UNIVERSITÀ
DEGLI STUDI
DI PALERMO

NDS
NETWORKING AND DISTRIBUTED SYSTEMS
RESEARCH GROUP

Short Report on Spam Detection in Twitter

Short report

O. A. Sencam

Tutor: Prof. Marco Morana

Short Report on Spam Detection in Twitter

Ömer Ayberk ŞENCAN[*]
Universita degli Studi di Palermo
Viale delle Scienze, ed. 6 - 90128 Palermo, Italy
omerayberk.sencan@you.unipa.it

Abstract

THe usage of social media has started increasing exponentially over the last decade. Currently, people started to share their information on social media platforms everyday using the services such as Facebook, Twitter, Instagram and LinkedIn. Because these services gives the opportunity of establishing interactions between users instantly regardless of their location etc. it was inevitable for these social networks to become a part of our daily life. The increase in the numbers of usage has also caused the volume and the importance of the information on social media networks to go up. Twitter, one of the most used social networks, can be considered as one of those networks which possesses all the above properties. And because it is used so commonly, the number of spammer in this network has increased in time parallel to the increase on the usage of the social network itself. This increase is affecting the experience and the performance users get from this social network in a negative way. For preventing it, various methods has been developed. In this study, some of those methods and the research papers of these methods has been analysed and reviewed. In addition to previously given information, a study focusing on retracting meaningful information with Social Sensing methods from Twitter data has also been reviewed in this study.

Keywords: Machine Learning, Social Networks, Social Network Analysis, Spam Detection, Social Sensing

[*] Erasmus Student

1. Introduction

With the exponential increase of social media usage in everyday life, some social networks has evolved into something that changes the way we live. The number of social media users increases every day, there were 3.6 billion social media users in 2020 and it is estimated in 2025 there will be up to 4.41 billion social media users worldwide[1]. There is various type of information uploaded and shared on social media in the form of text, videos, photos and audio via social media services [2]. One of them, Twitter, is the 42nd most popular website and the 3rd most popular social media network with 650 million registered, 326 million active users and 500 million post per day average [3, 4].

The exponential increase of social network usage can be explained with the created links between users via posts, comments, messages and likes, expressing their opinions through social media [5]. In Table 1 , the increase in the number of the usage rate of social media are given [6].

Table 1

Increase in the Social Media Usage

Year	Active social media users (Million)	Number of mobile device users (Million)	Active mobile social media users (Million)
2018	3419	2307	1968
2019	3751	2526	2251

With this increase in the usage, it was inevitable for the researchers and various organizations to start studies aimed at obtaining meaningful information at different levels from the existing data over social media networks.

Regarding previously given information, this review provides a comprehensive overview of the already-given articles given in Table 2, the methods used in these studies and outlines directions for future research.

Table 2

Reviewed Studies

Author(s)	Name of the Study	Year	F-Score (%)
Concone, Lo Re, Morana & Ruocco	Twitter Spam Account Detection by Effective Labeling	2019	95
Concone, Lo Re, Morana & Ruocco	Assisted Labeling for Spam Account Detection on Twitter	2019	81,2
Concone, De Paola, Lo Re & Morana	Twitter Analysis for Real-Time Malware Discovery	2017	NaN[1]

The remainder of the paper is organized as follows. In Section 2, we discuss and explain the methods used in the given articles in detail. In Section 3 we give the experimental results of the reviewed articles and give a comprehensive comparison between similar ones. Conclusions will follow in Section 4.

[1] No F-Score has calculated for this study

2. Methods

Communication, using the Social Media services, must have some security requirements. These requirements are:

- *Integrity*
- *Confidentiality*
- *Eligibility*
- *User communication confidentiality*

Due to the high usage numbers, Twitter, like the other popular social media services, is targeted by spam accounts and malicious softwares.

The Annual Cybersecurity Report for the year 2018, published by Cisco, shows that the popular social media services, such as Twitter, are containing misuse cases with the aim of malware network traffic. Spams are the most common type of the malware in social networks and can be considered as the most important threats to users in social media networks[7].

Some of the academic studies on ow to detect these threats are categorized and presented in this study.

2.1 Labeling

The developments in the Machine Learning made it easier for researchers to observing, finding patterns and understanding the organization of the data easier. According to that, approaches like Labeling has emerged. In the Labeling approach, the aim is to seperate spammers from the legitimate users.

The biggest obstacle to reach that aim is the constant evolving characteristics of malicious behavior. Because the behavior is changing constantly, it is not possible to define a fixed rule-set to distinguish malicious accounts from the trustworthy ones. Although, the behaviour is constantly changing, the goal of the spammers usually doesn't change. The biggest objective of the malicious accounts is reaching as many users as possible.

In order to achieve this goal, spammers often behave abnormally. Abnormal behavior can be defined as *"Behaving different than normal"*. And if the behavior is differing from the normal, it is considered dangerous[8]. The most important information here is that because the normal behavior is known, the abnormal behavior can be seperated from the normal one.

Regarding previously given information, the chracteristics of the malicious accounts can be determined with:

1. URL Analysis
2. Finding Similar Tweets
3. Finding Similar Users

2.1.1 URL Analysis

The first step of URL analysis is using blacklisting services. Blacklisting services provides information about the given link concerning if the link is malicious or not based on previously collected reports. Like all other systems, Blacklisting services also has their own weaknesses which can be defined as follows:

1. Four days needed to ad a malicious website to be added to a blacklist [9]
2. Blacklisting services can't detect malicious links that has been shortened twice or more
3. These services aim to find unsafe links, as a result, a spammer who shares safe links several times (still spamming) would never be found by URL blacklisting services

Regarding the given information, relying on those services is not enough by itself. As a consequence, the URL's shared on social networks needs more detailed examination.

To overcome this obstacle, researchers analysed more features. Both of the URL analyzers used in the reviewed studies uses the same features. In particular, three factors are considered:

1. The presence of maicious *urls* according to Google Safe Browsing (GSB)
2. The total number of *urls*
3. The ratio R_{UT} between number of unique *urls*, U and T.

2.1.2 Finding Similar Tweets

Another anomaly, observed on malicious accounts can be considered as sharing similar content multiple times. Since the purpose of malicious accounts is to maximize the number of legitimate users they reach, they widely use this method. To detect this kind of tweets, a detailed analysis of the given data is required.

Analysing data is not possible without the help of the machine learning methods. Because the data is so big and heterogeneous, it makes the analysing progress slower. To overcome this issue, researchers has developed various techniques. Regarding that, to be able to divide the given data into homogeneous clusters, several clustering techniques have been proposed in the literature[10].

For clustering the data into homogeneous clusters, *near duplicates clustering* method has been widely used amongst researchers. This method intends grouping given items, i.e., such as tweets like given in these studies, that are completely same or slightly different from each other by a few characters.

With the help of this methods, researchers aim is finding and measuring the D_oS [2] between the given items, which, in this case, tweets shared in the timeline of each user. For finding the near-duplicate tweets, MinHash and LSH [3] algorithms can be used[11].

Tweets are complex structures, containing, text, images, links, mentions etc. with a limitation of available characters. Even with the above-given techniques, some pre-processing work is required to minimize the time-span and maximize the efficiency of the given algorithms. On this purpose, the pre-processing methods used by researchers can be given as follows:

- Remove all non-english[4] tweets
- Remove Mentions
- Convert text to lower-case
- Remove hashtag symbol and other common symbols
- Expand the *urls*
- Remove stop-words
- Normalize accented characters

2.1.3 Finding Similar Users

With the help of another clustering algorithm, named as QTC[5], which aims reducing the manual annotation effort, the accounts which has not been labeled during the previous stages can be distinguished. The main goal in this technique is grouping the users with similar behavior with only performing

[2] Degree of Similarity

[3] Locality-Sensitive Hashing

[4] Or any other target language

[5] Quality Threshold Clustering

manual annotation for a small group of users and extending the label to the whole dataset.

The main difference between the QTC and other clustering algorithms is, other clustering algorithms require prior specification of the number of clusters to be found, where QTC does not. Instead of that, when using QTC algorithm, elements are progressively grouped while maintaining the quality of each cluster above a vertain threshold.

To be able to achieve the goal given above, two parameters must be defined which are given as follows:

- The maximum cluster diameter (d).
- The minimum number of elements a cluster has to contain (m).

2.2 Real-Time Malware Alerting

Real-time malware alerting method extracts tweets from the original source using the API provided by Twitter. After this, all posts received go through a searching process by looking out if the commonly used keywords in messages for the computer attacks exist. To be able to find and select related posts, a set of keywords are defined pre-search. With this method, researchers aim to prevent the spread of the messages containing a new malware.

Tweets targeting the spread of malware do not always appear in the same shapes and forms. In addition, tweets containing related keywords cannot be said to be malicious under all circumstances, for example, these tweets may be posts made for advertisements of anti-virus software.

Therefore, these posts need to be examined in more detail. In order to carry out this review process, first all "non-important" words are removed from the relevant tweet thread in order to improve the performance of the classifier. Afterwards, the remaining tweets are classified using a Naïve Bayes classifier.

3. Reporting the Result

BAsed on the papers reviewed, all of the papers demonstrated the usage of either Clustering method, hashing method or a mix of both methods.

The techniques rely on the quality of the dataset, features of the dataset and feature selection process.

According to the papers which has been reviewed in this study, it can be seen that the researchers has taken advantage of one of the mostly used methods for social media analyzing, which is Naïve Bayes model. Naïve Bayes is an algorithm know as producing satisfying results when applied on well-formed text corpus [12].

The first of these studies, (Concone, F., Re, G. L., Morana, M., & Ruocco, C. , 2019) [13] aims to detect malicious accounts that use OSNs for non-legit activities and reduce spam detection time. In this study, they developed two different algorithms and tested these algorithms with a data set that included almost 8 *million* tweets from more than 40 *thousand* users.

After testing their algorithm with different settings and using different feature sets, researchers has reported the proposed algorithm reaches 95% accuracy rate with genuine users and 70% accuracy rate with spammers.

The second study, (Concone, F., Re, G. L., Morana, M., & Ruocco, C. , 2019) [14] to solve the spam detection problem, they developed a method that performs URL inspection and tweet classification tasks. In this way, it was ensured that legit users and spammers were separated based on the behavior patterns frequently used by spammers.

As a result of their study, the researchers has reported, they have reached about 80% accuracy rates within the actual and spammer accounts labeling.

The third and the last study, (Concone, F., De Paola, A., Re, G. L., & Morana, M. , 2017) [15] offers a real-time malware warning system which has been developed by the researchers. This system has been tested with tweets obtained through the Twitter API using the Naive Bayes classifier. Afterwards, tweets with the same topics e.g, a new malware infection, summerized to create an alert.

After the testing phase, researchers has stated that with the involvement of the users in the process, a system that adapts itself to the situation and conditions has been created.

4. Conclusion

The results obtained after the examination of the relevant studies are examined in this section. Accordingly, it has been determined that there are studies that use different methods in the fields of both URL analysis and the discovery of spam accounts. The studies based on this research have been examined in detail. The results obtained after the examination of the studies found show that it is possible to perform URL analysis and spam account detection on social networks using both machine learning-based methods and other methods. In this context, it has been observed that the researchers have succeeded with different methods on the specific subjects they want to investigate. Accordingly, it cannot be said that a single method in the field of spam account detection or URL analysis is more successful than others, previous studies should be examined specifically on the area to be investigated, and if necessary, the previously used models should be arranged in accordance with the specific research subject or new models should be created according to the subject inspired by these models.

References

[1] Statista, "Number of Worldwide Social Network Users," tech. rep., Statista, 2020.

[2] K. J. Giri and T. A. Lone, "Big Data-Overview and Challenges," *International Journal of Advanced Research\in Computer Science and Software Engineering*, vol. 4, no. 6, 2014.

[3] S. Gaglio, G. Lo Re, and M. Morana, "A framework for real-time Twitter data analysis," *Computer Communications*, vol. 73, pp. 236–242, 2016.

[4] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications*, vol. 164, p. 114006, 2021.

[5] G. A. Ruz, P. A. Henriquez, and A. Mascareno, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems*, vol. 106, pp. 92–104, 2020.

[6] O. Çıtlak, M. Dorterler, and I. A. Dogru, "A survey on detecting spam accounts on Twitter network," *Social Network Analysis and Mining*, 2019.

[7] M. Verma and S. Sofat, "Techniques to detect spammers in twitter-a survey," *International Journal of Computer Applications*, vol. 85, no. 10, 2014.

[8] M. H. Bhuyan, D. Bhattacharyya, and J. K. Kalita, "An effective unsupervised network anomaly detection method," in *Proceedings of the international conference on advances in computing, communications and informatics*, pp. 533–539, 2012.

[9] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37, 2010.

[10] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[11] A. Gionis, P. Indyk, R. Motwani, *et al.*, "Similarity search in high dimensions via hashing," in *Vldb*, vol. 99, pp. 518–529, 1999.

[12] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 138–140, IEEE, 2017.

[13] F. Concone, G. Lo Re, M. Morana, and C. Ruocco, "Twitter Spam Account Detection by Effective Labeling," in *ITASEC*, 2019.

[14] F. Concone, G. Lo Re, M. Morana, and C. Ruocco, "Assisted Labeling for Spam Account Detection on Twitter," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 359–366, IEEE, 2019.

[15] F. Concone, A. De Paola, G. Lo Re, and M. Morana, "Twitter analysis for real-time malware discovery," in *2017 AEIT International Annual Conference*, pp. 1–6, IEEE, 2017.