An Investigation of Studies on Spam Filtering Based on Machine Learning

Yalu Wang[a]
Internet of Things, Beijing University of Technology, Beijing, 100000, China

Keywords: Machine Learning, Spam Email, Deep Learning.

Abstract: With the development of network communication, mail is one of the most widely used means of communication. But a lot of spam also followed, to the users and mail platform caused a lot of trouble. The common approach to filter spam is divided into six steps, and, respectively, the data collection, preprocessing, model building, model training and model testing, deployment. This paper also introduces logistic regression, decision tree, random forest, deep learning models and Atomic Orbital Search (AOS) algorithm in detail. Models such as Genetic Decision Tree Processing with Natural Language Processing (GDTPNLP) and Area Under curve (AUC) use machine learning to deal with spam. The logistic regression machine learning method is a machine learning technique and usually used to classify emails. Decision tree processing (DT) is a significant and effective machine learning model, widely employed as a predictive analysis technique for classification purposes. Random Forest is an ensemble learning classification and regression method for dealing with problems that involve grouping data into classes. In the research of electronic spam, some methods still have some limitations. The development of more efficient technologies is crucial in effectively addressing trends or advancements in spam features.

1 INTRODUCTION

Since E-mail is a cheap and fast communication tool in many fields and scenarios, it is used more and more widely (Magdy, 2022). Unwanted, unsolicited, mass-produced, useless email is called spam. The term "spammer" refers to the one who sends unsolicited messages (Dada, 2019). Spam can have a negative impact on mail server memory bandwidth, CPU efficiency, and user memory management (Dada, 2019). Spam on the Internet is a huge problem that needs to be solved. Machine learning and the development of a reliable tool that can filter spam are particularly important. Every year, the threat posed by spam email grows, accounting for more than 77% of all email traffic worldwide (Bahgat, 2016). Users who are the victims of network spam may suffer financial losses as a result of receiving spam.

Some well-known email sites, such as Yahoo, Gmail, and outlook, use different techniques in their spam filters through machine learning to reduce the threat and negative impact of spam (Dada, 2019). For example, Yahoo's spam filter can not only detect spam with pre-set instructions, but also generate new instructions and rules to identify and block spam messages and phishing messages through machine learning knowledge in spam detection. Google's machine learning model is 99.9% accurate (Dada, 2019). Moreover, Google's detection model integrates the tool of "Google Safe Browsing", which can be used to identify websites containing malicious URLs. Google's detection model also introduces a model that can delay the sending of emails for a while, so as to carry out comprehensive detection and investigation of network spam, because when they are detected together, Bad messages such as spam and phishing emails will be more easily detected. Many scholars and researchers have also proposed better research methods and ideas, such as probabilistic, decision tree, artificial immune system (Bahgat, 2016). Artificial Neural Networks (ANN) (Cao, 2004), case-based technique (Fdez, 2007) and Support Vector Machine (SVM) (Bouguila, 2009) to identify spam emails. Previous research papers demonstrate that for the filtering technology of information content, these methods can be applied to

[a] https://orcid.org/0009-0007-9068-9292

Wang, Y.
An Investigation of Studies on Spam Filtering Based on Machine Learning.
DOI: 10.5220/0012973400004508
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 790-793
ISBN: 978-989-758-713-9
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

the technology of spam filtering. This technology mainly filters spam through keywords, that is, the probability of the unique characteristics of spam. If an email exceeds this probability, then the information can be identified as spam and filtered out (Mason, 2003). The literature shows that the method of filtering spam using neural networks achieves moderate performance. Radial Base Function Neural Networks (RBFNN) and Multilayer Perceptron Neural Networks (MLPNNs) are two popular spam sorting techniques. Sanz Hidalgo goes into length on the problems with spam research, how it impacts users, and what they may do to lessen its effects. The study effort clarified the structure and operation of numerous machine learning techniques used to filter spam emails (Sanz, 2008).

Interest in filtering spam filters is growing rapidly among research groups worldwide. Due to some breakthroughs made by researchers in recent years, this work attempts to present a comprehensive examination of spam methods for identifying it. That is how the remaining portion of the chapter is structured. First the paper will recapitulate the organization, solutions as well as the main methods of spam detection in section 2. For example, how to use machine learning models to predict junk files and the details of their implementation. Then, in Section 3, the lessons learned from the approaches in these existing studies and possible challenges for the future. Section 4 concludes the article by reviewing the techniques covered here.

2 METHOD

2.1 Overview of Framework of AI-Based Algorithms for Spam Email Detection

In the field of spam filtering, it is usually detected by AI and machine learning methods, which usually includes the following six steps as shown in Figure 1. First, data can be collected from different email platforms or social platforms. A lot of mail messages include normal mail, junk mail, advertising mail etc. So based on that data, a classification of them can be implemented by categorizing what's junk and what's not junk, creating a good data set. Second, preprocess the data. For example, in order for the model to analyze the data, it is necessary to convert the string into a vector, and carry out some normalization operations. The creation of models for the application of machine learning, such as support vector machine learning and neural networks with artificial intelligence, is the third phase. The fourth step is to build a model. The fifth step is training, configuring some parameters to train the model. Then finally test the model on the test machine. For example, when the final prediction error is below a certain threshold, the model can be considered for deployment in the actual production environment for use, which is the sixth step, implementation of the model.

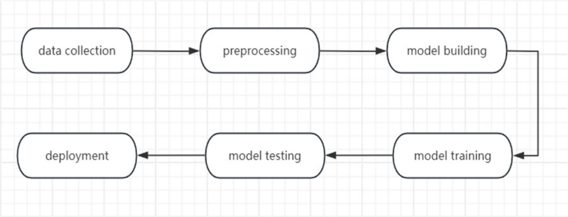

Figure 1: The workflow of spam email detection based on machine learning models (Photo/Picture credit: Original).

2.2 Traditional Machine Learning Models

2.2.1 Logistic Regression

The logistic regression machine learning method is a machine learning technique with efficiency, simplicity and high speed. It has been widely used to classify emails. An efficient spam filtering method, OAOS-LR, combines an improved Atomic Orbital Search (AOS) algorithm with an LR classification model (Manita, 2023). The detailed implementation process of this method is to seek to minimize the error by calculating the error at the output through the logic activation function (Manita, 2023). Orthogonal learning techniques are combined to enhance development and balance the trade-off between exploration and development, employing methodologies for preprocessing that incorporate frequency of terms and inverse text frequency (tf-idf) to reduce the limitations associated with feature-sensitive characters (Manita, 2023).

2.2.2 Decision Tree

Among some of the most notable and efficient artificial intelligence models is the decision-tree processing, a predictive analysis method that is commonly used for classification. It functions by first establishing a category, which in turn establishes subcategories, each of which establishes a subcategory of its own (Ismail, 2022). It is up to the programmer to decide how long this process lasts until the desired effect is achieved (Ismail, 2022). Its ultimate goal is to analyze textual information, which is influenced by a large number of features. It is called a decision tree because it is a structure similar to a flow tree, with the internal structure representing tests on attributes and each branch representing the test results (Ismail, 2022).

In this regard, the proposed Genetic Decision Tree Processing with Natural Language Processing (GDTPNLP) method utilizes the decision tree model. A confidence threshold is assigned to every single email message and is compared with the data set. The method that is suggested labels the message as spam if the measured value of the data set reaches the message threshold level which had been in place at the time (Ismail, 2022).

2.2.3 Random Forest

For challenges involving classifying data, Random Forest is an ensemble learning technique for classification and regression (Bazzaz Abkenar, 2021). During the training phase, the programmer defines a few decision trees that are then used to predict classes. This is accomplished by taking into account each individual tree's voting classes, with the class with the highest vote count being regarded as the output.Test datasets are classified into spammers and non-spammers using optimized RF(Bazzaz Abkenar, 2021). When the algorithm reaches the defined threshold of the number of iterations, the process is terminated, the evaluation parameters are calculated, and the results of Area Under curve (AUC) of the algorithm are calculated (Bazzaz Abkenar, 2021).

2.3 Deep Learning Models

Take, for example, Convolutional Neural Networks (CNNs), a well-known deep learning model that may identify pertinent characteristics in data in many tasks (Qiu, 2022). CNN is primarily split into three stages: first, building a word matrix; second, extracting text's hidden properties; and third, classifying the words into predetermined groups (Roy, 2020). Initially, messages vary in size because some messages have more words while others have fewer (Roy, 2020). Nevertheless, input of varying lengths is not accepted by the CNN network. Thus, post-fill approach is used to equalize the message lengths before building the message matrix (M). The classification task is completed by a fully linked multi-layer perceptron at the end of the proposed CNN model (Roy, 2020).

3 DISCUSSIONS

In the research of electronic spam, some methods still have some limitations. First, AI models are generally poorly interpretable, including spam sorting models. Some normal messages may be considered as spam, resulting in the recipient not receiving the message. How to improve the interpretability of spam classification model is a problem worth studying. Secondly, the universality of the model. For example, there may be different data on different platforms, as well as different data distributions, and it may be difficult to apply data patterns from one platform to another. Thirdly, when the content of the mail involves some fixed professional languages and professional terms, the corresponding keyword filtering may be missed or misjudged. Fourth, spam senders often change their email addresses, which poses a great challenge to the blacklist filtering technology. Fifth, a higher proportion of spam filters are susceptible to various types of attacks. For instance, copycat attacks can target Bayesian filters (Nelson, 2008). Additionally demonstrating the unending decline of control attacks against adversaries are Naive Bayes and AdaBoost (Dada, 2019).

It becomes imperative for establishing stronger technologies that can cope with newly developed trends or behaviors in spam that will allow it to stay out of detection by many spam filters. In the future, it is encouraged that the client email filter prioritizes the user's customized characteristics. This approach would enable the automatic identification of new spam organisms at any given moment, as well as the automated analysis and evaluation of those specimens based on their spam characteristics, so as to re-establish and upgrade the new spam feature code base. Furthermore, automated rule creation for mail screening could possibly be configured, culminating to the eventual automatic blocking of all forms of spam. As research into spam filtering technology continues, it is growing clear that feature extraction, rule development, detection, and forensics are all crucial components of the field, judgment and filtering measures of spam, the research focus has shifted from the single and single technology research to the research on the integration of multi-technology systems and collaborative spam filtering system. Additionally, create image trash filters that are more effective. The majority of spam filters are limited to sorting text-based spam (Dada, 2019). Nevertheless, a lot of cunning spammers transmit spam by hiding images in their text, which helps them avoid being picked up by filters (Dada, 2019). In addition, spam detection algorithms also need to rely on more advanced hardware or transmission mechanism (Deng, 2023; Liu, 2021; Sugaya, 2019) to achieve higher processing speeds and more accurate identification capabilities.

4 CONCLUSIONS

A comprehensive description of the machine learning and deep learning research in spam filtering is presented in this article. The approach predominantly investigates the logistic regression, logistic decision tree, and random forest strategies. After discussion and analysis, the main problems in the current field are not strong interpretability, not strong generality, wrong classification etc. In addition, this article also has some limitations, such as not taking into account the latest research methods, the summary is not comprehensive enough. Some methods are not covered. In the future, these new methods will be covered to form a complete method system.

REFERENCES

Bahgat, E. M., Rady, S., & Gad, W. 2016. An e-mail filtering approach using classification techniques. In The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt (pp. 321-331). Springer International Publishing.

Bazzaz Abkenar, S., Mahdipour, E., Jameii, S. M., & Haghi Kashani, M. 2021. A hybrid classification method for Twitter spam detection based on differential evolution and random forest. Concurrency and Computation: Practice and Experience, 33(21), e6381.

Bouguila, N., & Amayri, O. 2009. A discrete mixture-based kernel for SVMs: application to spam and image categorization. Information processing & management, 45(6), 631-642.

Cao, Y., Liao, X., & Li, Y. 2004, August. An e-mail filtering approach using neural network. In International symposium on neural networks (pp. 688-694). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6).

Deng, X., Li, L., Enomoto, M., Kawano, Y., 2019. Continuously frequency-tuneable plasmonic structures for terahertz bio-sensing and spectroscopy. Scientific reports, 9(1), p.3498.

Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R., & Corchado, J. M. 2007. SpamHunting: An instance-based reasoning system for spam labelling and filtering. Decision Support Systems, 43(3), 722-736.

Ismail, S. S., Mansour, R. F., Abd El-Aziz, R. M., & Taloba, A. I. 2022. Efficient E-mail spam detection strategy using genetic decision tree processing with NLP features. Computational Intelligence and Neuroscience, 2022.

Liu, Y. and Bao, Y., 2021. Review of electromagnetic waves-based distance measurement technologies for remote monitoring of civil engineering structures. Measurement, 176, p.109193.

Magdy, S., Abouelseoud, Y., & Mikhail, M. 2022. Efficient spam and phishing emails filtering based on deep learning. Computer Networks, 206, 108826

Manita, G., Chhabra, A., & Korbaa, O. 2023. Efficient e-mail spam filtering approach combining Logistic Regression model and Orthogonal Atomic Orbital Search algorithm. Applied Soft Computing, 144, 110478.

Mason, S. 2003. New Law Designed to Limit Amount of Spam in E-Mail.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I., Saini, U., ... & Xia, K. 2008. Exploiting machine learning to subvert your spam filter. LEET, 8(1-9), 16-17.

Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., & Guo, L. 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323.

Roy, P. K., Singh, J. P., & Banerjee, S. 2020. Deep learning to filter SMS Spam. Future Generation Computer Systems, 102, 524-533.

Sanz, E. P., Hidalgo, J. M. G., & Pérez, J. C. C. 2008. Email spam filtering. Advances in computers, 74, 45-114.

Sugaya, T., Deng, X., 2019. Resonant frequency tuning of terahertz plasmonic structures based on solid immersion method. 2019 44th International Conference on Infrared, Millimeter, and Terahertz Waves, p.1-2.