Public Opinion Spamming: A Model for Content and Users on Sina Weibo

Ziyu Guo
Shandong University
Jinan, China

Liqiang Wang*
Shandong University
Jinan, China

Yafang Wang†
Shandong University
Jinan, China

Guohua Zeng
Chinese Academy of Social Sciences
Beijing, China

Shijun Liu
Shandong University
Jinan, China

Gerard de Melo
Rutgers University – New Brunswick
USA

ABSTRACT

Microblogs serve hundreds of millions of active users, but have also attracted large numbers of spammers. While traditional spam often seeks to endorse specific products or services, nowadays there are increasingly also paid posters intent on promoting particular views on hot topics and influencing public opinion. In this work, we fill an important research gap by studying how to detect such opinion spammers and their micro-manipulation of public opinion. Our model is unsupervised and adopts a Bayesian framework to distinguish spammers from other classes of users. Experiments on a Sina Weibo hot topic dataset demonstrate the effectiveness of the proposed approach. A further diachronic analysis of the collected data demonstrates that public opinion spammers have developed sophisticated techniques and have seen success in subtly manipulating the public sentiment.

CCS CONCEPTS

• **Information systems → Spam detection**; • **Human-centered computing → Social media**; • **Applied computing → Sociology**;

KEYWORDS

Opinion Spam, Public Opinion, User Classification

ACM Reference Format:
Ziyu Guo, Liqiang Wang, Yafang Wang, Guohua Zeng, Shijun Liu, and Gerard de Melo. 2018. Public Opinion Spamming: A Model for Content and Users on Sina Weibo. In *WebSci '18: WebSci '18 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3201064.3201104

*The first two authors contributed equally.
†Corresponding author: yafang.wang@sdu.edu.cn

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201104

1 INTRODUCTION

Social network services provide platforms for massive information dissemination and sharing between hundreds of millions of users. Unfortunately, they also have led to new opportunities for malicious users. This is particularly true of the most well-known Chinese microblogging platform Sina Weibo, which reportedly has a larger base of daily active users than Twitter. Hot, trending topics on this platform attract remarkable public interest and have substantial significance for business and society. As a result, it has attracted spammers with malicious intent. Widespread spamming threatens the quality and credibility of the user-generated content on social media platforms, and erodes the publicness of these platforms. Thus, it is important to develop techniques to detect such spammers and to examine their impact on the formation of public opinion.

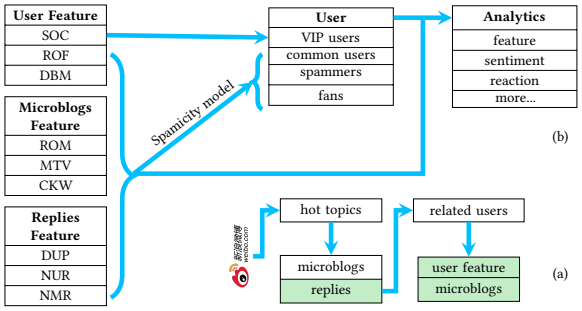


Figure 1: Data capture and model.

The publicness of social media platforms has long been a major concern in academia. Early research highlights the political and social polarization [13, 18], and the impact of the underlying algorithms [14]. More recent research focuses on the mechanisms and impact of "fake news" created and circulated on social media [1, 17]. However, how opinion spammers subtly steer and manipulate the public opinion on such platforms and what impact these micro-techniques may exert on society remains underexplored.

There has been ample research on detecting spammers, including specifically for Sina Weibo [2, 3, 9, 15], as well as several supervised learning methods [5, 12, 16] to detect instances of opinion spam. However, such models are largely based on a dichotomy of fake vs. non-fake labels. Unsupervised methods have as well been proposed [4, 6, 8, 10, 11, 20]. Despite this progress, such previous work has focused on identifying spammers seeking to place ads for products or services, as well as detecting imposters, extremists and the like.

This paper instead focuses on the unique problem of *public opinion spamming*, i.e. identifying spammers that seek to influence public opinion on hot topics. As we will explain in further detail, such actors operate quite differently both from traditional ad-like product promotion spammers [7, 19] as well as from the kind of opinion spammers that post fake product reviews. We propose a novel and principled model to detect public opinion spamming. The model is an unsupervised one that does not require labeled training data and overcomes the limitations of existing work discussed above. We adopt a fully Bayesian modeling approach. This setting allows us to model the *spamicity* of users as latent, while treating other observed behavioral features as known. The *spamicity* here refers to the degree to which the exhibited behavior can be regarded as public opinion spamming. Our key ideas hinge on the hypothesis that opinion spammers differ from others on behavioral dimensions. This creates a separation margin between the population distributions of three naturally occurring groups: spammers, fans, and regular users. The inference procedure enables us to learn the distributions of these groups by means of the behavioral features.

Figure 1(b) illustrates our approach. Based on pertinent social media data, we extract features from user profiles, postings, and replies under hot topics. For user classification, we identify VIP users based on the official certification on the platform, while other features are used to train our model to assess a user's degree of spamicity and categorize them with respect to the three remaining clusters. Subsequently, we proceed to explore how and to what extent spammers shape the public opinion on specific topics.

In summary, this paper makes the following contributions:

(1) It proposes a novel and principled method to exploit observed behavioral footprints to classify users and detect public opinion spammers in an unsupervised Bayesian framework, without the need for laborious manual labeling, which is both time-consuming and error-prone. Unlike existing work, this allows both detection and analysis to occur in a single framework, providing deeper insights into the data.
(2) We conduct a comprehensive set of experiments to evaluate the proposed model based on human expert judgments.
(3) On the basis of the spammer behavior detection, we conduct a diachronic analysis of a specific case, "Wang Baoqiang's divorce", to examine the effectiveness of the spammers' behavior in shaping public opinion. The results showcase that the spammers have developed rather sophisticated tactics in reshaping the public opinion, which calls for more attention in academia and industry to be paid to this underexplored, yet extremely important issue.

2 DATA

In the following, we consider an instructive example as a case study. On August 14, 2016, popular Chinese actor Wang Baoqiang issued a public statement accusing his wife Rong Ma, an actress, of having an extramarital relationship with Wang's agent Zhe Song and collusively transferring their mutual assets to Song. He went on to denounce their wedlock with a lawsuit against Rong Ma. This statement stirred up enormous, long-lasting attention in both digital and traditional media outlets. Sina Weibo was one of the major involved online platforms – The hashtag "#Wang Baoqiang divorced#" (in Mandarin) and related ones frequently emerged in the Top Topics lists.

Thus, we collected pertinent data from Sina Weibo from 14 August to 15 December 2016, a period when this event attracted massive public attention. Sina Weibo contains **Voting posts**, as well as **Topic posts**, in which certain keywords are marked with ##. We first crawled the 440 most popular microblog postings about the hot topic *#Wang Baoqiang divorced#* as seeds, as well as replies to them. We then retrieved all relevant users for these comments and replies. From their home pages, we then crawled their postings in the same time window. After data cleaning, we chose 2,000 users for our experiment, with data from December 2016. The posting features are computed for a user's postings posted from August to December 2016. In May 2017, we re-crawled the data again to check if these users were banned or the topic-related postings were deleted. Finally, in August 2017, we re-crawled the data once again to determine whether any such ban had been lifted.

3 MODEL

3.1 Observed Features

Users participating in a hot topic are categorized into four different subsets: *regular users*, *fans* (i.e., enthusiastic devotees or admirers of one of the parties), *spammers* (i.e., paid posters specifically seeking to sway public opinion), and *VIP users* (i.e., those verified by Sina Weibo). In the following, we propose some characteristics of abnormal behavior that may prove useful as observed features in our model to learn to distinguish these clusters of users.

User Reply Features: Replies here refer to a user's responses to hot topic postings. Spammers often post multiple replies that are duplicate or near-duplicate versions of previous replies or replies of others on the same topic (**DUP**). The number of user replies (**NUR**) and number of postings that a user responded to (**NMR**) are two important features to detect spammers, due to the more limited time and effort spent online by regular users.

Posting Features: Posting features are based on all postings that a user has made within a given time period, beyond just those pertaining to the hot topic under consideration. Regular users tend to express their personal opinion in original postings, while spammers tend to copy template postings for efficiency. To highlight their arguments, spammers also post or repost more topic postings and voting postings. They also tend to post more postings containing certain specific keywords to make the topic more hot. Correspondingly, for each user, we compute the ratio of original mircroblog postings (**ROM**), the ratio of that user participating in the postings about topics or with voting polls (**MTV**), and the ratio of postings containing keywords (**CKW**).

User Features: We select three features, taken from the user profile data, as features: whether the user deletes all of their postings or the user is banned a few months later (**DBM**), the ratio of followers to followees (**ROF**), and the Sina official certification (**SOC**). The SOC feature is not considered in the model, but instead serves as a marker to identify VIP users.

3.2 The Graphical Model

A number of factors may aid in spam detection, including replies on a particular hot topic, a user's postings on their microblog, and user features. Normalized continuous features in $[0, 1]$ are modeled as

following a Dirichlet distribution. This enables the model to capture more fine-grained dependencies between user behavior and spamming. θ_k^f for each feature $f_1, \ldots, f_8$ denote the per class/cluster (spam vs. non-spam) probability of emitting feature f. Latent variables π_U denote the spamicity of a user U. The objective of the model is to learn the latent behavior distributions for *spammer*, *fan*, and *common user* clusters along with spamicity scores of users as π_U. We detail the generative process in Algorithm 1. For model inference, we rely on Gibbs sampling with the following equations:

$$p(\pi_U = i \mid \pi_U = -i) \propto (n^{\pi_{Ui}} + \gamma) \prod_f \frac{n_f^{\pi_{Ui}} + \alpha^f}{n^{\pi_{Ui}} + U^f \alpha^f} \quad (1)$$

$$f \in \{DUP, NUR, NMR, ROM, MTV, CKW, DBM, ROF\}$$

Notations	Description
u; U	User u; set of all users U
π_U	Spam/Non-spam class label for users based on homepage
α^f	Dirichlet shape parameters (priors) for θ^f for each feature f
β	Dirichlet shape parameters (priors) for π_U of users
θ^f	Per class prob. of exhibiting the user behavior, for $f_1, \ldots, f_8$
π^f	The class each of a user's features belongs to, for $f_1, \ldots, f_8$
$n^{\pi_{ui}}$	Counts of user u being assigned to i
$n_f^{\pi_{ui}}$	Counts of feature f of user u being assigned to i
U^f	Total number of features f

Table 1: List of notational conventions.

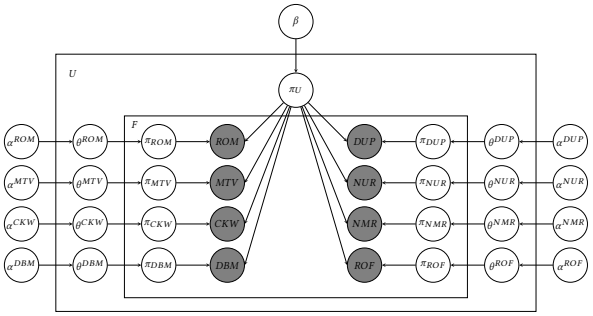


Figure 2: Plate notation

Algorithm 1 The generation process for users.

1: **for** each cluster π_U **do**
2: Draw a user type mixture distribution $\pi_U \sim$ Dirichlet(β).
3: **for** each user $u \in U$ **do**
4: **for** each feature $f \in \{1, \ldots, 8\}$ **do**
5: Draw a multinomial distribution $\theta^f \sim$ Dirichlet(α^f)
6: Draw user type assignment $\pi_f \sim$ Multinomial(θ^f)
7: Draw spamicity for feature f from distribution π_U with π_f

4 EXPERIMENT 1 – USER CLASSIFICATION

Our first experiment focuses on distinguishing between common users, fans, spammers, and VIP users for a given hot topic. For this, we are not aware of any gold-standard labeled data identifying public opinion spammers. Hence, we hired 15 students to label users manually. The judges are first briefed with many typical characteristics of public opinion spam: The content is not practical and full of praise or belittling words. The content is purely praise without counterarguments for one party or purely negative without counterarguments for the other party. The postings posted earlier and later do not match. Given a user, their postings, and their replies, the judges were asked to independently examine the entire profile and to provide a label so as to classify the user.

From these users, we selected 2,000 users, including 500 spammers, 500 fans, 500 common users, 70 VIP users, and some random users for our experiments. In our supervised experiments, among the spammers, fans, and common users, we use 300 for training and reserve 200 for testing. Among the VIP users, we use 50 for training and 20 for testing. In our unsupervised experiments, we considered the users with the top 150 spamicities as spammers, the ones with the lowest 150 spamicities as regular users, and the 150 users with spamicities closest to 0.5 as fans, as well as 60 VIP users.

Model with Estimated Priors (MEP). This setting estimates the hyperparameters α^f, $f_1, \ldots, f_8$, and β by a Monte Carlo EM algorithm, which learns hyperparameters α and β that maximize the model's complete log-likelihood L. Posterior estimates are drawn after 3,000 iterations with an initial burn-in of 250 iterations.

SVM. As we have manually annotated some users, we can use supervised support vector machines (SVM) as a baseline.

	MEP (unsupervised)			SVM (supervised)		
	B_1	B_2	B_3	B_1	B_2	B_3
P	0.87	0.87	0.85	0.65	0.65	0.64
R	0.83	0.77	0.99	0.8	0.42	0.68
F_1	0.85	0.81	0.91	0.72	0.51	0.66
kappa coefficient	0.7931			0.4542		

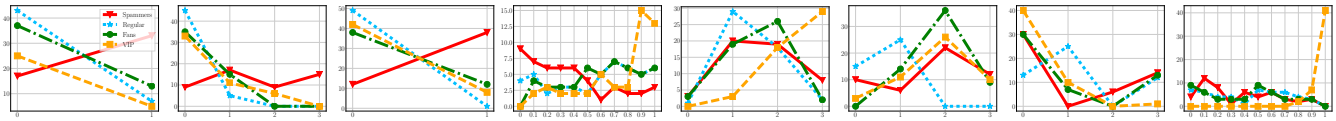
Table 2: P: Precision, R: Recall, and F_1: F_1-Score of using the average human evaluation results, B_1: spammers, B_2: fans, B_3: regular users.

Discussion. From the results in Table 2, we observe that the proposed MEP is remarkably effective at discriminating between the groups of users, despite being an unsupervised algorithm. Only the predictive accuracy for fans is comparably lower. On one hand, some fans behave much like spammers, especially when they continue expressing their views and arguing with others to defend the interests of one party. On the other hand, when some fans just present their opinions without continuously paying attention to the topic-related discussions, their behavior is quite similar to that of regular users. Furthermore we notice that the kappa coefficient of MEP is much higher, which indicates not only the reliability of the MEP results, but also that the SVM results are more difficult for the judges to estimate. Section 5 will explain the differences in behavior between fans, spammers, and regular users in more detail.

5 EXPERIMENT 2 – FEATURE ANALYSIS

Apart from generating a spamicity scores π_U for users, the model also estimates θ^f, the latent distributions of users' spamicity scores corresponding to each observed feature dimension f, as reflected in the spamicity. It is interesting to analyze the posterior on the learned distributions θ^f for each feature dimension f. We report the posterior on the latent spamicity distributions under each feature $f(\theta^f)$ estimated by MEP.

Duplicate/Near Duplicate Comments (π^{DUP}). From Figure 3(a), where 0 means *non-duplicate reply users* and 1 means *duplicate or*



(a) DUP (b) NUR (c) NMR (d) ROM (e) MTV (f) CKW (g) DBM (h) ROF

Figure 3: The frequency distribution of arranged events. Spammers (solid red), regular users (dotted blue), fans (dash-dotted green), and VIP users (dashed orange)

near-duplicate reply users, we note that many spammers post numerous duplicate or near-duplicate replies, while fans and regular users as well as VIP ones post very few such duplicates. However, compared with common users and VIP users, the number of duplicates for fans is somewhat higher. This feature is in line with expectations and contributes quite notably to the model.

The number of user replies (π^{NUR}). In Figure 3(b), there are four kinds of users with different amount of replies (N_r): 0 ($N_r = 0$), 1 ($0 < N_r \leqslant 5$), 2 ($5 < N_r \leqslant 10$), and 3 ($10 < N_r$). The density curve for non-spammers reaches its peak towards the left, evincing that non-spammers attain much lower values for NUR. Spammers yield an ascending curve, showing that they attain much higher values for NUR. In addition, the average number of replies is 9.5 for spammers, 1.1 for common users, 1.4 for fans, and 1.5 for VIP users.

The number of microblog postings that the user responded to (π^{NMR}). In Figure 3(c), 0 implies that the user responded to one, while 1 means that the user responded to more than one posting. This feature is very similar to DUP, with the difference that most regular users only reply to one posting and VIP users include some replying to more than one posting. Further analysis reveals some VIP users exhibiting spammer-like behavior for a given hot topic.

Ratio of Original Microblogs (π^{ROM}). In Figure 3(d), the scale of $0 \ldots 1$ refers to the ratio of original postings. Unlike for the aforementioned behaviors, the density curve for spammers has its peak towards the left, showing that spammers attain much lower values. Fans and regular users are very similar in their behavior. VIP users show a very high ratio of original postings, averaging about 0.95, which means that almost all of their postings are original.

Ratio of User participating in Topic and Voting Microblog Postings (π^{MTV}). In Figure 3(e), for spammers, the peak value is smaller and the extreme value is larger than for regular users or fans. This shows that the spammer's MTV values are not concentrated. Apart from VIP users, it is more difficult to distinguish between the remaining three categories of users. Relative to other characteristics, this feature's contribution to the model is rather low.

Ratio of Containing Keywords (π^{CKW}). In Figure 3(f), there are four kinds of users with different amounts of postings containing relevant keywords (N_p), 0 ($N_p = 0$), 1 ($0 < N_p \leqslant 5$), 2 ($5 < N_p \leqslant 100$) and 3 ($10 < N_p$). For this feature, regular users reach their peak on the left, whereas spammers, VIP users, and fans reach theirs on the right. This implies that except for regular users, most of the considered users post hot topic-relevant posts.

Whether the user deletes all their postings or is banned later (π^{DBM}). In Figure 3(g), there are four kinds of users, 0 (users who have never been banned or deleted all of their postings), 1 (who have been banned a few months later and the ban was not lifted), 2 (users who delete the postings on their home page), 3 (user who have been banned but the ban was soon thereafter lifted). Banned users for whom the ban was soon lifted are usually spammers rather than fans. Users deleting all their postings tend to be spammers. Yet, users who have been banned without the ban being lifted, interestingly, tend *not* to be spammers for particular topics.

Ratio of Followers (π^{ROF}). In Figure 3(h), values in $0 \ldots 1$ refer to the ratio of followers, as defined earlier. The curves for this feature are similar to those for ROM, in that the density curve for spammers attains it peak towards the left of the plot, while others attain their peak towards the right of the plot. That is to say, most spammers have fewer followers, and vice versa. In addition, VIP users have a very high ratio of followers.

6 EXPERIMENT 3 – SENTIMENT ANALYSIS

Sentiment Evaluation. Our final experiment assesses the sentiment, i.e. positive/negative attitude. This is computed based on a sentiment lexicon labeling words as "negative" (negative dictionary D_{neg}) or "positive" (D_{pos}). We also consider any boosting words appearing before a sentiment-bearing term to enhance the weight of that term, e.g. "very", "extremely", etc. The default for the boost score b_t is 1, while if a boosting word is encountered for a term t, it is set to 2. For each post, we first split it into words or phrases as a term set P. Then, we compute the sentiment as follows.

$$S_p = \frac{\sum_{t \in P \cap D_{\text{pos}}} b_t - \sum_{t \in P \cap D_{\text{neg}}} b_t}{|P|} \tag{2}$$

Figure 4 plots the diachronic trends of the sentiment towards this event, and considers the volume of spamming posts over time.[1] Spamming activities emerged right at the onset of this event and appears to have exerted a strong influence on the public attitude. Initially, spammers in favor of Wang aided in mobilizing a positive sentiment towards Wang. However, after reposting unverified claims alleging that Wang had as well had extramarital affairs and had exhibited domestic violence behavior, spammers supporting Rong Ma swayed back the public attitude, and wrestled with the opposing side for about a month. From September 10 through October 3, including the September 20–23 spamming surge, there was extensive publicity for Wang's new film *Buddies in India*, which indirectly contributed to steering the sentiment in a positive direction.

Two more indicative spamming periods emerged afterwards. The first one, from October 20 through November 15, was in favor of Wang. Having witnessed a negative trend lasting for several days, spammers supporting Wang initiated a strong fightback campaign

[1]Note that on some dates, e.g., October 17 and November 28, 2016, the number of posts genuinely relevant to the topic was lower than indicated in Fig. 4 (top), because on these dates, there were many other spamming posts lacking any identifiable stance towards this event, most of which were promoting other businesses, e.g., finance management, divorce consultancy, and private detective services

(including employing or coincidentally attracting spammers with an unidentifiable stand), which largely deterred the negative trend, until around October 28, when spammers in favor of Ma reversed the trend. Although Wang's spammers still effected a resurge on November 4, the second more decisive period was from November 5. Spammers supportive of Ma exhibited sophisticated manipulating skills and successfully remained in control for more than 20 days, despite several minor efforts from the opposing side.

Figure 4: Sentiment, posting amount, and post type of spammers for the two main protagonists involved in the event.

Another noteworthy result drawn from our diachronic analysis is that opinion spammers on Sina Weibo displayed a deliberate micro-tactic of hiding. Given that it is not possible anymore to analyze the impact of spamming "liking" behavior (Sina Weibo no longer displays the users who "liked" a posting), this work focuses on reposting of existing articles, posting of original articles, and replies to posts. The result reveals a different finding from Allcott & Gentzkow [1], and further works, which posit that posting thematic articles serves a vital role in mobilizing endorsement to a specific political opinion. The opinion spammers on Sina Weibo, in contrast, deliberately avoid posting or reposting articles. Instead, they preferred to reply to existing posts to avoid mention (@) of their client's names, trying to alter the general attitude towards a post (tweet) with overwhelmingly sentimental replies. Since Sina Weibo typically displays replies to a posting one by one under that tweet, this practice can often create an exclusive "bubble filter" [13] that repels users on the opposite side. The replies may evoke the feeling in other readers that the opinion reflected in the article is false (if replies denounce it) or true (if replies support it). Original writing is also an option less frequently used. This specific combination of spamming tactics, while being effective in shifting the public sentiment towards an event as analyzed, makes the spamming activity more challenging to detect (Sina Weibo deletes tweets or posts that are deemed spam or for which they receive heavy complaints of it being such), and therefore, more subtle and effective.

7 CONCLUSIONS

This paper proposes a novel and principled method to exploit observed microblog posting behavior to detect spammers in the special setting of public opinion spamming on Sina Weibo, and examine the impact it exerts on the public opinion. The precision of model affirmed the estimated characterization of spamming behavior. Based on the precise detection of public opinion spamming, a diachronic analysis about the impact of opinion spammers on a widely noted case in China demonstrates that such spammers subtly manipulated the public sentiment on Sina Weibo, one of the top social media platforms in China. This work, therefore, sets the path towards new research on public opinion spamming, and calls for a more detailed and nuanced analysis of the spammers' impact on public opinion, and potentially, on the social justice and well-being of the society.

8 ACKNOWLEDGEMENTS

The authors wish to acknowledge the support provided by the National Natural Science Foundation of China (61503217, 91546203), the Key Research and Development Program of Shandong Province of China (2017CXGC0605) and China Scholarship Council (201606220187). Gerard de Melo's research is funded in part by ARO grant W911NF-17-C-0098 (DARPA SocialSim).

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election.* Working Paper 23089. National Bureau of Economic Research.
[2] Hao Chen, Jun Liu, Yanzhang Lv, Max Haifei Li, Mengyue Liu, and Qinghua Zheng. 2017. Semi-supervised Clue Fusion for Spammer Detection in Sina Weibo. *Information Fusion* (2017).
[3] Hao Chen, Jun Liu, and Jianhong Mi. 2016. SpamDia: Spammer Diagnosis in Sina Weibo Microblog. In *MobiMedia 2016.* 116–120.
[4] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. 2013. Exploiting burstiness in reviews for review spammer detection. In *ICWSM.*
[5] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *ACL short.* 171–175.
[6] Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012. Distributional Footprints of Deceptive Product Reviews. In *ICWSM.*
[7] Kunal Goswami, Younghee Park, and Chungsik Song. 2017. Impact of reviewer social interaction on online consumer review fraud detection. *Journal of Big Data* 4, 1 (15 May 2017), 15.
[8] Ee-Peng Lim, Viet An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *CIKM.*
[9] Yingcai Ma, Niu Yan, Ren Yan, and Yibo Xue. 2013. Detecting Spam on Sina Weibo. *CCIS-13* (2013).
[10] Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *WWW.* 191–200.
[11] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, and Nitin Jindal. 2011. Detecting group review spam. In *WWW Companion.* 93–94.
[12] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. (2011), 309–319.
[13] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* The Penguin Group.
[14] Cornelius Puschmann and Jean Burgess. 2014. *The Politics of Twitter Data.* Peter Lang Publishing Inc.
[15] Yang Qiao, Huaping Zhang, Min Yu, and Yu Zhang. 2016. Sina-Weibo Spammer Detection with GBDT. In *Chinese National Conference on Social Media Processing.*
[16] Shebuti Rayana and Leman Akoglu. 2015. Collective Opinion Spam Detection: Bridging Review Networks and Metadata. In *SIGKDD.* 985–994.
[17] Laura Spinney. 2017. The Shared Past that Wasn't: Facebook, fake news and friends are warping your memory. 543 (2017), 168–170.
[18] Cass R Sunstein. 2009. *Going to extremes: How like minds unite and divide.* Oxford University Press.
[19] Zhuo Wang, Tingting Hou, Dawei Song, Zhun Li, and Tianqi Kong. 2016. Detecting Review Spammer Groups via Bipartite Graph Projection. *Comput. J.* 59, 6 (2016), 861–874.
[20] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. 2012. Review spam detection via temporal pattern discovery. In *SIGKDD.* 823–831.