Available online at www.asric.org

ASRIC Journal on Natural Sciences 1 (2021) 63-68

Using Machine Learning Techniques for predicting Email Spam

Hager Mohey Abohalfaya[1], Someya Mohsen[1]

[1]Department of Information System, Faculty of Computer & Information,
Al-minia University - Minia, Egypt

Corresponding Authors Email: Hagerabohalfaya95@gmail.com

Received 10 June 2021; revised 10 July 2021; accepted 22 August 2021

Abstract:
Email has become one of the most efficient and cost-effective methods of communication in recent years. However, as the number of email users grows, so does the number of spam emails. Email management has become a big and rising concern for both people and companies as a consequence of its sensitivity to abuse. Spam, or the unsolicited sending of unwanted email messages, is one example of misuse. Spam is defined as unsolicited bulk email, or email sent to a large number of people without their consent. Half of users receive 10 or more spam emails each day, while some users receive hundreds of unwanted emails per day. Online spiders are used by many spammers to discover email addresses on web pages. Because of spam emails can fill up the storage space of a file server quickly, they could cause a very severe problem for many websites with thousands of users for this in this study, we present a method for spam filtering using some machine learning techniques to predict whether an email is spam or no.

I. INTRODUCTION

Millions of individuals use email on a daily basis. Email is used by them for a number of purposes, including employment, research, and other activities. E-mail is a kind of electronic communication that allows two or more individuals who are connected to the Internet to communicate with each other. Due to the growing use of email and the incursions of online marketers, unwanted commercial email has become a problem on the internet. Unsolicited and undesired junk email delivered in bulk to an indiscriminate recipient list is known as spam email. Spam is typically sent for commercial objectives. Botnets, or networks of infected machines, may send it in large quantities. A spammer sends an email to millions of individuals with the expectation that just a small fraction of them will respond or interact with it. Email spam takes several forms, the most common of which is to advertise blatant frauds or shady business ventures. Emails are being utilized for more than simply communication; they are also used for work management and customer service. Email categorization was inspired by text classification in machine learning, and it is now accepted in a variety of forms, such as classifying emails into a spam folder, blocking spam email, and detecting the user's mood from the email body. Most recent email apps and services, such as Gmail and Hotmail, allow users to easily filter received emails based on the email subject and key tokens in the email body. This technique is suitable for individual work or home operators, as it eliminates the need to create token-based rules to sort emails into different folders. As the problem with which we are working is a classification problem, we not only need to have models that maximize the accuracy results of correct classified samples.

In e-mail filtering, two main techniques are used: knowledge engineering and machine learning. A set of rules must be established in the knowledge engineering method, according to which emails are classified as spam or ham. A collection of such rules should be developed either by the filter's user or by some other authority (e.g. the software company that provides a particular rule-based spam-filtering tool).Because the rules must be continually updated and maintained, which is a waste of time and inconvenient for most users, this technique yields no promising outcomes. Machine learning is more efficient than knowledge engineering since it does not need the specification of any rules. Instead, a set of training samples is used, which consists of a collection of pre classified e-mail messages. The categorization criteria are subsequently learned from these e-mail messages using a particular algorithm. Machine learning has been extensively researched, and there are several algorithms that may be utilized in e-mail filtering, Support vector machines, Neural Networks, K-nearest neighbour, Rough sets, and the artificial immune system are among them.

[4]. Idris, I. and Selamat, A., 2014. Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22, pp.11-27.
[5]. Scholar, M., 2010. Supervised learning approach for spam classification analysis using
[6]. Abu Naser, S., Zaqout, I., Ghosh, M. A., Atallah, R., & Alajrami, E. (2015). Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology. International Journal of Hybrid Information Technology, 8(2), 221-228.
[7]. Elzamly, A., Abu Naser, S. S., Hussin, B., & Doheir, M. (2015). Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods. Int. J. Adv. Inf. Sci. Technol, 38(38), 108-115.
[8]. Abu Naser, S. S., & Alhabbash, M. I. (2016). Male Infertility Expert system Diagnoses and Treatment. American Journal of Innovative Research and Applied Sciences, 2(4).
[9]. Qwaider, S. R., & Abu Naser, S. S. (2017). Expert System for Diagnosing Ankle Diseases. International Journal of Engineering and Information Systems (IJEAIS), 1(4), 89-101.
[10]. Abu Naser, S. S., & Al-Hanjori, M. M. (2016). An expert system for men genital problems diagnosis and treatment. International Journal of Medicine Research, 1(2), 83-86.
[11]. Kumaresan, T. and Palanisamy, C., 2017. E-mail spam classification using S-cuckoo search and support vector machine. International Journal of Bio-Inspired Computation, 9(3), pp.142-156.
[12]. Sharma, A. and Suryawanshi, A., 2016. A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. International Journal of Computer Applications, 136(6), pp.28-35. https://doi.org/10.5120/ijca2016908471
[13]. Shah, N.F. and Kumar, P., 2018. A Comparative Analysis of Various Spam Classifications. In Progress in Intelligent Computing Techniques: Theory,Practice, and Applications(pp. 265-271). Springer, Singapore.
[14]. Bassiouni, M., Ali, M. and El-Dahshan, E.A., 2018. Ham and Spam E-Mails Classification Using Machine Learning Techniques. Journal of Applied Security Research, 13(3), pp.315-331.
[15]. Sah, U.K. and Parmar, N., 2017. An approach for Malicious Spam Detection In Email with comparison of different classifiers.
[16]. Spambase.documentation at the UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLReposito ry.html, May 01, 2018, 06:54:55 pm.
[17]. Mohamad, M. and Selamat, A., 2015, April. An evaluation on the efficiency of hybrid feature selection in spam email classification. In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on (pp. 227-231). IEEE.
[18]. Hall, M.A., Practical Machine Learning Tools and Techniques. United State: Morgan Kauffman, 2011.
[19]. DeBarr, D. and Wechsler, H., 2012. Spam detection using random boost. Pattern Recognition Letters, 33(10), pp.1237-1244.
[20]. Zhang, Y., Wang, S., Phillips, P. and Ji, G., 2014. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based systems, 64, pp.22-31. https://doi.org/10.1016/j.knosys.2014.03.015.
[21]. Hamsapriya, T., D. Karthika Renuka, and M.Raja Chakkaravarthi. "Spam Classification based on Supervised Learning using Machine Learning Techniques." DIGITAL WORLD 2.04 (2011).
[22]. Dada, Emmanuel Gbenga, et al. "Machine learning for email spam filtering: review, approaches and open research problems." *Heliyon* 5.6 (2019): e01802.