IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, February 2024

Spam Detection in Social Networks Using Machine Learning

Miss. Sneha Bajirao Sahane
Master of Computer Applications
CHMES Society's, Dr. Moonje Institute, Nashik, India

Abstract: *Many social media platforms have emerged as a result of the online social network's (OSN) rapid expansion. They have become important in day-to-day life, and spammers have turned their attention to them. Spam detection is done in two different ways, such as machine learning (ML) and expert-based detection. The expert-based detection technique's accuracy depends on expert knowledge, and the manual process is a time-consuming task. Thus, ML-based spam detection is preferred in OSN. Spam identification on social networks is a difficult operation involving a variety of factors, and spam and ham have resulted in an imbalanced data distribution, which gives an advantage to spammers for corrupting our devices. Spam detection based on ML algorithms like Logistic Regression (LR), K-Nearest Neighbour (KNN), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), and XGB, Voting Classifier (VC), and many other algorithms are used to design the address balance and to attain high assessment accuracy. There is a non-balance issue. Text is vectorized by vectorizers and all the relative results are stored. The experimental result shows that, as compared to KN, NB, ETC, RF, SVC, LR, XGB, and DT, the proposed VC provides a higher classification accuracy rate of 97.96%. The proposed methods are effective in identifying balanced and imbalanced datasets, as evidenced by the validation results. The website was created to detect messages as spam or not.*

Keywords: economic organization

I. INTRODUCTION

Unsolicited commercial email (UCE), a.k.a. spam, is not a new problem causing complaints from many Internet users. Spamming, i.e., the act of sending UCE, involves the sending of nearly identical emails to thousands or even millions of recipients without the recipients' prior consent or even violates recipients' explicit refusal [9, 30, 34]. Unsolicited bulk email (UBE) is another category of emails that can be considered spam. As suggested in recent reports by Spam Haus [4] and Symantec [31], spam is increasingly being used to distribute virus, spyware, links to phishing web sites, etc. The problem of spam is not only an annoyance, but is also becoming a security threat. There is an increasing trend for both UCE and UBE. For instance, Symantec has detected a 44% increase in phishing attempts from the first half of 2005 to the second half. Statistics from the Distributed Checksum Clearinghouse (DCC) project [24] shows that 54% of the email messages checked by the DCC network in 2005 are likely to be from bulk email. Also, statistics from MX Logic [22] shows that on average 80.78% of the email messages delivered to their clients during the week of March 24–30, 2007 are considered spam, with peaks at more than 90%. Various legal means of anti-spam attempts have been discussed in [16, 23]. Legislations specifically targeted at email spam as well as unwanted messages in general have been introduced in some countries, such as the United States of America. Before targeted legislations are introduced, some existing laws are sought for fighting spam. Possible approaches are based on laws and statutes that combat fraud, antiracketeering, trespassing and anti harassment.

II. PROBLEM STATEMENT

Given a set of n email accounts A = {a1, a2, ..., an}, a sender set S ⊆ A is defined as the set of email accounts that have sent at least one email and a receiver set R ⊆ A is the set of email accounts that received at least one email. Within the set of senders, t of them are initially labeled as follows:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, February 2024

$$y_i = \begin{cases} 1 & \text{if } a_i \text{ is a legitimate sender,} \\ -1 & \text{if } a_i \text{ is a spammer,} \end{cases}$$

for ai ∈ S and t < n. We call this set of t labeled sender the training set ai∈Sl⊂ S. Although the training set may contain all the senders Sl⊆ S, such a scenario will not be of interest to us as all senders are already labeled. Logs of events in email transactions L = {li} between accounts are available as a tuple of attributes:
(ai, aj1 , aj2 , . . . , aji , x1, x2, . . . , xm)
where ai ∈ S and {aji∈ R} are the sender and the corresponding set of receiver accounts, respectively, and x1 through xm are other attributes that the log may have, such as time of transaction, message size, event type, sender's host IP, authentication status, etc. In particular, the possible event types can be {accepted, delayed, rejected sender address, rejected recipient address, unexpected connection termination, other errors}. The goal is to assign the remaining accounts {ak+1, ..., an} with a score yi in [−1, 1], where the sign of the score classifies a sender as either a spammer when negative or a legitimate sender otherwise. Moreover, the magnitude of yi reflects the confidence of the classification. The score can also be interpreted as the extent of legitimacy of the sender. In this paper, we limit our focus on the two categories: account that spams (spammer) and account that does not (legit./non-spammer).

III. LITERATURE SURVEY

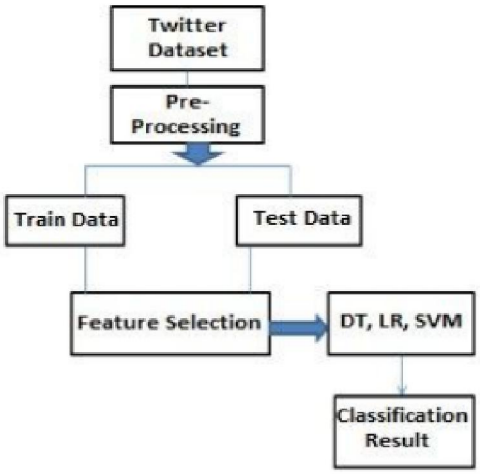
In research literature, many spam detection methods have been studied to provide various malware detection schemes and improve the performance, stability, scalability. The [1] have worked on evaluation of the spam detection performance on dataiu0hu 8i90set by using machine learning algorithms. The process of Twitter spam detection is done by using machine learning algorithms efficiently. Before classification, a classifier that contains the knowledge structure should be trained with the pre-labelled tweets. The [2] has proposed A Review on Spam Detection this study we conclude that there are different approaches in spam detection. Some approaches used features to detect spam with the Machine learning algorithms. Futures involved were content features, user-based features, URL-based features, features based on social graph. The [3] has worked on Identification of the Human or Bots Twitter Data that aiming at to identify the malicious activities in social contact using machine learning engineered techniques. The has worked on Twitter Spammers Detection. The proposed method compares the performance of three different Machine Learning algorithms in tackling this spam detection task. The experimental session involves a publicly available dataset. The [4] have proposed Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network that presents a methodology to detect and associate fake profiles on Twitter social network which are employed for defamatory activities to a real profile within the same network by analysing the content of comments generated by both profiles. [3] in their paper 'Using Social Network Analysis for Spam Detection' describes about the use of centrality in the social graph of a social networking site to predict spam detection such as the probability of a user is likely to post spam in a social network. In another research about Twitter, Wang mentioned about another technique which is the use of graph-based metrics to improve spam classification on a microblogging platform. [2] presented a scheme utilized for identifying spam URLs in social sites which have been used to protect users from links that are related with malware and other low-quality suspicious text. The behaviour has been analysed using two different schemes (i) initially, study the links posted by public on Twitter; (ii) secondly is how these links are accessed by the user..

IV. PROPOSED SYSTEM

Spam detection is a machine learning task where we want to determine which the general detection of a given document is using machine learning techniques and natural language processing, we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative. It is a really useful analysis since we could possibly determine the overall opinion about a selling object, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar) but it is also why it is very interesting to working on. In this project I choose to try to classify tweets from Twitter into "positive" or "negative" detection by building a model based on probabilities. Twitter

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, February 2024

is a microblogging website where people can share their feelings quickly and spontaneously by sending a tweet limited by 140 characters. You can directly address a tweet to someone by adding the target sign "@" or participate to a topic by adding a hashtag "#" to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.



V. EXPERIMENTAL RESULT

In practice, most websites use proprietary algorithms forspam detection. Decision making process is usually done with combination of spam filtering software and human analysis. As software for detection of untruthful reviews is not publicly available, results obtained by using a popular program for detecting bot spam are presented. Akismet is the most popular plugin for spam detection in blog comments. Its reported stats as of today are over 400 billion removed spam messages and accuracy over 99%. Although its algorithm is not publicly available, some information about its decision-making process is known. When comment is posted, Akismet compares its content to known spam messages in its database. If matches found, that comment is marked as spam. Setup for this experiment included creating five WordPress websites on a local machine and installing Akismet on each of them.To post a comment, unlogged user is required to enter his name and email address. Six types of comments aredefined based on their content:

1.Comments without alphanumerical characters
2.Random combination of characters
3.Random combination of words
4.Common phrases

Conducted tests show that comment is marked as spam if it contains random combination of non-alphanumerical characters (dashes, dots, commas…).Random combination of characters or words that were tried were not marked as spam immediately. However, if the same comment was posted more than five times in a short time interval it was marked as spam in all next cases. Common phrases and links were posted tens of times consecutively without being marked as spam. Common phrases are defined as short sentences or combination of words that are likely to be found on web. Comments containing text with links were marked as spam if posted more than five times in a relatively short amount of time. For example, comments "http://www.example.com" and "Example test" were not marked as spam when posted tens of times. However, comment "Example test":http://www.example.com" was marked as spam when posted over five times consecutively.

Experiment included posting these comments multiple times on the same website as well as on different websites. Comments on different websites were posted using the same email address, as well as different email addresses. Further, cases when times between posting comments were five seconds, one minute and ten minutes were tested. It was shown that in all these cases the same results were obtained. After a comment is marked as spam, using slight variations of it will also be blocked by Akismet. If one comment of a user was marked as spam manually, all further comments posted on the same website by that user were marked as spam automatically. However, other users were still

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, February 2024

able to post comments with that same content. This shows that manual labelling spam comments is used for detecting spam users, rather than determining if comment's content represents spam.

V. CONCLUSION

Nowadays, Spam Detection is a hot topic in machine learning. We are still far to detect the spam of the corpus of texts very accurately. In this project I tried to show the basic way of classifying tweets into real or spam category using a s baseline and how language models are related to the algorithm and can produce better result. The python code was used to run the machine learning algorithms which are support vector machine (SVM), decision tree, and the logistic regression to find out which algorithm is the suitable for Twitter to identify the spam and real accounts. From the results received, it is identified that the support vector machine algorithm provides more accurate results compared to other algorithms. Moreover, the python program identified eight best features that can be used to identify the account, whether it is spam or not they are ID, Created At, Viewed At,Friends Count, Followers_Count, Staus_count, Len_Screen_Name, Len_Profile.

VI. ACKNOWLEDGMENT

We express our heartfelt gratitude to our esteemed mentors and professors, especially, for their invaluable guidance in our academic and project endeavours. We also extend our thanks to the *Information technology*Department and its staff for their continuous support. Our sincere thanks go to Principal of Matoshri Aasarabai Polytechnic, Eklahare, Nashik for his support and permission to complete this project. We appreciate the assistance of our department's support staff, and we're grateful to our parents, friends, and all those who supported us throughout this project.

REFERENCES

[1] RohitV.Adagale, AniketC.Sanap, Anil V.Gitte, Prof. R. H. Kulkarni, "A Survey on Statistical Twitter Spam Detection Demystified: Performance, Stability and Scalability", International Journal of Interdisciplinary Innovative Research & Development (IJIIRD), ISSN: 2456-236X Vol. 02 Issue 02 | 2018.

[2] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," Inf. Sci., vol. 260, pp. 64–73, Mar. 2014.

[3] NambouriSravya, ChavanaSaipraneetha, S. Saraswathi, "Identify the Human or Bots Twitter Data using Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03 | Mar 2019.

[4] CLAUDIA MEDA, FEDERICA BISIO, PAOLO GASTALDO, RODOLFO ZUNINO DITEN, "Machine Learning Techniques applied to Twitter Spammers Detection", Recent Advances in Electrical and Electronic Engineering, ISBN: 978- 960-474-399-5, AUGUST, 2016.

[5] PatxiGal´an-Garc´ıa, Jos´eGaviria de la Puerta, Carlos LaordenG´omez, Igor Santos and Pablo Garc´ıaBringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Cyberbullying", IET Software, Vol. 6, Iss. 6, MAY 2014.

[6] 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.

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

[8] Verma, M., &Sofat, S. (2014). Techniques to detect spammers in twitter-a survey. International Journal of Computer Applications, 85(10).

[9] Wang, D., Irani, D., & Pu, C. (2011, September). A social-spam detection framework. In Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (pp. 46-54). ACM.

[10] Isa Inuwa-Dutse∗ , Mark Liptrott, IoannisKorkontzelos "Detection of spam-posting accounts on Twitter" 6, AUGUST , 2018.