ISSN No:-2456-2165

Spam Profile Detection on Instagram Using Machine Learning Algorithms on WEKA and RapidMiner

Usman Rasheed[1], Muhammad Hashim Hameed[2]
Master of Sciences, Department of Computer Science,
University of Agriculture Faisalabad, Pakistan

Akmal Rehan[3]
Lecturer, Department of Computer Science,
University of Agriculture Faisalabad, Pakistan

Abstract:- With every passing second social network community is growing rapidly, because of that, attackers have shown keen interest in these kinds of platforms and want to distribute mischievous contents on these platforms. With the focus on introducing new set of characteristics and features for counteractive measures, a great deal of studies has researched the possibility of lessening the malicious activities on social media networks. This research was to highlight features for identifying spammers on Instagram and additional features were presented to improve the performance of different machine learning algorithms. Performance of different machine learning algorithms namely, Multilayer Perceptron, Random Forest, K-Nearest Neighbor and Support Vector Machine were evaluated on machine learning tools named, RapidMiner and WEKA. The result from this research tells us that Random Forest outperformed all other selected machine learning algorithms on both selected machine learning tools. Overall, Random Forest provided best results on RapidMiner. These results are useful for the researchers who are keen to build machine learning models to find out the spamming activities on social network communities.

Keywords:- *Malicious Activities, Spammers, Machine Learning Algorithms, Multilayer Perceptron, Random Forest, K-Nearest Neighbor, Support Vector Machine, Rapidminer, WEKA, Social Network Communities.*

I. INTRODUCTION

Online platform which is used by people to build social relationships or social networks with other people is known as social networking service (or also known as social media, or social media networking or SNS). These social networks or social relationships with other people are based on similar activities, real life connections or backgrounds, personal interests, career interests. These social networks are spread across numerous computer networks. These social networks are basically computer networks that links knowledge, organizations and most importantly people. Social life has been drastically changed by social networks in the recent past and web is change into social web where users and their networks are the points of online development, commerce, growth, and data sharing. Facebook is currently at the top of the chain with being the first social media platform that has 1 billion registered accounts and over 2 billion active monthly users. To spread contents, not only ordinary people but also the politician's, public figures, people of interest and celebrities use social media platforms. Instagram, the photo sharing app is ranked 6th with over 1 billion active monthly accounts. Linked Inn is the most famous social media professional network with over 300 million active monthly users. Key elements that is the base for the components being shared on these social media are its users (Ahmed & Abulaish, 2012).

Online social platforms have emerged as easily accessible, cheap, and effective social media, that facilitates worldwide users for information sharing and communication. Although the basic purpose of social networking sites is online communication and interaction, but the pattern of usage and specific goals are different on different services. In the recent years, social media platforms like Facebook, Instagram and Twitter have become worldwide sensation and one of the quickest emerging e-services, as stated by (Ellison, 2008).

Users are usually recognized by a profile at these social media sites. It generally comprises of name and picture, birthday, possibly an address and other personal information. However, these social media sites do not strictly identify that the one creating the profile and using it, is actually the same person as stated in the profile. Someone else is using somebody else's identity, if that is not the case, this is known as false identity. One can easily create profile with fabricated names and other information that is not associated with any person living in any part of the world. In these kinds of cases the identity is known as faked identity. Pictures in these kind of profile can still be of real persons that can be easily taken from the internet (Romanov et al., 2017).

5% to 6% of the registered accounts on the Facebook are fake accounts, as stated by Facebook. It is clearly stated in the Terms and Conditions of Facebook that the users must not provide false information and their information must be kept up to date. False and wrong information puts in danger the supportability of Facebook's business model. This clearly shows that for Facebook's business model it is important that user data should be correct and accurate (Krombholz et al., 2012).

There are number of mischievous activities executed by the spam profiles, which consists of phishing, following great number of users with little followers, overflowing the social media platforms with fake profiles, random link connection, spreading malware and endanger existing valid accounts etc. In spite of the benefits got from the social relationships, user's profile has turned out to be one of the focused resource by the spammer's, who among user's, influence the trust relationship to acquire more unfortunate victims (Hanif et al., 2018).

By using machine learning algorithms and graphs to detect spammers on the social networks, present studies have proposed a variety of methods to control spam profiles on the social networks. Consequently, it is very important to control the extension of spam profiles on the social networks as it has become an effective way of action for spammers. Studies have shown that, by using social engineering attacks, spammers can effectively compromised existing valid accounts (Egele et al., 2017).

II. LITERATURE REVIEW

(Grier et al., 2010) utilized blacklist-based method to decrease spamming activities and to identify spammers on Twitter. Clickthrough data brought about by the URLs that were posted on Twitter, was analyzed by the researchers to identify these spamming activities. According to their findings, on Twitter, phishing assaults have been used successfully. All kinds of uninvited posts that makes an appearance trending topics, user's wall, personal feeds, and comments are represented by word spam.

(Chu et al., 2012) proposed Random Forest algorithm, to detect spam movement on Twitter, in the domain of supervised machine learning. The open structure and the popularity of Twitter have pulled in countless automated programs, that are known as bots. Genuine bots produce a lot of favorable tweets updating feeds and conveying news, on the other hand malicious contents or spam is spread by malicious bots. In distinguishing spammers on social networks, major role has been played by the use of machine learning methods.

(Aggarwal et al., 2012) applied different features, which can recognize tweets with malicious URLs, to build a real-time detection system. Tweet, link, campaign, and account property features were combined by them to train Random Forest algorithm. URL based features along with the specific twitter features demonstrate to be a solid system to identify the phishing tweets. They used machine learning classification methods to detect and identify the legitimate and phishing tweets on Twitter.

(Saini, 2014) stated that initially, certain researchers paid attention on the improvement of honey pots to distinguish spams. By utilizing honey pots, researchers managed in gathering misleading spam profiles present in the social media networks that were dependent on some unknown behavior of the users, to recognize the spams. This made unique user profiles with discrete data like geographic location as in locality, sex, age, and date of birth and sent it into Myspace (social network) people group. Spammers send friend requests for a long time, following one of the systems. By allocating bots, honey profile inspects the actions of spammers. The bots store the spammers profile whenever the spammer sends friend requests and skids through the web pages to identify the objective page where advertisement arise.

(Washha et al., 2016) utilized Random Forest algorithm to detect the spammers. They proposed new time-based features and advanced the design of some existing features that were used before. Their design of features was divided into behavioral features that identifies the pattern of posting behavior and statistical features that incorporates time attribute. Using the time property and proposed features they were able to correctly classify the spammers and legitimate users with higher accuracy.

(Almaatouq et al., 2016) found that there are two key classes of spam accounts present, that displays the different spamming patterns and use distinct tactics for spreading the spam material and targeting victims. While distinguishing the legitimate user from the spammers, the results of their analysis emphasized on the value of social interaction features. They categorized the features in three categories, that are profile properties, social interactions, and content attributes, to detect the spam profiles.

(X. Zhang et al., 2016) expressed that a number of researchers excerpt a wide range of users features like, network, content feature and profile feature etc. and after that pick various classification algorithms to detect spammers on online social networks. As a result of spammer identification in traditional stages like web and e-mail, some of the work has been devoted in detecting spammers in different social media sites, for example, Twitter, Facebook, Sina Weibo.

(W. Zhang & Sun, 2017) presented a scheme that applies supervised learning techniques and features based method to detect the spam posts on Instagram. To find the finest pair of parameters of the model and supervised learning model, they used K-fold cross validation. To label the media posts swiftly, they used two-pass clustering method i.e., K-medoids clustering and Minhash clustering, to group the almost duplicate posts into the same clusters and based on the results of the clustering, marked the posts as spam or non-spam.

(Liu & Hu, 2017) stated that they can create classifiers that can work consequently to reveal spammers, by using machine learning techniques on online social networks. Depending on detection method and specific scenario, the originators of these spams may be identified as an individual or in the form of the groups. There are distinct forms of spam in different social communities and after some time they might change, that is why as to filter them out few standards must be made. Still, designing these set of standards one by one and establishing these rules consumes a large amount of time and is error prone.

(Setiawan et al., 2017) aimed to offer the solutions to decrease the effect from the spammers using Markov Clustering algorithm to identify the spam profile. They labeled the gathered profiles with Yes and No, where the spam profiles were represented by Yes and normal profiles were represented by No, to see that how good an algorithm performs to identify the spam profiles. They used F-Measure method and BCubed metrics to examine that how the process of clustering performs.

(Sohrabi & Karimi, 2018) stated that as the trend of social media is increasing day by day, the systems have become a highlighted instrument for spammers by spreading spam.

ISSN No:-2456-2165

Various spamming operations, like, spreading malware, acquiring significant data of users, false publicity, and various other tasks. A large number of spammers divert the users to other pages where these spammers want to, and different kinds of spam are spread. Because of the false information present, it is very difficult to analyze manually.

(Concone et al., 2019) applied MinHash signature algorithm and Locality Sensitive Hashing LSH to detect spam profiles on twitter by effective labeling. Some of the common behaviors of spammers were captured by the proposed system, i.e., patterns and habit of sharing malicious URLs. By compressing every single document into a signature, MinHash signature algorithm resolves the problem of associating large datasets. By maximizing the probability of similar documents to be hashed into the same bucket, LSH performs pairwise comparisons efficiently.

III. RESEARCH OBJECTIVE

The main focus of this research is to find the different features among the Instagram user's profile data that will help to detect and identify the spam profiles on Instagram. The performance of machine learning algorithms is checked on machine learning tools, to find which of the machine learning algorithm performs better than the other machine learning algorithms. The accuracy level of the machine learning tools is also being measured.

IV. RESEARCH METHODOLOGY

A. Data Collection and Processing

For this research purpose data of social platform Instagram is collected manually by visiting this platform and carefully observing and recording the selected features of the Instagram, that are used in this research. After basic dataset was collected from Instagram profiles, new features were created using that dataset. After that metrics were applied, that were further evaluated to create our target variable. After removing unnecessary features and creating the new ones the final data set contained the features discussed in the table below.

Table I : Data Set Features

Name	Description
Name	Records the presence and absence of name. 1 represents the presence of name and 0 absence.
Num Name	Records the presence and absence of numeric character in name. 1 represents the presence of numeric character in name and 0 absence.
Username	Records the presence and absence of username. 1 represents the presence of username and 0 absence.
Num Username	Records the presence and absence of numeric character in username. 1 represents the presence of numeric character in username and 0 absence.
Name=Username	Records whether the name and username are same or not. 1 represents same and 0 represents different.
Phishing	Records the result from Phishtank. 1 represents that the URL is phishing and 0 represents it's not.
Blocked	Records the result from Sitechecker. 1 represents that the URL is safe and 0 represents it's not.

B. Metrics

Three metrics were created after carefully observing the users behavior and used in this research and the results from these metrics were further used to label our target variable as spam and non-spam. Formulas were applied in excel sheet in which the data was recorded in the first place and also the new features that were created were also present in the same file. Results of all three metrics were recorded in binary form i.e., in the form of 0,1.

In M1 or metric one, 2 features were used that are Phishing and Blocked. In this metric these 2 were checked and if either of the feature was 1, then its result was recorded as 1.

M1 = IF(OR(Phishing=1,Blocked=1),"1","0")

In M2 or metric two, 2 features were used that are Num Name and Verified. This metric checks that if the Num Name is 1, that means the Name of the user contains a numeric character in it and Verified is 0, which indicates that the user account is not verified by the Instagram, then its result was recorded as 1.

M2 = IF(AND(Num Name=1,Verified=0),"1","0")

In M3 or metric three, 4 features were used that are Profile Pic, Posts, Description and Verified. This metric checks that if all of these 4 features are 0, then its result was recorded as 1. This means that there is no profile picture, no short description is given in profile, no post is shared, and the account is also not verified by Instagram.

M3=IF(AND(Profile Pic=0,Post=0,Description=0,Verified=0),"1","0")

C. Target Variable

Target variable was created by using the results of the three metrics that are discussed before. If the result of any metric is recorded as 1, then it means this profile is behaving differently from the legitimate users. So, we applied the formula in the excel sheet which uses the results from these three metrics and if any of them was recorded as 1 then it labelled the target variable Spam as 1(represents spam profile) and if all of them were recorded as 0 then it labelled the target variable as 0(representing not spam/genuine profile).

ISSN No:-2456-2165

V. RESULTS AND DISCUSSIONS

After the results of four selected classifiers on both selected tools are gathered, we evaluated the performance of these classifiers on selected tools. Accuracy of the four selected classifiers on both selected tools is given in the table below.

Table II : Results on WEKA & RapidMiner

	WEKA	RapidMiner
KNN	99.78	98.47
MLP	99.45	97.38
SVM	98.8	98.91
RF	99.89	100

Generally, WEKA provided better results. RF gave best results in both tools. RF was on top with 99.89% accuracy in WEKA and 100% accuracy in Rapid Miner. Overall RF gave the best results in RapidMiner with 100% accuracy. KNN gave second best results in WEKA and SVM gave second best results in RapidMiner. MLP gave the third best results in WEKA and KNN gave the third best results in RapidMiner. Least best results are given by SVM in WEKA and MLP in RapidMiner. Overall RF proved to be the best amongst the four selected Machine Learning algorithms in both selected Machine Learning tools.

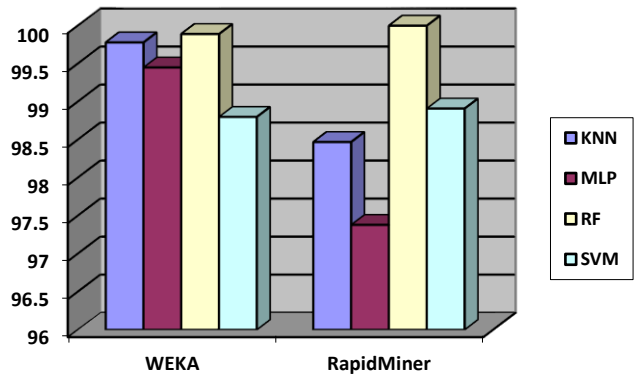

Fig. 1: Results Comparison

In figure above, accuracy of all four selected Machine Learning algorithms in both of the selected Machine Learning tools i.e., WEKA and RapidMiner, is shown in the form of graph. In the figure above we can see clearly that RF proved to be the best amongst the four selected algorithms.

REFERENCES

[1]. Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: Automatic realtime phishing detection on twitter. ECrime Researchers Summit, ECrime, 1–12. https://doi.org/10.1109/eCrime.2012.6489521

[2]. Ahmed, F., & Abulaish, M. (2012). An MCL-based approach for spam profile detection in online social networks. Proc. of the 11th IEEE Int. Conference on Trust, Security and Privacy in Computing and Communications, TrustCom-2012 - 11th IEEE Int. Conference on Ubiquitous Computing and Communications, IUCC-2012, 602–608. https://doi.org/10.1109/TrustCom.2012.83

[3]. Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V. K., Alsaleh, M., Alarifi, A., Alfaris, A., & Pentland, A. 'Sandy.' (2016). If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. International Journal of Information Security, 15(5), 475–491. https://doi.org/10.1007/s10207-016-0321-5

[4]. Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting social spam campaigns on Twitter. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7341 LNCS, 455–472. https://doi.org/10.1007/978-3-642-31284-7_27

[5]. Concone, F., Re, G. Lo, Morana, M., & Ruocco, C. (2019). Twitter spam account detection by effective labeling. CEUR Workshop Proceedings, 2315.

[6]. Egele, M., Stringhini, G., Kruegel, C., & Vigna, G. (2017). on Social Networks. 14(4), 447–460.

[7]. Ellison, N. B. (2008). Social Network Sites : Definition , History , and Scholarship. 13, 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

[8]. Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @Spam: The underground on 140 characters or less. Proceedings of the ACM Conference on Computer and Communications Security, 27–37. https://doi.org/10.1145/1866307.1866311

[9]. Hanif, M. H. M., Adewole, K. S., Anuar, N. B., & Kamsin, A. (2018). Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner. Advanced Science Letters, 24(2), 1043–1046. https://doi.org/10.1166/asl.2018.10683

[10]. Krombholz, K., Merkl, D., & Weippl, E. (2012). Fake identities in social media: A case study on the sustainability of the Facebook business model. Journal of Service Science Research, 4(2), 175–212. https://doi.org/10.1007/s12927-012-0008-z

[11]. Liu, N., & Hu, X. (2017). Spam Detection on Social Networks. 1–9. https://doi.org/10.1007/978-1-4614-7163-9

[12]. Romanov, A., Semenov, A., Mazhelis, O., & Veijalainen, J. (2017). Detection of fake profiles in social media: Literature review. WEBIST 2017 - Proceedings of the 13th International Conference on Web Information Systems and Technologies, Webist, 363–369. https://doi.org/10.5220/0006362103630369

[13]. Saini, J. S. (n.d.). A Study of Spam Detection Algorithm on Social Media Networks. 195–202. https://doi.org/10.1007/978-81-322-1680-3

[14]. Setiawan, E. I., Susanto, C. P., Santoso, J., Sumpeno, S., & Purnomo, M. H. (2017). Preliminary study of spam profile detection for social media using Markov clustering: Case study on Javanese people. 20th International Computer Science and Engineering Conference: Smart Ubiquitos Computing and Knowledge, ICSEC 2016, 1–4. https://doi.org/10.1109/ICSEC.2016.7859942

[15]. Sohrabi, M. K., & Karimi, F. (2018). A Feature Selection Approach to Detect Spam in the Facebook Social Network. Arabian Journal for Science and Engineering, 43(2), 949–958. https://doi.org/10.1007/s13369-017-2855-x
[16]. Washha, M., Qaroush, A., & Sedes, F. (2016). Leveraging time for spammers detection on Twitter. 8th International Conference on Management of Digital EcoSystems, MEDES 2016, 109–116. https://doi.org/10.1145/3012071.3012078
[17]. Zhang, W., & Sun, H. (2017). Instagram Spam Detection. https://doi.org/10.1109/PRDC.2017.43
[18]. Zhang, X., Bai, H., & Liang, W. (2016). A social spam detection framework via semi-supervised learning. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9794(61272374), 214–226. https://doi.org/10.1007/978-3-319-42996-0_18