# Detecting Spammers in Social Networks

**Eman A. Altameem[1†] and Mehmet Sabih Aksoy [2††],**

[1†] Imam Mohammad Bin Saud University, College of Computer and Information Sciences, Riyadh, Saudi Arabia
[2††]King Saud University, College of Computer and Information Sciences, Riyadh, Saudi Arabia

**Abstract**
Social networking sites involve millions of users all over the wor ld for communicating, sharing, storing and managing significant information. Twitter has become one of most widely used social networking platforms. Unfortunately, this huge popularity also attracts spammers who misuse the valuable information and threatened normal users' personal privacy and information security. A lot of research has been developed for detecting spammers on social networking sites. However, social spammers frequently change their spamming strategies to overcome the detection system. In this paper, we perform a review of techniques for detecting spammers on Twitter. Features for the detection of spammers are also described.

***Key words:***
*Online Social Networks, Spam Detection, Machine Learning, Twitter.*

## 1. Introduction

Online social networking (OSNs) has grown tremendously and changing the way people keep in contacts with each other. These social networking sites , such as Twitter ,Facebook and LinkedIn become a very important tool as they allow people to conveniently communicate with friends and family, share posts about their lives freely, and follow hot topics immediately [1,2]. One of the most popular OSNs, Twitter, a platform whereby people share ideas with their followers using short messages known as 'tweets'. It becomes an online source for acquiring real-time information about users. Twitter has 330 million monthly active users, has experienced the fastest growth over the past few years [3].

Social networking sites allow people to share information anonymously and as such, have contributed significantly toward free expression of ideas and opinions, without fear of intimidation or punishment. However, an increase in the number of social media users has created opportunities for spammers. Spammers use social networking platforms as an avenue for disseminating spam messages to achieve their malicious goals. Additionally, spread large amounts of harmful information that seriously threatened normal users' personal privacy and information security [5]. Twitter has been one of the targets for attackers as evidenced from a previous research results found that in well-defined periods, more than 9.9 million spammy or automated accounts per week were identified by twitter [6].

Spam messages constrain the network resources and data mining processes, thereby increasing the network's operational burden. Indeed, some spammers profit from interfering with the normal marketing promotion through the use of malicious replies, votes, comments and likes, and this ultimately harms a network's credibility as well as the trust between the network and its users. In this light, timely detection of spam messages distributed in social networks is indispensable for the users' privacy and credibility of social networks [7].

Multiple studies have been conducted on the characteristics of social network spammers. Although these studies have formulated ways of blocking spamming, spammers have consistently devised ways of bypassing spam detection systems [4]. Thus, continuous exploration of new spamming techniques, features, and characteristics is necessary for consistent detection of spamming activities on social networking sites.

In this regard, this paper aims to provide a review of the academic research and work conducted by other researchers. Moreover, presenting the techniques available for detection of spammers in Twitter. The rest of this paper

Manuscript received November 5, 2020.
Manuscript revised November 20, 2020.
https://doi.org/10.22937/IJCSNS.2020.20.11.19

is organized as follows. Section 2 Literature Review. Section 3 Features Used in Spammer Detection. Section 4 presents the Existing techniques for spammers detection. Finally, section 5 concludes the paper.

## 2. Literature Review

Detection of spamming is not a new field of research. It has been under analysis for a decade. It has concentrate d predominantly on email and web spam detection [8,9] in the past, while it has been an increasingly trending topic for OSNs in the last few years.

Amongst the fast growing social networking sites, Twitter has attracted millions of daily active users. This has not been without challenges since the environment has currently received threats mostly as a result of the malicious activities propagated by spammers. Users are greatly concerned about their privacy with a number of them leaving the sites completely despite the measures installed in place to address the issue. It has, in fact, become a highly contested topic in both academic and corporate fields consequently offering research opportunities for spam detection and identification and methods of preventing spammers' activities.

Most studies of spam detection are overly based on exploiting users' features as training for machine-learning classifiers. This, for instance, depends on user's number of followers over followings, the ratio of followers over followings, tweets generated on daily basis, and hashtags per given tweet amongst other factors [10]. In addition and based on the study [11] that conducted by Chao and Vern, the detection can be done by automatically analyzing the creation time of spam messages.

According to Benevento et al [12], They developed their model by using the Support Vector Machine (SVM) and involving over 80 million user accounts classifies account type based on their contents and behavior. For instance, in terms of content based, this includes such attributes like the text of tweets, the number of URLs, number of words per tweet, number of mentions and so on. Approximately 70% of spammers and 96% of non-spammers were correctly classified. Moreover, this experiment also identified the most important features for spam detection on Twitter.

In M. McCord et al, tweets were collected and such features like the timing of tweets posted, number of retweets, mentions, and keywords were extracted. Their model depends on different classification algorithms with random forest that show the best results [13]. Based on the 100 most recent tweets also show that spam detection based on their suggested features achieve 95.7% precision and 95.7% F-measure. Other techniques used for spam detection in Twitter are as outlined in [14]. In this case, around 26,000 users and their tweets were investigated for spam detection. This involved the extraction of features such as URL rate and interaction ratio and using J48 classifier for spam classification.

In other studies, the graph algorithms alongside the necessary features were used for spam detection. In each case, every Twitter account was viewed as a node while the followings relations were treated as directed edges. The assumptions made were that spammers usually had many followings and fewer followers. As such, the bi-directional edges were low. The developed graph algorithms detected the relationships in the essence that the edges for spam accounts tend to differ from those of legitimate user accounts [15] [16].

Further successful studies capitalized on clustering algorithms in streaming data for spam detection indicating that the algorithms could offer similar results compared to other classification algorithms only that if they were properly used [17].

## 3. Features Used in Spammer Detection

A high prevalence of spam actions promoted many OSNs to adopt spam detection features. Identification of the

features and characteristics of spammers is grounded on the differences between the behaviour of spammers and non-spammers. Literature shows that there are a number of features that are unique to spammers' accounts. Spam detection features on Twitter can be categorized as User-based, Content- based, or by the relationship between sender and receiver.

### 3.1 User Based Features

The use of features linked to social network user profile is an attribute used in earlier for spam detection. These features include the number of followers, the number of friends and number of followings [12]. By using such features, the ratios of followers to followings, and that of followers to friends as well as the reputation scores are obtained. In this case, reputation score is the number of followers over the total number of the people on the user network [18]. That is:

$$Reputation\ score = \frac{num\ (followers)}{num(folowers + following)}$$

### 3.2 Content Based Features

Content-based features depend on the content or messages that users write. Spammers use to post tweets to spread misinformation and mostly advertise their products. Accordingly, their tweets' contents have distinguished features that differentiate them from legitimate users. One such feature is the total number of tweets a user posted on the site as spammers have high tweeting frequency compared to legitimate users [12]. On the tweet level, the minimum, maximum, mode ,median ,standard deviation and average number of words per tweet are computed [25] .Others include the proportion of tweets with URLs, hashtags and mentions and are taken into consideration based on the fact that their distribution usages are different between the spammers and non-spammers [14,25].They are obtained by:

$$\frac{num\ of\ tweets\ \ with\ URL\ or\ mentions\ or\ hashtags}{total\ num\ of\ tweets}$$

In the case of the URLs, for instance, a legitimate user is expected to use a unique URL only once. On the other hand, a spammer tends to have a higher average usage of unique URL [25]. This is calculated as follows:

$$\frac{Num\ (URLs)}{Num\ (unique\ URLs)}$$

### 3.3 Graph Based Features

As Twitter is a network of users with edges between them and tweets. Graph-based features modulate a user's relationship with tweets by a structure that maps as a graph. On the graphical representation, nodes represent users and tweets while the links between nodes indicate relationships. Mapping the interactions enables twitter to track relationships between the sender and mentions in order to reveal the possibility of a spam connection. This method has been successfully applied for anomaly detection in a wide array of applications [19].

Unlike account-based and user-based functions, graph-based features are difficult to manipulate. Extracting these features, however, involves an in-depth analysis of the enormous and complex Twitter graph, which is time- and resource-intensive [20].

## 4. Existing Techniques for Spammers Detection

Different techniques have been used by researchers to find out the spammers in various OSNs.The classification of spam is commonly handled by machine learning algorithms intended to distinguish between spam and non-spam messages. machine learning techniques discover hidden structures and patterns from the data and can detect anomalies in the data like spam messages or network intrusion. Machine learning algorithms achieve this by using an automatic and adaptive technique. In the following parts, we will review some of

the most popular machine learning methods for spamming detection.

*Clustering*

Clustering analysis is widely used in many fields and most common form of Unsupervised Learning. It classifies objects or opinions into similar collections known as clusters [26]. This technique is an effective way used for spam detection, involves classifying different spammers and spamming strategies. Once spamming patterns are identified, it is easier to detect new spammers or new tweets containing spammed messages based on the categories in predefined clusters [21]. There are several types of clustering techniques that solve many problems, for spam classification, two types that are primarily used. Density-based clusters are dense areas in the data space that are isolated from each other by sparser areas [32]. The other method is that K-nearest neighbors (KNN) where tweets are classified based on the class of their nearest neighbors and this method is widely used for spam filtering [21].

*Support Vector Machine*

Supervised machine learning is a category of machine learning that uses labelled datasets in training algorithms to predict outcomes or classify data correctly. Supervised machine learning enables organizations such as twitter to classify spam messages or social media posts [22] . Support Vector Machine (SVM) is one of the most well-known supervised machine learning that uses statistical theory to classify the dataset. They have supervised learning models that evaluate knowledge, define categorization trends, and investigate the relationship between interest variables [21]. Many academic researchers have tended to use SVM, because it gives state-of-art performance on various pattern recognition applications. SVM classifier has been used commonly to distinguish between spammers and non-spammers with high accuracy due to its ability to model multidimensional boundaries that are not sequential [23]. Furthermore, SVM is considered a major example of "kernel methods," which is one of the key machine learning fields that help solve social network spams. However, SVM's power and effectiveness decline with time for high-dimensional data due to the technical complexities of the data processed [21].

*Naïve Bayes algorithm*

Naïve Bayes algorithm is standard probabilistic approaches that have been used effectively by various machine learning techniques. It is a fast learning algorithm that can be used for classification in real time. Naïve Based classifier is a method that is used to detect spam through the principle of class conditional independence from the Bayes theorem. In this classifier, the presence of one feature does not have a sufficiently significant impact on the presence of another in the determination of the probability of a given outcome [20]. As a result, each predictor has an equal effect on the provided result. Naïve Bayes classifier is important in spam identification as it uses the Bayes theorem rule to allow algorithms to classify each object by looking at all the existing features in an individual manner which enables filtering of spam and givens a low false positive spam detection to users [24].

*Decision-Tree*

Decision-trees (D-Trees) is a supervised machine learning algorithm whose shape looks like the structure of a tree and made up of decision nodes and prediction nodes. Through learning decision rules generated from features, D-trees predicts the values of responses. This method has been successfully implemented in the area of spam identification. D-trees make use of feature selection or variable analysis of the trained data. However, the operation of D-trees is not dependent on the relationships which exist amongst parameters. D-trees provides a capacity to assign unambiguous values to different decisions, problems, and

the results of every decision [2]. D-trees are one of the most effective spam filtering techniques due to the decreased vagueness experienced in the decision-making process.

*Deep Learning*

Deep learning is a subfield of machine learning that utilized multiple layers of data to progressively extract higher level-features from the raw data input. Deep learning algorithms are used to detect if the tweet is spammer or not. One form of deep learning is deep neural network which is a technique used in spam identification that involves training the data set on the neural network [27]. Deep neural networks (DNNs) are primarily leveraged for deep machine learning algorithms which train data sets by mimicking the interconnectivity that exists in the human brain through the many layers of nodes. The utilization of deep neural networks in spam identification has been found to have higher levels of accuracy as compared to other spam detection techniques [28].

One of the other widely used algorithm in deep learning is Convolutional Neural Network (CNN), which is a form of deep learning technology that works different from the traditional neural networks. CNN operates precisely on messages to mine valuable, essential features for classification [3]. CNN is used for spam detection by adding a semantic layer which is composed of the training of random word vectors. The semantic convolutional neural network (SCAN) can identify spam in social media text regardless of the exponential increase of spam volume over social media networks especially on Twitter [29].

## 5. Discussion

Twitter uses intelligence techniques to detect spamming. Different machine learning techniques have been applied for classifying massages as either spam or not. Spam detection systems observe user behavior as well as the content of tweets. In that regard, the spam detection systems can be graph-based, content-based, or user-based data. As much as Twitter may not be considered to detect all spammers, continuous improvement of its spam detection systems shows that it is approaching optimal levels [30]. The constant evolution of attacks is evident since spammers are also becoming technologically wiser. For instance, some malicious-minded people use autonomous agents such as bots to test the vulnerability of the spam detection systems. For this reason, there is a need for continuous improvement of the spam detection system than the currently used supervised and unsupervised machine learning techniques [31]. The current machine learning strategies are considered not fully-capable of detecting the dynamic behavior of spammed content. In most cases, social networks permit legitimate users to report suspicious activities so as to let administrators confirm whether a particular account is malicious or not. Machine learning spam detection systems should incorporate learning capabilities so that they can adapt to user behavior.

## 6. Conclusion

Twitter attracts a significant volume of spam that is on a steady increase into the future. Current machine learning spam detection technology has proved to be useful in the creation of effective spam filtering systems. Machine learning techniques such as clustering, SVM, Naïve Bayes, decision-trees and deep learning algorithms have been found to be effective in spam detection regardless of the increasing volume of spam experienced in twitter and other social media platforms. Although weaknesses do exist, it is important that current machine learning spam detection techniques are able to evolve with the changing needs given the increased technical knowledge displayed by spammer in Twitter. Features for detecting spammers has been reviewed on the basis of user based features or content based features or graph base feature.

## References


[1] C. Yang , R. Harkreader , G. Gu , Empirical evaluation and new design for fighting evolving twitter spammers, IEEE Trans. Inf. Forensics Secur. (2013) 1280–1293.

[2] M. Fazil , M. Abulaish , A hybrid approach for detecting automated spammers in twitter, IEEE Trans. Inf. Forensics Secur. (2018) 2707–2719.

[3] Oberlo ,10 Twitter Statistics Every Marketer Should Know ,(2020), Accessed on 26 Sep 2020. [Online]. Available: https://www.oberlo.com/blog/twitter-statistics.

[4] H .Sundaram, Y. R .Lin, M. De Choudhury and A. Kelliher, Understanding community dynamics in online social networks: a multidisciplinary review, IEEE Signal Processing Magazine, (2016), 33-40.

[5] H. Shen, X. Liu, Detecting Spammers on Twitter Based on Content and Social Interaction, International Conference on Network and Information Systems for Computers,(2015).

[6] Y. Roth, D. Harvey, How Twitter is fighting spam and malicious automation, (2018).

[7] A. Sanzgiri, A. Hughes, S. Upadhyaya, Analysis of malware propagation in Twitter, In Reliable Distributed Systems (SRDS), 2013 IEEE 32nd International Symposium,(2015).

[8] I. Idris , A. Selamat , N.T. Nguyen , S. Omatu , O. Krejcar , K. Kuca , M. Penhaker , A combined negative selection algorithm–particle swarm optimization for an email spam detection system, Eng , (2015).

[9] L. Araujo , J. Martinez-Romo , Web spam detection: new classification features based on qualified link analysis and language models, IEEE Trans. Inf. Forensics Secur. 5 (3) (2010) 581–590.

[10 ] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, Who is tweeting on Twitter: human, bot, or cyborg?, In Proceedings of the 26th Annual Computer Security Applications Conference (ACM), (2010).

[11] c. M. Zhang, and Y. Paxson, Detecting and analyzing automated activity on Twitter, Proceedings of the 12th international conference on Passive and active measurement,(2011).

[12] F. Benevenuto, G. Magno, T.Rodrigues, and Y. Almeida, Detect spammers on Twitter, Proceedings of the 7th Annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference (CEAS) , (2010).

[13] M. McCord and M. Chuah, Spam detection on twitter using traditional classifiers, Autonomic and Trusted Computing. Springer, pp. 175–186, (2011).

[14] P.-C. Lin and P.-M. Huang , A study of effective features for detecting long-surviving twitter spam accounts, Advanced Communication Technology (ICACT), 15th International Conference on. IEEE, pp. 841–846, (2013).

[15] C. Yang, R. C. Harkreader, and G. GU, Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers, Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection (RAID), (2016).

[16] J. Song, S. Lee, and J. Kim , Spam Filtering in Twitter using Sender Receiver Relationship, Proceedings of International Symposium on Recent Advances in Intrusion Detection (RAID),(2011).

[17] Z. Miller, B. Dickinson, W. Deitrick ,W. Hu , Twitter spammer detection using data stream clustering. Information Sciences, 260: p. 64-73, (2014).

[18] E. Abozinadah, A. Mbaziira and J. Jones Jr, Detection of abusive accounts in Twitter, J Knowl Eng, 1(2), 113-119, (2015).

[19] R. Paudel, P. Kandel , W. Eberle, Detecting Spam Tweets in Trending Topics Using Graph-Based Approach, (2020).

[20] A. Talha, R. Kara, A survey of spam detection methods on Twitter. International Journal of Advanced Computer Science and Applications, 8(3), 29-39 , (2017).

[21] E. Dada, J. Bassi, H. Chiroma, S. Abdulhamid, A. Adetunmbi, O. Ajibuwa , Machine learning for email spam filtering: review, approaches and open research problems ,(2019).

[22] N. Sun, G. Lin, J. Qiu, P. Rimba, Near real-time Twitter spam detection with machine learning techniques. International Journal of Computers and Applications ,(2020).

[23] Z. Torabi, S. Nadimi, H. Mohammad, A. Nabiollahi, Efficient Support Vector Machines for Spam Detection: A Survey, (IJCSIS) International Journal of Computer Science and Information Security, (2015).

[24] S. Hershkop, S. Stolfo, Combining email models for false positive reduction,(2005).

[25] N. El-Mawass, S. Alaboodi, Detecting Arabic spammers and content polluters on Twitter, 2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC), (2016).

[26] T. Madhulatha, An Overview on Clustering Methods. IOSR Journal of Engineering , (2012).

[27] T. Wu, S. Wen, S. Liu, J. Zhang, Y. Xiang, M. Alrubaian, M. Hassan, Detecting spamming activities in Twitter based on deep-learning technique, (2017).

[28] S. Nilizadeh, F. Labrèche, A. Sedighian, A. Zand, J. Fernandez, C. Kruegel, G. Stringhini, G. Vigna, Poised: Spotting Twitter spam off the beaten paths. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, (2017).

[29] M. Verma, D. Divya, S. Sofat, Techniques to detect spammers in twitter- A survey ,International Journal of Computer Applications, 85(10), (2014).

[30] R. Katpatal, A. Junnarkar, Spam detection techniques for twitter, International Research Journal of Engineering and technology, (2018).

[31] I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos, Detection of spam-posting accounts on Twitter. Neurocomputing, (2018).

[32] M. Ester , Density-based Clustering. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, (2009).