Detecting "Smart" Spammers On Social Network: A Topic Model Approach

Linqing Liu,[1] Yao Lu,[1] Ye Luo[1,*], Renxian Zhang[2,*], Laurent Itti[1,3] and Jianwei Lu[1,4,*]
[1] iLab Tongji, School of Software Enginnering, Tongji University
[2] Dept. of Computer Science and Technology, Tongji University
[3] Dept. of Computer Science and Neuroscience Program, University of Southern California
[4] Institute of Translational Medicine, Tongji University
likicode@gmail.com, {95luyao,rxzhang}@tongji.edu.cn,
{kennyluo2008,jwlu33}@hotmail.com, itti@usc.edu

Abstract

Spammer detection on social network is a challenging problem. The rigid anti-spam rules have resulted in emergence of "smart" spammers. They resemble legitimate users who are difficult to identify. In this paper, we present a novel spammer classification approach based on Latent Dirichlet Allocation (LDA), a topic model. Our approach extracts both the local and the global information of topic distribution patterns, which capture the essence of spamming. Tested on one benchmark dataset and one self-collected dataset, our proposed method outperforms other state-of-the-art methods in terms of averaged F1-score.

1 Introduction

Microblogging such as Twitter and Weibo is a popular social networking service, which allows users to post messages up to 140 characters. There are millions of active users on the platform who stay connected with friends. Unfortunately, spammers also use it as a tool to post malicious links, send unsolicited messages to legitimate users, etc. A certain amount of spammers could sway the public opinion and cause distrust of the social platform. Despite the use of rigid anti-spam rules, human-like spammers whose homepages having photos, detailed profiles etc. have emerged. Unlike previous "simple" spammers, whose tweets contain only malicious links, those "smart" spammers are more difficult to distinguish from legitimate users via content-based features alone (Ferrara et al., 2014).

*Corresponding Author

There is a considerable amount of previous work on spammer detection on social platforms. Researcher from Twitter Inc. (Chu et al., 2010) collect bot accounts and perform analysis on the user behavior and user profile features. Lee et al. (2011) use the so-called social honeypot by alluring social spammers' retweet to build a benchmark dataset, which has been extensively explored in our paper. Some researchers focus on the clustering of urls in tweets and network graph of social spammers (Yang et al., 2012; Wang et al., 2015; Wang, 2010; Yang et al., 2011), showing the power of social relationship features.As for content information modeling, (Hu et al., 2013) apply improved sparse learning methods. However, few studies have adopted topic-based features. Some researchers (Liu et al., 2014) discuss topic characteristics of spamming posts, indicating that spammers are highly likely to dwell on some certain topics such as promotion. But this may not be applicable to the current scenario of smart spammers.

In this paper, we propose an efficient feature extraction method. In this method, two new topic-based features are extracted and used to discriminate human-like spammers from legitimate users. We consider the historical tweets of each user as a document and use the Latent Dirichlet Allocation (LDA) model to compute the topic distribution for each user. Based on the calculated topic probability, two topic-based features, the Local Outlier Standard Score (LOSS) which captures the user's interests on different topics and the Global Outlier Standard Score (GOSS) which reveals the user's interests on specific topic in comparison with other users', are

extracted. The two features contain both local and global information, and the combination of them can distinguish human-like spammers effectively.

To the best of our knowledge, it is the first time that features based on topic distributions are used in spammer classification. Experimental results on one public dataset and one self-collected dataset further validate that the two sets of extracted topic-based features get excellent performance on human-like spammer classification problem compared with other state-of-the-art methods. In addition, we build a Weibo dataset, which contains both legitimate users and spammers.

To summarize, our major contributions are two-fold:

- We extract topic-based features (GOSS and LOSS) for spammer detection, which outperform state-of-the-art methods.
- We build a dataset of Chinese microblogs for spammer detection.

In the following sections, we first propose the topic-based features extraction method in Section 2, and then introduce the two datasets in Section 3. Experimental results are discussed in Section 4, and we conclude the paper in Section 5. Future work is presented in Section 6.

2 Methodology

In this section, we first provide some observations we obtained after carefully exploring the social network, then the LDA model is introduced. Based on the LDA model, the ways to obtain the topic probability vector for each user and the two topic-based features are provided.

2.1 Observation

After exploring the homepages of a substantial number of spammers, we have two observations. 1) social spammers can be divided into two categories. One is content polluters, and their tweets are all about certain kinds of advertisement and campaign. The other is fake accounts, and their tweets resemble legitimate users' but it seems they are simply random copies of others to avoid being detected by anti-spam rules. 2) For legitimate users, content polluters and fake accounts, they show different patterns on topics which interest them.

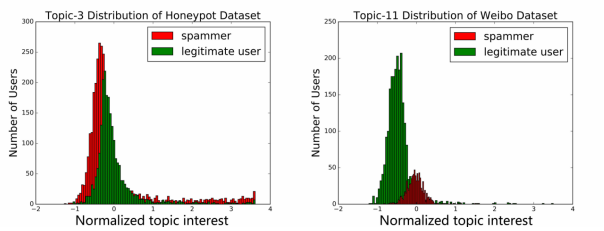

Figure 1: The topic distribution of legitimate users and social spammers on Honeypot dataset (left) and on Weibo dataset (right), respectively.

- Legitimate users mainly focus on limited topics which interest him. They seldom post contents unrelated to their concern.
- Content polluters concentrate on certain topics.
- Fake accounts focus on a wide range of topics due to random copying and retweeting of other users' tweets.
- Spammers and legitimate users show different interests on some topics e.g. commercial, weather, etc.

To better illustrate our observation, Figure. 1 shows the topic distribution of spammers and legitimate users in two employed datasets(the Honeypot dataset and Weibo dataset). We can see that on both topics (topic-3 and topic-11) there exists obvious difference between the red bars and green bars, representing spammers and legitimate users. On the Honeypot dataset, spammers have a narrower shape of distribution (the outliers on the red bar tail are not counted) than that of legitimate users. This is because there are more content polluters than fake accounts. In other word, spammers in this dataset tend to concentrate on limited topics. While on the Weibo dataset, fake accounts who are interested in different topics take large proportion of spammers. Their distribution is more flat (i.e. red bars) than that of the legitimate users. Therefore we can detect spammers by means of the difference of their topic distribution patterns.

2.2 LDA model

Blei et al.(2003) first presented Latent Dirichlet Allocation(LDA) as an example of topic model.

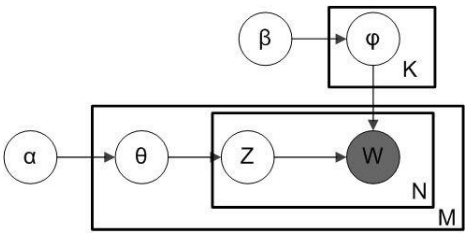

Figure 2: The generative model of LDA

Each document i is deemed as a bag of words $W = \{w_{i1}, w_{i2}, ..., w_{iM}\}$ and M is the number of words. Each word is attributable to one of the document's topics $Z = \{z_{i1}, z_{i2}, ..., z_{iK}\}$ and K is the number of topics. ψ_k is a multinomial distribution over words for topic k. θ_i is another multinomial distribution over topics for document i. The smoothed generative model is illustrated in Figure. 2. α and β are hyper parameter that affect scarcity of the document-topic and topic-word distributions. In this paper, α, β and K are empirically set to 0.3, 0.01 and 15. The entire content of each Twitter user is regarded as one document. We adopt Gibbs Sampling (Griffiths and Steyvers, 2004) to speed up the inference of LDA. Based on LDA, we can get the topic probabilities for all users in the employed dataset as: $X = [X_i; X_2; \cdots; X_n] \in R^{n \times K}$, where n is the number of users. Each element $X_i = [p(z_1)\, p(z_2) \cdots p(z_K)] \in R^{1 \times K}$ is a topic probability vector for the i^{th} document. X_i is the raw topic probability vector and our features are developed on top of it.

2.3 Topic-based Features

Using the LDA model, each person in the dataset is with a topic probability vector X_i. Assume $x_{ik} \in X_i$ denotes the likelihood that the i^{th} tweet account favors k^{th} topic in the dataset. Our topic based features can be calculated as below.

Global Outlier Standard Score measures the degree that a user's tweet content is related to a certain topic compared to the other users. Specifically, the "GOSS" score of user i on topic k can be calculated as Eq.(1):

$$\begin{aligned} \mu(x_k) &= \frac{\sum_{i=1}^{n} x_{ik}}{n}, \\ GOSS(x_{ik}) &= \frac{x_{ik} - \mu(x_k)}{\sqrt{\sum_i (x_{ik} - \mu(x_k))^2}}. \end{aligned} \tag{1}$$

The value of $GOSS(x_{ik})$ indicates the interesting degree of this person to the k^{th} topic. Specifically, if $GOSS(x_{ik}) > GOSS(x_{jk})$, it means that the i^{th} person has more interest in topic k than the j^{th} person. If the value $GOSS(x_{ik})$ is extremely high or low, the i^{th} person showing extreme interest or no interest on topic k which will probably be a distinctive pattern in the fowllowing classfication. Therefore, the topics interested or disliked by the i^{th} person can be manifested by $f^i_{GOSS} = [GOSS(x_{i1}) \cdots GOSS(x_{iK})]$, from which the pattern of the interested topics with regarding to this person is found. Denote $f^i_{GOSS} = [GOSS(x_{i1}) \cdots GOSS(x_{iK})]$ our first topic-based feature, and it hopefully can get good performance on spammer detection.

Local Outlier Standard Score measures the degree of interest someone shows to a certain topic by considering his own homepage content only. For instance, the "LOSS" score of account i on topic k can be calculated as Eq.(2):

$$\begin{aligned} \mu(x_i) &= \frac{\sum_{k=1}^{K} x_{ik}}{K}, \\ LOSS(x_{ik}) &= \frac{x_{ik} - \mu(x_i)}{\sqrt{\sum_k (x_{ik} - \mu(x_i))^2}}. \end{aligned} \tag{2}$$

$\mu(x_i)$ represents the averaged interesting degree for all topics with regarding to i^{th} user and his tweet content. Similarly to $GOSS$, the topics interested or disliked by the i^{th} person via considering his single post information can be manifested by $f^i_{LOSS} = [LOSS(x_{i1}) \cdots LOSS(x_{iK})]$, and $LOSS$ becomes our second topic-based features for the i^{th} person.

3 Dataset

We use one public dataset Social Honeypot dataset and one self-collected dataset Weibo dataset to validate the effectiveness of our proposed features.

Social Honeypot Dataset: Lee et al. (2010) created and deployed 60 seed social accounts on Twitter to attract spammers by reporting back what accounts interact with them. They collected 19,276 legitimate users and 22,223 spammers in their datasets along with their tweet content in 7 months. This is our first test dataset.

Our Weibo Dataset: Sina Weibo is one of the most famous social platforms in China. It has implemented many features from Twitter. The 2197 legitimate user accounts in this dataset are provided

Feature	Method	Weibo Dataset			Honeypot Dataset		
		Precision	Recall	F1-score	Precision	Recall	F1-score
GOSS	SVM	0.974	0.956	0.965	0.884	0.986	0.932
	Adaboost	0.936	0.929	0.932	0.874	**0.990**	0.928
	RandomForest	0.982	0.956	0.969	0.880	0.969	0.922
LOSS	SVM	0.982	0.958	0.97	0.887	0.983	0.932
	Adaboost	0.941	0.929	0.935	0.878	0.976	0.924
	RandomForest	0.986	0.956	0.971	0.882	0.965	0.922
GOSS+LOSS	SVM	0.986	0.958	0.972	0.890	0.988	**0.934**
	Adaboost	0.938	0.931	0.934	0.881	0.976	0.926
	RandomForest	**0.988**	**0.958**	**0.978**	**0.895**	0.951	0.922

Table 1: Performance comparisons for our features with three baseline classifiers

by the *Tianchi Competition*[1] held by Sina Weibo. The spammers are all purchased commercially from multiple vendors on the Internet. We checked them manually and collected 802 suitable "smart" spammers accounts.

Preprocessing: Before directly performing the experiments on the employed datasets, we first delete some accounts with few posts in the two employed since the number of tweets is highly indicative of spammers. For the English Honeypot dataset, we remove stopwords, punctuations, non-ASCII words and apply stemming. For the Chinese Weibo dataset, we perform segmentation with "Jieba"[2], a Chinese text segmentation tool. After preprocessing steps, the Weibo dataset contains 2197 legitimate users and 802 spammers, and the honeypot dataset contains 2218 legitimate users and 2947 spammers. It is worth mentioning that the Honeypot dataset has been slashed because most of the Twitter accounts only have limited number of posts, which are not enough to show their interest inclination.

		Predicted	
		Polluter	Legitimate
Actual	Polluter	TP	FN
	Legitimate	FP	TN

Table 2: Confusion matrix

[1]Tianchi Site http://tianchi.aliyun.com

[2]Jieba Project Page https://github.com/fxsjy/jieba

Feature	Description
UFN	standard deviation of following
	standard deviation of followers
	the number of following
	following and followers ratio
UC	|links| per tweet
	|@username| in tweets / |tweets|
	|unique @username| in tweets / |tweets|
	|unique links| per tweet
UH	the change rate of number of following

Table 4: Honeypot Feature Groups

4 Experiment

4.1 Evaluation Metrics

The evaluating indicators in our model are show in Table 2 . We calculate precision, recall and F1-score (i.e. F1 score) as in Eq. (3). Precision is the ratio of selected accounts that are spammers. Recall is the ratio of spammers that are detected so. F1-score is the harmonic mean of precision and recall.

$$precision = \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN}$$
$$F1-score = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

4.2 Performance Comparisons with Baseline

Three baseline classification methods: Support Vector Machines (SVM), Adaboost, and Random Forests are adopted to evaluate our extracted features. We test each classification algorithm with scikit-learn (Pedregosa et al., 2011) and run a 10-fold cross validation. On each dataset, the em-

Features	SVM			Adaboost		
	Precision	Recall	F1-score	Precision	Recall	F1-score
UFN	0.846	0.919	0.881	0.902	0.934	0.918
UC	0.855	0.904	0.879	0.854	0.901	0.877
UH	0.906	0.8	0.85	0.869	0.901	0.885
UFN+UC+UH	0.895	0.893	0.894	0.925	0.920	0.923
LOSS+GOSS	0.890	0.988	0.934	0.881	**0.976**	0.926
UFN+UC+UF+LOSS+GOSS	0.925	0.920	0.923	**0.952**	0.946	**0.949**

Table 3: Comparisons of our features and Lee et al.'s features

ployed classifiers are trained with individual feature first, and then with the combination of the two features. From Table 1, we can see that GOSS+LOSS achieves the best performance on F1-score among all others. Besides, the classification by combination of LOSS and GOSS can increase accuracy by more than 3% compared with raw topic distribution probability.

4.3 Comparison with Other Features

To compare our extracted features with previously used features for spammer detection, we use three most discriminative feature sets according to Lee et al. (2011)(Table 4). Two classifiers (Adaboost and SVM) are selected to conduct feature performance comparisons. Using Adaboost, our LOSS+GOSS features outperform all other features except for UFN which is 2% higher than ours with regard to precision on the Honeypot dataset. It is caused by the incorrectly classified spammers who are mostly news source after our manual check. They keep posting all kinds of news pieces covering diverse topics, which is similar to the behavior of fake accounts. However, UFN based on friendship networks is more useful for public accounts who possess large number of followers. The best recall value of our LOSS+GOSS features using SVM is up to 6% higher than the results by other feature groups. Regarding F1-score, our features outperform all other features. To further show the advantages of our proposed features, we compare our combined LOSS+GOSS with the combination of all the features from Lee et al. (2011) (UFN+UC+UH). It's obvious that LOSS+GOSS have a great advantage over UFN+UC+UH in terms of recall and F1-score. Moreover, by combining our LOSS+GOSS features and UFN+UC+UH features together, we obtained another 7.1% and 2.3% performance gain with regard to precision and F1-score by Adaboost. Though there is a slight decline in terms of recall. By SVM, we get comparative results on recall and F1-score but about 3.5% improvement on precision.

5 Conclusion

In this paper, we propose a novel feature extraction method to effectively detect "smart" spammers who post seemingly legitimate tweets and are thus difficult to identify by existing spammer classification methods. Using the LDA model, we obtain the topic probability for each Twitter user. By utilizing the topic probability result, we extract our two topic-based features: GOSS and LOSS which represent the account with global and local information. Experimental results on a public dataset and a self-built Chinese microblog dataset validate the effectiveness of the proposed features.

6 Future Work

In future work, the combination method of local and global information can be further improved to maximize their individual strengths. We will also apply decision theory to enhancing the performance of our proposed features. Moreover, we are also building larger datasets on both Twitter and Weibo to validate our method. Moreover, larger datasets on both Twitter and Weibo will be built to further validate our method.

References

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Chu et al.2010] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM.

[Ferrara et al.2014] Emilio Ferrara, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2014. The rise of social bots. *CoRR*, abs/1407.5225.

[Griffiths and Steyvers2004] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

[Hu et al.2013] Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. 2013. Social spammer detection in microblogging. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2633–2639. AAAI Press.

[Lee et al.2010] Kyumin Lee, Brian David Eoff, and James Caverlee. 2010. Devils, angels, and robots: Tempting destructive users in social media. In *ICWSM*.

[Lee et al.2011] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *ICWSM*. Citeseer.

[Liu et al.2014] Yu Liu, Bin Wu, Bai Wang, and Guanchen Li. 2014. Sdhm: A hybrid model for spammer detection in weibo. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 942–947. IEEE.

[Pedregosa et al.2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Wang et al.2015] Bo Wang, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2015. Making the most of tweet-inherent features for social spam detection on twitter. *arXiv preprint arXiv:1503.07405*.

[Wang2010] Alex Hai Wang. 2010. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE.

[Yang et al.2011] Chao Yang, Robert Chandler Harkreader, and Guofei Gu. 2011. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Recent Advances in Intrusion Detection*, pages 318–337. Springer.

[Yang et al.2012] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM.