Journal of Global Information Management
Volume 29 • Issue 2 • Bi-Monthly 2021

Spammer Group Detection Using Machine Learning Technology for Observation of New Spammer Behavioral Features

Li-Chen Cheng, National Taipei University of Technology, Taiwan
Hsiao-Wei Hu, Soochow University, Taiwan
Chia-Chi Wu, Tamkang University, Taiwan

ABSTRACT

Recently, the rapid growth in the number of customer reviews on e-commence platforms and in the amount of user-generated content has begun to have a profound impact on customer purchasing decisions. To counter the negative impact of social media marketing, some firms have begun hiring people to generate fake reviews which either promote their own products or damage their competitor's reputation. This study proposes a framework, which takes advantage of both supervised and unsupervised learning techniques, for the observation of behaviors among spammers. Then, based on the behavior of participants on web forums, the authors build up a post-reply network. The main focus is on the behavior-related features of the reviews, their propagation, and their popularity. The primary objective of this study is to build an effective online spammer detection model and the method detailed in this work can be used to improve the performance of spammer detection models. An experiment is carried out with a real dataset, the results of which indicate that these new features are important for identifying spammers. Finally, random walk clustering is applied to investigate the post-reply network. Some interesting and important features are observed in the interactions between a group of spammers which could be subjected to further research.

KEYWORDS

EC, Fake Review, Machine Learning, Spammer, Word of Mouth

1. INTRODUCTION

The recent emergence of social media as a means of social communication has had profound effects on general communication structures and the interactions between businesses, communities and individuals. Social media gives organizations the opportunity to target a wider audience and establish connections within a short span of time using limited resources (Chen, De, & Hu, 2015). These changes have also meant that organizations now have to consider new ways of marketing their products and services (Trapp, 2016).

The development of social media has led to rapid growth in the amount of user-generated content which has not had a big impact on purchasing behavior, but affects the public perception of products/services, and thus the business development landscape. Naturally this had drawn the attention of researchers and marketers. Online consumer reviews have proven particularly influential in shaping the purchase decisions of potential customers. Positive reviews can ensure the success of a product while negative reviews can doom it to failure (Zhang, Zhou, Kehoe, & Kilic, 2016).

DOI: 10.4018/JGIM.2021030104

Most of the research on social media marketing has focused on the opportunities and advantages of these developments. Relatively little work has been done examining the negative ramifications (Shirish 2018). The negative impact of social media marketing is illustrated by a report appearing on BBC about the fake web reviews of Samsung products. The article made clear that Samsung was paying people to write negative reviews about HTC products on several web forums in Taiwan. This action was judged to violate fair trade practices and thus resulted in Samsung having to pay a 350 million USD to Taiwan's Fair Trade Commission (FTC). The case only came to light in 2013, when a hacker released confidential marketing documents which they had obtained from Samsung Taiwan (Elmer-DeWitt, 2013).

It has been shown that this was not an isolated incident, that other firms, in efforts to cultivate a positive company image and improve sales, have taken steps to manufacture positive (fake) reviews of their products/services (Wang, Day, and Lin, 2016). In short, fake reviews are a growing problem that seriously undermines consumer trust in the review system.

Although these fake reviews are skillfully crafted to avoid detection, advances in machine learning technology are opening the door to automated detection (Jindal & Liu, 2008). Zhang, Zhou, Kehoe, and Kilic (2016) examined the predictive features that an automated system could use to detect which reviews are fake and which are not. They categorized these predictive features as either verbal or nonverbal. They defined verbal features as those extracted from the text of the review. Verbal features dominate the set of predictive features used in existing fake review detection models. In contrast, the nonverbal features are defined as the review posting behaviors and social interactions of reviewers with other reviewers on social media, especially on online review platforms. The focus in the detection of fake content has been on verbal focus features. Ott, Choi, Cardie, and Hancock (2011) built a prediction model using content-related features. Xie, Wang, Lin, and Yu (2012) focused upon identifying fake quantitative social information such as fake product rankings and ratings. Mukherjee, Liu, and Glance (2012) carried out experiments to identify fake review data which had been posted on Yelp. Past studies have proven that it is very hard to detect spammers (in this case the people who write fake reviews) simply by reviewing the content features because of the subtle way such reviews (opinions) are produced. This has motivated many researchers to strive to develop machine-learning methods which can be applied to examine the nonverbal aspects of posted reviews based on the reviewers' behavior-related characteristics (Lim, Nguyen, Jindal, Liu, & Lauw, 2010; Li, Huang, Yang, & Zhu, 2011). For example, it has been found that fake reviews can be distinguished by their temporal patterns (Xie et al., 2012).

According to Mukherjee, Venkataraman et al. (2013b), behavioral features are far more effective than linguistic n-grams in terms of detection performance. When examining nonverbal features, it is important to observe patterns in the way spammers work. The earnings of spammers are usually based on the number of reviews they post. Thus, many of the fake reviews they produce (in particular, replies) do not necessarily even express an opinion about the product under discussion. It is often the case that fake review posts are meant only to keep the discussion alive or attract attention to the threads pertaining to the objectives of their campaign.

Most spammers will write many product reviews and use multiple accounts to disseminate them (Chen & Chen, 2015). However, spam-based threads tend to be more active than non-spam threads because they are written to draw attention in the form of replies (both spam and non-spam). The interactions between spammers and other posters also provide data that can be used for the detection of spammers. Unfortunately, it is difficult to acquire ground truth data pertaining to online reviews.

To remedy the difficulty in acquiring ground truth data, this study uses a real dataset acquired from posts on Mobile01.com for the period from 2011 to 2012, during which the Samsung spamming attack on HTC occurred. It should be noted that this event has not been studied so far (Cheng, Tseng, & Chung, 2017). We focus on the following three points:

1. Reviewer behavior-related feature extraction: we propose three new nonverbal features that can be used to quantify different aspects of the spammer's behavior. These include threads (including the length of a thread, the number of posts and the number of replies contributed by each account), propagation (the average number of replies in a thread from a specific account which relates to the information propagation ability of that account), and popularity (the average length of the threads in which a user participates and the extent to which a user participates in popular threads).
2. We seek to improve the performance of fake-review spammer detection through the use of supervised learning techniques, such as support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), Logistic Regression (LR) and random forest (RF) techniques. These methods are chosen because they are the frequently used classification algorithms.
3. We also employ an unsupervised learning technique (i.e., the random walk clustering algorithm) to observe the interaction among groups of spammers. It is important to provide evidence that spammers work together to achieve their goals.

To the best of our knowledge, this is the first study to address the detection of fake-review spammers in this way. The rest of the paper is structured as follows. The next section provides an overview of the relevant literature and lays out the techniques in the proposed framework. We then discuss the development of our framework and provide a detailed description of each module. For experimental evaluation, reviews were collected from Mobile01 from which to provide a dataset. We describe the preparation of the dataset, experimental setup, and the evaluation results. In the final section, we conclude with a discussion of the implications of our findings.

2. RELATED WORK

2.1 Fake Reviews

Jindal and Liu (2008) proposed a supervised method for the detection of fake reviews. The goal of automated classification methods is to apply machine learning techniques to automatically learn hidden knowledge or recognize patterns based on training data or previous experience. There are several approaches for the detection of fake reviews. Past research has been aimed at distinguishing between spam posts and legitimate posts as well as spammer detection. The importance of reviewing the content features of consumer evaluations has been emphasized by Huang et al. (2013). Others have focused on the patterns shown in numerical rating product reviews. However, many have argued that there are subtle linguistic cues in review posts which can be used to distinguish them from each other. These kinds of features include the length of the review, the n-grams, the subjectivity of the review content, the number of nouns, verbs, and adjectives, etc. (Hu & Liu, 2004). Wang et al. (2011) described the relationshops between reveiwers, their reviews and the online merchants being reveiewed using a heterogeneous review graph. They also developed an effective computation method for quantification of the trustworthiness and honesty of the reviewers, and the reliability of the online vendor.

Supervised methods have been used most often for the detection of fake reviews, by identifying them as a special kind of post (Jindal & Liu, 2008; Ott et al., 2011; Dou, 2019). For example, Jindal and Liu (2008) manually labeled 470 spam reviews to build a set of tranining data. Ohters have focused on analyzing the contents of the online reviews then used the results to build a supervised model for identifying fake reviews. However, the manually labelling of a spam review dataset is a massive undertaking. Thus, the functioning of the proposed supervised models has often been verified based on data sets comprsied of pseudo-fake reviews which have either been manually annotated or generated by Amazon Mechanical Turk (AMT) (Jindal & Liu, 2008; Ott et al., 2011; Mukherjee et al., 2013a).

Some previous studies focusing on identifying spammers include those by Lim et al. (2010), and Wang et al. (2011). Mukherjee et al. (2013b) analyzed the reviews filtered by Yelp but found it difficult to differentiate reviews identified as fake from truthful reviews (Mukherjee et al., 2013b). They

proposed a supervised linguistic and behavioral feature-based model to test and evaluate the ground truth dataset. However, their detection results were not good enough but did show that behavioral feature based methods performed better than linguistic feature based methods (Mukherjee et al., 2013b). Recently, Zhang et al. (2016) proposed a model for the detection of fake reviews which is based on new features including both verbal and nonverbal features. They evaluated the model using the same Yelp dataset. The results reveal nonverbal features to be more important for fake review detection than verbal features.

Mukherjee et al. (2012) defined the Group Spam Ranking (GSRanking) function which was based on the frequent item-set mining method. The method is aimed at finding a set of fake reviewer candidate groups from which to build a labeled dataset. In an earlier study, Yardi, Romero, and Schoenebeck (2009) used the degree of centrality in a social network to detect twitter spam. Dou (2019) gave us some reviews of tackles cold start problem on spam detection by graph model and deep model.

Chen and Chen (2015) also conducted a study of real Chinese spam dataset. Their method could be potentially helpful in detecting spam opinions in various threads. They found that the first spam posts placed more focus on certain topics and generally used more words and pictures in an attempt to impress the reader. Wang et al. (2016) proposed a supervised model which was based on using the same dataset indices for some features for the analysis of a social network. They found that the relationship between the authors of product review posts and their replies could be used for spammer group detection.

2.2 Ground Truth Acquisition

The biggest challenge in the study of methods for the detection of spam reviews is the difficulty in building the ground truth dataset. Ott et al. (2011) was the first to crowdsource anonymous online workers, the so-called Turkers, using Amazon Mechanical Turk (AMT). They hired these Turkers to write hotel reviews which portrayed some hotels in a positive light. Although the proposed method could detect fake reviews with an accuracy of 90% for this dataset, the performance of proposed method was not so good when applied to another real dataset. It was concluded that the reason was that Turkers were not familiar with the knowledge of the domain and the payoff was not large enough for them to put their heart into writing fake reviews.

In another study, the authors built a small labeled dataset on data acquired from Amazon. They then hired people to write reviews and used the frequency-pattern mining technique to identify spammers from the candidate group (Mukherjee et al., 2012). However, it is ineffective to use datasets based on hiring people to produce fake reviews. Real spammers are motivated to always do their best to write convincing reviews in order to satisfy those who pay them to write those reviews, whether to promote the specified product or damage a competitor.

Popular frequently used websites like Yelp.com commonly use algorithms to filter out suspicious reviews (Mukherjee et al., 2013b). The online review platform collects and analyzes a lot of information when designing their filter algorithms, including the reviewers' posting behavior, their interactions with other reviewers and the average number of reviews posted by an individual reviewer per day (Mukherjee et al., 2013; Mukherjee et al., 2013b).

The real case of fake web reviews of Samsung products reported by BBC in April 2013, occurred in Taiwan. It was uncovered when a hacker released confidential documents, including leaked spreadsheets which contained detailed histories and information including the poster's username, the time of posting, and some other details for the period from 2011-2012. In their analysis of this dataset Chen and Chen (2015) suggested that spammers actively posted in various threads doing their best to catch the reader's eye. Wang et al. (2016) analyzed this dataset using social network analysis techniques. In a related study we obtained information by crawling the Taiwan website mobile 01 for comments posted on the Samsung board (Cheng et al., 2017). According to the account information contained in the leaked spreadsheets, we defined spammers as anyone who had ever submitted a spam post. Five classification algorithms (support vector machine (SVM), Naïve Bayes (NB), decision tree

Table 1. List of features considered in developing our prediction models

Features	Definition
np(*x*)	The number of first posts submitted by user *x*
nr(*x*)	The number of replies submitted by user *x*
na(*x*)	The number of articles submitted by user *x*
nt(*x*)	The number of topic threads participated in by user *x*
npr(*x*)	The number of topic threads in which user *x* plays the roles of poster and replier simultaneously
nat(*x*)	The total length of topic threads in which *x* is a participant
anat(*x*)	The average length of topic threads in which *x* is a participant
nre(*x*)	The total length of topic threads in which *x* is a poster
anre(*x*)	The average length of topic threads in which *x* is a poster
degree(*x*)	The degree of centrality of user x in the post-reply network
betweenness(*x*)	The betweenness centrality of user x in the post-reply network
closeness(*x*)	The closeness centrality of user x in the post-reply network
spammer (*x*)	The class of user *x* ("spammer" or "normal reviewer")

(DT), Logistic Regression (LR) and random forest (RF)) are selected for building the fake reviewer detection models.

3. RESEARCH DESIGN

In this study, we propose a model which is able to detect active spammers who post on web forums. After reviewing the literature, we build a nonverbal feature-based model. In addition, the centrality measures used in network analysis are also applied to extract interactions among users. The SNA features defined include *degree*(*x*), *closeness*(*x*), and *betweenness*(x). The features considered in developing our model are listed in Table 1.

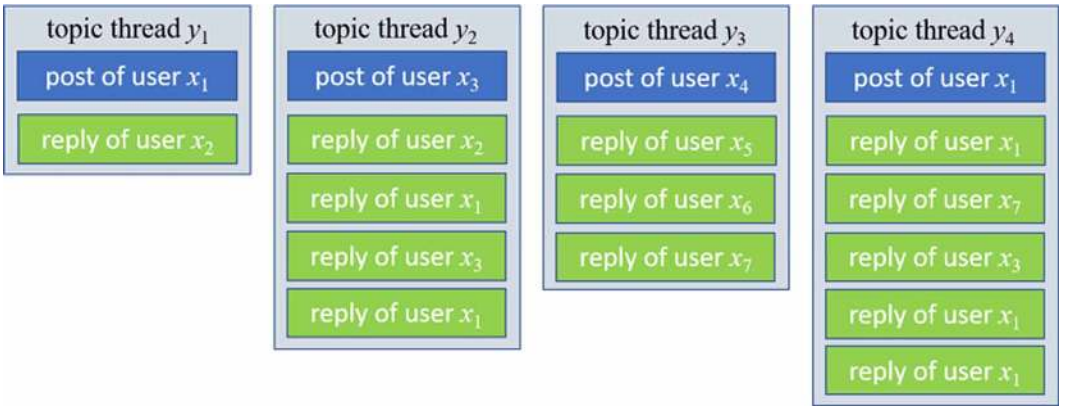
In a web forum, there are many topic threads, as shown in Fig. 1. In this figure, it can be seen that there are four topic threads, each of which include several separate statements.

Each topic thread is initiated with a first post written by a user. Then, this user, or other users reply to this first post, for discussion within the same topic thread. It is interesting that in most online web forums, the first posts and the replies within threads play different roles. The first posts tend to be relatively richer in content which is necessary to attract readers' attention. The first posts function to initiate a discussion on a specific topic. In contrast, replies tend to be quite concise and to support the first post. Sometimes they do not really carry any opinion, as manifested in some of the examples of spam replies listed in this section. There are some platforms, such as Amazon and TripAdvisor, on which users can only recommend or vote, but cannot reply a post. The proposed model may not work on these platforms, which is a limitation of this study.

In this study, $np(x)$ and $nr(x)$ denote the number of first posts and the number of replies submitted by a user x, respectively. In addition, the total number of posts submitted by user x is denoted by $na(x)$, while $na(x) = np(x) + nr(x)$. For example, in Figure 1, $np(x_1) = 2$, $nr(x_1) = 5$, and $na(x_1) = 7$.

The user who writes the first post is considered the poster of the corresponding topic thread, while others are repliers. We use $TP(x)$ to denote the set of topic threads in which user x is a poster, and $TR(x)$ is the set of topic threads in which user x is a replier. In Fig. 1, $TP(x) = \{$topic thread y_1, topic thread $y_4\}$ and $TR(x) = \{$topic thread y_2, topic thread $y_4\}$.

Figure 1. Topic threads in a web forum



A user may participate in a topic thread as a poster or a replier. For user x, $TT(x) = TP(x) \cup TR(x)$ is the set of topic threads participated in by x, and $nt(x) = \left|TT\left(x\right)\right|$.

A user may write a first post, and then submit replies within the same topic thread. For a user x, $TPR(x) = TP(x) \cap TR(x)$ is the set of topic threads in which user x plays the roles of poster and replier simultaneously, and $npr(x) = \left|TPR\left(x\right)\right|$. For example, in Fig. 1, $nt(x_1) = 3$ and $npr(x_1) = 1$.

We define $l(y)$, the length of topic thread y, as the number of replies in y. In Fig 1, the lengths of the four topic threads are 1, 4, 3, and 5, respectively. We use $nat(x)$ and $nre(x)$ to denote the total length of threads in which x is a participant and a poster, respectively, while $nat\left(x\right) = \sum_{y \in TT\left(x\right)} l\left(y\right)$ and $nre\left(x\right) = \sum_{y \in sTP\left(x\right)} l\left(y\right)$.

Thus, $anre(x)$ can be measured as the propagation ability of user x 's post and $anat(x)$ can be measured as the popularity of user x 's post.

Definition: popularity

$$anat\left(x\right) = \frac{nat\left(x\right)}{\mathrm{nt}\left(\mathrm{x}\right)} \tag{1}$$

Definition: propagation

$$anre\left(x\right) = \frac{nre\left(x\right)}{\mathrm{np}\left(\mathrm{x}\right)} \tag{2}$$

Example:

In Fig. 1, $nat(x_1) = 10$ and $nre(x_1) = 6$. On the other hand, $anat(x)$ and $anre(x)$ are the average length of topic threads in which x is a participant and a poster, respectively, while $anat(x_1) = 10/3$ and $anre(x_1) = 6/2$.

We also build a post-reply network to extract interactions among users in a web forum. In this post-reply network, a vertex is a user, and there is a directed edge from user x_1 to user x_2, if x_1 submits a reply in a topic thread initiated by x_2. In Table 1, *degree*(x), *betweenness*(x), and *closeness*(x) indicate the values of the degree centrality, betweenness centrality, and closeness centrality of user x, respectively.

4. EXPERIMENTS AND EVALUATION

4.1 Data Set

Reviews crawled from the website Mobile01.com were used to build a new ground truth dataset which provides a rich posting history (Cheng et al., 2017). The crawled data contained both fake reviews and non-fake reviews. In this study we used Chen and Chen's (2015) spammer list. The number of spammers was 293 and the number of nonspammers was 21358. We analyzed the posts-replies in threads contained our ground truth dataset. In the forum, the reviewers' behaviors consisted of first posts and the subsequent replies. Any poster account listed as a spammer by Chen and Chen (2015) was considered to be a spammer. The summarized results appear in Table 2.

Table 2.

	This study		**Chen & Chen (2015)**	
	Spammer	**Non-spammer**	**Spammer**	**Non-spammer**
Post	1140	17630	3116	632234
Reply	10035	235987		
User accounts	293	21358	300	58231

4.2 Evaluation Metrics

After construction of the training and testing datasets, five distinct classification methods were used, namely, support vector machine (SVM), Naïve Bayes (NB), decision tree (DT), Logistic Regression (LR) and random forest (RF). We used fivefold cross-validation for evaluation. The average performance of the five experimental results was reported for each of the classification methods. The performance of the constructed spammer detection models was evaluated using the five metrics commonly used for evaluating classification models, including precision (P), recall (R), F-measure (F) and ROC (Receiver Operating Characteristic) and Matthews Correlation Coefficient (MCC).

Precision is determined by the ratio of spammers correctly classified as spammers that are indeed spammers. Recall is the percentage of total spammers in the data set that are correctly identified. The relevant formulas are expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \tag{4}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{ Recall}}{\text{Precision} + \text{ Recall}}, \tag{5}$$

where *TP* represents the number of spammers successfully predicted by the classifier; *FP* represents the number of nonspammers incorrectly identified by the classifier as spammers; and *FN* represents the number of f spammers that are not detected. The values of these four measures range from 0 to 100 percent.

The ROC Curve is used to examine the performance of a binary classifier, by creating a graph of *TP* vs. *FP* for every classification threshold. An ROC area with a value greater than 0.5 indicates better classifier performance.

The MCC ranges from -1 to 1, where -1 indicates a completely wrong binary classifier while 1 indicates a completely correct binary classifier.

$$\text{MCC} = 2 \times \frac{\text{TP} \times \text{ TN} - \text{FP} \times \text{FN}}{\sqrt[2]{\left(TP + FP\right) \times \left(FN + TN\right) \times \left(FP + TN\right) \times \left(TP + FN\right)}} \tag{6}$$

4.3 Results and Analyses

The primary objective of this study is to build an effective online spammer detection model. Two different experiments are designed to test the performance of the model. First, the performance is evaluated with detection models designed using the data containing all the features defined in Table 1. Next, the performance is evaluated but the SNA features are excluded. By comparing the detection performance obtained using datasets with, and without SNA variables, we are able to determine whether or not incorporating SNA behavioral features can improve spammer detection performance. Finally, the random walk clustering algorithm is used to observe the post-reply spammer relationships. The results of each of the experiments are reported in detail below.

4.3.1 Classification Models

The ratio of normal reviewers to spammers in Mobile01 is approximately 100:1. This provides a highly imbalanced set of training data, therefore steps must be taken to avoid producing poor models with such an unbalanced data set. We use the techniques commonly used in machine learning to build a good model from imbalanced data, oversampling and underdamping.

In the underdamping technique similar numbers of normal users and spammers are chose to avoid the common data imbalance problem. Some of the users are randomly removed from the normal user's group to form a balanced class distribution data set. The classification results obtained using balanced (50:50) data, from DT classifiers both with and without SNA features are reported in Table 2 and illustrated in Figs. 2 and 3. F

Observe that ***anat*** is a root in both Figs. 2 and 3, which suggests that popularity is an important feature. Propagation is also another important feature, with ***npr*** and ***anre*** found at level 2 in the DT tree. The results illustrate that spam-based threads tend to be more active than non-spam threads.

The inclusion of the SNA features does not seem to be important. Table 2 shows that DT performed better using data without SNA. Both Figs. 2 and 3 show that SNA features rarely appeared in the decision trees.

To address the extremely skewed data distribution, we resampled the data set using SMOTE (Synthetic Minority Over-Sampling Technique). SMOTE is an integrated oversampling and underdamping method, well-known for its simplicity and effectiveness (Bhagat & Patil, 2015). The

Figure 2. The results of DT classification with SNA features

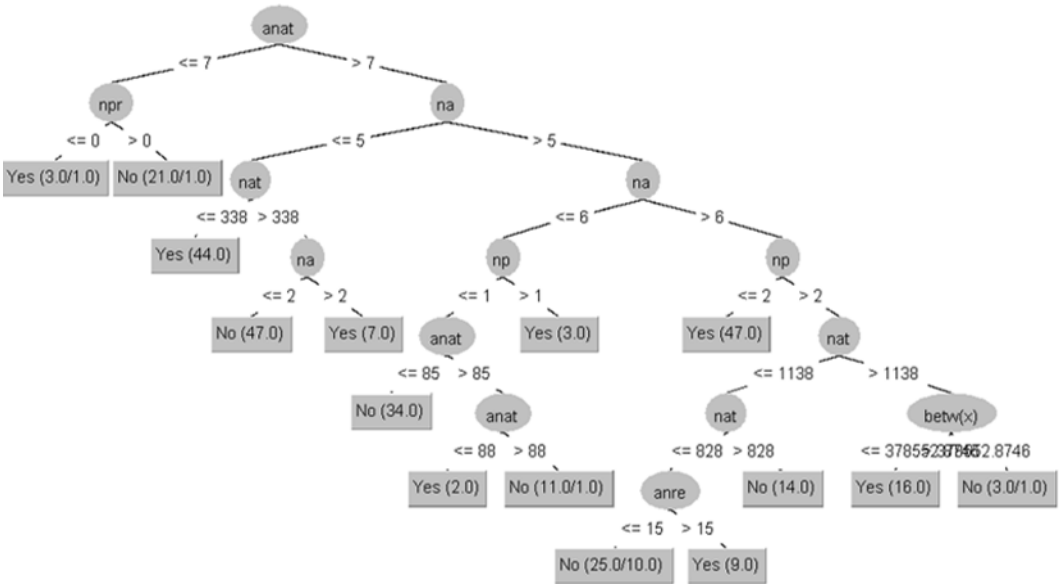


Figure 3. The results of DT classification without SNA features

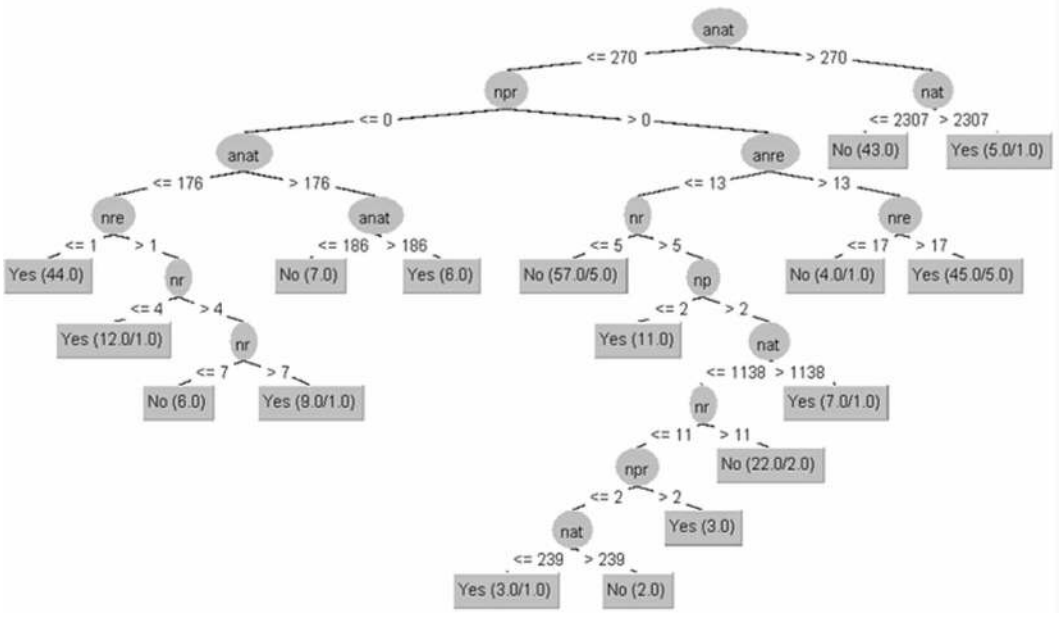


Table 2. The metrics of the classification results

	Precision	Recall	F measure	MCC	Roc area
DT with SNA	0.798	0.797	0.797	0.595	0.820
DT without SNA	0.805	0.804	0.804	0.609	0.834

Table 3.

Classification algorithms	Precision	Recall	F measure	MCC	Roc area
DT without SNA	0.965	0.964	0.964	0.929	0.985
DT with SNA	0.977	0.977	0.977	0.954	0.989
NB without SNA	0.711	0.610	0.558	0.305	0.801
NB with SNA	0.713	0.622	0.577	0.323	0.802
LR without SNA	0.783	0.780	0.780	0.564	0.863
LR with SNA	0.795	0.791	0.790	0.586	0.867
SVM without SNA	0.862	0.852	0.851	0.714	0.852
SVM with SNA	0.811	0.748	0.735	0.556	0.748
RF without SNA	0.979	0.979	0.979	0.959	0.993
RF with SNA	0.988	0.988	0.988	0.975	0.998

SMOTE algorithm not only replicates the minority class but also generates new instances in the minority class data.

An examination of the results show that the performance of the RF and DT methods is significantly better than that of the other classification algorithms. It can also be seen that the RF model produces the best performance in terms of precision, recall, F-measure, MCC and ROC value.

Comparison of the results obtained with the RF model with SNA features extracted from reviewers (P: 98.8 percent, R: 98.8 percent, F: 98.8), with the results obtained using the other classifiers with SNA features shows that the performance of the RF model with SNA features is significantly better across all classification methods. This result suggests that the data with SNA features with the RF model are more effective for fake reviewer detection. Table 3 shows that the SNA features have no significant impact on model performance for fake reviewer detection.

4.3.2 Clustering Approach For Group Spammers

Sometimes manufacturers or online retailers will hire spammers to promote their own products or damage the reputation of a competitors' products (Liu, 2012). Most spammers responsible for writing many product reviews will use multiple accounts to disguise their activity while satisfying the requirements of promoting the target (Chen & Chen, 2015). In this study aimed at the detection of spammer groups we use as an example data crawled from the Taiwan website Mobile01. However, given that there is no labeled dataset for Mobile01 it is impossible to use a supervised learning framework for the detection of groups of spammers.

Thus, after extracting data about reviews and threads from the Mobile01forum we built a post-reply network illustrating interactions among users. The post-reply network represents the reviewers' social network. The network consists of a vertex poster or a replier. There is an undirected edge from user A to user B, if both users are found to participate in the same thread.

The random walk algorithm was applied to cluster users in the post-reply network according to their behavioral associations (Pons & Latapy, 2005). The results indicate associations between normal users and spammers. The clustering results listed in Table 4 provide evidence of the identifiable interaction of spammers in the post-reply behavior of the community.

As can be seen in Table 4, most spammers may be found in the same groups, for example, groups 8, 10 and 34. Some spammers only communicate with themselves because there is only one member in the group, for example, groups 178, 252, 264, 442 and 513.

Table 4. Clustering results obtained with the random walk algorithm

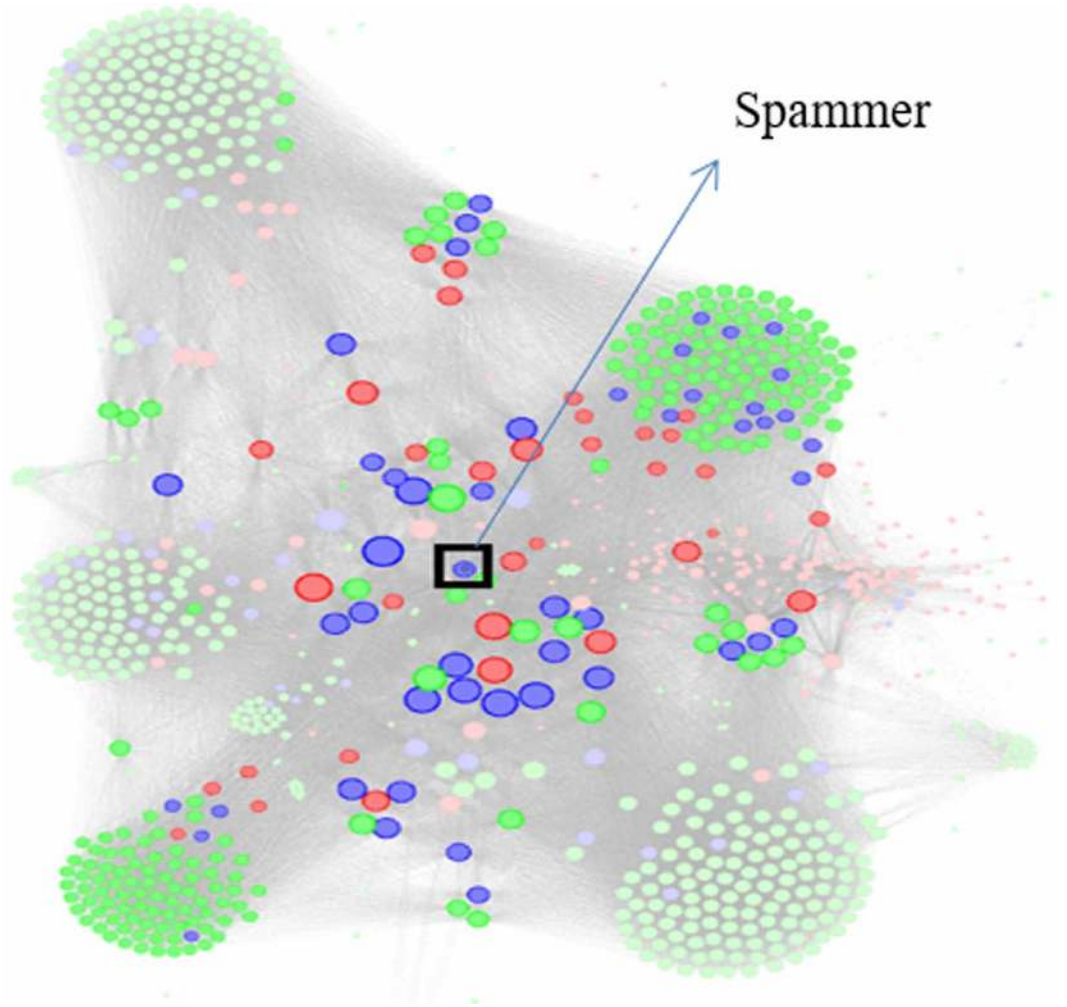
Groups	Number of fake reviewers	Number of normal reviewers	Faking rate
8	86	676	0.1129
9	4	2700	0.0015
10	108	9717	0.0110
16	2	1588	0.0013
23	3	2613	0.0011
32	1	22	0.0435
34	71	3692	0.0189
58	1	32	0.0303
60	1	27	0.0357
69	1	58	0.0169
81	6	189	0.0308
102	1	3	0.2500
178	1	35	0.0278
252	1	1	0.5000
264	1	2	0.3333
442	1	2	0.3333
513	1	1	0.5000
582	1	0	1.0000
658	1	0	1.0000
1285	1	0	1.0000

Figure 4 shows the users' interaction network for group 8. In this group, the number of spammers is 86 and the number of normal reviewers is 676. The green nodes represent normal reviewers and the blue node represent spammers in the same group. The red nodes represent spammers in other groups. An examination of Fig. 4 shows that most blue nodes are closer together and bigger than the green nodes. The spammers in the same group are always in communication with each other. If two users are always found on the same threads, these two nodes will be closer together than the other nodes. The spammers seemed to collaborate, to enhance the popularity of each other's posts. Nodes that are bigger than other nodes have an indegree higher than any other nodes.

Figure 5 shows the users' interaction network for group 34. The number of spammers in this group is 71 and the number of normal reviewers is 3692. However, observe that the blue nodes and the red nodes are closer together than the other nodes. It is interesting that the spammers are communicating with other groups of spammers.

Figure 6 shows the users' interaction network for group 10. The number of spammers is 108 and the number of normal reviewers is 9717 in this group. Fig. 6 and Fig. 5 are very similar. The blue nodes and red nodes are closer than the others. The interactions of spammers are similar to the activities of normal reviewers.

Figure 4. Users' interaction network for group 8

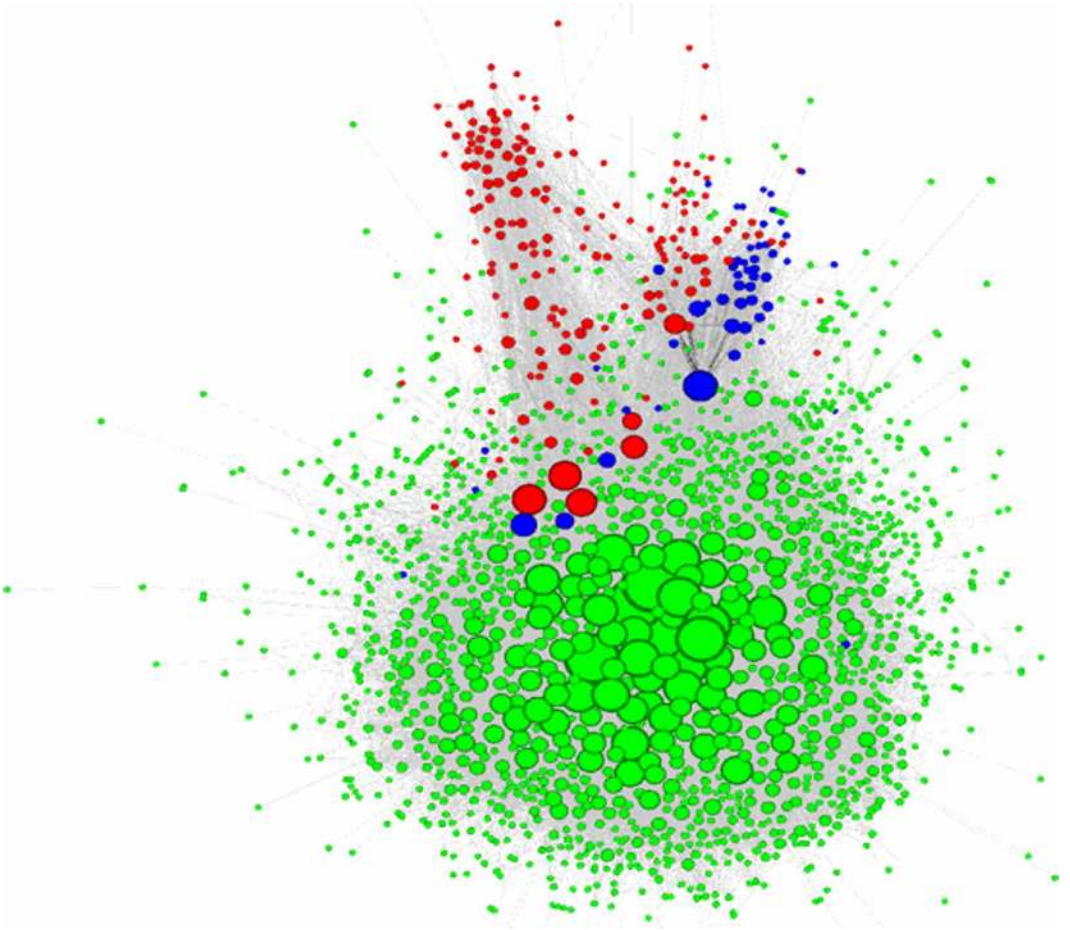


5. CONCLUSION AND FUTURE WORKS

Along with the growth of e-commence, user generated content has proven to be an efficient way for reviewers to influence customers' purchasing decision. Some firms have even gone so far as to hire people to spread fake reviews for the purpose of promoting their products. These fake reviews are designed to mislead customers and influence their purchasing decisions. The present study proposes algorithms to detect fake reviews, focusing on observing the activities of spammers. To the best of our knowledge, our detection model is the first to focus on identifying spammer behavior by examining the quantitative effects using a ground truth dataset. Although there have been several methods proposed for the identification of spammers, grouping of the reviewers' behavior has rarely been investigated.

Therefore, our study makes several interesting contributions both to theory and practical applications. First, reviewer behavior-related features, defined as review posting behaviors and social interactions with other reviewers, are proven to be an important index for identifying spammers. We defined several features and used five classifiers to identify spammers. Our analysis clearly shows that the new features of popularity and propagation, are quite useful for identifying spammers. Every

Figure 5. Users' interaction network for group 34



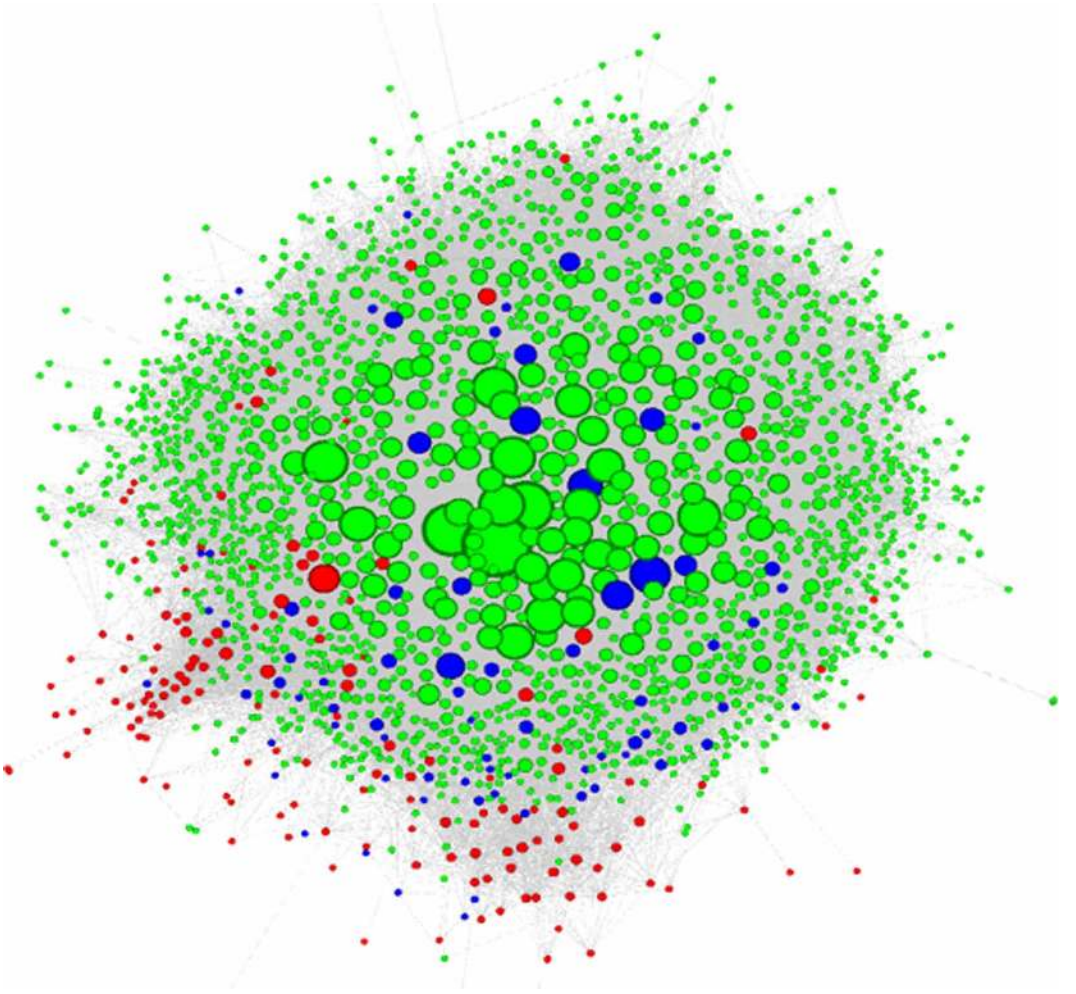
supervised method with new features of in the experiments outperforms methods without new features, except for SVM. The propagation feature measures the activeness of a thread by counting the number of posts in that thread. An explanation for this might be that spammer generated threads are meant to attract the readers' attention. The obvious stratagem is that the first post should be written so as to attract replies regardless of whether they are spam or non-spam replies. It is therefore reasonable to expect that spammers would work together to make the threads longer.

Our findings do suggest that spammers work together with other spammers to increase the length of threads to attract other users' attention. The use of random walk clustering for examination of the post-reply network present a valuable new tool for future research to detect groups of spammers.

There are some practical implications that should be considered by platform providers. Our study provides insights into the activities of the spammers themselves and interaction among groups of spammers. The designers of social media platforms need to develop filtering systems to reduce the posting of fake reviews based upon the activities of the reviewers. Our findings suggest that the interactions among and between reviewers should be considered as features for the detection of fake reviews.

Although our study provides important insights for both theory and practice, we acknowledge certain limitations that have to be considered when interpreting the results. The experimental dataset is limited, containing only data crawled from one platform, moreover the labeled reviewers listed

Figure 6. Users' interaction network for group 10



in leaked data are limited. In the future, researchers could apply algorithms to enlarge spammers list. Applying other methods like reinforcement learning, deep learning and graph model with the proposed method "reviewer behavior-related feature extraction" is also an interesting future research direction. Some reviewer features included in our models may not be available in other online review platforms. As a result, the findings of this study may not be directly applicable to fake detection on other platforms. Thus, we expect that the importance of the interaction of reviewers to fake review detection discovered in this study may not necessarily be generalizable to other online review platforms, but this could be worthy of future investigation.

ACKNOWLEDGMENT

This study was supported in part by the Ministry of Science and Technology of Taiwan under grant numbers MOST 105-2410-H-031-035-MY3, 108-2410-H-027-020 and 107-2218-E-007-045

REFERENCES

Bhagat, R. C., & Patil, S. S. (2015). *Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest.* Paper presented at the Advance Computing Conference (IACC), 2015 IEEE International. doi:10.1109/IADCC.2015.7154739

Chen, H., De, P., & Hu, Y. J. (2015). IT-enabled broadcasting in social media: An empirical study of artists' activities and music sales. *Information Systems Research*, *26*(3), 513–531. doi:10.1287/isre.2015.0582

Chen, Y. R., & Chen, H. H. (2015). Opinion spam detection in web forum: a real case study. *Proceedings of the 24th International Conference on World Wide Web*. doi:10.1145/2736277.2741085

Cheng, L.-C., Tseng, J. C., & Chung, T.-Y. (2017). Case Study of Fake Web Reviews. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. doi:10.1145/3110025.3110119

Dou, Y. (2019). *A Review of Recent Advance in Online Spam Detection*. Retrieved from https://www.researchgate.net/profile/Yingtong_Dou

Elmer-DeWitt, P. (2013). Samsung Fined $340,000 for Astroturfing in Taiwan. *Fortune*. Retrieved from https://fortune.com/2013/10/24/ samsung-fined-340000-for-astroturfing-in-taiwan/

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Huang, L., Tan, C.-H., Ke, W., & Wei, K.-K. (2013). Comprehension and assessment of product reviews: A review-product congruity proposition. *Journal of Management Information Systems*, *30*(3), 311–343. doi:10.2753/MIS0742-1222300311

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *Proceedings of the 2008 International Conference on Web Search and Data Mining*.

Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. *Proceedings of the 19th ACM international conference on Information and knowledge management*. doi:10.1145/1871437.1871557

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1-167.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Spotting opinion spammers using behavioral footprints. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. doi:10.1145/2487575.2487580

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*. doi:10.1145/2187836.2187863

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013a). *Fake review detection: Classification and analysis of real and pseudo reviews*. Academic Press.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013b). *What yelp fake review filter might be doing?* Paper presented at the ICWSM.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1*.

Pons, P., & Latapy, M. (2005). *Computing communities in large networks using random walks.* Paper presented at the International symposium on computer and information sciences. doi:10.1007/11569596_31

Shirish, A. (2018). *Shadow Side of Social Media Marketing: A User's Perspective. In Social Media Marketing*. Springer.

Trapp, R. (2016). *Digital Transformation Drives Business for Social Media Managers*. Retrieved from https://www.forbes.com/sites/ rogertrapp/2016/10/12/digital-transformation-drives-business-for-social- media-managers/#3102bc9a3547

Wang, C.-C., Day, M.-Y., & Lin, Y.-R. (2016). *A Real Case Analytics on Social Network of Opinion Spammers.* Paper presented at the Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on. doi:10.1109/IRI.2016.89

Wang, G., Xie, S., Liu, B., & Philip, S. Y. (2011). *Review graph based online store review spammer detection.* Paper presented at the Data mining (ICDM), 2011 IEEE 11th international conference on. doi:10.1109/ICDM.2011.124

Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012). Review spam detection via temporal pattern discovery. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. doi:10.1145/2339530.2339662

Yardi, S., Romero, D., & Schoenebeck, G. (2009). Detecting spam in a twitter network. *First Monday*, *15*(1).

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, *33*(2), 456–481. doi:10.1080/07421222.2016.1205907

Li-Chen Cheng is currently an Associate Professor of Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan. She received her Ph.D. degree in information management from National Central University. Dr. Cheng serves as an editorial board member and a reviewer for more than 10 academic journals. Her research interests include deep learning, opinion mining, financial technique, AI in internet marketing, business intelligence and decision-making models. She has published papers in in well-recognized SSCI and SCI journals including Decision Sciences, Decision Support Systems, Electronic Commerce Research and Applications, Journal of Information Science, European Journal of Operational Research, Applied Soft Computing and many others She also served as an associate editor of Electronic Commerce Research and Applications.

Hsiao-Wei Hu is currently an Associate Professor at the School of Big Data Management, Soochow University. She received her Ph.D. degree in Information Management from National Central University. Her research interests include big data mining, social network mining, information retrieval, decision support systems, social network analysis, marketing technology, digital transformation and EC technologies. Currently she is focusing on big data mining and social network mining. Her research has appeared in Decision Support Systems, IEEE TKDE, IEEE SMCB and Expert Systems with Applications.

Chia-Chi Wu is an Assistant Professor at Department of Management Sciences, TamKang University. He graduated with a Ph.D. degree in Information Management from National Central University of Taiwan and received a M.S. degree in Management Information Systems from National Chengchi University of Taiwan. Fields of research interests include data mining, machine learning, and social network analysis.