Astroturfing Detection and Analysis: A Literature Review

Yalun Wu*

***Abstract*—With the development of Internet, user comments produced unprecedented impact on information acquisition, goods purchase and other aspects. For example, user comments can quickly make a topic become the focus of discussion in social networks, it can promote the sales of goods in e-commerce and it can influence the ratings of books, movies or albums. Among these network applications and services, "astroturfing", a kind of online suspicious behavior, can generate abnormal, damaging, even illegal behaviors in cyberspace, which mislead the public perception and bring a bad effect on both Internet users and society. Thus, how to detect and combat the astroturfing behavior becomes highly urgent, which draws lots of interests from both information technology and sociology researchers. In this paper, we restudy it mainly from the information technology perspective, summarize the latest research results of astroturfing detection, analysis the astroturfing feature, classify the machine learning-based detection methods and evaluation criteria and introduce the main applications. Different from previous survey, we also discuss new future direction of astroturfing detection, such as cross-domain astroturfing detection and user privacy protection.**

***Index Terms*—Astroturing, online suspicious behavior, astroturfing detection, machine learning.**

I. Introduction

WITH the rapid development of Internet, more and more people are communicating with each other through network applications and services. Recently, the vigorous development of e-commerce, especially the emergence of the social network, such as Facebook, Twitter, Wechat and Weibo, produced unprecedented significant impact on the way that people work, social contact, information acquisition and goods purchase. Among these network applications and services, astroturfing, a kind of suspicious online behavior, is widely existing, which mostly appears in business and politics events. That is because the truth that an individual impression on a particular subject are often influenced by others'opinions [1]. Accordingly, the attitude or opinions of online users are easily influenced by other users. Several recent breaking news are exposed to show the distinct evidence for astroturfing. In October 2010, Mengniu, the dairy products company, hired astroturf organizations to harm the reputation of their competitor Yili, another dairy products company, which has been widespread concerned in society. It has also been reported that a group of astroturfing could launch well-coordinated attacks, and generate positive or negative opinions to attract public attention or trigger curiosity. Such a practice is referred to as "cyber-gossip", which can mislead the public, and put the competing business at serious risk. In October 2015, Amazon sued the 1114 "Internet Water Army", accusing them of providing inveracious reviews for goods and services on Amazon.com, which was in violation of the laws of the United States [2]. In April 2016, a technology social networking site in U.S. announced that the "Internet Water Army" on twitter become a secret weapon during the US presidential election [3]. We can see that, the astroturfing still makes a worldwide development in a fast speed.

Obviously, astroturfing can generate abnormal, damaging, even illegal behaviors in cyberspace, which may mislead the public perception and bring a bad effect on both Internet users and society. Unfortunately, it is very difficult for users to aware the astroturfing and distinguish truth from falsehood. Thus, how to detect and combat the astroturfing behavior becomes very urgent, which draws lots of interests from both IT and sociology researchers. In this work, we will study it mainly from the IT perspective rather than sociology perspective. In other words, to design algorithms to detect online astroturfing effectively and help users identify potential online astroturfers quickly, is our main focus.

Unfortunately, as a kind of suspicious online behaviors, astroturfing, has so much close relation and easily-confused differences with other known suspicious online behaviors: traditional spam [4], fake review [5], social spam [6], [7] and link farming [8]. Thus, we should firstly try to find the essential characteristic belonging to astroturfing. Then, how to design and realize semi-automatic or automatic computer algorithm will be naturally made out to suit different data size. To wrap up, from the IT perspective, we hope that the computer has the learning ability to effectively detect unknown astroturfing among big social data. This is our survey view and we will summarize the astroturfing detection and analysis research work from astroturfing feature, machine learning-based detection method, evaluation, to applications.

The structure of this article is organized as follows. Section 2 will discuss the astroturfing feature. Section 3 will present learning-based astroturfing detection methods. Section 4 describes effective detection evaluations. Section 5 shows the astroturfing detection applications. Finally, the future directions are envisaged in section 6.

II. Astroturfing Feature Characterization

A. Feature in Definition

The group or individuals who participate in "astroturfing" can be called as "astroturfer" or "internet water army". It is

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received October 1, 2022; revised October 5, 2022. (*Corresponding author: Yalun Wu*)

Yalun Wu is with the Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China (e-mail: wuyalun1@bjtu.edu.cn).

very difficult to give a quantified and accurate definition on astroturfing. Thus, only descriptive explanations on "astroturfing" are given and listed as following:

- The practice of masking the sponsors of a message or organization (e.g., political, advertising, religious or public relations) to make it appear as though it originates from and is supported by a grassroots participant[1] .
- The artificial grassroots campaigns created by public relations firms [9]. Organizations that engage in astroturfing activities usually hire public relations or lobbying firms to simulate grassroots campaigns [10]. In other words, astroturfing occurs when groups of people are hired to present certain beliefs or opinions, which these people do not really possess, through various communication channels. In most cases, the hired groups and individuals support arguments or claims for their employers favor while challenging critics and denying adverse claims [11].
- A form of costly state falsification [12]. Astroturfing creates falsified impressions among decision makers or the general public and achieves the goal of persuasion. Traditionally, the scope and influence of astroturfing are limited by the strength of financial support behind the effort since hiring public relation firms to generate and disseminate these false messages can be costly [13].

From these explanations on "astroturfing", we generalized the following feature keywords:

- *Money Business*: if the astroturfing exists, there must be an employer give the money indirectly to the lobbying firms or directly to the grassroots.
- *Content Effect*: any astroturfing should aim to achieve the astroturfing effect in short time, i.e., making the employer-requested topic to attract more public awareness, though posting massive content or increasing the content ranking. The content are consist of post and comments.

From IT perspective, the *money business* feature is very difficult to mine, that's because lots of evidence are missed if we only rely on open Web data. While the *content effect* feature can be found among public Web data. As a result, all the IT researches on astroturfing belong to the category of mining and utilizing the *content effect* feature.

B. Feature in Suspicious Behavior Category

In behavior category of suspicious online behaviors in social network, astroturfing has both close relation and differences with other known suspicious online behaviors: traditional spam [4], fake review [5], social spam [6], [7] and link farming [8]. We make a comparison for them(See Table I).

As Table I shows, five dimensions are selected for comparison.

Main application refers the main carrier that suspicious online behaviors occur. The astroturfing mainly appears in B2C and C2C e-commerce application (e.g. Amazon, Netflix, Taobao), and popular social contact application (Facebook, Twitter, Weibo, LinkedIn), while traditional spam only occurs in Email or SMS.

For participant type, interestingly, in order to perform requiring effect, the astroturfing executes both through AI program and human, which can be employed at online crowdsourcing platform or other online way. Similarly, the fake review, social spam and link farming also utilize AI program or human to perform.

Behavior time refers participating time in a task period. For an astroturfing task, the astroturfing behaviors usually last very short time within several hours or days to execute and generate enough influence. While link farming behavior usually need more time months to months.

Participant visual means whether the participant can be shown to common users or not. Astroturfing must expose their behavior and opinion for falsification.

Operation account scale refers real number of participating application account. Usually, at least hundreds of astroturfing accounts execute an astroturfing task.

Thus, the astroturfing main features from the suspicious online behavior category are that ones, which can cover those ones in fake review and social spam behaviors. This can be simply represented as follows

$$f_{Astroturfing} \supset f_{Fake\ Review} \supset f_{Social\ Spam} \qquad (1)$$

Thus, we can exploit the detection approach to fake review and social spam to help detecting astroturfing.

III. Learning-Based Astroturfing Detection Approach

To completely portray the character of astroturfing completely, many people started to use ensemble models, which using many kinds of information to train a classifier for astroturfing detection. Mainly divided into three categories: supervised, unsupervised, semi-supervised learning.

A. Supervised Learning-Based Astroturfing Detection

1) Content Feature Learning:

- *Expectation-Maximization(EM) Model.* EM model is used for text classification using labeled and unlabeled documents. A Naive Bayes classifier is firstly constructed by utilizing the existing labeled tweets. In the E-Step, this classifier is used to predict the labels of the remaining unlabeled tweets. In the M-Step, the probabilities of the word features are reestimated based on the predicted labels. Then the classifier in the E-Step is updated accordingly and is used to re-classify the tweets in the M-Step. This process continues till the number of changes in the predicted labels is below 0.01% of the total unlabeled tweets.

2) Behavior Feature Learning:

- *AdaBoost.* Adaboost is an iterative algorithm, the core idea of which is to train different classifiers (weak classifiers) for the same training set, and then combine these weak classifiers to form a stronger final classifier (strong classifier). It is achieved by changing the distribution

[1] https://en.wikipedia.org/wiki/Internet_Water_Army

TABLE I
COMPARISONS AMONG SUSPICIOUS ONLINE BEHAVIORS.

	Traditional Spam	Fake Review	Social Spam	Link Farming	Astroturfing
Main Application	Email and SMS	B2C e-commerce (Amazon, Netflix), C2C e-commerce (Taobao)	Social Contact (Facebook, Twitter, Weibo, LinkedIn)	Search Engine, Social Contact	B2C and C2C e-commerce, Social Contact
Participant Type	AI Program	AI Program or Human	AI Program or Human	AI Program or Human	AI Program or Human
Behavior Time	No Requirements	No Requirements	No Requirements	Long Time	Short Time
Participant Visual	Not Visible	Visible	Visible	Not Visible	Visible
Account Scale	Large	No Requirements	No Requirements	/	Large

of data, and it determines the weights of each sample according to if the classification of each sample in the training set is correct and the accuracy of the previous overall classification. The new data set with the modified weight value will be sent to the lower classifier for training. Finally, the trained classifiers are merged to be the final decision classifier.

- *K-Nearest*. The k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. In the classification phase, k is a user-defined constant, and an unlabeled vector is classified by assigning the label which is most frequent among the k training samples nearest to that query point. In the detection of astroturfing based on behavior, the goal is to distinguish between tweets receiving retweets from astroturfing accounts and tweets receiving retweets from normal accounts.

  Song and Lee et al. [14] find four new retweet-based features that allow us to distinguish astroturfing tweets from others: (i) retweet time distribution,(ii) the ratio of the most dominant application, (iii) the number of unreachable retweeters, and (iv) the number of received clicks. Next, they build the classification model, k-nearest neighbors, by using the retweet-based features and evaluate it with their ground-truth dataset.

3) Structure Feature Learning:

- *Support Vector Machine(SVM)*. Chen et al. [15] explore the characteristics of opinion spams and spammers in a web forum to obtain some insights, including subtlety property of opinion spams, spam post ratio, spammer accounts, first post and replies, submission time of posts, activeness of threads, and collusion among spammers. And then build classifier SVM with RBF kernel to detect spammer.

4) Multiple-Feature Learning:

- *Random Forest(RF)*. Lee and Webb et al. [16] propose a comprehensive analysis of astroturfing. Then trained classifier Random Forest-based classifier to detect astroturfing. They first find some valuable features such as profile features, content features and social network features. Then they computed feature values for each user in the training and testing sets, according to the previously described features. The authors selected the popular classification algorithm: Random Forest. Using the machine learning toolkit's implementation of the algorithm, they developed Random Forest-based classifier to predict whether a user is a astroturfing worker or a legitimate user.
- *Neural Autoencoder Decision Forest*. Dong et al. [17] proposed an end-to-end trainable unified model to leverage the appealing properties of autoencoder and random forest. In this model, they use autoencoder to generate the hidden representations of the features, and take them as the input of the random forest. The entire model is trained jointly via the stochastic and differentiable decision tree model, and the decision forest generates the final prediction.

B. Semi-Supervised Learning-Based Astroturfing Detection

In other domains, it has been found that using unlabeled data in conjunction with a small amount of labeled data can considerably improve learner accuracy compared to completely supervised methods.

1) Content Feature Learning:

- *PU-Learning Model*. Liu et al. [18] propose a semi-supervised learning approach to learn a few positive examples and a set of unlabeled data, the model is named PU-Learning and achieved an F1 Score of 83.7%.
- *FakeGAN Model*. In general, the main problem with applying classification methods for detecting deceptive reviews is the lack of substantial ground truth datasets. Aghakhani et al. [19] propose FakeGAN, a first system which using semi-supervised neural network-based learning methods for detecting deceptive fraudulent reviews. Unlike standard GAN models which have a single Generator and Discriminator model, FakeGAN uses two discriminator models and one generative model. The generator is modeled as a stochastic policy agent in reinforcement learning (RL), and the discriminators use Monte Carlo search algorithm to estimate and pass the intermediate action-value as the RL reward to the generator.

2) Behavior Feature Learning:

- *C4.5*. For the given data set, each tuple can be described by a set of attribute values, and each tuple belongs to a class in a mutually exclusive category. The goal of C4.5 is to find a mapping from attribute values to categories by learning, and this mapping can be used to categorize new entities with unknown classifications.

  In the detection of astroturfing, Xu et al. [20] proposed an analysis of the entire ecosystem that underlies the astroturfing attacks from multiple perspectives, and the behavioral discrepancies between the astroturfing accounts and the legitimate users in community QA. Then they use the Profile attributes, QA attributes and SN attributes of users as feature to train a classifier (C4.5) for detecting the astroturfing.

C. Unsupervised Learning-based and Other Astroturfing Detection

Because of the difficulty of producing accurately labeled datasets of astroturfing, the use of supervised learning is not always applicable. A novel unsupervised text mining model was developed.

1) Content Feature Learning:

- *LDA-based model*. Latent Dirichlet Allocation(LDA) is a document theme generation model and it is used to analyze the topics that are discussed by each cluster of users. It takes a corpus of documents as in-put, and outputs K topics, each of which is a list words sorted by the strength of their association with that topic.

  Dong et al. [21] propose an unsupervised topic-sentiment joint probabilistic model (UTSJ) based on Latent Dirichlet Allocation (LDA) model. This model first employs Gibbs sampling algorithm to approximate parameters of maximum likelihood function offline and obtain topic-sentiment joint probabilistic distribution vector for each review. Secondly, a Random Forest classifier and a SVM (Support Vector Machine) classifier are trained offline, respectively. Experimental results on real-life datasets show that UTSJ model is better than baseline models.
- *MF Model*. A matrix factorization model is employed to learn lexicon information from external spam resources. Instead of learning knowledge at word-level, we propose to capture the external knowledge from topic-level. The proposed method is built on the orthogonal nonnegative matrix tri-factorization model (ONMTF). The basic idea is to cluster data instances based on distribution of features, and cluster features according to the distribution of data instances. The ONMTF can be formulated by optimizing:

  $$\begin{aligned} &\min_{U,H,V\geq 0} \; \|X - UHV^T\|_F^2, \\ &s.t. \quad U^TU = I, \; V^TV = I \end{aligned} \tag{2}$$

  where X is the content matrix, and U and V are nonnegative matrices indicating low-dimensional representations of words and users, respectively. m is the size of vocabulary, c is the number of classes, d is the number of users. H provides a condensed view of X. The orthogonal and nonnegative conditions of U and V provide a hard assignment of class label to the words and users. With the ONMTF model, we project the original content information from the other media into a latent topic space. By adding a topic-level least squares penalty to the ONMTF.
- *Semantic Language Model(SLM)*. SLM is a novel unsupervised text mining model which was developed and integrated into a semantic language model for detecting untruthful reviews by Raymond [22]. They create an approximation method for calculating the degree of untruthfulness for reviews based on the duplicate identification results by estimating the overlap of semantic contents among reviews using a Semantic Language Model. In addition to performing unsupervised review spam detection, they also developed a high-order concept of association mining to extract context-sensitive concept association knowledge. Their model follows the assumed logic that if the semantic content of a review is close to those of another review, it is likely that the two reviews are duplicates and thus examples of spam reviews. For their experiment, they built a dataset from real-world reviews collected from Amazon and achieve an AUC of 0.9987.

2) Structure Feature Learning:

In the detection of astroturfing based on structure, Fakhraei and Foulds [23] proposed a method to detect spammers in multi-relational social networks, they model the social network as a time-stamped multi-relational graph where vertices represent users, and edges represent different activities between them. First, extract graph structure features for each of the relations. Second, consider the activity sequence of each user across these relations and extract k-gram features and employ mixtures of Markov models to label spammers. Finally, propose a statistical relational model based on hinge-loss Markov random fields to perform collective reasoning using signals from an abuse reporting system in the social network.

3) Graph-Model Based:

Numerous methods are based on graph models, especially the structure based ways in the previous paper. [6], [8], [23] The graph-model based technology can be widely used in the detection of fake review, social spam [6], [7]and link farming [8] etc.

Ratkiewicz and Conover et al. [24] put forward a machine learning framework which combines topological, content-based and crowdsourced features of information diffusion networks on Twitter to detect the early stages of viral spreading of political misinformation. To represent the flow of information through the Twitter community, a directed graph is constructed, whose nodes are individual user accounts and edges are retweet and mention operation.

Liu et al. [25] propose a complex probabilistic graph classification approach to address the problem of opinion astroturfing detection. To obtain an initial effective estimation for the nodes (reviews, authors, and products) in the graph, they first train a neural network with attention mechanism to learn the multimodal embedded representation of nodes by leveraging both textual and rich features. Then based on the node prior computation, a heterogeneous graph is constructed to capture the relationships among different kinds of nodes, and the beliefs are further updated through iterative message propagation.

Due to the past astroturfing detection techniques only consider one or two types of astroturfing entities such as review, reviewer, group of reviewers or product. Noekhah [26] proposed a novel graph-based model called “Multi-iterative Graph-based opinion Spam Detection” (MGSD). The model is able to evaluate the ‘spamicity’ effects of entities more efficiently given it applies a novel multi-iterative algorithm which considers different sets of factors to update the spamicity score of entities. Besides, to enhance the accuracy of the MGSD detection model, a higher number of existing weighted features along with the novel proposed features from different categories were selected using a combination of feature fusion techniques and machine learning (ML) algorithms. The output of the MGSD model showed that the feature selection and feature fusion techniques showed a remarkable improvement in detecting astroturfing.

IV. Evaluation Criterion

There are many metrics that can be used to evaluate the performance of the algorithm for spammer detection. Such as Accuracy, Precision, Recall, F1 Score, AUROC, FPR. All these metrics are used for classification models.

1) Precision, Recall and F1 Score:

Precision and Recall are frequently used as indicators for classification. Precision measures the fraction of examples classified as positive that are truly positive. Recall measures the fraction of positive examples that are correctly labeled. And F1 Score is the weighted harmonic mean of Precision and Recall, it tradeoff the Precision and Recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

TP means true positives (i.e. items correctly labeled as belonging to the positive class), FP means false positives (i.e. items incorrectly labeled as belonging to the positive class), FN means false negatives (i.e. items which were not labeled as belonging to the positive class but should have been).

In the area of spammer detection, many approach used Precision, Recall and F1 Score as the appropriate metric to measure the performance of models, such as Hu [27], Hu [28], Hu [29], Sedhai [30], Liu [31], Dong [21], Liu [32], Liu [25], Dong [17] and You [33].

2) AUROC and AUPR:

AUROC means the area under the ROC, it draws how TPR (True Positive Rate) changes according to the changes of FPR (False Positive Rate). Among them, The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labeled. AUPR means the area under the Precision-Recall. In the work of Fakhraei [23], in order to avoid over-optimistic estimates of the PR curve and ROC, Fakhraei used the AUROC and AUPR to estimate the performance of their method. But in the work of Song [14] and Wang [34], they only use the AUROC as evaluate metric.

3) TPR, FPR and FNR:

TRP and FPR havee been explained above. FNR indicates the proportion of a sample that is classified as negative but is actually positive. In the actual experiment, we hope the TPR is as large as possible, while the FPR and FNR are as small as possible.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

$$FNR = \frac{FN}{TP + FN}$$

In their works, Lee et al. [16], Barushka [35] and Lee [36], they both used the FPR and FNR as metrics to evaluate the classifier for astroturfing detection. In the work of Sun [37] and Xu [20], both of them are used the TPR and FPR as measures to evaluate the detection performance of classifier.

4) Accuracy and ER:

Accuracy measures the fraction of examples that correctly labeled account for all examples. But the ER (Error Rate) is measures the fraction of misclassified examples over all examples. They are also the most used evaluation metric.

$$Accuracy = \frac{TP + TN}{P + N}$$

$$ER = \frac{FP + FN}{P + N}$$

Sun [37] and Wang [38] are used the ER to evaluate the detection performance of method they proposed. But in work of Lee [16], Yang [39], Li [40], Zhang [41], Aghakhani [19], Dhingra [42], You [33] and Dong [17], they are both used accuracy as measure to evaluate the classifier they build.

V. Applications

Astroturfing can disrupt the normal order of the network and bring many negative effects to society and people's life. It is necessary to design schemes to help normal users, administrators, or even law enforcers quickly identify potential astroturfing. Astroturfing detection can help users distinguish truth from falsehood and get the information that they really need. In social networks, astroturfing detection can be used for various applications; here we will address some typical applications, such as detection of astroturfing in social networks, astroturfing account recognition, and deceptive reviews identification.

A. Single Astroturfing Detection

- *Detecting Astroturfing in Twitter*. Lee and Webb et al. [16] present a comprehensive analysis of astroturfing on Fiverr.com and Twitter, and find some valuable features such as profile features, activity features, content features, social network features, personality features and temporal features. Then they compute feature values for each user in the training and testing sets, and they train the classifier (SVM-based classifier and Random Forest-based classifier) to detect astroturfing on Fiverr.com and Twitter.
- *Automatic Review Synthesis Model*. Sun and Morales et al. [37] leveraged the difference of semantic flows between synthetic and truthful reviews to identify the review spam, they used SVM and Naive Bayes as classifiers to tell if one review is truthful or synthetic. They develop a model to automatically generate positive reviews by mixing up those existing reviews.

B. Group Astroturfing Detection

- *Real-Time Astroturfing Detection System*. Detecting astroturfing through establishing classifier is a typical application for astroturfing detection. Chen et al. [43] discuss fundamental architecture and design of a detection system that can identify malicious behavior and potential paid posters in real time. The goal of the system is to identify potential paid posters and locate their user IDs during the process of collecting information. This system will automatic collect data from different resources/websites and generate reports of the behavior of potential paid posters. The system will provide valuable information for the analysts and online users to differentiate on various aspects.
- *Multi-agent System*. Analyze the distribution and behavior characteristics of astroturfing is also helpful to better understanding and monitoring the astroturfing accounts. The work investigated the behavior patterns and strategies of astroturfing in online forums [44]. They construct a multi-agent system and conduct an experiment using the real world dataset of online forum to study the impact factors of the astroturfing' ability to exert influence. They found that Internet Water Army dynamically adjusts their behavior strategy to maximize their influence and the effectiveness of strategy of Internet Water Army is closely related to the features of the users.

VI. Conclusion

Although numerous efforts have been made in astroturfing detection, there are still many potential future challenges, and some new open problems require further study. Here, we address some possible future research challenges on the astroturfing detection problem:

- *Crossing Domain*. Detection of astroturfing involves a number of areas and how to apply the model trained in the source field to the target field is an important research direction. Li et al. trained the model in hotel domain, and tested in restaurant and doctor domains, but experimental results showed that the accuracy and F1 score droped seriously compared with the performance in the same filed [45]. So more in-depth exploration and research are needed for cross-domain deceptive opinion astroturfing detection [46].
- *Missing Datasets*. Public astroturfing generally works underground, and the available dataset are not enough for researching their behavior. Thus, the research work for the astroturfing detection is difficult to be carried out.
- *Complexity of the Internet*. The astroturfing work is always cross platform, cross channel, and have numerous sock puppets. This makes their behavior information become fragmented, and difficult to link to each other.

Acknowledgments

The author is grateful to Dr. Tong Chen and Dr. Yingxiao Xiang for their help with the preparation in this paper.

References

[1] H. C. Kelman, "Compliance, identification, and internalization: Three processes of attitude change," *Journal of Conflict Resolution*, vol. 2, no. 1, pp. 51–60, 1958.

[2] J. Li, https://www.sohu.com/a/37167071_114774.

[3] Y. Cui, http://news.ifeng.com/c/7fbHXR4A41x.

[4] M. G. R. T. A. V. Benevenuto, F., "Detecting spammers on twitter," *In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 2, p. 12, 2010.

[5] M. Jiang, A. Beutel, P. Cui, and B. Hooi, "A general suspiciousness metric for dense blocks in multimodal data," in *International Conference on Data Mining*, 2015, pp. 781–786.

[6] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July*, 2010, pp. 435–442.

[7] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," in *IEEE International Conference on Data Mining*, 2014, pp. 180–189.

[8] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *International Conference on World Wide Web*, 2012, pp. 56–61.

[9] J. C. Stauber and S. Rampton, "Toxic sludge is good for you : lies, damn lies, and the public relations industry," *Journalism & Mass Communication Educator*, vol. 52, no. 3, pp. 314–317, 1995.

[10] J. G. McNutt, "Researching advocacy groups: Internet sources for research about public interest groups and social movement organizations," *Journal of Policy Practice*, vol. 9, no. 3, pp. 308–312, 2010.

[11] C. H. Cho, M. L. Martens, H. Kim, and M. Rodrigue, "Astroturfing global warming: It isn't always greener on the other side of the fence," *Journal of Business Ethics*, vol. 104, no. 4, pp. 571–587, 2011.

[12] T. P. Lyon and J. W. Maxwell, "Astroturf: Interest group lobbying and corporate strategy," *Journal of Economics & Management Strategy*, vol. 13, no. 4, pp. 561–597, 2004.

[13] J. Hoggan and R. Littlemore, ""climate cover-up: The crusade to deny global warming"," *Energy & Environment*, vol. 21, no. 3, pp. 363–364, 2010.

[14] J. Song, S. Lee, and J. Kim, "Crowdtarget: Target-based detection of crowdturfing in online social networks," in *ACM Sigsac Conference on Computer and Communications Security*, 2015, pp. 111–114.

[15] Y. R. Chen and H. H. Chen, "Opinion spam detection in web forum: A real case study," *in Proceedings of the 24th International Conference on World Wide Web.International World Wide Web Conferences Steering Committee*, 2015.

[16] K. Lee, S. Webb, and H. Ge, "Characterizing and automatically detecting crowdturfing in fiverr and twitter," *Social Netw. Analys. Mining*, vol. 5, pp. 2:1–2:16, 2015.

[17] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognition Letters*, vol. 132, pp. 21–29, 2020.

[18] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Data Mining, IEEE International Conference on*, 2003, p. 179.

[19] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 89–95.

[20] A. Xu, X. Feng, and Y. Tian, "Revealing, characterizing, and detecting crowdsourcing spammers: A case study in community q and a," in *INFOCOM*, 2015.

[21] L.-y. Dong, S.-j. Ji, C.-j. Zhang, Q. Zhang, D. W. Chiu, L.-q. Qiu, and D. Li, "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Systems with Applications*, vol. 114, pp. 210–223, 2018.

[22] R. Y. K. Lau, S. Y. Liao, C. W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," *Acm Transactions on Management Information Systems*, vol. 2, no. 4, pp. 1–30, 2011.

[23] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor, "Collective spammer detection in evolving multi-relational social networks," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1769–1778.

[24] J. Ratkiewicz, M. Conover, M. Meiss, B. GonÃ§alves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media."

in *International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July*, 2011.

[25] Y. Liu, B. Pang, and X. Wang, "Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph," *Neurocomputing*, vol. 366, pp. 276–283, 2019.

[26] S. Noekhah, N. binti Salim, and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," *Information Processing & Management*, vol. 57, no. 1, p. 102140, 2020.

[27] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *International Joint Conference on Artificial Intelligence*, 2013, pp. 1709–1714.

[28] X. Hu, J. Tang, and H. Liu, "Leveraging knowledge across media for spammer detection in microblogging," in *SIGIR*, 2014.

[29] ——, "Online social spammer detection," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[30] S. Sedhai and A. Sun, "Hspam14: A collection of 14 million tweets for hashtag-oriented spam research," in *The International ACM SIGIR Conference*, 2015, pp. 223–232.

[31] Y. Liu, Y. Liu, M. Zhang, and S. Ma, "Pay me and i'll follow you: Detection of crowdturfing following activities in microblog environment," in *IJCAI*, 2016.

[32] Y. Liu and B. Pang, "A unified framework for detecting author spamicity by modeling review deviation," *Expert Systems with Applications*, vol. 112, pp. 148–155, 2018.

[33] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang, "Integrating aspect analysis and local outlier factor for intelligent review spam detection," *Future Generation Computer Systems*, vol. 102, pp. 163–172, 2020.

[34] X. Wang, B. Zhou, Y. Jia, and S. Li, "Detecting internet hidden paid posters based on group and individual characteristics," in *WISE*, 2015.

[35] A. Barushka and P. Hajek, "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4239–4257, 2020.

[36] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media," in *in ICWSM*, 2013.

[37] H. Sun, A. Morales, and X. Yan, "Synthetic review spamming and defense," in *KDD*, 2013.

[38] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao, "Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers," in *The Usenix Security Symposium*, 2014.

[39] X. Yang, Q. Yang, and C. Wilson, "Penny for your thoughts: Searching for the 50 cent party on sina weibo," in *ICWSM*, 2015.

[40] L. Li, B. Qin, W. Ren, and T. Liu, "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, vol. 254, pp. 33–41, 2017.

[41] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "Dri-rcnn: An approach to deceptive review identification using recurrent convolutional neural network," *Information Processing & Management*, vol. 54, no. 4, pp. 576–592, 2018.

[42] K. Dhingra and S. K. Yadav, "Spam analysis of big reviews dataset using fuzzy ranking evaluation algorithm and hadoop," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2143–2162, 2019.

[43] C. Chen, K. Wu, S. Venkatesh, and X. Zhang, "Battling the internet water army: Detection of hidden paid posters," *CoRR*, vol. abs/1111.4297, 2013.

[44] K. Zeng, X. Wang, Q. Zhang, X. Zhang, and F. Y. Wang, "Behavior modeling of internet water army in online forums," in *Ifac World Congress*, 2014, pp. 9858–9863.

[45] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1566–1576.

[46] Y. Ren and D. Ji, "Learning to detect deceptive opinion spam: A survey," *IEEE Access*, vol. 7, pp. 42 934–42 945, 2019.