actual performance in reality.

2. HASHTAG RECOMMENDATION

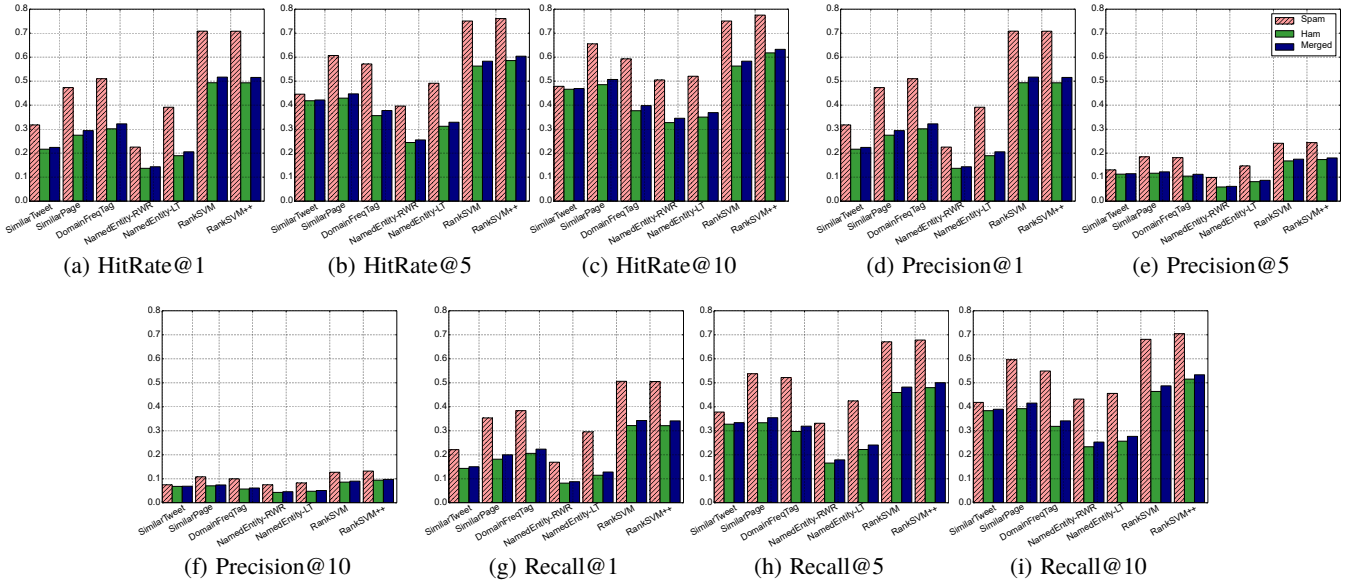
Hashtag recommendation for hyperlinked tweets is a task to recommend a list of hashtags that are most relevant to a hyperlinked tweet t containing link ℓ to a Web page. We brief the 7 methods used to evaluate the effect of spam (see [3] for details of the methods).

- **SimilarTweet/SimilarPage** recommends most frequently used hashtags in similar tweets/Web pages, based on the assumption that similar content is likely to be annotated with similar hashtags.
- **DomainFreqTag** recommends most frequently used hashtags of a given domain. The domain of a tweet is obtained from the link ℓ contained in the tweet.
- **NamedEntity-RWR** recommends hashtags based on the named entities in the linked Web page using Random Walk with Restart (RWR) algorithm. The intuition is that hashtags which are in the close neighbourhood of entities in the Web page linked by a tweet is likely to be appropriate hashtags for the tweet.
- **NamedEntity-LT** recommends hasthags based on the named entities present in the linked Web page using Language Translation (LT) model.
- **RankSVM/RankSVM++** recommends hashtags using learning to rank. RankSVM uses the above 5 candidate methods as features. RankSVM++ uses four additional features.

3. EXPERIMENTS AND ANALYSIS

Dataset. HSpam14 consists of 14.07 million tweets in English, that are labeled as ham and spam [4] . There are 1.07 million hyperlinked tweets ignoring the deadlinks and links requiring authentication (*e.g.,* Facebook links). Among 1.07 million hyperlinked tweets,

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889404.



(a) HitRate@1 (b) HitRate@5 (c) HitRate@10 (d) Precision@1 (e) Precision@5

(f) Precision@10 (g) Recall@1 (h) Recall@5 (i) Recall@10

Figure 1: HitRate, Precision, and Recall of the 7 methods on spam, ham and merged datasets

0.95 million linked pages have at least one named entity in the page based on StanfordNER.[1] Out of the 0.95 million tweets, 0.220 million tweets have at least one hashtag.

In the following experiments, we use these 0.220 million hyperlinked tweets. Among them, 22 thousand are spam tweets and 198 thousand tweets are ham tweets. To learn the RankSVM models, we randomly select 40% of the tweets as training and the remaining 60% as test from spam and ham tweets, respectively.[2] The training and test tweets from the spam and ham datasets respectively are merged to create the training and test for the merged dataset. In candidate hashtag selection for both training and testing, we limit the search for similar tweets (resp. Web pages) posted one day before the currently processing tweet to simulate inaccessibility of future data in reality.

Evaluation Metric. We use three metric to evaluate hashtag recommendation accuracy: *Precision@k*, *Recall@k*, and *HitRate@k* (*Pr@k*, *Re@k*, and *HR@k* for short); k={1, 5, 10} is the number of top-ranked recommended hashtags. Let H_k be the set of top-k recommended hashtags and H_g be the set of ground-truth hashtags of a tweet t. ***Pr@k*** for tweet t is $|H_k \cap H_g|/k$; ***Re@k*** for tweet t is $|H_k \cap H_g|/|H_g|$; and ***HR@k*** for t is 1 if $|H_k \cap H_g| \geq 1$ and 0 otherwise. The values reported for each method are the averaged values over the test tweets.

Experimental Results. Figure 1 reports *HR@k*, *Pr@k* and *Re@k* respectively of all the 7 methods. We make following observations.

- Hashtag recommendation is much easier for spam tweets than ham. *HR@1*, *HR@5* and *HR@10* of RankSVM++, the best performing method, are at least 15% higher on spam tweets than ham tweets. Similar observations hold for precision and recall.
- Simple methods (SimilarPage and DomainFreqTag) achieve very good recommendation accuracy for spam tweets, probably due to the fact that many spam tweets are posted to promote Web pages from a limited number of domains. For instance, a large number of tweets linking to game-insight.com contain 3 hashtags from a pool of hashtags (*e.g.,* android, androidgame, ipad, ipadgames, iphone, iphonegames). Another simple method, SimilarTweet, also achieves good results due to the presence of near-duplicate tweets in spam.
- Even though there are fewer than 10% of spam tweets, the accuracies on the merged dataset are at least 2% higher than ham tweets on all measures. Without removing spam in a data collection, the results obtained for a hashtag recommendation method may be (misleadingly) better than its actual performance in reality.
- RankSVM++ is the best performing method on all datasets by all evaluation measures. However, the **percentage of improvement** by using RankSVM++ over simple methods (*e.g.,* SimilarPage, DomainFreqTag) on spam dataset is much smaller than that on ham dataset. That is, features/methods that are effective on ham may not necessarily be effective on spam tweets.

[1] `http://nlp.stanford.edu/software/CRF-NER.shtml`

[2] Results of training/test datasets created by random selection are found to be consistent with the results from partitioning of tweets by time. Due to space limitation we only present results by random selection of training and test tweets.

4. CONCLUSION

Spam has adversely affected many applications. Our experiment shows that spam on Twitter skews the performance evaluation. It is observed that simple methods such as DomainFreqTag gives very good accuracy for hashtag recommendation for spam tweets. Due to the presence of near-duplicate tweets, methods like similar tweets and similar pages also achieve good results. In the literature, similar item based recommendation has been considered as a strong baseline, which maybe heavily affected by the presence of spam. We have also observed that features/methods which are effective for ham tweets may not be equally effective for spam. Hence, presence of spam in the dataset may affect feature selection process in some applications. It is necessary to perform spam filtering before conducting any analysis or evaluation on Twitter dataset.

5. REFERENCES

[1] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *CCS*, pages 27–37, 2010.

[2] T. Jones, D. Hawking, P. Thomas, and R. Sankaranarayana. Relative effect of spam and irrelevant documents on user interaction with search engines. In *CIKM*, pages 2113–2116, 2011.

[3] S. Sedhai and A. Sun. Hashtag recommendation for hyperlinked tweets. In *SIGIR*, pages 831–834, 2014.

[4] S. Sedhai and A. Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *SIGIR*, pages 223–232, 2015.