Exploring and Detecting Opinion Spam on Social Media

Yang Xiao
Peking University
Beijing, China
xiaoyangpku@gmail.com

JieFan Qiu
Zhejiang University of Technology
Hangzhou, China
qiujiefan@zjut.edu.cn

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327093

***Abstract*—In recent years, microblogging service such as Twitter and Weibo attracts a large number of users. Unlike traditional media by which user can only accept information, social media allows users to share their opinions. Among the social media users, there exist a group of users called opinion spam. They are well organized and post a large number of purposed comments to misdirect the public opinion. In this way, they significantly magnify the impact of their employers. We conduct quantitative analysis to study and understand the characteristics of opinion spam. We analyze the psycholinguistic styles of opinion spam, explore their behavior patterns and network structure. Finally, based on the analysis, context based collective classification is proposed to detect opinion spam and the model can achieve 91% F1 score.**

***Index Terms*—social media, opinion spam, analysis, detection**

I. Introduction

Microblogging social network like Twitter and Weibo is social media that diffuses information fast through the link of trust. Among the social media users, a group of people is paid to post purposed comments in order to mislead public opinion. Asch conformity experiments [1] show that the majority have a great influence on people's opinions. It finds out that people want to be conform with the majority even when they know the majority is wrong. The opinion spam make use of this mechanism and use the overwhelmingly large number of comments to influence others. Their employers' purposes can be classified into two categories: to promote some viewpoints, or to crack down on opponents. Some popular stars hire the opinion spam to post supportive comments when they become focus of controversy; If the opinion spam are used to post deceptive point of view, the consequence can be severe. To illustrate this, a dairy company hires the opinion spam to spread rumors on its competitors and makes the latter's stock plunged.

Previous work on opinion spam [2] mainly focused on deceptive comments detection for electronic commerce websites like Amazon. These posters are paid to give undeserved high comment scores or intentional low scores to products. E-commerce opinion spam try to change the real average score, while the social media opinion spam post well-designed comments to change others' attitudes. Since there is no score along with the tweets, it is more difficult to automatically detect opinion spammers on social media. In this paper, we analyze the characteristics of the opinion spam on social media and develop a model to detect the opinion spam. Our main contributions can be summarized as follows: Firstly, we conduct empirical quantitative analysis on opinion spam. We look close into the psycholinguistic features, the behavior pattern, and social network structure of opinion spam. Secondly, we propose a context based classification method to detect opinion spam. We analyze the distinguishing ability of each feature dimension and find that the psycholinguistic dimension have the most powerful ability to distinguish opinion spam.

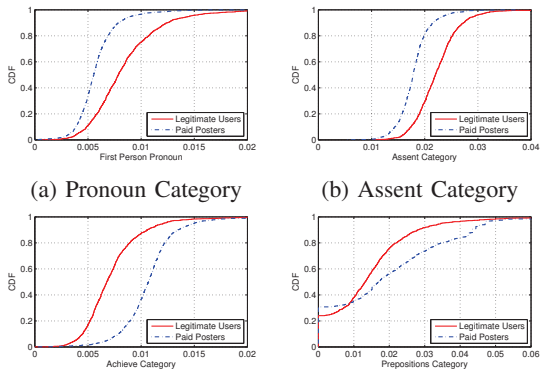
II. Related work

Researchers have analyzed the behavior and structure characteristics of fake accounts on social network [3, 4] and social media [5, 6]. Some existing works on spam detection exploit the network structure features. The nodes that are better connected to trusted nodes are more likely to be legitimate nodes, and how well a node is connected to trusted node can be taken as the credibility of the node [7, 8]. There are also some works that employ machine learning methods to detect the social network spam[9]. CrowdTuring is a special form of crowdsourcing whose purpose is to spread malicious url or product promotions. Since the message is posted by users distributed on the site, it appears like grassroots campaign. Wang[10] explore the crowdturfing systems in China and evaluate the effectiveness of product promotions. Social network spam are fake accounts that send commercial content to legitimate accounts, eg. message containing links to product. While ad messages sent by social network spam are easy to detect, comments posted by the opinion spam are closely related to a specific topic and are harder to distinguish.

III. Dataset

Previous work [11] concludes that it is hard to classify the concept of spam. Research work on opinion spam[12] and social spam [4] use manually annotation dataset as ground truth [2]. For opinion spam dataset constitution, we also choose manually annotations to construct ground truth dataset.

We interview those who have once worked as opinion spam, reference the guidelines that are used to detect opinion spam. Two assessors independently identified suspicious opinion spam using guidance as follows: 1)Users who post comments that are not relate to the topic. 2)Users who post comments that are incomplete. eg, some of the comments only contain part



(a) Pronoun Category (b) Assent Category

(c) Achievement Category (d) Prepositions Category

Fig. 1: Linguistic Category Distribution. Legitimate users tweet more pronoun and less achievement words, opinion spam tweet less achievement words and more prepositions. More prepositions and less social words are signals of complex sentences.

of words, which seem like that the opinion spam miss some characters when they copy from text. 3)Users whose comments are obviously fake. 4)Users whose tone are exaggerate, which are not consistent with the context. e.g., use the serious written language under a tweet on entertainment or use the extremely intimate words which are abnormal considering the context. 5)Users who post a large number of duplicate comments in a short time. Similar to previous work, this is not a critical rule for opinion spam. As some legitimate users have similar behavior, we also consider the semantic of the comments to judge the user. 6)Users who post same comments with others. Our dataset contains 19628 tweets and it is impossible to annotate all the corresponding comments. We choose a topic that is on a controversy event relates to a TV show under which we find evidence of opinion spam existence. We filter tweets that are related to the topic by using some keywords. In this way, we manually label 75228 distinct users.

IV. Characteristic Analysis

In this section, we empirically study the opinion spam in our ground truth dataset, including linguistic styles, tweet content and network structure.

A. Linguistic Styles

Linguistic inquiry and word count(LIWC) is a dictionary that classifies English words into psychological meaningful categories. LIWC demonstrates its ability to detect psychological meaning in a wide range of applications, eg, emotionality investigation, personality prediction. We analyze the LIWC features for both legitimate users and opinion spam to explore the linguistic style difference. As illustrated in Figure 1, legitimate users and opinion spam show very different distribution on some linguistic dimensions. Legitimate users use more first person pronouns than opinion spam in their tweets. Opinion spam use less spoken language words. However, on achievement category(eg, earn, hero, win), we find that opinion spam have much higher value compared with legitimate user. Opinion Spam tend to use more prepositions(eg, to, with, above), less 'you' and less social words. Research on psychology argues that sentences with more prepositions and less social words are complex. When opinion spam are paid to post comments, they tend to use the *complex* sentences, which are quite different from the *simple* oral comments on social media.

B. Content Analysis

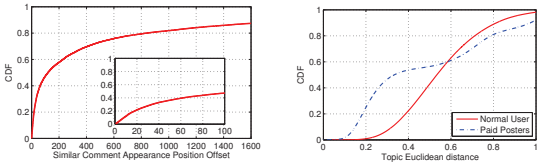
In order to better understand the content characteristic of opinion spam, we dig deeper to look into the topic distribution difference between legitimate users and opinion spam. As shown in Table I, the most dominant topic of opinion spam is lucky draw and promotion. In contrast, legitimate users are more interested in discussing the trending topics, such as politics, economy and constellation. We also compare the topic similarity within users in each group. For every pair of users within a group, we calculate the similarity between them. The similarity between user content is measured by Euclidean distance of two topic distributions. As shown in Figure 2b, topic Euclidean distances between opinion spam are much smaller than that of legitimate users. We also calculate the position offset of consecutive comments. As shown in Figure 2a, 80% of similar comments position offset is larger than 20, therefore, it is hard for legitimate users to be conscious of the duplicate comments existence.

TABLE I: Frequent Words of Top topics

User	Topic Name	Top Words
Opinion spam	Lucky draw	Match, Get, Phone, music, Philips, Collect, Address, Star, Celebrate
	Promotion	Shoppe, Voice, Price Lose weight, Skin, Advertisement, Link, Trend
Legitimate users	Entertainment	concert, Spring festival gala, contest, Student, Fan, club, Singer, Sing, Tutor
	Daily life	Me, Learn, Give up, Habit, Quotation, Think, Forget, Understand, Sleep

C. Social Network Analysis

We construct a graph based on all the users in our dataset. If two users in our dataset have following relationship, an edge is established. The user graph contains $75,228$ nodes and $68,019$ edges. The average degree of the whole graph is 1.81. For opinion spam, the average degree is 4.78, while for legitimate users, the value is 1.16. The degree of opinion spam are much higher than that of the legitimate users. We investigate edges whose two endpoints are in our ground-truth dataset. We find that 8.48% edges are formed by a legitimate user and a opinion spam, and 15.17% edges are formed by two legitimate users and 76.35% edges are formed by the two opinion spam. The edge between users represents homogeneity of the two endpoint users. Hence, we can exploit the relationship between users as context to better predict the label.



(a) Position offset (b) Topic Euclidean distance

Fig. 2: Content Metric Distribution

V. Opinion Spam Detection

We have already identify the content and structure characteristic of opinion spam. In this section, we study the opinion spam detection. As described in data description section, we collect the information about a user that is public available. In this paper, we categorize the user centric features into five categories as follows.

a) Comment linguistic styles: For every user, we calculate the count of words W that fall into every LIWC category, and get a 71 dimensions vector. The vector is then normalized by the sum of the all values in the vector. Each value in the vector represents the users' usage preference on the linguistic category.

b) Tweet linguistic styles: Tweet linguistic features construction is similar to comment linguistic features.

c) Comment behavior features: Comment behavior features include amount of comments, amount of reply to others' comments and the amount of duplicate comments.

d) Tweet Behavior patterns: Tweet behavior patterns features includes the ratio of tweets that contains hashtags, urls, emoticon, mentions, relationship between user and retweet author, average tweet amount a day and average length of a tweet.

e) User profile features: Profile features includes location, account created time, gender, amount of tweets, amount of followers, amount of friends, amount of bilateral friends, location, length of self description, length of username, and whether username includes number.

f) Context Information: Context information is very useful for predicting the characteristic of user. If a large quantity of the users' friends are opinion spam, it is highly probable that the user is also a member of opinion spam. To exploit the context information, we need to break the cyclic dependency between users. We build a collective classification model to combine the label of parent and the label of child. We take the context information as part of input, which is given in detail in the following section.

A. Classifier

In order to combine the features and the context information, we follow the framework proposed in previous work. As shown in Algorithm 1, we firstly construct a graph based on the training set in which every user is a node, and if one user follows another, then there is a direct edge between them. For each node in the training set, the node has a label that indicates whether it is opinion spam. We add two features based on the graph, one feature represents the user's opinion spam followers percentage, the other represents the user's opinion spam followees percentage. From the very start, we build a logistic regression classifier using all the features including comments, tweets, user profile related features and the two context features. For the test procedure, each user in the test dataset is initialized using the all the features except the context features and get a label. Then the following iteration starts, as the context information is known for the instances in the test dataset, we can apply the classifier proposed in the training dataset and get a probability of being opinion spam. If the probability is larger than a threshold, the label of the instance is updated. Similar to the technique used in previous work, the threshold decreases as the number of iteration grows. In this way, nodes with low confidence will not affect its neighbors until the end of the algorithm.

Algorithm 1 Context Based Classification

Input: Users features(including comment, tweet, profile related features) X, user labels in the training set Y, social graph between users G

Output: Label inferred for every user in the test dataset.

1: Construct context features C based on G and Y for every user in the training dataset.
2: Build a logistic regression classifier $LR_{context}$ based on user features X and context features C
3: Build a logistic regression classifier LR_{user} based on user features X only.
4: Assign initial probability to every user in test set using LR_{user} classifier
5: **for** iter=1 to 60 **do**
6: threshold $\theta = \max(100 - iter, 50)$
7: **for** j=1 to sizeof(training set) **do**
8: $\vec{x} = [X_j, C_j]$
9: $p = LR_{context}(\vec{x})$
10: **if** $p > \theta$ **then**
11: update the label of the instance
12: **end if**
13: **end for**
14: **end for**

B. Experiment Results Analysis

We setup experiments using the ground truth dataset that we have manually built. All the experiments are done in the 10-fold cross validation. The performance of the classifier is evaluated by precision, recall and F1 value.

As shown in Table II, behavior features and profile features work reasonably well for detection, i.e., comment behavior alone can get 0.711 F1 value, tweet behavior can achieve 0.758 F1 value and profile features can reach 0.712 F1 value. When context-based classifier is applied, both kinds of behavior features get improved F1 value. The context information can relieve the sparsity problem to some extent. For example, if a user only post one comment, the classifier is not confident with

the result. Given that a large portion of the users' friends are opinion spam, the confidence of the classifier become larger.

TABLE II: Experiment results using different methods

Types	Classifier	Features	Precision	Recall	F1
Behavior	LR	C-Behavior	0.587	0.902	0.711
	LR	T-Behavior	0.815	0.709	0.758
	LR-Context	C-Behavior	0.689	0.856	0.763
	LR-Context	T-Behavior	0.795	0.793	0.794
Profile	LR	Profile	0.77	0.663	0.712
	LR-Context	Profile	0.828	0.75	0.787
Content	LR	C-Content	0.782	0.78	0.781
	LR	T-Content	0.87	0.843	0.856
	LR-Context	C-Content	0.794	0.83	0.811
	LR-Context	T-Content	0.885	0.858	0.871
ALL	LR-Context	All	**0.901**	**0.919**	**0.91**

Linguistic features are powerful in predicting the label of the user. Both tweets LIWC and and comment LIWC outperform the corresponding behavior features. Tweets LIWC can achieve 0.856 F1 value and Comment LIWC can get 0.781 F1 value. When context based classifier is applied, Comment LIWC features performs 3.0% better and tweet LIWC features performs 1.5% better. Comment features experience more sparsity compared with the tweet features, therefore, when we add context information, comment LIWC gets a larger improvement. From this we can conclude that the psycholinguistic style between opinion spam and legitimate users are quite different. In order to better understand the detailed difference of the psycholinguistic style. We extract the most important tweet LIWC features by calculating the mutual information between label and every feature. Mutual information is defined as the reduction of entropy of labels Y achieved by learning the state of the features X_i. The top features selected is shown in Table III.

$$I(Y; X_i) = H(Y) - H(Y|X_i)$$

As shown in Table III, opinion spam have higher positive emotions, lower negative emotions and express less anger. Social psychology theory shows willingness to express negative emotions promote the relationships and indicate trust between people. The opinion spam accounts express much more positive emotions not because the people behind the accounts are much happier than legitimate users. Legitimate users use social media to report updates of themselves and express negative emotions frequently to their followers and seek for comfort. While for opinion spam, the account is maintained not to express themselves but to make money, and the follower of the account is highly probable of being opinion spam, hence, less negative emotions are expressed. Assent and nonfluency belong to the LIWC spoken categories. Both nonfluency and assent occur less frequently in opinion spam's tweets. Compared with legitimate users, opinion spam use less words that fall into spoken language. Deceptive statements compared with truthful statements are distanced from self [13] and use more motion words, which is consistent to the phenomena we discover in this case.

TABLE III: Top Linguistic Categories Ranked by Mutual Information

Category	Legitimate User Avg	Paid Poster Avg
achieve	0.72%	1.07%
assent	2.22%	1.79%
motion	1.2%	1.5%
negative emotion	0.79%	0.56%
leisure	0.53%	0.76%
insight	1.72%	1.44%
nonfluency	0.38%	0.23%
positive emotion	1.84%	2.06%
anger	0.19%	0.11%

VI. Conclusion

In this paper, we focus on exploration and detection of opinion spam. Opinion spam are paid to post comments and misdirect public opinion. We manually annotate ground truth dataset, quantitatively analyze the content and structure character of opinion spam. Internet Water Army tend to form a close knit and have different linguistic usage convention compared with legitimate users. Finally, to detect opinion spam, collective classification that combine the context information with the user centric features is applied, and empirical evaluation proves its effectiveness in distinguishing opinion spam.

References

[1] S. Asch, "Studies of independence and conformity: I. a minority of one against a unanimous majority," *Psychological Monographs: General and Applied*, 1956.

[2] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of WSDM*, 2008.

[3] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of IMC*. ACM, 2010.

[4] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," in *Proceedings of IMC*. ACM, 2011.

[5] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *Proceedings of CCS*. ACM, 2010.

[6] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of twitter spam," in *Proceedings of IMC*. ACM, 2011.

[7] B. Viswanath, A. Post, K. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," in *ACM SIGCOMM Computer Communication Review*, 2010.

[8] H. Yu, P. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," in *IEEE Symposium on SP*, 2008.

[9] N. Tran, B. Min, J. Li, and L. Subramanian, "Sybil-resilient online content voting," in *Proceedings of NSDI*. USENIX Association, 2009.

[10] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Zhao, "Serf and turf: Crowdturfing for fun and profit," in *Proceedings of WWW*. ACM, 2012.

[11] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, "A reference collection for web spam," in *ACM Sigir Forum*, vol. 40, no. 2. ACM, 2006, pp. 11–24.

[12] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of WWW*. ACM, 2012.

[13] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, 2010.