A Survey of Spam Detection Methods on Twitter

Abdullah Talha Kabakus
Abant Izzet Baysal University
IT Center
Bolu, Turkey

Resul Kara
Duzce University
Faculty of Engineering,
Department of Computer Engineering
Duzce, Turkey

***Abstract*—Twitter is one of the most popular social media platforms that has 313 million monthly active users which post 500 million tweets per day. This popularity attracts the attention of spammers who use Twitter for their malicious aims such as phishing legitimate users or spreading malicious software and advertises through URLs shared within tweets, aggressively follow/unfollow legitimate users and hijack trending topics to attract their attention, propagating pornography. In August of 2014, Twitter revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates. Thus, detecting and filtering spammers from legitimate users are mandatory in order to provide a spam-free environment in Twitter. In this paper, features of Twitter spam detection presented with discussing their effectiveness. Also, Twitter spam detection methods are categorized and discussed with their pros and cons. The outdated features of Twitter which are commonly used by Twitter spam detection approaches are highlighted. Some new features of Twitter which, to the best of our knowledge, have not been mentioned by any other works are also presented.**

Keywords*—Twitter spam; spam detection; spam filtering; mobile security

I. INTRODUCTION

Twitter is one of the most popular social media platforms which provide a social network of users post messages up to 140 characters called as "tweet". Twitter lets users share their messages about everything related to the real life including news, events, celebrities, politics [1–5]. According to Twitter, Twitter has 313 million monthly active users that post 500 million tweets per day which equal 350,000 tweets per minute [6–8]. Thanks to this huge social network, users are able to stay connected with the topics they are interested in. Twitter provides a list of most talked topics at a given point in time called "Trending Topics (TT)" to let users be aware of most popular topics on Twitter. "Hashtag" is a term which starts with "#" character is commonly used to mention the topic of the tweet and let users track the topics they are interested in [9] . Thanks to its popularity and design, Twitter immediately reflects noteworthy events in real-time. This structure of Twitter lets real-time search systems and meme-tracking services mine real-time tweets to find out what is happening in the world with minimum delay [10,11]. Sentiment analyzing services are able to make a conclusion about topics in Twitter which turns Twitter into a real-time poll system [12–16]. The success of those services completely relies on filtering spammers from legitimate users. Consumers tend to use Twitter to learn ideas of others about the products they are going to buy. Similarly, companies use Twitter to measure the satisfaction of their customers for their products [17–21]. However, this popularity and practicalness also attract the attention of spammers. In April of 2014, Twitter was flooded by an avalanche of malicious tweets that were sent by thousands of compromised user accounts [22]. In August of 2014, Twitter revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates [23,24]. A report shows that 83% users of social networks have received at least one unwanted friend request or message [25]. Most common definition of spam is unsolicited one [26–28]. Spammers share links within their tweets in order to spread advertise to generate sales, propagate pornography, share malicious links which direct users to malicious software, hijack trending topics for their purposes, abuses reply or mention functions to post unsolicited messages to legitimate users to attract their attention, and phish legitimate users [1,21,28–37]. According to the report by statista, 80% of Twitter users access Twitter via their mobile devices [38]. Thus, users who access Twitter via their mobile devices should more care about spam than the users who access Twitter via web browsers since it may (1) collect excessive amount of personal information such as user location, call history, SMS, bank account details, calendar events, (2) access the data located in the device's memory or SD card, (3) send premium-rate SMS messages, (4) capture key-strokes by key logging, (5) make calls, and (6) detect user's location via Internet or GPS and share [39–45]. Another issue with users of social media is that according to the reports, users of social media do not show an adequate understanding of the threats of social media as much as they are on other platforms. Bilge et al. [46] report that 45% of users on a social media platform readily click on links posted by their "friends", even though they may not know that person in real life. Content-filtering approaches are not effective for Twitter since spammers tend to share shorten URLs in order to (1) overcome the character limitation defined by Twitter, and (2) manipulate spam filtering methods based on URL blacklisting [28,36,47–52]. The major contributions of this paper are given as follows:

- Features of Twitter which can be used to detect spam are presented with discussing their effectiveness,
- A comprehensive review of Twitter spam detection methods are discussed with considering their pros and cons in order to give a clear idea to the researchers who are interested in spam detection in Twitter,
- The new features of Twitter which, to the best of our knowledge, have not been used by any spam detection

approaches yet that can be used to detect spam are presented,

- The outdated features of Twitter which are commonly used by spam detection approaches in literature are presented.

The rest of the paper is structured as follows: Section 2 describes the background including features of Twitter and how Twitter deals with spam. Section 3 presents the features of Twitter spam detection. Section 4 presents the Twitter spam detection methods. Section 5 presents discussion. Finally, Section 6 concludes the paper.

II. BACKGROUND

In this section, features of Twitter and the way Twitter deals with spam are presented.

A. Features of Twitter

Twitter lets accounts to "follow" other accounts which they are interested in. Unlike other social media platforms, the relationship between users is bi-directional instead of unidirectional links which mean one user may not be following one of his followers. The user can "like" or "retweet (RT)" a tweet which means sharing that tweet with his "followers". The relationship between users in Twitter is presented in Fig. 1. Each user has a unique Twitter username, and users can post tweets that refer others by adding their usernames with starting "@" character which is called as "mention" on Twitter. Users are immediately informed with notifications when a mention, like, or RT happens to one of his tweets.

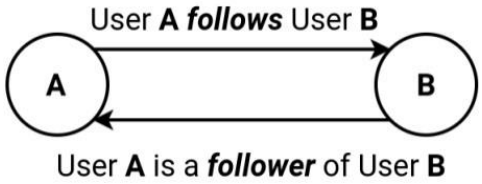

Fig. 1. The relationship between users in Twitter

Another feature of Twitter is letting users create user public or private lists in order to organize their interests by grouping users whose interests are same or similar [53–55]. Similarly, it is possible to manage lists by adding users to the lists or removing users from the lists which the user is the owner of. The lists the user subscribed are categorized as "subscribed to" while the lists the user is added by their owners are categorized as "member of" which are presented in Fig. 2.

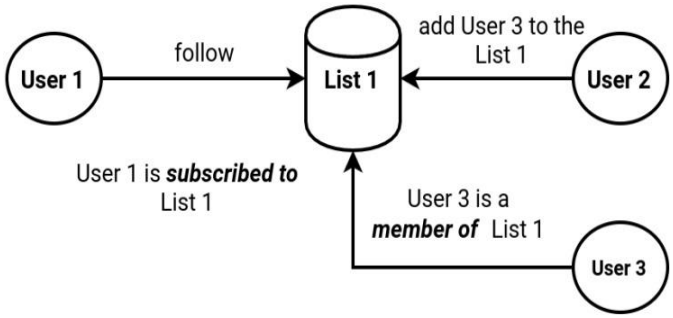

Fig. 2. The relationships between lists and users

B. How Twitter Deals with Spam

Twitter uses both manual and automated services to compete spammers in order to provide a spam-free environment. The manual way is that Twitter lets users report spammers through the spammers' profile pages. Twitter provides a user interface as it is presented in Fig. 3 to report the account by selecting the reason. Another way which is commonly reported in the literature is mentioning spammers to the official "@spam" account [28,29,37,56–58] but according to the recent report by Twitter, this method of reporting spam is outdated [30]. Also, Wang reports that this method is abused by both hoaxes and spam [29]. These manual approaches are labor-intensive and would not be enough to detect all spammers considering billions of users. Twitter uses various factors such as (1) posting duplicate messages over multiple accounts or multiple duplicate messages on one account, (2) following/unfollowing large number of accounts in a short time period, (3) having large number of spam complaints filed against the account, (4) aggressively liking, following, and retweeting, (5) posting malicious links, (6) posting tweets which mainly consist of links instead of also posting personal updates, and (7) posting unrelated tweets to a trending topic to determine what conduct is considered to be spamming [59].

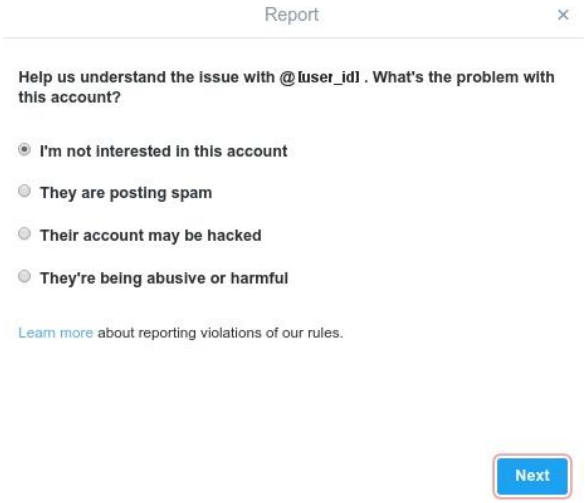

Fig. 3. The user interface of Twitter which is used to report an account by selecting the reason

III. FEATURES OF TWITTER SPAM DETECTION

The features of Twitter spam detection are categorized as follows: (1) Account-based features, (2) tweet-based features, and (3) relationship between the tweet's sender and receiver. These features are the mainframes of the features used by the related works in literature. Each feature category is discussed in the following subsections.

A. Account-based Features

Spammers can be detected by analyzing their Twitter accounts which contain the features listed in Table 1. Since some of these features such as biography, location, homepage, and creation date are user-controlled, they are useless in term of spam detection

TABLE. I. ACCOUNT-BASED SPAM DETECTION FEATURES

Feature	Description	Is User-controlled?
Username	The unique identifier of the account	Yes
Biography	The biography of the account	Yes
Profile photo	The profile photo of the account	Yes
Header photo	The header photo of the account which is displayed at the top of the profile	Yes
Theme color	The theme color choice of the account	Yes
Birth date	The birth date information of the account	Yes
Homepage	The website of the account	Yes
Location	The location of the account	Yes
Creation date	The date the account is created	Yes
Number of tweets	Total number of tweets the account has	No
Number of following	Total number of accounts the account follows	No
Number of followers	Total number of followers the account has	No
Number of likes	Total number of likes the account's tweets have	No
Number of retweets	Total number of retweets the account's tweets have	No
Number of lists	Total number of lists the account has	Yes
Number of moments	Total number of moments the account has	Yes

When the behaviors of spammers are analyzed within the scope of account-based features, these facts are observed:

- Since spammers tend to follow too many legitimate accounts in order to attract attention, the number of following is expected to be high compared to legitimate users.
- Since spammers are not followed by legitimate users, the number of followers is expected to be less compared to legitimate users.
- Since spammers' tweets are unsolicited, the number of likes and retweets for their tweets are expected to be less compared to legitimate users.
- Since spammers tend to post lots of tweets to attract the attention of legitimate users, the number of tweets sent by the account is expected to be high compared to legitimate users.
- Spammers' tweets mostly contain links and hashtags to attract the attention of legitimate users.
- Since spammers' tweets are ignored by legitimate users, the number of replies and mentions spammers get are expected to be low compared to legitimate users.
- Spammers tend to post same or similar tweets which are posted by one or more controlled accounts.
- Legitimate users tend to be added to the lists unlike spammers unless bots under the command and control (C&C) architecture add them to the lists they intentionally created in order to manipulate spam detection approaches.

B. Tweet-based Features

Spammers tend to post lots of unsolicited tweets to legitimate users to attract attention. Spammers can be detected by analyzing their tweets. This is necessary to filter spam tweets from legitimate ones and provide users a spam-free environment which is the aim of Twitter [60]. Each tweet contains the information listed in Table 2.

TABLE. II. TWEET-BASED SPAM DETECTION FEATURES

Feature	Description	Is User-controlled?
Sender	The sender of the tweet	Yes
Mentions	The mention(s) used in the tweet	Yes
Hashtags	The hashtag(s) used in the tweet	Yes
Link	The link used in the tweet	Yes
Number of likes	The number of likes the tweet has	No
Number of retweets	The number of retweets the tweet has	No
Number of replies	The number of replies the tweet has received	No
Sent date	The date tweet is sent	Yes
Location	The detected location of the place the tweet is posted	Yes

When the behaviors of spammers are analyzed within the scope of tweet-based features, these facts are observed:

- Spammers tend to use links to direct legitimate users to their malicious purposes.
- Spammers tend to use lots of mentions to attract the attention of more legitimate users.
- Spammers tend to use lots of hashtags (especially the trending ones) to reach more users.
- Since spammers' tweets are unsolicited, the number of likes and retweets their tweets have received are much lower compared to legitimate users.

C. Graph-based Features

Twitter is a network of users with relationships between them and tweets. This structure can be represented as a graph. For the graph model, users and tweet can be represented as nodes and relationships can be represented links between nodes. These relationships show how the tweet's sender and mentions are connected to each other. Also, these relationships are clear indicators of legitimate conversations. By constructing a graph model to represent users and their relationships, the distance between the tweet's sender and mentions can be calculated for spam analysis. Graph-based features are listed in Table 3.

TABLE. III. GRAPH-BASED FEATURES

Feature	Description	Is User-controlled?
Distance	The length of the shortest path between users	No
Connectivity	The strength of the connection	No

When the behaviors of spammers are analyzed within the scope of graph-based features, these facts are observed:

- The distance between a spammer and a legitimate user is further than the distance between two legitimate users.
- The connectivity between a spammer and a legitimate user is more robust than the connectivity between two legitimate users.
- Graph-based features provide the most robust performance to detect spam and spammers since they are hard to manipulate and not user-controlled.

IV. TWITTER SPAM DETECTION METHODS

In this section, Twitter spam detection methods in literature are presented and discussed. The proposed methods are categorized as follows: (1) Account-based spam detection methods, (2) tweet-based spam detection methods, (3) graph-based spam detection methods, and (4) hybrid spam detection methods.

A. Account-based Spam Detection Methods

Account-based spam detection methods are based on the features (or a combination of them) of Twitter account which are listed in Table 1. Lee et al. [61] propose a honeypot-based approach to detect spam in social media platforms. The features they consider detecting spam are the longevity of the account on Twitter, the average tweets per day, the ratio of the number of following and number of followers, the percentage of bi-directional friends, the ratio of the number of URLs in the 20 most recently posted tweets, the ratio of number of unique URLs in the 20 most recently posted tweets, the ratio of the number of usernames in the 20 most recently posted tweets, and the ratio of the number of unique usernames in the 20 most recently posted tweets. Lin and Huang [62] propose a method to detect spam in Twitter on the basis of two features: (1) URL rate which defines the ratio of the number of tweets with URL in the total number of tweets, and (2) interaction rate which defines the ratio of the number of tweets interacting over the total number of tweets. Gee and Hakson [58] propose a method based on account-based features such as followers-to-following ratio, the number of tweets to account lifetime ratio, the average time between posts, posting time variation, max idle hours, and link fraction. The limitation of this work is that they utilize the manual way of reporting spam in Twitter which is outdated as it is discussed before. Many Twitter spam detection methods use account-based features but alongside with other spam detection features in order to provide more robust spam detection methods which are called as "hybrid" spam detection methods in this paper.

B. Tweet-based Spam Detection Methods

Tweet-based spam detection methods are based on the features (or combinations of them) of a tweet which are listed in Table 2. URL filtering approaches use static or dynamic crawlers to investigate newly observed URLs. Also, they use URL or domain blacklisting in order to detect suspicious URLs from a knowledge base. These approaches use several features such as URL and DNS information, URL redirections, and the landing website's source code (HTML). McGrath and Gupta [47] present a phishing detection method based on lexical features of an URL. The features they consider detecting phishing are the length of URL and the domain name, the character composition of the domain name, the presence of brands in URLs, and misuse of URL-aliasing and free web hosting services. Ma et al. [63] propose a method to detect malicious websites by analyzing their URLs. The features they use detecting malicious websites contain WHOIS properties such as who is the registrar of the website, who is the registrant of the website, when the website is registered, domain name properties such as the time-to-live (TTL) value for DNS records, and geographic properties such as in which country does the IP address belong, the speed of the uplink connection alongside lexical features of URL. Prophiler [64] is a filter that uses static analysis techniques to detect the malicious content of a website. The features Prophiler considers are derived from (1) the HTML content of the website such as the number of elements with small area, the number of elements contain suspicious content, the number of included URLs, and the number of known malicious patterns, (2) the associated JavaScript code such as keywords-to-words ratio, the number of long strings presence of decoding routines, probability of shellcode presence, and the number of DOM-modifying function, and (3) the corresponding URL such as the number of suspicious URL patterns, presence of subdomains or IP addresses in URLs, and the TTL value for DNS A and NS record. Since Prophiler uses static analysis techniques, it is not able to detect malicious URLs embedded into dynamic content such as part of JavaScript which is currently the most commonly used programming language [65,66], Flash, and Java applets. Methods based on dynamic analysis techniques [67–70] use virtual machines and automated web browsers such as Selenium for in-depth content analysis. Chhabra et al. [49] present a phishing detection method based on URL analysis. Their method is specially designed to be able to analyze shortened URLs which are commonly used in Twitter to manipulate spam tweets as it is discussed before. The features the proposed method use detecting phishing through an URL are the number of clicks, geographical spread, temporal spread, and web popularity. WarningBird [71] is a suspicious URL detection system for Twitter which investigates correlations of URL redirect chains. WarningBird uses 14 features to detect suspicious URL such as the length of URL redirect, the number of different landing URLs, the relative number of different Twitter accounts, the similarity in the account creation dates, the similarity in the number of followers and following, the similarity in the follower-following ratio, and the similarity of tweets. Martinez-Romo

and Ajauro [72] propose a tweet-based spam detection method which focuses on the analysis of the language used in tweets. Specifically, the language models they use are (1) the language model of the tweets related to a trending topic, (2) the language model of the tweet, and (3) the language model of the page linked by the tweet. Similar to the account-based spam detection methods, many Twitter spam detection methods use tweet-based features alongside with other spam detection features in order to provide more robust spam detection.

C. Graph-based Spam Detection Methods

Graph-based spam detection methods are based on the features (or combinations of them) of a tweet which are listed in Table 2. Song et al. [28] extract the distance and connectivity between the tweet's sender and mentions. While distance defines the length of the shortest path between the tweet's sender and mentions, connection defines the strength of the connection between users. Graph-based spam detection methods use graph data structures to model features of Twitter as nodes and edges. Graph data models are the perfect solution to represent the data where information about data interconnectivity or topology is at least as important as the data itself [73]. Thus, graphs are commonly used by social networks such as Facebook, Twitter [74–81] which are mostly built on users, topics, and bi-directional interactions. Despite that graph-based features provide the best performance in terms of accuracy and sensitivity to differentiate spammers from legitimate users, other graph-based spam detection methods are presented in hybrid spam detection methods since they are combined with other spam detection methods.

D. Hybrid Spam Detection Methods

Hybrid spam detection methods use a combination of spam detection methods described in previous subsections in order to provide more robust spam detection which investigates the possibility of spam in a more comprehensive way. Stringing et al. [51] propose an approach based on both account-based and tweet-based features which are the ratio of the number of friend requests that the user sent to the number of friends she has, the ratio of the number of tweets which contain URLs to the total number of tweets the user has, the similarity of tweets sent by the user, the number of tweets sent by the user, the number of friends the user has, and the possibility of whether an account likely used a list of names to pick its friends or not. Gao et al. [82] propose a tweet-based spam detection approach based on the social degree of the tweet's sender, the history of interaction, the size of the cluster, the average time interval, the average number of URL in tweets, and the unique number of URL in tweets. Chen et al. [83] present a real-time spam detection method for Twitter based on 12 lightweight features which are extracted from a dataset contains 6.5 million spam tweets. The features they consider detecting spam on Twitter are age of the account, the number of followers, the number of following, the number of likes the account received, the number of the account's lists, the number of tweets of the account, the number of retweets of the tweet, the number of hashtags used in the tweet, the number of mentioned users in the tweet, the number of URLs used in the tweet, the number of characters used in the tweet, and the number of digits used in the tweet. Wang [29] proposes a hybrid Twitter spam detection method based on graph-based and tweet-based features. The graph-based features considered in the proposed method are the number of followers, the number of following, a reputation score which is calculated as the ratio between the number of followers over the total sum of the number of followers and following, and the number of following. The tweet-based features considered in the proposed method are tweet similarity, the number of tweets which contain URLs in the most recent 20 tweets, the number of tweets contains mentions in the most recent 20 tweets, and the number of tweets contains hashtags. Yang et al. [84] propose a Twitter spam detection method based on a combination of graph-based, tweet-based, and account-based features. The proposed method uses more robust features including the number of bi-directional links, the ratio of bi-directional links, betweenness centrality, clustering coefficient alongside tweet-based and account-based features such as the number of followers, the number of following, the number of tweets sent by the account, the age of the account, the ratio of the number of tweets contain URL, the ratio of the number of tweets contain hashtags, the number of duplicate tweets, the ratio of spam word, the ratio of the number of tweets used to reply to others, and the ratio of the number of retweets. Benevenuto et al. [1] propose a hybrid spam detection method based on account-based features such as the number of followers, the number of following, the ratio between followers over following, the number of tweets sent by the account, the number of mentions the account received, the number of replies, and the ratio of tweets received from the account's followers. The tweet-based based features of the proposed method are the number of words in each tweet, the number of URLs per word, the number of words of each tweet, the number of characters of each tweet, the number of hashtags on each tweet, the number of mentions on each tweet, the number of URLs of each tweet, and the number of times the tweet is retweeted. Chu et al. [48] present a method to categorize Twitter accounts as human, bot, and cyborg which is based on both account-based and tweet-based features. The features they consider categorizing the Twitter account into human, bot or cyborg are the number of the ratio of tweets contain URLs, device makeup, the number of the ratio of followers to friends, link safety, and whether the account is verified. Amleshwaram et al. [85] propose a hybrid Twitter spam detection method based on both account-based and tweet-based features. They categorize spammers into two: (1) users centric, and (2) URL-centric. The features they consider for spam analysis are the number of unique mentions, unsolicited mentions, hijacking trends, intersection with famous trends, variance in tweet intervals (VaTi), variance in number of tweets per unit time (VaTw), ratio of VaTi and VaTw, tweet sources, duplicate URLs, duplicate domain names, IP/domain fluxing, tweet's language dissimilarity, similarity between tweets, URL and tweet similarity, followers-to-following ratio, and profile description’s language dissimilarity. Chakraborty et al. [86] propose a hybrid method based on account-based and tweet-based features which use some new features such as spam score of profile description, name, and screen name, presence or absence of profile image and average same hashtag count. McCord and Chuah [9] present a hybrid method based on account-based and tweet-based features to facilitate spam detection. The features they use in the proposed method are the distribution of tweets over a

24-hour period, the number of URLs, the total number of replies/mentions in the most 100 recent tweets, the number of retweets in the 20-100 most recent tweets, the total number of hashtags in the 100 most recent tweets. Wang et al. [87] propose a spam detection method based on account-based, tweet-based, natural language processing (NLP), and sentiment features. Some unique features they use while detecting spam are length of the profile name, automatically or manually created sentiment lexicons, the number of exclamation marks, the number of question marks, maximum word length, mean word length, the number of capitalization words, the number of white spaces, and part of speech (POS) tags per tweet. Outline of the related works including their methodologies, the categories their metrics are based on, and accuracies are listed in Table 4.

TABLE. IV. OUTLINE OF THE RELATED WORKS INCLUDING THEIR METHODOLOGIES, THE CATEGORIES THEIR METRICS ARE BASED ON, AND ACCURACIES

Title	Methodology	Metrics Based on	Accuracy
"Uncovering Social Spammers: Social Honeypots + Machine Learning" [61]	Decorate, LogitBoost, HyperPipes, Bagging, RandomSubSpace, BFTree, FT, SimpleLogistic, LibSVM, ClassificationViaRegression	Account	99.21%
"Beyond blacklists: learning to detect malicious web sites from suspicious URLs" [63]	Naive Bayesian, SVM with linear kernel, SVM with an RBF kernel, l1-regularized logistic regression	Tweet	95-99%
"Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages" [64]	Naive Bayesian, Random Forest, Decision Tree, Logistic Regression	Tweet	90.41%
"WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream" [71]	LIBLINEAR	Tweet	0.9028
"Spam Filtering in Twitter using Sender-Receiver Relationship" [28]	Bagging, LibSVM, Decision Tree, Bayes Network, FT	Graph	99.7%
"Towards Online Spam Filtering in Social Networks" [82]	Decision Tree	Hybrid	TPR with 80.8%, FPR with 0.32%
"6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection" [83]	Random Forest, Decision Tree, Bayes Network, Naive Bayesian, k-NN, SVM	Hybrid	TPR with 90%
"Don't follow me: Spam detection in Twitter" [29]	Naive Bayesian, Neural Network, SVM, Decision Tree	Hybrid	93.5%
"Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers" [84]	Random Forest, Decision Tree, Decorate, Naive Bayesian	Hybrid	88.6%
"Detecting spammers on Twitter" [1]	SVM	Hybrid	87.6%
"Who is Tweeting on Twitter: Human, Bot, or Cyborg?" [48]	Bayesian	Hybrid	TPR with 90.47%
"CATS: Characterizing Automation of Twitter Spammers" [85]	Random Forest, Decision Tree, Decorate, Naive Bayesian	Hybrid	93.6%
"SPAM: A Framework for Social Profile Abuse Monitoring" [86]	Random Forest, Decision Tree, SVM, Naive Bayesian	Hybrid	89%
"Spam Detection on Twitter Using Traditional Classifiers" [9]	Random Forest, Decision Tree, Naive Bayesian, k-NN	Hybrid	95.7%
"A study of effective features for detecting long-surviving Twitter spam accounts" [62]	Decision Tree	Account	Precision with 86%
"Twitter Spammer Profile Detection" [58]	Naive Bayesian, SVM	Account	89.6%
"Detecting Spammers on Social Networks" [51]	Random Forest	Hybrid	90.93%
"Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter" [87]	Naive Bayesian, k-NN, SVM, Decision Tree, Random Forest	Hybrid	Precision with 94.6%
"Detecting malicious tweets in trending topics using a statistical analysis of language" [72]	Decision Tree, Naive Bayesian, Logistic Regression, SVM, Decorate, Random Forest	Tweet	94.5%

V. DISCUSSION

Spam detection in Twitter needs different ways from traditional spam detection methods for email and the web since (1) spammers tend to use shortened URLs instead of the full form of URL, and (2) Twitter is based on a huge and detailed network which is built on tweets, accounts, lists, moments, and the relationships between them. Thus, a more robust approach is required to detect spam in Twitter to with considering the variety of legitimate users who may behave similarly to spammers under certain circumstances. Even Twitter itself has false positive (spammers which are classified as legitimate users) detections as it is reported that Twitter has recommended a legitimate user to follow bots instead of related accounts [88]. In this paper, the features of Twitter spam detection are presented with discussing their effectiveness in detecting spam. Then, the proposed works in literature are categorized into four: (1) Account-based, (2) tweet-based, (3) graph-based, and (4) hybrid spam detection methods which use a combination of others.

Methods based on account-based features analyze account by using features related with accounts which some of them can be manipulated by spammers such as the number of following, the number of tweets sent by the account, the number of lists created by the account, the number of moments created by the account which is a brand new feature and, to the best of our knowledge, it has not been used by any works in

literature yet [89–91], the number of mentions the account received, the number of likes received by the tweets of account, and the number of retweets received by the tweets of account. Similarly, the number of followers, the ratio between the number of followers over the number of following, the ratio of the number of tweets liked by others, the ratio of the number of tweets retweeted also can be slightly manipulated by using a group of bots. Bots use various tools to do automated tasks such as following a user, sending a tweet. Some works investigate a number of last tweets of an account in order to reveal if the account posts spam tweets whose contents are almost identical to the tweets recently posted which is useful to detect spam distributed by bots, a set of accounts under the command and control (C&C) infrastructure. Account-based features are lightweight enough to be used detecting real-time spam which requires instant analysis. The number of lists the user is a member of can be considered a useful metric to detect spammers since it is an obvious sign of the user's impact on others but it is open to manipulation by creating fake lists and adding the fake accounts which are under the C&C infrastructure into these lists. Account-based features are lightweight enough to be used detecting real-time spam which requires instant analysis but they can be easily manipulated by spammers [37].

Tweet-based spam detection methods use parts of a tweet such as mentions, hashtags, the number of likes the tweet received, the number of retweets the tweet received, the content of tweet, lexical analysis of the tweet, the URL of the tweet, the location of the tweet, the post date of the tweet. Since the most common way to spread spam is sharing via a malicious URL [92], URLs of tweets are needed to be inspected. Therefore, almost all Twitter spam detection methods inspect URLs of tweets. The traditional ways to filter spam are based on IP blacklisting [93], domain and URL blacklisting [94]. Since spammers tend to use shortened URLs, traditional URL or IP blacklisting methods are not able to filter malicious URLs in Twitter. Also, Grier et al. [36] show that methods based on blacklisting are too slow to protect users since there is a delay before the malicious URLs are included in the database. Similar to account-based features, tweet-based features are lightweight enough to be used detecting real-time spam which requires instant analysis.

Graph-based spam detection methods use features of relationships between the sender and the mentions of a tweet such as connectivity and distance to analyze how these accounts are connected each other and to measure strengths of their connections in order to reveal the possibility of a spam connection. Graph-based features are hard to be manipulated [21], unlike account-based and tweet-based features. However, extracting of these features require in-depth analysis on the huge and complex Twitter graph which is time and resource intensive. Therefore, unlike account-based and tweet-based features, graph-based features are not lightweight enough for real-time spam detection. Another limitation of the graph-based approaches is that they assume that tweets come from friends are benign regardless of their content [21] which is not valid when attackers steal the accounts of legitimate users for their malicious aims.

VI. Conclusion

Twitter is the most popular microblogging platform which provides easy-to-use user experience thanks to its architecture. This popularity attracts the attention of spammers who post tweets to phish legitimate users by directing them to malicious websites through the URLs shared in tweets, spread malicious software and advertises through URLs shared within tweets, aggressively follow/unfollow legitimate users and hijack trending topics to attract their attention, propagate pornography. In August of 2014, Twitter has revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates. Since Twitter has unique characteristics from email services and websites, traditional spam filtering methods are not able to detect spam in Twitter. Thus, a more robust spam detection approach which is specially designed for Twitter is needed. In order to provide a spam-free environment, tweets of spammers are needed to be detected and filtered as well as the owners. By doing this, it is critical to reduce false positive detections in order to prevent legitimate users to be classified as spammers. In this paper, the features of Twitter spam detection and proposed approaches in the literature are discussed with considering their advantages and disadvantages. Also, the outdated features of Twitter which are commonly used by Twitter spam detection approaches are highlighted. Some new features of Twitter which, to the best of our knowledge, have not been mentioned by any other works are also presented.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on Twitter, in: CEAS 2010 - Seventh Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf., Redmond, Washington, USA, 2010: pp. 12–21. doi:10.1.1.297.5340.

[2] N.K. Alex Cheng, Mark Evans, Inside the Political Twittersphere, Sysomos. (2009). https://sysomos.com/inside-twitter/political-twittersphere (accessed February 5, 2017).

[3] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, in: Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media, Washington, DC, USA, 2010: pp. 178–185. doi:10.1074/jbc.M501708200.

[4] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining strengths, emotions and polarities for boosting Twitter sentiment analysis, in: Proc. Second Int. Work. Issues Sentim. Discov. Opin. Min. (WISDOM '13), Chicago, IL, USA, 2013: pp. 1–9. doi:10.1145/2502069.2502071.

[5] A. Pak, P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Computer (Long. Beach. Calif). 2010 (2010) 1320–1326. doi:10.1371/journal.pone.0026624.

[6] Company | About, Twitter. (2017). https://about.twitter.com/company (accessed February 5, 2017).

[7] Twitter Usage Statistics - Internet Live Stats, InternetLiveStats. (2017). http://www.internetlivestats.com/twitter-statistics/ (accessed February 5, 2017).

[8] D. Sayce, Number of tweets per day?, (2016). http://www.dsayce.com/social-media/tweets-day/ (accessed February 5, 2017).

[9] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, in: Auton. Trust. Comput. - 8th Int. Conf. (ATC 2011), Banff, Canada, 2011: pp. 175–186. doi:10.1007/978-3-642-23496-5_13.

[10] B. Stone, Google Adds Live Updates to Results, New York Times. (2009). http://www.nytimes.com/2009/12/08/technology/companies/08google.html (accessed February 5, 2017).

[11] A. DuVander, Which APIs Are Handling Billions of Requests Per Day? | ProgrammableWeb, Program. Web. (2012). http://www.programmableweb.com/news/which-apis-are-handling-billions-requests-day/2012/05/23 (accessed February 5, 2017).

[12] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent Twitter Sentiment Classification, in: Comput. Linguist., 2011: pp. 151–160.

[13] R.P. Schumaker, A.T. Jarmoszko, C.S. Labedz, Predicting wins and spread in the Premier League using a sentiment analysis of twitter, Decis. Support Syst. (2016). doi:10.1016/j.dss.2016.05.010.

[14] A. Go, L. Huang, R. Bhayani, Twitter Sentiment Analysis, Entropy. (2009) 17. doi:10.1007/978-3-642-35176-1_32.

[15] A. Montejo-Ráez, E. Martínez-Cámara, M.T. Martín-Valdivia, L.A. Ureña-López, Ranked WordNet graph for Sentiment Polarity Classification in Twitter, Comput. Speech Lang. 28 (2014) 93–107. doi:10.1016/j.csl.2013.04.001.

[16] S. Liu, X. Cheng, F. Li, F. Li, TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets, IEEE Trans. Knowl. Data Eng. 27 (2015) 1696–1709. doi:10.1109/TKDE.2014.2382600.

[17] E. Haddi, X. Liu, Y. Shi, The Role of Text Pre-processing in Sentiment Analysis, Procedia Comput. Sci. 17 (2013) 26–32. doi:10.1016/j.procs.2013.05.005.

[18] W. Chow, S. Shi, Investigating customers' satisfaction with brand pages in social networking sites, J. Comput. Inf. Syst. 55 (2015) 48–58.

[19] H. Saif, Y. He, M. Fernandez, H. Alani, Semantic Patterns for Sentiment Analysis of Twitter, in: Proc. 13th Int. Semant. Web Conf., Trentino, Italy, 2014: pp. 324–340.

[20] M.H.M. Sharif, I. Troshani, R. Davidson, Public Sector Adoption of Social Media, J. Comput. Inf. Syst. 55 (2015) 53–61. doi:10.1017/CBO9781107415324.004.

[21] C.D. Gowri, V. Mohanraj, A Survey on Spam Detection in Twitter: A Review, Int. J. Comput. Sci. Bus. Informatics. 14 (2014) 92–102. http://ijcsbi.org/index.php/ijcsbi/article/view/418.

[22] D. Goodin, Mystery attack drops avalanche of malicious messages on Twitter, Ars Tech. (2014). http://arstechnica.com/security/2014/04/mystery-attack-drops-avalanche-of-malicious-messages-on-twitter/ (accessed February 5, 2017).

[23] Z.M. Seward, Twitter admits that as many as 23 million of its active users are automated, Quartz. (2014). http://qz.com/248063/twitter-admits-that-as-many-as-23-million-of-its-active-users-are-actually-bots/ (accessed February 5, 2017).

[24] L. Whitney, Twitter says as many as 23 million accounts connect with automated services, CNET. (2014). https://www.cnet.com/news/twitter-reveals-23-million-of-accounts-active/ (accessed February 5, 2017).

[25] A Study of Social Network Scams, 2008.

[26] H. Drucker, D. Wu, V.N. Vapnik, Support Vector Machines for Spam Categorization, IEEE Trans. Neural Networks. 10 (1999) 1048–1054. doi:10.1109/72.788645.

[27] E. Blanzieri, A. Bryl, A Survey of Learning-Based Techniques of Email Spam Filtering, Artif. Intell. Rev. 29 (2008) 63–92.

[28] J. Song, S. Lee, J. Kim, Spam Filtering in Twitter using Sender-Receiver Relationship, in: RAID'11 Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., Menlo Park, CA, USA, 2011: pp. 301–317. doi:10.1007/978-3-642-23644-0_16.

[29] A.H. Wang, Don't follow me: Spam detection in Twitter, in: SECRYPT 2010 - Proc. Int. Conf. Secur. Cryptogr., Athens, Greece, 2010: pp. 1–10. doi:978-989-8425-18-8.

[30] Reporting Spam on Twitter, Twitter. (2017). https://support.twitter.com/articles/64986 (accessed February 5, 2017).

[31] X. Zhang, S. Zhu, W. Liang, Detecting Spam and Promoting Campaigns in the Twitter Social Network, in: IEEE Int. Conf. Data Min. (ICDM 2012), IEEE, Brussels, Belgium, 2012: pp. 1194–1199. doi:10.1109/ICDM.2012.28.

[32] D. Boyd, J. Heer, Profiles as Conversation: Networked Identity Performance on Friendster, in: HICSS '06 Proc. 39th Annu. Hawaii Int. Conf. Syst. Sci., Kauai, Hawaii, USA, 2006. doi:10.1109/HICSS.2006.394.

[33] T.N. Jagatic, N.A. Johnson, M. Jakobsson, F. Menczer, Social Phishing, Commun. ACM. 50 (2007) 94–100. doi:10.1145/1290958.1290968.

[34] K. Lee, B.D. Eoff, J. Caverlee, Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter, in: Fifth Int. AAAI Conf. Weblogs Soc. Media, AAAI Press, Barcelona, Spain, 2011: pp. 185–192.

[35] C.M. Zhang, V. Paxson, Detecting and Analyzing Automated Activity on Twitter, in: PAM'11 Proc. 12th Int. Conf. Passiv. Act. Meas., Atlanta, GA, USA, 2011: pp. 102–111. doi:10.1007/978-3-642-19260-9_11.

[36] C. Grier, K. Thomas, V. Paxson, M. Zhang, @spam: The Underground on 140 Characters or Less, in: Proc. 17th ACM Conf. Comput. Commun. Secur., Chicago, IL, USA, 2010: pp. 27–37. doi:10.1145/1866307.1866311.

[37] P. Kaur, A. Singhal, J. Kaur, Spam Detection on Twitter: A Survey, in: 2016 Int. Conf. Comput. Sustain. Glob. Dev., IEEE, New Delhi, India, 2016: pp. 2570–2573.

[38] M. Brandt, 80% Of Twitter's Users Are Mobile, Statista. (2015). https://www.statista.com/chart/1520/number-of-monthly-active-twitter-users/ (accessed February 5, 2017).

[39] A.P. Felt, M. Finifter, E. Chin, S. Hanna, D. Wagner, A Survey of Mobile Malware in the Wild, in: SPSM '11 Proc. 1st ACM Work. Secur. Priv. Smartphones Mob. Devices, Chicago, IL, USA, 2011: pp. 3–14. doi:10.1145/2046614.2046618.

[40] M. Chandramohan, H.B.K. Tan, Detection of Mobile Malware in the Wild, Computer (Long. Beach. Calif). 45 (2012) 65–71. doi:10.1109/MC.2012.36.

[41] Y. Zhou, Z. Wang, W. Zhou, X. Jiang, Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets, in: Proc. 19th Annu. Netw. Distrib. Syst. Secur. Symp., San Diego, California, USA, 2012. http://www.csd.uoc.gr/~hy558/papers/mal_apps.pdf.

[42] G. Delac, M. Silic, J. Krolo, Emerging Security Threats for Mobile Platforms, in: 2011 Proc. 34th Int. Conv. MIPRO, Opatija, Croatia, 2011: pp. 1468–1473.

[43] T. Cannon, Android Data Stealing Vulnerability, (2010). https://thomascannon.net/blog/2010/11/android-data-stealing-vulnerability/ (accessed February 5, 2017).

[44] X. Wei, L. Gomez, I. Neamtiu, M. Faloutsos, Malicious Android Applications in the Enterprise: What Do They Do and How Do We Fix It?, in: ICDEW '12 Proc. 2012 IEEE 28th Int. Conf. Data Eng. Work., IEEE, Arlington, Virginia, USA, 2012: pp. 251–254.

[45] A.T. Kabakus, I.A. Dogru, A. Cetin, APK Auditor : Permission-based Android malware detection system, Digit. Investig. 13 (2015) 1–14. doi:10.1016/j.diin.2015.01.001.

[46] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, S. Antipolis, All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks, in: 18th Int. World Wide Web Conf. (WWW '09), Madrid, Spain, 2009: pp. 551–560. doi:http://doi.acm.org/10.1145/1526709.1526784.

[47] D.K. McGrath, M. Gupta, Behind Phishing: An Examination of Phisher Modi Operandi, in: LEET'08 Proc. 1st Usenix Work. Large-Scale Exploit. Emergent Threat., San Francisco, CA, USA, 2008: p. 4. http://portal.acm.org/citation.cfm?id=1387713.

[48] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Who is Tweeting on Twitter: Human, Bot, or Cyborg?, in: 26th Annu. Comput. Secur. Appl. Conf. (ACSAC 2010), Austin, Texas, USA, 2010: pp. 21–30. doi:10.1145/1920261.1920265.

[49] S. Chhabra, A. Aggarwal, F. Benevenuto, P. Kumaraguru, Phi.sh/$oCiaL: The phishing landscape through short URLs, in: 8th Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf. (CEAS 2011), Perth, Australia, 2011.

[50] F. Klien, M. Strohmaier, Short Links Under Attack: Geographical Analysis of Spam in a URL Shortener Network, in: 23th ACM Conf. Hypertext Soc. Media (HT 2012), Milwaukee, WI, USA, 2012: pp. 83–87. doi:10.1145/2309996.2310010.

[51] G. Stringhini, C. Kruegel, G. Vigna, Detecting Spammers on Social Networks, in: 26th Annu. Comput. Secur. Appl. Conf. (ACSAC 2010), Austin, Texas, USA, 2010: pp. 1–9.

[52] D. Antoniades, E. Athanasopoulos, T. Karagiannis, we.b: The web of short URLs, in: WWW '11 Proc. 20th Int. Conf. World Wide Web, Hyderabad, India, 2011: pp. 715–724. doi:10.1145/1963405.1963505.

[53] D. Kim, Y. Jo, I.-C. Moon, A. Oh, Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, in: CHI 2010 Work. Microblogging What How Can We Learn From It, Atlanta, Georgia, USA, 2010. doi:10.1.1.163.7391.

[54] Y. Yamaguchi, T. Amagasa, H. Kitagawa, Tag-based User Topic Discovery Using Twitter Lists, in: 2011 Int. Conf. Adv. Soc. Networks Anal. Min. (ASONAM 2011), Kaohsiung, Taiwan, 2011: pp. 13–20. doi:10.1109/ASONAM.2011.58.

[55] Using Twitter lists, Twitter. (2017). https://support.twitter.com/articles/76460 (accessed February 5, 2017).

[56] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M.M. Hassan, A. AlElaiwi, M. Alrubaian, A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection, IEEE Trans. Comput. Soc. Syst. 2 (2016) 65–76. doi:10.1109/TCSS.2016.2516039.

[57] M. Verma, S. Sofat, Techniques to Detect Spammers in Twitter - A Survey, Int. J. Comput. Appl. 85 (2014) 27–32. doi:10.5120/14877-3279.

[58] G. Gee, T. Hakson, Twitter Spammer Profile Detection, Stanford, California, USA, 2010. http://cs229.stanford.edu/proj2010/GeeTeh-TwitterSpammerProfileDetection.pdf.

[59] The Twitter Rules, Twitter. (2017). https://support.twitter.com/articles/18311 (accessed February 5, 2017).

[60] C. Yang, R. Harkreader, G. Gu, Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers, IEEE Trans. Inf. Forensics Secur. 8 (2013) 1280–1293. doi:10.1109/TIFS.2013.2267732.

[61] K. Lee, J. Caverlee, S. Webb, Uncovering Social Spammers: Social Honeypots + Machine Learning, in: Proc. 33rd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., Geneva, Switzerland, 2010: pp. 435–442. doi:10.1145/1835449.1835522.

[62] P.-C. Lin, P.-M. Huang, A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts, in: 2013 15th Int. Conf. Adv. Commun. Technol., PyeongChang, Korea, 2013: pp. 841–846. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6488315&matchBoolean=true&rowsPerPage=30&searchField=Search_All&queryText=(%22twitter+spam%22)%5Cnpapers3://publication/uuid/60707410-4AE4-4FBE-A667-C91C41C51802.

[63] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, in: KDD '09 Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Paris, France, 2009: pp. 1245–1253. doi:10.1145/1557019.1557153.

[64] D. Canali, M. Cova, G. Vigna, C. Kruegel, Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages, in: WWW '11 Proc. 20th Int. Conf. World Wide Web, Hyderabad, India, 2011: pp. 197–206. doi:10.1145/1963405.1963436.

[65] Developer Survey Results 2016, 2016. http://stackoverflow.com/research/developer-survey-2016 (accessed February 5, 2017).

[66] D. Rowinski, It's Official: JavaScript Is The Most Commonly Used Programming Language On Earth, Appl. Resour. Cent. from Applause. (2016). https://arc.applause.com/2016/03/22/javascript-is-the-worlds-dominant-programming-language/ (accessed February 5, 2017).

[67] C. Whittaker, B. Ryner, M. Nazif, Large-Scale Automatic Classification of Phishing Pages, in: 17th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '10), San Diego, California, USA, 2010. http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf%5Cnhttp://research.google.com/pubs/pub35580.html.

[68] Y. Wang, D. Beck, X. Jiang, R. Roussev, Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities, in: 13th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '06), San Diego, California, USA, 2005: pp. 1–11.

[69] K. Thomas, C. Grier, J. Ma, V. Paxson, D. Song, Design and evaluation of a real-time URL spam filtering service, in: SP '11 Proc. 2011 IEEE Symp. Secur. Priv., Oakland, California, USA, 2011: pp. 447–462. doi:10.1109/SP.2011.25.

[70] M. Cova, C. Kruegel, G. Vigna, Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code, in: WWW '10 Proc. 19th Int. Conf. World Wide Web, Raleigh, North Carolina, USA, 2010: pp. 281–290.

[71] S. Lee, J. Kim, WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream, IEEE Trans. Dependable Secur. Comput. 10 (2013) 183–195. doi:10.1109/TDSC.2013.3.

[72] J. Martinez-Romo, L. Araujo, Detecting malicious tweets in trending topics using a statistical analysis of language, Expert Syst. Appl. 40 (2013) 2992–3000. doi:10.1016/j.eswa.2012.12.015.

[73] R. Angles, C. Gutierrez, Survey of graph database models, ACM Comput. Surv. 40 (2008) 1–39. doi:10.1145/1322432.1322433.

[74] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, P. Alto, The Anatomy of the Facebook Social Graph, Arxiv Prepr. arXiv. abs/1111.4 (2011) 1–17. doi:10.1.1.31.1768.

[75] J. Weaver, P. Tarjan, Facebook Linked Data via the Graph API, Semant. Web. 4 (2013) 245–250. doi:10.3233/SW-2012-0078.

[76] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about twitter, in: Proc. 1st Work. Online Soc. Networks, Seattle, WA, USA, 2008: pp. 19–24. doi:10.1145/1397735.1397741.

[77] S. Myers, A. Sharma, P. Gupta, J. Lin, Information Network or Social Network? The Structure of the Twitter Follow Graph, in: WWW'14 Companion Proc. 23rd Int. Conf. World Wide Web, Seoul, Korea, 2014: pp. 493–498. doi:10.1145/2567948.2576939.

[78] M. Gabielkov, A. Legout, The Complete Picture Of the Twitter Social Graph, in: Conex. Student '12 Proc. 2012 ACM Conf. Conex. Student Work., Nice, France, 2012: pp. 20–21. doi:10.1145/2413247.2413260.

[79] L. Zou, L. Chen, J.X. Yu, Y. Lu, A novel spectral coding in a large graph database, in: EDBT '08 Proc. 11th Int. Conf. Extending Database Technol. Adv. Database Technol., Nantes, France, 2008: pp. 181–192. doi:10.1145/1353343.1353369.

[80] R. a Hanneman, M. Riddle, Introduction to Social Network Methods, University of California Press, Riverside, CA, USA, 2005. doi:10.1016/j.socnet.2006.08.002.

[81] U. Brandes, T. Erlebach, Network Analysis, Springer Berlin Heidelberg, Heidelberg, Germany, 2005. doi:10.1007/b106453.

[82] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. Choudhary, Towards Online Spam Filtering in Social Networks, in: 19th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2012), San Diego, California, USA, 2012.

[83] C. Chen, J. Zhang, X. Chen, Y. Xiang, W. Zhou, 6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection, in: 2015 IEEE Int. Conf. Commun., IEEE, London, UK, 2015: pp. 7065–7070. doi:10.1109/ICC.2015.7249453.

[84] C. Yang, R.C. Harkreader, G. Gu, Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers, in: RAID'11 Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., Menlo Park, CA, USA, 2011: pp. 318–337. doi:10.1007/978-3-642-23644-0_17.

[85] A.A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, C. Yang, CATS: Characterizing Automation of Twitter Spammers, in: 2013 5th Int. Conf. Commun. Syst. Networks (COMSNETS 2013), Bangalore, India, 2013. doi:10.1109/COMSNETS.2013.6465541.

[86] A. Chakraborty, J. Sundi, S. Satapathy, SPAM: A Framework for Social Profile Abuse Monitoring, in: CEAS '11 8th Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf., Perth, Australia, 2011: pp. 46–54. http://www.cs.sunysb.edu/~aychakrabort/courses/cse508/report.pdf.

[87] B. Wang, A. Zubiaga, M. Liakata, R. Procter, Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter, in: Proc. 5th Work. Mak. Sense Microposts, Florence, Italy, 2015: pp. 10–16.

[88] D. Hernandez, Why can't Twitter kill its bots?, Fusion. (2015). http://fusion.net/story/195901/twitter-bots-spam-detection/ (accessed February 5, 2017).

[89] A. Read, Everyone Can Now Create Twitter Moments: Here's All You Need to Know, Buffer. (2016). https://blog.bufferapp.com/twitter-moments (accessed February 5, 2017).

[90] J. Roettgers, Twitter Now Lets Everyone Create and Share Moments, Variety. (2016). http://variety.com/2016/digital/news/twitter-moments-1201872731/ (accessed February 5, 2017).

[91] T. Huddleston, Now Twitter Wants You to Create Your Own "Moments," Fortune. (2016). http://fortune.com/2016/09/28/twitter-create-your-own-moments/ (accessed February 5, 2017).

[92] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, COMPA: Detecting Compromised Accounts on Social Networks, in: 20th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2013), San Diego, California, USA, 2013.

[93] Z. Qian, Z.M. Mao, Y. Xie, F. Yu, On Network-level Clusters for Spam Detection, in: 17th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '10), San Diego, California, USA, 2010.

[94] Y. Xie, F. Yu, R. Panigrahy, Spamming Botnet: Signatures and Characteristics, in: ACM SIGCOMM 2008, Seattle, WA, USA, 2008.