SpamDia: Spammer Diagnosis in Sina Weibo Microblog

Hao Chen
SPKLSTN Lab
Department of Computer Science and Technology
Xi'an Jiaotong University
lechenhao@gmail.com

Jun Liu
SPKLSTN Lab
Department of Computer Science and Technology
Xi'an Jiaotong University
liukeen@mail.xjtu.edu.cn

Jianhong Mi
SPKLSTN Lab
Department of Computer Science and Technology
Xi'an Jiaotong University
m189111@stu.mail.xjtu.cn

ABSTRACT
Microblogs open opportunities for social spammer accounts. These widespread spammers, are threatening for microblog services and normal users. Therefore, detecting spammers should be conducted to fight and stop them. In this paper, we propose an approach to diagnose user accounts in China's most popular microblog site *Sina Weibo*. Unlike existing approaches, which can hardly discover sophisticated spammers and only give a simple conclusion as spammer or not lacking of detail information, but our work provides a more responsible way that reveals the clues to verify a spammer account by using classifier-level fusion and feature-level comparison. Distinct discriminative features are used to train basic classifiers. Then a fusion model is learned to combine the outputs of the basic classifiers and make the final prediction. Comparing basic classifiers outputs with the final prediction offers the insights of spammer identification. Experiments show that our approach significantly improves the classification performance and this approach can point out spammers' specific spam action in a detail way that helps us strike spammers accurately.

Keywords
microblog, spammer, detection

1. INTRODUCTION
Social network services give us a wonderful platform to share and enjoy life. Unfortunately, it also opens opportunities for spammers, such as sina weibo microblog, in which spamming activities picked up in numbers and varieties [1] [2]. "Spammer" is the user account who introduces undesirable information to normal users. In detail, spammers spread and promote mass advertisement posts; mislead readers with fake content; lurk as predators for some personal gains [3]. Figure 1 shows examples of social spammers. We note that opinion leaders will not be regarded as spammers although they may sometimes share the above characters naturally or half unconsciously.

Figure 1. Examples of social spammers.

Widespread spammers bring are harmful to microblog and normal users [4]. The rise of social spamming can significantly hinder the use of microblog for effective information dissemination. Detecting spammers is the basis work for study of the mechanism of spamming. Moreover, spammer detection can provide an evidence to judge the trustworthiness of posts, which is crucial for all opinion based applications. Due to its seriousness, spammer detection has attracted a lot of attention in research.

Spammer detection has been studied with different techniques in various social network platforms. Wang [5] proposed a graph based method to capture intricate relationships among different entities for detecting spammers in a product review site. Their method is well complementary to approaches based on text analysis. It required building a heterogeneous review graph from raw data, however, data from microblog do not support this requirement for privacy. In [3] and [4], they introduced optimization methods to fix the problem by solving the object function. The deficiencies of this method is that ideal training data is required, so it is labor intensive or even may not work well. Spammer detecting is usually cast as a binary classification problem using content, behavior or relationship features. However, existing methods usually work with too much data or features which cannot be simply sampled from microblog services. In addition, as a great challenge of machine learning, clues used to identify a spammer are not released, making it hard to evaluate the results or do some more detail study.

In this work, we try to solve the following two problems:

- How to detect spammers in *sina weibo* microblog?
- How to reveal insightful evidences to verify a spammer?

To resolve this dilemma, we studied the features of microblog users to find out some discriminative features and the meanings of them. Based on these features we applied a hierarchy classification method to identify spammer accounts, in which the results of predictive functions learned from distinct features on the same instance. Each feature represents a suspect clue for

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China

DOI 10.4108/eai.18-6-2016.2264143

spammer identification. Then, a fusion model is learned by integrating basic classifiers to give the final prediction. Comparing the basic classifiers predictions with final prediction we can know which features are the clues for spammer detection. Our experiments and case study show that our method is a proper way to detect spammers.

The rest of the paper is organized as follows. Section 2 gives a brief survey of spam and spammer detection in Internet applications. Section 3 introduces the construction of our real-world dataset and ground truth. In section 4, we proposed a hierarchy classification method with classifier-level fusion and feature-level comparison. Then we conducted experiments and analyzed the proposed method in section 5. Finally, section 6 concludes this work and points out plans.

2. RELATED WORK

Spamming has been a long existing problem with Internet applications. We briefly give a survey of spam and spammer detection research here.

2.1 Web/Email Spam detection

Gyöngyi et al. [6] separated useful webpages from spam with TrustRank. Bencz´ur et al.[7] proposed to detect nepotistic links using language models. In this method, a link is down-weighted if the language models from its source and target page have a great disagreement. Gomes et al. [8] analyzed an e-mail workload consisting of the messages received in a university network, pointing out a number of features that can be used to differentiate spam from legitimate messages. [9] proposed the SpamHINTS project, which develops measurement techniques that, based on the analysis of the e-mail protocol packets. ISnotSPAM [10], a tool for Email spam detection, has been opened to the public for several years by utilizing text mining and links analysis. However, these solutions are not sufficient for microblog spammer accounts detection. Because microblog posts are short, noisy, unstructured and with links conducting a heterogeneous network, traditional methods are not fit this social network problem.

2.2 Social spammer detection

In social spammer detection, User Generated Content (UGC), behaviors and relationships are key indicators of spamming verifying. Sarita et al. [11] studies the behavior of spammers in Twitter, and find that spammers is different from legitimate users in the field of posting tweets, followers, following friends and so on. Zhu et al. [3] proposes a matrix factorization based spam classification model to collaboratively induce a succinct set of latent feature learned through social relationship for each user in RenRen. Wang [12] proposes a naïve Bayesian based spammer classification algorithm to distinguish suspicious behavior from normal ones in Twitter. Hongyu et al. [13] adopts a set of novel feature for effectively reconstructing spam messages into campaigns rather than examining them individually.

3. DATASET

To have a deep look at into the problem, we need to examine large amount of real-world cases. And in order to evaluate our proposed method, we need a labeled collection of users, classified into spammers and non-spammers. However, to the best of our knowledge, there is no proper public dataset available. Therefore, we constructed one and made it public in the world most famous machine learning dataset repository UCI named microblogPCU. Below we introduce the tool we used to collect data and the way we labeled the instances. Then we evaluate our labeled data and construct the ground truth.

3.1 Microblog Crawling

We developed a crawler to collect data from *sina weibo*. We launched our crawler for one year from July 2014 to June 2015. Our crawler can be download at [14]. The crawler is capable to get any public content and user behavior messages. It first gain raw HTML data from web and then extracts metadata from these HTML documents. Finally, it builds a network with relationships among users and posts.

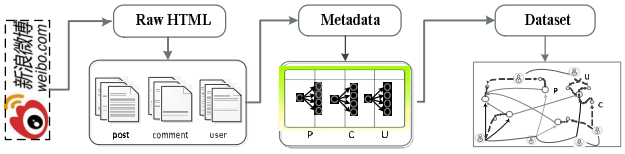

Figure 2. Workflow of our crawler.

Figure 3. Our *microblogPCU* crawler tool interface.

3.2 Ground Truth Acquisition

We invited five human judges to identify spammers individually. Judging suspicious spammers is a complex task for human. These five judges give their labels according to intuition, background knowledge and searching for additional information. In totally, every judge labeled 16,236 users.

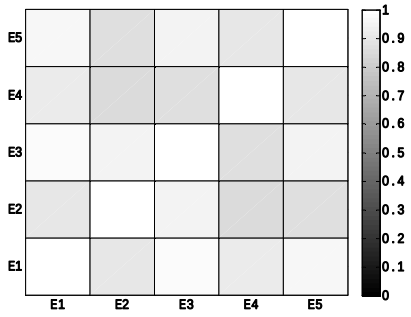

Figure 4. Human Labeling Results Agreement Matrix.

There is no “true” ground truth and the ground truth dataset we taken was constructed by the following. First, we evaluated the agreement of labeling result. We assume that if the labels of a specific instance given by different judges have acceptable

agreement, the human judgment is trustworthy. In order to evaluate their agreement, we used *Fleiss' kappa* [15]:

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$$

where $\bar{P}$ is the agreement probability and $\bar{P}_e$ is the probability of chance agreement. And among our five judges the kappa is 0.71, which represents almost substantial agreement. The human judgments on these instances are acceptable. Figure 4 shows the agreement of human judges on dataset. Second, we selected instances to construct the ground truth dataset. We assume that if one user is labeled as spammer by five judges, it would have a high probability to be spammer and it is the same for non-spammer. We selected the consistent labeled data as the ground truth. As the result, 3,000 spammers and 3,000 non-spammers are randomly selected as the ground truth from same-labeled 4,531 spammers and 9,427 non-spammers. Figure 5 concludes the data prepare process and the size of our data set.

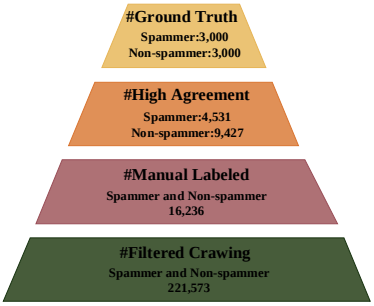

Figure 5. Our real-world dataset construction process.

4. METHODS

4.1 Spammer indicators analysis

We discuss some discriminatory features organized as content-based feature and behavior-based feature.

Content-based features are used. We first analyze content-based features from URLs it contains, and hashtags it refers. Spammers post with URLs to refer readers to their target web pages aiming at promoting more information, so spammers usually post with more URLs than normal users. Spammers post with hashtags especially the hashtag of hot topics to increase the probability of being searched by others. We found that spammers may post with many non-related hashtags once as well as many of them never post with any hashtag because they only focus on their content. We calculate the fraction of posts with URLs and hashtags. We also consider the fraction of posts with similar content. The high fraction of similar posts indicates that the spammers strongly promote their content. We then analyze content-based features from another aspect. We investigate the average number of being reposted and commented of one's posts and the fraction of reposted posts in all of one's posts. We found that spammer's posts have a low chance to be reposted and commented, for normal users are usually not interested in spammers. On the other hand, spammers generally repost others' posts with a quite low frequency. But we should note that in our sampled dataset, some of the spammers repost others' posts with a quite high frequency.

Behavior-based features are considered. We first study the user mention behavior. We found that a plenty of spammers trend to never mention others as well as some spammers mention a lot but normal users mention with a frequency neither low nor high. We also consider determining whether a user posts in a regular way. Because we found that some spammers are agent machine which works with predetermined rules. We analyzed users posting time series and we propose a feature regular posting to describe whether user's behavior is regular. We found that none normal users post in a regular way and about 1/3 spammers post in a regular way. One of the obviously motivational features is the ratio of follower and followee. Microblog allows establishing some connections between two users without mutual consent, which makes it easier for spammers to imitate normal users by quickly accumulating a large number of "human" friends. Nevertheless, few normal users actively follow these spammers. We then analyzed the posting style of users. We found that the distribution of posting numbers in one day for all users statistically. From it, we can see that spammers and normal users are quite different. During 1 to 6 o'clock, called midnight, spammers post considerable quantity of posts but normal ones almost not. And during 7 to 24 o'clock, called daytime, spammers post number looks quite stable but normal one's quite oscillatory. For midnight behavior we propose a feature midnight activity to measure how frequent a user posts during midnight. We found that spammers are more active than non-spammers during midnight. Spammers work hard even at night. For daytime we propose a feature named variance $\mathcal{V}$, which is formally defined as $\mathcal{V} = \sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$ whereas $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and x_i is the number of post posted in i-*th* time space, to verify whether a user posts different quantities of posts within a day. Spammers behave indeed in a more stable frequency way because spammers do spamming as work, but normal ones trend to post at some time when they are free out of work.

In total, we have 10 features related to content of the posts and 8 features related to behavior of microblog users. We evaluate the independence of all the features by PPMCC and the result indicates that they are independent.

4.2 Hierarchy spammer detection

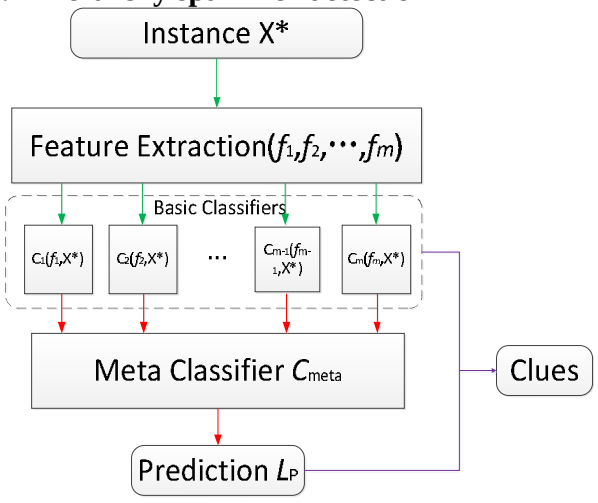

Figure 6. Framework of our approach

From the above analysis, we know that spammers may use different strategies to do spamming,. It means that a spammer may show its real role on a few of features but not all of the features. This makes accuracy detection hard. To solve this

problem, we propose a hierarchy spammer detection approach with four components. First, extract features from raw data; second, train basic classifiers on each feature of instances; third, combine the basic classifiers results and give the final prediction; And at last, do simple XNOR operation on each basic classifier and meta classifier to gain clues that expose a spammer or not. Figure 6 shows the framework of our spammer detection approach.

Feature extraction part generates 10 content-based features and 8 behavior-based features. Basic classifier training part trains *m* classifiers on *m* features of instances. We use Naïve Bayes as the basic classifier. All these classifiers generate probability estimates. The basic classifiers are quite different because they are trained on different data according to distinct features. In meta classifier learning part, outputs of the basic classifiers are combined to make the final prediction, as follow:

$$\mathcal{L}_{\mathcal{P}} = C_{meta}(C_1(f_1, x^*), C_2(f_2, x^*), ..., C_m(f_m, x^*))$$

We use decision tree style algorithm C4.5 as our meta classifier. Clues are a subset of features, which are selected by using XNOR operation between basic classifier output and the final prediction.

5. EXPERIMENT

In this section, we first conducted experiments to evaluate the performance of our proposed method compared with the state-of-art methods as well as basic and meta learners used in our hierarchy approach. Second, to further illustrate that our approach can give detail information about classification, we did some case studies and demonstrated the developed spammer detection software named SpamDia, which gives measures of every feature to point out what the spammer do with significant insights. We also detected spam posts by extending our spammer detection method to verify the effectiveness of our approach.

5.1 Performance evaluation

We compare the proposed method with three different classification methods and two of which are used in our hierarchy classification method. We also compare our method with one of the outperforming social spammer detection methods. The following is details about baselines.

- **SVM** has been widely used in spammer detection [16][17] when research focus on features.
- **Naïve Bayes** (NB) is a probability style classification method. It has verified that NB can achieve significant improvement compared with heuristic methods in many fields.
- **C4.5** is a tree-based classification method. It builds decision trees from a set of training data using the concept of information entropy.
- **SSDM** is proposed in [4], which collectively use network and content information to do social spammer detection in microblog.

In our experiments, the parameters are turned well. We used 10-fold cross validation for validating the performance. The AUC (Area under ROC Curve) is employed to evaluate the classification results. We conduct experiments on three subsets of our real-world dataset separated by topics such as sports, music and politics. The results are shown in figure 7.

We can find that our approach outperforms than other baselines. SSDM almost performed worst because lacking of enough meta features especially users relationships which are key indicators in this approach. These features in need are so ideal that it is difficult to achieve in reality. C4.5 performs better than SVM and naïve Bayes slightly. SVM and Naive Bayes generate almost the same effect. Our approach takes the advantage of combining different classifiers on different features and achieves the best performance.

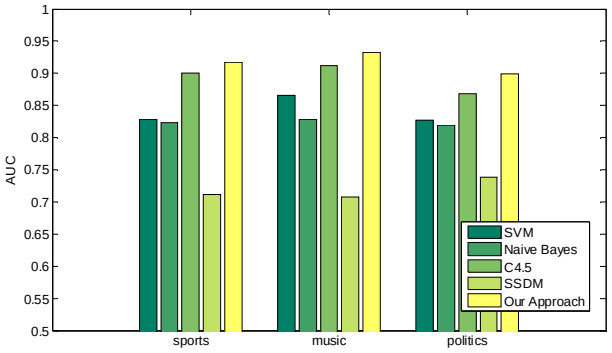

Figure 7. Performance evaluation results

5.2 Case study

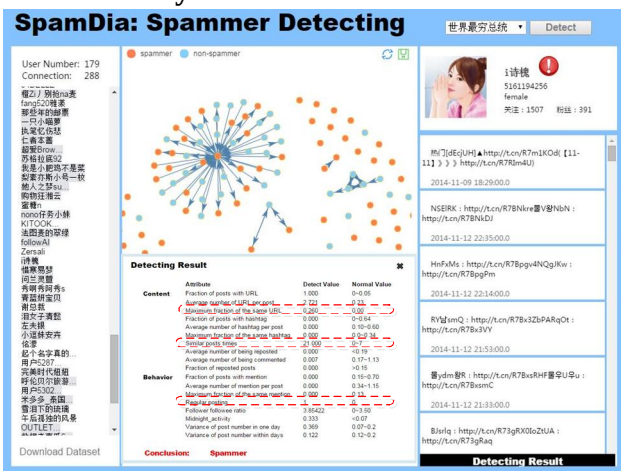

Figure 8. Interface of our spammer detection software.

We take a close look at users identified as highly suspicious with picking one candidate. The candidate is the user ‘ i 诗槐 ’, whose posts are written in Korea but not Chinese, making it hard for our judges to identify it. Our method does not rely on language, but according to its content and behavior features, so our method correctly discovered this account. In addition, our method gave the evidence for identifying it as a spammer. ‘i 诗槐’ is quite different with normal users in the features similar post times and fraction of posts with URLs as showed in figure 8. Instead of pure experiment upon specific dataset, a prototype software is specifically developed to help us understand spammer detection visually. We give normal value of normal users for each feature based on statistical technology. Then we can learn insights of spammer identification by comparing detected value with normal value just as if we watch a disease diagnose sheet. For example, normal value of normal users in feature similar

posts times is 0 to7, but ‘i 诗槐’ has a quite big value 21. This clue tells us that ‘i 诗槐’ often post same content again and again to strength what it wants to say.

5.3 Detect posts instead of users

Our approach for the spam problem in *Sina weibo* focuses on the detection of spammers instead of posts containing spam. The detection of the spam posts can be useful for filtering spam on real time search, so detection of spam posts is required. Intuitively the detection of spam post is associated with the detection of existent spammer accounts. We briefly give a rule that a post is a spam for a topic if a spammer posts it. We make use of a confusion matrix to show the experiment results in table 1 and the results indicate that this approach is acceptable. Although it is useful to have simple forms of spam detection, other techniques are still required when dealing with some scenarios in which this simple method may fail. Recently, considering spammer and spam detection together has been an interesting trend in research, Wu [18] combined social spammer detection with spam message detection to boost the performance of each task. However, their assumption is too strong in real-world data.

Table 1. Detect spam post instead of spammers

	Predicted	
	Spam Post	**Non-spam Post**
True Spam Post	79.2%	20.8%
Non-spam Post	17.3%	82.7%

6. CONCLUSION AND FUTURE WORK

In this paper, we have studied the problem of detecting spammers in *sina weibo*. We discovered that spammers are widespread now and sophisticated spammers sometimes are hard to distinguish from normal users. We propose to detect spammers in *sina weibo* microblog by a hierarchy classification approach. Experiments on our constructed real-world dataset illustrate that the method we applied is efficient. We also try to offer reasonable clues as evidence for spammer identifying by measure and compare feature values. For future work, we plan to study image features as many canny spammers only spam in photos instead of text or behavior.

7. ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation of China (Grant No. 91118005, 91218301, 91218301, 61428206, 61572399), Ministry of Education Innovation Research Team (IRT13035), Creative Project in Shanghai Science and Technology Council under Grant No. 12dz1506200, Key Projects in the National Science and Technology Pillar Program under Grant No. 2011BAK08B02.

8. REFERENCES

[1] Nitin, J. and Bing, L. Review Spam Detection. In *WWW POST*, 2007.

[2] Chengfeng, L., Yi, Z., Kai, C., Jianhua, H., Li, S. and XiaoKang, Y. Analysis and Identification of Spamming Behaviors in Sina Weibo Microblog. In *SNAKDD*, 2013.

[3] Yin, Z., Xiao, W., Erheng, Z., Nanthan, L., He, L. and Qiang, Y. Discovering Spammers in Social networks. In Proc. of the 26th *AAAI Conference on Artificial Intelligene.* 2012.

[4] Xia, H., Jiliang, T., Yanchao, Z. and Huan, L. Social Spammer Detection in Microblog. In *IJCAI*, 2013.

[5] Guan, W., Sihong, X., Bing, L. and Philip S, Y. Review Graph based Online Store Review Spammer Detection. In *ICDM*, 2011.

[6] Gyöngyi, Z., Garcia-Molina, H., Pedersen, J. Combating web spam with trustrank. In Proc. of the Thirtieth *international conference on Very large data bases*. Volume 30, pages 576-587, 2004.

[7] Benczúr, A., Bíró, I., Csalogány, K and Uher, M. Detecting nepotistic links by language model disagreement. In Proc. of the 15th *international conference on World Wide Web*, pages 939-940, 2006.

[8] Luiz Henrique, G., Cristiano, C., Jussara M, A., Virgílio, A. and Wagner, M. Workload models of spam and legitimate e-mails. In *Performance Evaluation.* Volume 64, pages 690-714, 2007.

[9] Richard, C. Using Early Results from the 'spam HINTS' Project to Estimate an ISP Abuse Team's Task. In *CEAS*, 2006

[10] ISnotSPAM: http://isnotspam.com

[11] Sarita, Y., Daniel, R. and Grant, S. Detecting spam in a twitter network. In First Monday 15(1), 2009.

[12] Alex Hai, W. Don't follow me: Spam detection in twitter. In Proc. of the 2010 International Conference on Security and Cryptography (SECRYPT). Pages 1-10.

[13] Hongyu, G., Yan, C., Kathy, L., Diana, P. and Alok N, C. Towards Online Spam Filtering in Social Networks. In Proc. of the symposium on network and distributed system security (NDSS), 2012.

[14] http://sd.skyclass.net:8080/Spammer/frame.jsp

[15] Joseph L, F. and Jacob, C. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 1973.

[16] Xianghan, Z., Zhipeng, Z. Zheyi, C., Yuanlong, Y. and Chunming, R. Detecting spammers on social networks. In Neurocomputing, 2015.

[17] Benevenuto, F. and Magno, G. Detecting Spammers on Twitter. In *CEAS*, 2010.

[18] Fangzhao, W., Jinyun, S., Yongfeng, H. and Zhigang, Y. Social Spammer and Spam Message Co-Detection in Microblogging with Social Context Regularization. In Proc. of the 24th ACM International on Conference on Information and Knowledge Management, 2015.