Social Network Analysis and Mining (2019) 9:35
https://doi.org/10.1007/s13278-019-0582-x

REVIEW ARTICLE

A survey on detecting spam accounts on Twitter network

Oğuzhan Çıtlak[1] · **Murat Dörterler**[1] · **İbrahim Alper Doğru**[1]

Received: 6 August 2018 / Revised: 20 May 2019 / Accepted: 15 July 2019 / Published online: 19 July 2019

Abstract
Social networks have become an inseparable part of our lives today. Services such as Facebook, Twitter, Instagram, Google + and LinkedIn in particular have had a significant place in Internet use in recent years. People establish instant interactions between each other over the Internet using these social services. They get many advantages such as creating their own groups, being informed about different interest areas and being able to make many contacts. Twitter is one of the mostly used platforms among the social networks. A social network that is being used so commonly has become a target for the vicious people (spammers). There is an increase in the number of spammers on Twitter too. Malicious content and messages (spams) prepared by the spammers do threat the security as well as performance. The first and most important condition to protect against this threat is to know the harmful methods of spam. Thus, this will make it easier to detect and protect. In this study, prominent detection methods of spams are analyzed. How the real users and fake users are distinguished as well as weak and strong aspects of the methods for these processes are compared and evaluated.

Keywords Social network · Spam accounts · Spam detection on Twitter · Spam analysis on Twitter

1 Introduction

Today, people are able to use the Internet widely regardless of time and space. For Internet users, social networks have become areas where people spend most of their time. These social platforms, where individuals can express themselves easily and share information, have, too become an indispensable part of everyday life (Erdoğan and Bahtiyar 2014). Today, when the Internet has entered our lives completely, the use of social media has increased considerably and all people have reached the possibility of communicating with a person in a globalizing world. Statistics show that the average usage rate of social networks has exceeded the usage rates of other sites (Stringhini et al. 2010). Individuals' purpose of using social media changes from one person to another. The expectations of each individual from social media tools are different, which can lead to different uses as well. While, for some users, social media is a medium on which individuals could escape from socializing, be alone and where they are more like an audience, for others it can emerge as a medium that they can socialize, be appreciated within the community and followed up, and express themselves comfortably. From this point of view, social media is defined as a structure built on technology that enables a deep social interaction, group formation and cooperation (Akar 2010). The fact that people prefer the use of technology instead of face-to-face communication in human relationships has caused social media to be used in every field. The fact that social media has begun to be effectively used in the education of the individuals (Öztürk and Talas 2015). How these messages and the sharing they make and emotions reflected can be used to predict the outcomes in the real world (Stephen and Galak 2012), and even the use of these emotions, thoughts and behaviors therein in different areas such as determining the value of the reputation of the individual in the real world is the greatest indication of this (Peleja et al. 2014). In a study by Hacıefendioğlu (2011), social networks are often used as an advertising medium, while shared advertisements are viewed by users of social media at a rate of 75.8%, and 59.2% of users offer the products in advertising to other users. In Table 1, the increase in the number of internet users from January 2015 to January 2016 and the increase in the number of social media usage and the number of mobile device users and the increase in

✉ Oğuzhan Çıtlak
oguzhan.citlak@gazi.edu.tr

[1] Department of Computer Engineering, Faculty of Technology, Gazi University, Teknikokullar, Ankara, Turkey

Table 1 Rate of increase in active social media users (http://www.webcitation.org/78VP5lDmx)

# Internet users (Million)	# Active social media users (Million)	# Number of mobile device users (Million)	# Active mobile social media users (Million)
3419	2307	3790	1968
Increase compared to the previous year		Increase compared to the previous year	
332	219	141	283
9.71%	9.49%	3.72%	9.29%

Table 2 Average monthly time spent on social media (Sites 2019)

Social media	Average monthly time spent (days)	Average daily time spent (hours)	# Pages viewed per day per visitor
Facebook	15	14.41	5.48
Instagram	11	5.44	3.80
Twitter	7.5	3.86	3.86
Google+	3.2	8.55	8.82

the usage rate of social media are shown (http://www.webcitation.org/78VP5lDmx).

In addition, when the most frequently used social networking sites and numbers of monthly visitors are compared according to data on eBizMBA Web site dated (Sites 2019), Facebook is ranked first with 1500 million monthly unique users, YouTube second with 1499 million, Twitter third with 400 million, Instagram fourth with 275 million and LinkedIn fifth with 250 million unique users (Sites 2019). Table 2 shows how high the social media usage is, the monthly average number of days and how many hours a day is spent by individuals on a social media platform (Sites 2019).

The rapid growth of social networks has led to a dramatic increase in the number and spread of malware (Kabakus and Kara 2017). In a virtual environment with such a heavy use area, there are some responsibilities and attention to be taken by the user. Some security factors that all Internet users should pay attention to are valid for social networks as well. It also introduces some security vulnerabilities in social networks that, unlike a Web site, are caused by the highest level of instantaneous interaction: for example, the facts that users do not pay attention to privacy principles, that they cannot control and manage their own accounts completely and most importantly that they make themselves a target by easily sharing their personal information in these media (Yavanoğlu et al. 2012). Spam can be defined as attacks over the Internet and affects the safety of social accounts on social networks such as Twitter. Spam, which allows malicious people to interact easily and quickly, has provided an opportunity that includes nonobjective, misleading, harmful and negative elements on Twitter (Şahinaslan et al. 2010). Those who manage malicious software on social networks use special programs (BOT) that act as human beings to leak user's personal information and to spread false information. BOTs can affect the users by sending friend requests, sending messages, and transmitting false information on their Web sites, which can be done very quickly and automatically (Watts and Dodds 2007). The action of computers, behaving like humans, is first pointed out by Reeves and Nass in their work in 1996 (Reeves and Nass 1996). It was recognized that these BOTs exhibiting human behaviors had been perceived as trustworthy, attractive and competent by the users (Edwards et al. 2014). These BOTs run at random and communicate with those who accept their requests. The BOTs using these methods often exhibit social engineering by abusing the emotional states and weaknesses of real users and thus reach their goals many times over and over (Mateen et al. 2017). Social network services have to consider the risks they may encounter for the security of user data, but the rapidly increasing spams users in recent years seriously threaten social network services. The best measure against these threats, which are abundantly available on the Internet and use plenty of social media, is to know in what ways spammers are threatening users and take personal precautions against them (Yıldırım and Varol 2013).

In this study, the focus was targeted on the detection methods of spams that are commonly used in the social media platform Twitter. It focuses on popular social networks and explains where it comes from. In the second part, the studies conducted in the literature are examined. In addition, information about the most used spam detection methods is given and a detailed explanation is made about what should be taken into consideration in the detection of spam. In the third part, the advantages and disadvantages of these studies, which are examined in detail, are compared with each other. In the fourth chapter, the reader is given advice and information to shed light on studies that can be carried out in the future.

2 Twitter spam detection methods available in the literature

There are some security requirements that must be present in communication using Internet social networking services. These are:

- *Confidentiality*: It means that information is hidden from third parties. Essentially, confidentiality is to provide access to personal and sensitive information by the right people while preventing access by the wrong people.

- *Integrity*: It means that the information owned or shared is not changed by third parties and that integrity is not corrupted (Timm and Perez 2010).
- *Eligibility*: Providing access to the information for the required and reliable people.
- *User communication confidentiality*: There also have to be some security measures present provided by social networking services. One of these is the fact that the user's information is not available to network operators. Depending on the user's privacy requirements, the IP address, messages and profile information that he/she wishes to be hidden should not be accessed by network operators so that user communication confidentiality is maintained.

The legitimate Internet services, such as Twitter, are misused by malware network traffic in the Internet, according to Cisco's "annual cybersecurity report" for the year 2018. The behavior of real account in spam behavior on Twitter makes the detection impossible. This malware traffic becomes impossible for security teams to identify owing to the behavior of legitimate network traffic (Cisco 2018). Spams are the most important threats to users in social media networks, and many researches have been conducted to detect spam profiles (Verma et al. 2013). The academic studies on how to detect these threats are categorized and presented in this study. When these methods are classified, the most recent and most used methods are taken into consideration. The following list is the most commonly used spam detection methods on Twitter.

Anomaly detection model does not only build the profiles of normal behavior very precisely, but also uses the known attack information indirectly on Twitter. Link analysis approach analyzes whether there is a malicious link on Tweet or not. Comparison and contrasting approaches are a model that compares the behavior of Twitter users. Deceptive information detection method detects fake announcements or ads on Twitter streaming. In addition, trend-topics analysis method controls hashtags on Twitter. Following and follower comparison method compares the numbers of followers of Twitter user accounts. Ensemble learning method extracts a common algorithm from a few different algorithms to detect spam behaviors on Twitter. Moreover, account creation time-based method takes time criterion that Twitter users create on time of their accounts. Short message analysis method refers to direct messages on Twitter. Honeypot-based Twitter spam detection method attracts the attention of spam on Twitter. Lastly, methods for using spammer detection tools uses of external software to find spam accounts.

These approaches listed have different methods. They are frequently mentioned in the literature when spam research is made on Twitter. Users interacting on the social network create a new identity in a virtual environment and communicate with each other, share and create social relationships with friends and improve them. Social networks are frequently visited throughout the world, where people of all ages and especially young users spend most of their time (Palfrey and Gasser 2008; Sevli and Küçüksille 2016).

2.1 Anomaly detection method

Anomaly means to behave differently from normal behavior. Behavior is considered dangerous if it differs from normal (Bhuyan et al. 2012). Behaviors of a normal user on Twitter are similar to statistically predictable mathematical values. What is important here is that normal behavior is known and the abnormal behavior can be distinguished from the normal one. Behavior is observed on the basis of the pattern we have obtained because of normal behavior, and if there is an anomaly, this behavior is detected compared to normal behavior and distinguished (Liu et al. 2012). The advantage of the anomaly detection method is the possibility of discovering previously unknown spam actions. The biggest problem in the anomaly detection method is the high number of false alarms (negative/positive alarm) in spam detection.

Spam-generating accounts on Twitter display anomaly behavior on social networks, in follow-up groups of friends, or in popular titles created. Dini et al., in a study conducted in 2012 (Dini et al. 2012), developed an effective anomaly detection system for spam threats to mobile Android users. In this system, using the machine learning methods, the core level and the user level of the Android mobile OS system are monitored and the spam is distinguished from the normal behavior. Pursuant to the results they obtained, they put forward the success of this method they had proposed. The anomaly spam detection method is based on the behavioral techniques exhibited by the Twitter user. The content and number of messages sent by the user, the number of likes received, the comments made on the message and the evaluations of the time spent here are evaluated. In the anomaly detection method, the basic characteristics of the users are examined, and their normal or abnormal behavior is checked and this is used in determining the friendship relations between different users in social networks only. It tries to identify the next behavior by testing some groups and networks that the user participates in it. Social networks can capture a variety of relationships between participants. For example, a network formed by family members and behaviors they exhibit are important characteristics. For this reason, creating a behavioral prediction by overlapping the connections and relationships among users, friendship and family ties within the social network will be a very useful study in every aspect. In a study that takes account of all this (Zheleva et al. 2008), it appears that the relationships between friendships and family ties in social networks

have led to a 15–30% higher predictability than traditional features.

Egele et al., in a study conducted in 2013 (Egele et al. 2013), developed an anomaly spam detection tool called COMPA. When the COMPA tool was created, a dataset of over 1.4 million open Twitter messages and more than 106 million Facebook messages was used. The purpose of this tool was to detect abnormal behavior and establish a statistical model and to identify social network accounts with sudden behavior change. Datasets on Twitter and Facebook are tagged with the SMO of Weka software (Holmes et al. 1994).

In Table 3, behavioral profiles of randomly selected individual users were generated using the COMPA tool tagged in Weka. These behavior profiles created were sent from individual accounts and popular mass applications. Ones that were harmful were tried to be identified using these behavior profiles created. These values are given in Table 3, in text and URL basis. Abnormal behaviors can be also observed on network traffic. The abnormal behaviors in real-time network traffic are tried to be determined by establishing expensive systems. In this way, harmful behaviors are detected and attempted to be prevented (Yılmaz and Gönen 2018). Abnormal behavior similar to this can occur on Twitter social network. Therefore, detection of abnormal behavior is important wherever it is.

2.2 Link analysis approach

Today, Web sites are exposed to a variety of attacks by malicious people using security vulnerabilities. A Web site seized in this way is used as a "zombie" for other attacks and is served for the evil intentions of the attackers. For instance, URL redirection mechanisms are widely used as a way to secretly execute Web-based attacks; that is, an attacker is able automatically to redirect a visitor to a malicious software distribution site by adding a redirecting code to a captured Web site. Although many defensive mechanisms have been developed against malicious Web sites that work in this manner, one can still encounter a large number of active malicious Web sites today. A honeypot-based tracking system is being developed by monitoring malicious URL redirects. Akiyama et al. (2017) recommend such a system in their study; A URL tracking system they developed was brought into life and gathered data for 4 years. During the course of this time, more than 100,000 malicious redirecting URLs from 776 different Web sites were collected. The results of these collected data can be summarized as follows:

1. URL redirection is used by the vast majority of attackers who commit fraud by means of clicks.
2. Using domain genetic algorithms (DGA), these URLs are blacklisted to prevent redirecting URLs.
3. The simultaneous use of domain and IP addresses of the sites in the Web routing chain indicates the robustness of Web redirection.

Based on these results, the most practical measure against malicious URL redirects is as follows. It is the removal of the security or network operator's useful information obtained from the honeypot-based monitoring system from reachability. Thus, the infrastructure of the Web-based attacks is deteriorated and prevented. In addition, tracking identities of Web advertising information are collected and used to identify and prevent attackers. Millions of Twitter users on the planet, through real-time search systems and different types of data mining tools, are able to track echoes of events and news on Twitter. Nevertheless, news spreading in a short period of time and social networks that allow instant status notifications pave the way for some suitable media for new spam types. For instance, the most talked-about (trend topic) items on Twitter are seen as opportunities for traffic and revenue generation. Spammers shorten the tweets that contain typical words of a trending topic to resemble a URL and direct users to Web sites that are not related to each other. If there are no blocking or security precautions against such spammers created by the user, they will inadvertently contribute to reducing the efficiency of real-time search services. Benevenuto et al. (2010) in their work on this subject address the ways in which spammers can be detected on Twitter. The dataset is a collection of more than 54 million users, 1.9 billion links and a large Twitter database of about 1.8 billion tweets. Three famous trend-topic tweets from 2009 were used, and a large collection of tagged users was created, classified as spam and non-spam. Later, they created a number of features from this collection. These features were used as attributes of machine learning. Attributes have taken an important role to detect spam senders, tweet content

Table 3 Evaluation results, Twitter text and URL (Zheleva et al. 2008)

Network and similar features	Twitter text (Twitter message)		Twitter URL (link shared on Twitter)	
	Groups	Accounts	Groups	Accounts
#Groups and accounts with hazard detected	9362	343,229	1236	54,907
#Popular group application with hazard detected	1647	178,557	251	8254
#Client applications with hazard detected	7715	164,672	985	46,653

that can potentially be used for and to produce results related to user social behavior.

Link sharing on Twitter may sometimes like as in Fig. 1. Spam user has identified one hashtag #musicmonday. Then, a message is shared to attract the attention of users in hashtag. Spam user with this message in Fig. 1 shares one malicious Web site link. Other users do not know this link is malicious. Users who believe in the message click on it. In addition, it may sometime be a short link format. It does not matter which one. Short link analysis approaches look at whether these links are malicious at first. In this approach, there should be a link in tweet no matter what else. It is intended for users to click on the links. While most of the spam senders were managed to be identified with the methods they suggested, only small ratio of non-spam users were misclassified. Approximately 70% of spam senders and 96% of non-spam senders were correctly classified. In Fig. 1 (Benevenuto et al. 2010), there is a spam tweet example, which appears to be an advertisement but contains a malicious link, sent to a popular hashtag, which was used in the study.

2.3 Comparison and contrasting approaches

This approach, which we can call "Comparison Contrast," is the analysis of real and non-real users with the classification method used in machine learning system (support vector machine SVM). Comparison of the messages sent from robotic (BOT) accounts and the real messages is an effective method for spam detection. In a study conducted in 2015 (Fernandes et al. 2015), using this approach, similar F1 accuracy scores of 90% (F1 score is the harmonic mean of precision) were obtained. However, there were some issues in classifying abnormal behaviors exhibited by real users. In order to prevent this, another classification method was implemented and F1 accuracy with a mean of 74% was obtained after collecting information on the brands real users carried, on whether they were famous or not and on promotional and private information. These accuracies were obtained by reducing the size of the feature field using categorical balance resulting from the gradual feature selection and individual checking of category results.

Clark et al. (2016) used some typical features (time between tweets, number of followers, etc.) to define existing detection algorithms and robotic accounts in the method they developed. Here, they have introduced a powerful classification scheme that uses the natural language structure of real users with a criterion for defining accounts that send automatic messages. This scheme, as it only works on text, is flexible, and it could be applied to other social media services that contain any text, not just Twitter. In another study (Wu et al. 2016), the use of "microblogging" method is recommended in detection of social media users and spam messages all together. In this approach, the relationship between users and messages is examined by combining spam sender identification and spam message identification in the social media. In addition, links between social relationships among users and messages are derived to refine the identification results. In addition, an efficient algorithm of this method is derived and an accelerated method is proposed in which how to take steps in the shortest time and at the most and how to make a success of this are explained. Extensive experiments conducted in a "microblog" dataset in the real world have shown that the proposed approach can both successfully and efficiently detect social spam senders and detect spam messages.

In terms of number of users, the most common social networks used for different purposes on the Internet are Facebook, Twitter, Myspace, LinkedIn, Google+ and Instagram. Day by day, the use of social networks in our country is constantly increasing. According to the Survey on Household Use of Information Technologies conducted by the Turkish Statistical Institute (TUIK) for January–March 2018, 82.4% of the users between the ages of 16–74 who social media users in every platform can access the Internet have shared contents such as sending a message, sharing a link, writing on a comment, and so on (http://www.webcitation.org/78VOZMhpT).

2.4 Deceptive information detection method

Another sample of a widely used spam in social networks is deceptive spams. These spammers generally spread deceptive misinformation and content. Users are redirected to malicious sites or addressed through fake messages (Fig. 2) which allure the users, are attractive and apparently contain no harmful elements.

Spam detection is carried out by means of regional analyses of the responses to these deceptive messages, which sites they are being redirected and what type of information is

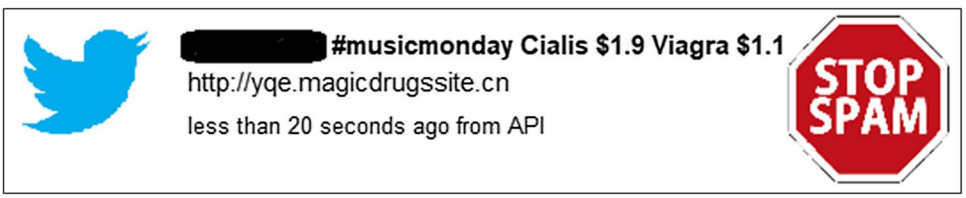

Fig. 1 A spam sample on Twitter for the hashtag #musicmonday

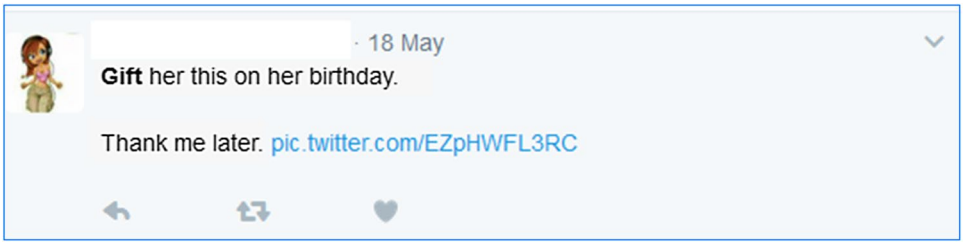

Fig. 2 A sample of deceptive and false information spam

requested from the user. Fake information sent to the user is detected and analyzed, and spam address is thus reached (Chen et al. 2017). The user is misled with fake information. This method can be used with malicious links. Figure 2 is a good example. However, the users can be taken by lying to other deceptive profiles or hashtags without link sharing. Spam accounts are frequently used some words; these are free, follow me, bonus, gift, and so on, on Twitter.

2.5 Trend-topics analysis method

Many previous studies on spam detection on Twitter seem to focus on identifying malicious user accounts and Honeypot-based approaches. However, two new methods have begun to be used. These are isolation of spam detections without the user's knowledge and application of linguistic statistical analysis to detect spam in trend topics. It captures trend topics, emerging Internet trends and discussion topics, paying attention to specific structure of the sentences. Moreover, in this approach proposed by Martinez-Romo and Araujo (2013), they tried to detect spam tweets in real time using the language as the primary tool. For the experimental study, a large dataset with 34 thousand trend topics and 20 million tweets was collected. In addition to this set, spam has also produced a reduced table of certain features that have not been modified by senders. They have also developed a machine learning system with some orthogonal features that can be combined with other features to analyze the features exhibited by spam in social networks. In addition, they made a comprehensive assessment demonstrating that the established system was able to achieve successful performance on the same level as the most advanced technology systems based on the detection of spam accounts, according to the *F*-measure metric. Because of this assessment, this system has been shown to be useful for detecting spam in real-time trend topics by analyzing tweets instead of user accounts.

On the electronic information exchange system (EIES), developed by Freeman and his team at the New Jersey Institute of Technology in 1978, which is considered to be the ancestor of social networks, members could send e-mails between themselves, could create bulletin boards and could prepare task lists on their own. The first network that resembles today's social networks is Sixdegrees.com, which was launched in 1997 (Turoff 1978). The name "Sixdegrees" comes from a study carried out by sociologist Stanley Milgram in 1960 to determine the ways people communicate. On this site, users were able to create profiles and list friends. Sixdegrees.com has the feature of being the first site to combine them, although some features, such as profiling, are also available on other social network sites and virtual groups. This was the first acknowledged social networking service to be closed in 2000, despite serving millions of users (Ellison 2007).

2.6 Following and follower comparison method

Twitter, one of the fastest growing social networks in recent years, has also hosted many spammers. Many researchers have proposed spam detection methods to identify these suspicious users on Twitter. One of these methods is to analyze well the "follower" and "friend" relationships among the users. Based on Twitter's existing spam policy, new content-based features and graph-based features also facilitate spam detection.

In a study conducted on this subject (Wang 2010a, b), a new method has been developed with an API provided by Twitter and a Web browser. A total of 25 thousand users, 500 thousand tweets and 49 M followers and friend relations were collected from the publicly available data on Twitter. The Naive Bayesian classification algorithm is applied in the machine learning system to distinguish suspicious behaviors from normal ones. The dataset was analyzed, and the performance of the detection system was compared with traditional assessment metrics and various classification methods. The results show that the "*F*-measure" values of the Naive Bayesian classification algorithm have the best overall performance. When the entire trained set is tested, the result shows that the spam detection system can achieve 89% precision.

Jeong et al. (2016) address ways to identify spam on Twitter. To identify these malicious messages, Spam senders are kept track of and at the same time, a legitimate user as well is kept track of. Thus, they proposed a classification scheme based on comparison and analysis of the active relationships of the spam senders with the active relationships of the legitimate users. For these features, it was "focused on cascading social relationships" and two plans were developed: TSP filtering and SS filtering, each using a triple significance

profile (TSP) and social status (SS) in a two-staged central subnet. In addition, they have proposed a cascading filtering method that combines both TSP and SS features and is an "ensemble method." True Twitter datasets were used in their study, and the three approaches suggested in the experimental studies were found to be very easy to use. The advantage of this method is that the proposed schemas are scalable and that rather than analyzing the entire network, it examines user-oriented "two-hop" social networks.

2.7 Ensemble learning method

Many new studies in the literature aimed at reducing user spam threats. In these studies, machine learning techniques are applied to classify Twitter spam and satisfactory results are obtained. In addition, this sort of classification removes the class imbalance in Twitter data. Ensemble learning methods are meta-algorithms that combine various machine learning techniques in a single estimation model to reduce variance, bias or increase estimates. An example, Weka, R, and YALE tools are different data mining methods. The desired ensemble learning model can be developed by selecting the algorithm from these machine learning techniques. Liu et al. (2017) proved in their study that the unbalanced distribution of spam and non-spam classes had a major influence on the spam detection rate. To solve this problem, they applied the "fuzzy-based information decomposition" algorithm. They proposed a fuzzy oversampling (FOS) method, which produces a synthetic dataset from limited observed samples. They also developed an ensemble learning approach that learned with a more accurate classifier than the data that seemed unstable in three stages. In this method, the class distribution in the unbalanced dataset was first set using "random oversampling," "random undersampling" and "FOS" together. Secondly, a classification model was built on each of the reclassified datasets. In the last stage, however, a majority voting system was developed to combine the results from all classification models. For purposes of evaluation, the results obtained from experiments on Twitter data show that the proposed learning approach can significantly increase the spam detection rate in spam and unstable data clusters.

2.8 Account creation time-based method

Spammers, who are active on the Internet, take advantage of numerous short-term malicious accounts to perform large-scale simple attacks such as spam distribution on social networks. However, conventional detection methods based on account or message information take too much time to collect such information before running detection algorithms, so Spammers try to keep their accounts running until they are suspended.

In their work, Lee and Kim (2014) proposed a new detection scheme to filter potentially malicious account groups in terms of their creation time. For this purpose, using similar algorithms, the differences between account names are created "algorithmically" and real account names are used to identify malicious accounts. For accounts that were created in a short time, they implemented a separate classification algorithm to classify group accounts and malicious account clusters sharing similar username properties. As a dataset in their work, they used 4.7 million user accounts collected from Twitter. The generated scheme only achieves an acceptable accuracy value, although it is based on user account names and creation times. This method can be used as a quick filter to perform a detailed analysis against malicious account groups.

2.9 Short message analysis method

In social networks, the presence of a large number of mass messages is one of the most common situations encountered. Although these unacknowledged mass messages can be effectively distinguished by existing spam filters, they mislead the spam filters and continue to function by changing message instances. However, the weak aspects of available spam filtering techniques for short messages (SMS) have not been investigated thoroughly. Unlike other spam applications, there are only a few keywords available in text message applications, and the character length usually has an upper limit. In this case, the existing contrast learning algorithms may not work effectively in short-message spam filtering. Users can send messages via Twitter. They can start a private chat or create a group chat with everyone who follows. Moreover, everyone in the chat can send short messages in the group created before. Even if people do not follow each other, everyone in the group can see all messages. Some accounts on Twitter, especially businesses, have enabled the direct message-receiving setting from everyone on Twitter.

In a study conducted in 2015 in this subject (Chan et al. 2015), short message spam filtering, a good word attack and a counter-attack method were searched to see how efficiently a lengthy message could work efficiently and how closely their relations with each other were. In this study conducted, a good word attack strategy that maximizes the effect of a classifier based on weight values and the length of words is proposed. On the other hand, a new scaling function was also presented that minimizes the significance of a feature that represents a short word, requiring a reassessed character and increased number of characters for successful avoidance. The success of the method was evaluated using a dataset consisting of SMS and comment spams. The results confirm that short-word spam filtering is a critical factor in the robustness against word attack. The rapid growth of

Twitter has led to a dramatic increase in spam capacity and complexity. Spam senders effectively exploit certain Twitter components, such as "hashtags," "mentions" and abbreviated URLs. However, similar features appear as an important factor in determining new spam accounts, as shown in previous studies. Miller et al. (2014) first stated that previous studies had regarded spam detection as a classification problem, but that they regarded it as an anomaly detection problem. Secondly, they have identified the characteristics of user information and tweet texts analyzed in previous studies as dataset. Finally, they used two flow clustering algorithms, "StreamKM++" and "DenStream," to make it easier to identify spam by effectively using the flow characteristics of tweets. Both algorithms group regular Twitter users and use the anomalous names as spam senders. These algorithms demonstrate good performance when tested separately. With StreamKM++, 99% recall and 6.4% false positive rate, and with DenStream, 99% recall and 2.8% false positive rate results were obtained. When these algorithms were used together, it was found that the system detected only 2.2% of normal users incorrectly, whereas it could correctly detect 100% of spam senders.

2.10 Honeypot-based Twitter spam detection method

Honeypot methods are a practical and easy method to use in spam detection and especially in data collection. Dagon et al. (2004) in their study conducted in 2004 remarked that social networks would be followed on a global scale and early detection of spams would be possible. In this study, they created a "honeypot network" (HoneyStat) that used modified honeypots to reach high detection rates and to create a correct warning mechanism. Unlike traditional interactive honeypots, its advantage involves directing the HoneyStat nodes through a script and covers a large user network. HoneyStat nodes generate three different warning categories:

1. Memory warnings (buffer overflow detection and process management based)
2. Disk write warnings (such as writing in registry keys and critical files)
3. Network warnings

With these nodes, data collection is automated and the timing of the previous traffic can be analyzed when the node is given a warning. With a log maintained, the situation describing the previous network activity is determined. The result shows whether the user has an automatic or worm attack. In this study, it was shown that building HoneyStat is more advanced than previous malware detection techniques. First, it demonstrates how to detect "zero-day worms" (emerging malware) using tracking files from malicious attacks on small networks. Secondly, it shows how multiple malwares are detected at the ports of attack. In addition, warnings from HoneyStat can be used with traditional information collected, such as attack information and rates. Yang et al. (2014) in their work, they proposed that, spam senders' likes (unwanted spam targets), to create new ways to create more effective social honeypots and defend them against social spam senders, and set some "criteria" for these ways. Spam senders create exciting honeypots with various social behavior patterns to entrap. After a 5-month data collection phase, a detailed analysis of how Twitter spammers find their goals was conducted. According to the results of the analysis, what needs to be done to create an advanced and effective social honeypot has been considered. In particular, these advanced honeypots, used in the same time period, are about 26 times faster than "traditional" honeypots in determining spammers. In the second part of this study, a new data collection approach of honeypots that attracted spammers was examined. The goal here is to develop a strategy for effective screening and sampling prioritization (for later spammer analyses) instead of scanning all Twitter accounts to get the possible samples, taking limited resources and limited time into consideration. Two new, effective and at the same time valid sampling approaches have been created by collecting data on the pleasure of spam accounts found in the extensive Twitter network.

2.11 Methods for using spammer detection tools

Malicious accounts and messages threatening users in social networks can be detected in many ways, and some of the external software is used to detect spam users and messages. Some of these software programs are given below.

- Integro: This software is a scalable defense system. It tries to detect fake Twitter accounts by using an user-ranking scheme. It starts by estimating victim accounts from user-level activities on Twitter. Then, it integrates these estimates into weights. At last, it ranks user accounts and compares them to the known real account. Integro warrants that most real accounts rank higher than fake accounts on Twitter. Low-ranking fake accounts can easily be detected by so (Boshmaf et al. 2015).
- SybilRank: It is deployed for Tuenti, the largest online social network in Spain. It can be also used for detecting spam accounts on Twitter like Tuenti. Both social media platforms have almost same characteristic property to use SybilRank. It analyzes social friendship graph in Tuenti. It tests the social connection among users. Fake accounts show various behaviors according to SybilRank (Cao et al. 2012).
- NodeXL, (Network Overview, Discovery and Exploration for Excel). It is a social network analysis tool for

Twitter. The network visualization is created by using it. Trusted accounts can be understood when the visual network is analyzed. Fake accounts have anomaly relationship between social users (Hansen et al. 2010).

- Pajek: It is a noncommercial tool for analysis and visualization of large social networks. It can work real time for Twitter dataset and can create a visualization to see connections between social users (Wang 2010a, b).
- ReDites: It brings social data together with monitoring, tracking and visualization into a one system of situation awareness. It works real time and interprets events in social media (Osborne et al. 2014).
- Canary honeypots: It is a system that mimics a production system and serves as an early detection mechanism for network. It is placed inside the network and mimics existing systems and details alerting. These honeypots may not seem to be benefit for Twitter spam detection. However, it can compare to a compromised system sing legitimate credentials when malicious users try to log in a network system with their Twitter accounts (Sanders and Smith 2013).

Out of these programs, Integro (Boshmaf et al. 2015) software compares the validity/authenticity of real accounts according to fake accounts and scores fake/spam accounts according to these scores.

In this section, a large literature review is completed. In addition, spam detection methods in social networks frequently are mentioned in the literature. Moreover, more information is given about the most commonly used spam detection methods on social networks. In the third section, these methods are compared in detail. The advantages and disadvantages of these methods are discussed.

3 Comparison of spam detection methods

In this study, the studies in the literature and the ones conducted in this field are compared with each other to show the ways in which the spam threatens the personal data of the users and which methods they do it and the results obtained are examined in detail. The focus is the differences among the current studies in the literature for spam detection in social networks. These studies and characteristics are seen in Table 4.

In Table 4, the techniques, algorithms, datasets, evaluation metrics and used methods are mentioned. The methods of detecting spam in malicious software spreading on social networks are examined. In the studies examined, which methods are used and what results are obtained are revealed. According to the comparisons, it is understood that different methods are used to detect spam accounts in social networks. The vast majority of studies are based on machine learning methods used.

In our study, we have examined the detection methods of spams, which rapidly spread malicious software on social networks. In the studies examined, which methods are used and what results are obtained are revealed. In this study, spam account detection methods in the literature are analyzed. These methods have had high success in their time when the examples in the literature are examined. Table 5 shows accuracy rates with high scores methods mentioned. It's focused on the accuracy results obtained rather than the technique and linguistics in this study. These results in very close results (Miller et al. 2014), using the machine learning-based and flow-based classification algorithms, which have the highest success from the studies. It is observed that the study conducted in Table 5. It is seen in Table 5 that the results obtained are very close to each other and that seven studies have very high values, over 90% in particular. It is seen, out of these studies with very close scores, that the study conducted by Miller et al. (2014) using the machine-based and flow-based classification algorithms is the one that reached the highest success rate—though it is not by much. The most prominent and important feature of this study is the combination of two classification algorithms. Again, it has been confirmed that the studies of Clark and Chen, who obtained a score of more than 95%, are very similar to those of Z. Millerin and that they also use a method based on machine learning and the only difference is that they used "standard classification algorithms." It is also seen that datasets consisting mainly of tweet messages are used in these studies (Table 6).

First, it has been observed that the vast majority of the studies addressed machine learning-based methods. In addition, there are some studies, albeit it is little, where external software is used. The only exception is seen in the study of Akiyama, and it only works with URL information as a dataset instead of messages. With the system and the algorithm he used, he produced a very different study from other studies and achieved a satisfactory success with the result. In fact, this is the most prevailing indication that we will be able to achieve higher results with more datasets and changing algorithms in the coming years. Among the studies examined, only two studies achieve values below 80%. It is seen that the study that achieved the lowest result out of the studies that are below 80% was realized with external software. Therefore, it is not recommended to use external software in such studies. It seems that the most effective system to use in social networks that have a very heavy traffic with millions of users is the system based on machine learning. It is also clear that the results and success rates of the studies vary depending on the algorithms and datasets used. In the methods in which machine learning methods are used, the fact that the dataset in the training part of

Table 4 Table of characteristics of spam detection studies in the literature

Article	Technique	Algorithm	Dataset	Evaluation metric	Methods
Akiyama et al. (2017)	Monitoring system	Domain generation algorithm (DGA)	Injected with redirect codes over URL dataset	Performance ratio	Link analysis approach, Honeypot-based Twitter spam detection method
Fernandes et al. (2015)	Machine learning system	Classification and clustering algorithms	Twitter dataset	*F*-score	Anomaly detection method
Clark et al. (2016)	Machine learning system	Traditional classification algorithms	Twitter Bot dataset	ROC-AUC	Short message analysis method, deceptive information detection method
Wu et al. (2016)	Machine learning system	Algorithm based on ADMM	Real Microblog dataset	Parameter λ	Trend-topics analysis method, short message analysis method
Chen et al. (2017)	Machine learning system	Graphical-based algorithm	Twitter and URL datasets	True positive	Deceptive information detection method
Martinez-Romo and Araujo (2013)	Machine learning system	Traditional classification algorithms	Twitter datasets	*F*-measure	Trend-topics analysis method
Jeong et al. (2016)	Machine learning system	TSP-SS filtering cascaded filtering	Real Twitter datasets	True positive	Following and follower comparison method, ensemble learning method
Liu et al. (2017)	Machine learning system	ROS, RUS, FUS algorithms	Tweets and URL dataset	*F*-measure	Ensemble learning method
Lee and Kim (2014)	Machine learning system	Creation and SVM algorithms	User account dataset	FNR	Account creation time-based method
Miller et al. (2014)	Machine learning system	Den Stream and Stream KM++	Real Twitter datasets	*F*-measure, recall	Ensemble learning method
Boshmaf et al. (2015)	External software	Indigo-RF algorithm	User account dataset	ROC	Methods for using spammer detection tools
Wang (2010a, b)	Machine learning system	Naive Bayes algorithms	Real Twitter datasets	*F*-measure	Comparison and contrasting approaches

Table 5 Table of characteristics of spam detection studies in the literature

Studies	Accuracy rates
Akiyama et al. (2017)	96.50
Fernandes et al. (2015)	90.00
Clark et al. (2016)	95.21
Wu et al. (2016)	93.00
Chen et al. (2017)	95.00
Martinez-Romo and Araujo (2013)	94.50
Wang (2010a, b)	89.00
Jeong et al. (2016)	96.30
Liu et al. (2017)	78.00
Lee and Kim (2014)	86.53
Miller et al. (2014)	97.10
Boshmaf et al. (2015)	76.00

the system is very rich and various provides more accurate results to be obtained on the dataset that the system tests. In the studies that were yet examined, real Twitter datasets, URL information and profile information in social media were used for spam detection. This information shows that the studies on which the actual Twitter datasets are used have achieved higher success rates. In studies conducted by the classification method, it is seen that accounts with spam messages and some messages belonging to real users who do not normally produce spam are also regarded as spam. However, since there is a very small error in the negligible level, the classification method is still the most effective and most important method.

4 Conclusion

This article explores the ways to detect spam and spammers on Twitter. In Twitter, a comparative methodology of studies in the literature approaching the identification of spam

Table 6 Some explanations of evaluation metrics

	Metrics	Explanation
1	Correctly classified instance	Accurate classification
2	Incorrectly classified instance	Incorrectly classified instance
3	True positive (TP)	True
4	False positive (FT)	False
5	Precision	Precision = TP/(TP + FP)
6	Recall	Recall = TP/(TP + FN)
7	*F*-measure	$F = 2DK/(D + K)$ K is precision and D is recall
8	Receiver operating characteristic (ROC)	Gives an idea of how the ROC curves generally perform only for classifiers
9	Precision recall (PRC)	The PRC is only concerned with how the classification behaves in a class. For example, it is successful in classifying patients as diseased or healthy.
10	Accuracy error rate	Accuracy = (TP + TN)/(TP + FP + FN + TN)

messages and accounts from different aspects is presented. When these studies are examined, the methods, algorithms, datasets and evaluation metrics they use are taken into consideration. Given the fact that it will shed light on to the studies to be conducted on growing social media usage in the future. These studies show that the datasets used in the detection of spam are adequate and varied and that they are included in the machine learning system and evaluated with an appropriate algorithm, and produce very effective and accurate results. This, too, increases the validity and reliability of the study conducted. In the literature part of this study, the studies examined have advantages over each other and the disadvantages, the accuracy values they obtained, the metrics they use, the algorithm and techniques are discussed in detail. Also measuring metrics, advantages and the disadvantages are compared to the accuracy values obtained from the algorithm, technical details. The advantages and disadvantages of the studies investigated against each other, the accuracy values obtained, the measurement metrics they use, and the algorithms and the techniques are examined in detail. The number of spammers on Twitter increases day to day. Spams and malicious messages do threat the security of Twitter users as well as performance. Another very useful and effective method is the method by which friendship relations are analyzed. This method is seen as a more realistic structure according to the spam user detection. Social media users should be suspicious of spam accounts on Twitter. There are innocent users with good intentions as well as malicious users on social media. Twitter users should review the Twitter spam policy before using this social platform (http://www.webcitation.org/78VPLwy11). It takes precautions such as blocking account, against spam accounts in every day. However, the most important precaution is to create conscious users again spams. All spam detection methods have important details in order to aware be of social media users. All of these studies and results have shown us that in our future studies, we plan to implement hybrid new methods, focusing on mobile devices, combining different classification methods. Besides, different and new forms of attacks that the ever-evolving and changing Internet technology will encounter should already take its place among research topics as well. It is extremely difficult for the Twitter users to be able to understand spam, using malicious methods, on Twitter. The most important condition to protect against spams and malicious messages threat is to know the harmful methods of spam. In this study, malicious spam ways to readers are shown. Moreover, prominent detection methods of spams are analyzed. How the real users and fake users are distinguished. In addition to this, weak and strong aspects of the methods for spam detection methods are compared and are evaluated. When spam targets to Twitter users, they can convey the malicious intentions without noticing. With our study, we found spam detection methods on Twitter and we aim to inform the users of social networks to be aware of these spams.

References

Akar E (2010) Sosyal medya pazarlaması: Sosyal webde pazarlama stratejileri. Efil Yayınevi

Akiyama M, Yagi T, Yada T, Mori T, Kadobayashi Y (2017) Analyzing the ecosystem of malicious URL redirection through longitudinal observation from honeypots. Comput Secur 69:155–173

Benevenuto F, Magno G, Rodrigues T, Almeida V (2010) Detecting spammers on Twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), Washington, pp 12–15

Bhuyan MH, Bhattacharyya DK, Kalita JK (2012) An effective unsupervised network anomaly detection method. In: Proceedings of the international conference on advances in computing, communications and informatics. ACM, pp 533–539

Boshmaf Y, Logothetis D, Siganos G, Lería J, Lorenzo J, Ripeanu M, Beznosov K (2015) Integro: leveraging victim prediction for robust fake account detection in OSNs. NDSS 15:8–11

Cao Q, Sirivianos M, Yang X, Pregueiro T (2012) Aiding the detection of fake accounts in large scale social online services. In: Proceedings of the 9th USENIX conference on networked systems design and implementation. USENIX Association, pp 15–16

Chan PP, Yang C, Yeung DS, Ng WW (2015) Spam filtering for short messages in adversarial environment. Neurocomputing 155:167–176

Chen C, Wen S, Zhang J, Xiang Y, Oliver J, Alelaiwi A, Hassan MM (2017) Investigating the deceptive information in Twitter spam. Future Gen Comput Syst 72:319–326

Cisco (2018) Annual cybersecurity report—download PDF-Cisco. http://www.webcitation.org/78VNO1Ova. Accessed 20 May 2019

Clark EM, Williams JR, Jones CA, Galbraith RA, Danforth CM, Dodds PS (2016) Sifting robotic from organic text: a natural language approach for detecting automation on Twitter. J Comput Sci 16:1–7

Dagon D, Qin X, Gu G, Lee W, Grizzard J, Levine J, Owen H (2004) Honeystat: local worm detection using honeypots. In: International workshop on recent advances in intrusion detection. Springer, Berlin, pp 39–58

Dini G, Martinelli F, Saracino A, Sgandurra D (2012) MADAM: a multi-level anomaly detector for android malware. In: International conference on mathematical methods, models, and architectures for computer network security. Springer, Berlin, pp 240–253

Edwards C, Edwards A, Spence PR, Shelton AK (2014) Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. Comput Hum Behav 33:372–376

Egele M, Stringhini G, Kruegel C, Vigna G (2013) Compa: detecting compromised accounts on social networks. In: The network and distributed system symposium, San Diego, CA, pp 1–8

Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput Med Commun 13(1):210–230

Erdoğan G, Bahtiyar Ş (2014) Sosyal Ağlarda Güvenlik. Akademik Bilişim Konferansı, Mersin, pp 1–6

Fernandes MA, Patel P, Marwala T (2015) Automated detection of human users in Twitter. Proc Comput Sci 53:224–231

Hacıefendioğlu Ş (2011) Reklam ortamı olarak sosyal paylaşım siteleri ve bir araştırma. Bilgi Ekonomisi ve Yönetimi Dergisi 6(1)

Hansen DL, Shneiderman B, Smith MA (2010) Analyzing social media networks with NodeXL: insights from a connected world. Morgan Kaufmann, Burlington

Holmes G, Donkin A, Witten IH (1994) Weka: a machine learning workbench. In Proceedings of the 1994 second Australian and New Zealand conference on intelligent information systems. IEEE, pp 357–361

Jeong S, Noh G, Oh H, Kim CK (2016) Follow spam detection based on cascaded social information. Inf Sci 369:481–499

Kabakus AT, Kara R (2017) A survey of spam detection methods on Twitter. Int J Adv Comput Sci Appl 8(3):5

Lee S, Kim J (2014) Early filtering of ephemeral malicious accounts on Twitter. Comput Commun 54:48–57

Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. ACM Trans Knowl Discov Data 6(1):3

Liu S, Wang Y, Zhang J, Chen C, Xiang Y (2017) Addressing the class imbalance problem in Twitter spam detection using ensemble learning. Comput Secur 69:35–49

Martinez-Romo J, Araujo L (2013) Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Syst Appl 40(8):2992–3000

Mateen M, Iqbal MA, Aleem M, Islam M A (2017) A hybrid approach for spam detection for Twitter. In: 14th International Bhurban conference on applied sciences and technology (IBCAST). IEEE, pp 446–471

Miller Z, Dickinson B, Deitrick W, Hu W, Wang AH (2014) Twitter spammer detection using data stream clustering. Inf Sci 260:64–73

Osborne M et al (2014) Real-time detection, tracking, and monitoring of automatically discovered events in social media. In: 52nd annual meeting of the association for computational linguistics: system demonstrations, Baltimore, MD, USA, pp 37–42

Öztürk MF, Talas M (2015) Sosyal medya ve eğitim etkileşimi. Zeitschrift für die Welt der Türken/Journal of World of Turks 7(1):101–120

Palfrey J, Gasser U (2008) Opening universities in a digital era. N Engl J High Educ 23(1):22–24

Peleja F, Santos J, Magalhães J (2014) Ranking linked-entities in a sentiment graph. In: Proceedings of the 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT). IEEE Computer Society, pp 118–125

Reeves B, Nass C (1996) Media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, New York

Şahinaslan Ö, Şahinaslan E, Borandağ E, Can E (2010) Güvenlik Tehdidi Oluşturan Spam Saldırılarına Karşı Önlemler

Sanders C, Smith J (2013) Using canary honeypots for detection. In: Applied network security monitoring: collection, detection, and analysis. Elsevier

Sevli O, Küçüksille EU (2016) Türkçe Paylaşım Yapan Kullanıcılar İçin Sosyal Ağ Tabanlı Analiz ve Tavsiye Sistemi. Akademik PlatformMühendislik ve Fen Bilimleri Dergisi 4(3)

Spam Policy Online Search on Twitter (2019) http://www.webcitation.org/78VPLwy11. Accessed 20 May 2019

Stephen AT, Galak J (2012) The effects of traditional and social earned media on sales: a study of a microlending marketplace. J Mark Res 49(5):624–639

Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceedings of the 26th annual computer security applications conference, USA, pp 1–9

Timm C, Perez R (2010) Seven deadliest social network attacks. Syngress

Top 15 Most Popular Social Networking Sites (2019). http://www.webcitation.org/78VNU1R4J. Accessed 20 May 2019

Türkiye İstatistik Kurumu Web sayfalarına Hoş Geldiniz (2019). http://www.webcitation.org/78VOZMhpT. Accessed 20 May 2019

Turoff M (1978) The EIES experience: electronic information exchange system. Bull Am Soc Inf Sci 4(5):9–10

Verma M, Divya D, Sofat S (2013) Techniques to detect spammers in Twitter-a survey. Int J Comput Appl. https://doi.org/10.5120/14877-3279

Wang AH (2010a) Don't follow me: spam detection in Twitter. In: Proceedings of the 2010 international conference on security and cryptography (SECRYPT). IEEE, pp 1–10

Wang AH (2010b) Machine learning for the detection of spam in Twitter networks. In: International conference on E-business and telecommunications. Springer, Berlin, pp 319–333

Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. J Consum Res 34(4):441–458

We Are Social Singapore (2019) Digital yearbook. http://www.webcitation.org/78VP5lDmx. Accessed 20 May 2019

Wu F, Shu J, Huang Y, Yuan Z (2016) Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. Neurocomputing 201:51–65

Yang C, Zhang J, Gu G (2014) A taste of tweets: reverse engineering Twitter spammers. In: Proceedings of the 30th annual computer security applications conference. ACM, pp 86–95

Yavanoğlu U, Sağıroğlu Ş, Çolak İ (2012) Sosyal ağlarda bilgi güvenliği tehditleri ve alınması gereken önlemler. Politeknik Dergisi 15(1):8–12

Yıldırım N, Varol A (2013) Sosyal ağlarda güvenlik: Bitlis Eren ve Fırat Üniversitelerinde gerçekleştirilen bir alan çalışması. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 6(1)

Yılmaz EN, Gönen S (2018) Attack detection/prevention system against cyber attack in industrial control systems. Comput Secur 77:94–105

Zheleva E, Getoor L, Golbeck J, Kuter U (2008) Using friendship ties and family circles for link prediction. In: International workshopon social network mining and analysis. Springer, Berlin, pp 97–113