International journal of innovation in Engineering

journal homepage: www.ijie.ir

Research Paper

Using Support Vector Machine For Classification And Feature Extraction Of Spam In Email

Anuradha Reddy[a1], **M. Uma Maheswari**[b], **A. Viswanathan**[a], **G. Vikram**[a]

[a] *Department of Computer Science and Engineering, MRITS, Maisammaguda, Secunderabad, India*

[b] *Department of Information Technology, MRCET, Maisammaguda, Secunderabad, India*

ARTICLE INFO

Received: 05 March 2022

Reviewed: 20 March 2022

Revised: 25 March 2022

Accepted: 03 April 2022

Keywords:

Email, Spam, SVM, Classification, Feature Extraction

ABSTRACT

We provide an overview of recent and successful content-based e-mail spam filtering algorithms in this article. Our main focus is on spam filters based on machine learning and variants influenced by them. We report on significant ideas, methodologies, key endeavors, and the field's current state-of-the-art. The initial interpretation of previous work demonstrates the fundamentals of spam filtering and feature engineering in e-mail. We finish by looking at approaches, procedures, and evaluation standards, as well as exploring intriguing offshoots of recent breakthroughs and proposing directions of future research.

[1] **Corresponding Author**
anuradhareddy.anu@gmail.com

1. Introduction

Spam, or unwanted commercial or bulk e-mail, has recently become a big issue on the Internet. Spam is inefficient in terms of time, storage space, and communication bandwidth. Spam and fraudulent e-mail have been on the rise for years. According to recent statistics, spam accounts for 40% of all email, sending 15.4 billion emails every day and costing Internet consumers $ 355 million per year. At the moment, automatic e-mail filtering is the most effective way to deal with spam, and spammers and spam filtering technologies are in a vicious rivalry. Spammers have started employing a variety of devious techniques to get over filtering, including utilising strange sender addresses and/or inserting arbitrary characters to the beginning or end of the message subject line.

In e-mail filtering, two common approaches are knowledge engineering and machine learning. The knowledge engineering technique entails defining a set of rules that determine whether an email is spam or ham. A collection of such rules must be created either by the filter's user or by someone else with permission (such as software company that provides a special rule spam filtering tool). Because the rules should be necessary, none of the results obtained using this strategy are encouraging. Constant updates and maintenance are performed, which is inefficient and inconvenient for the majority of consumers (Krishna et al., 2019). Knowledge engineering methodologies are less efficient than machine learning approaches. There are no regulations that must be specified (Patan et al., 2020). Instead, a set of training samples, consisting of pre-classified e-mail messages, is used (Ghantasala et al., 2020a). Machine learning techniques have been extensively researched, and several methods can be applied to e-mail filtering. Naive Bayes, Support Vector Machines, Neural Networks, K-nearest neighbours, rough sets, and artificial immune systems are examples of these (Bhowmik et al., 2021; Sreehari & Ghantasala, 2019; Chandana et al., 2020; Kishore et al., 2021; Ghantasala et al., 2021a; Reddy et al., 2021; Mandal et al., 2021).

2. Initial Statement of the Problem

Spamming is a serious and popular assault that entails spreading unwanted messages, malware, and phishing over email to a large number of infected workstations. We chose this project because many people are attempting to deceive you by sending you phoney e-mails claiming that you have won $1,000 and that you must transfer the money as soon as you open this link. They will then stalk you and attempt to hack your information. Relevant e-mail is sometimes mistaken for spam.

Unwanted email annoys Internet users in a variety of ways, including:

- *Important email communications were ignored and/or delayed.*
- *Consumers are constantly looking for ISPs that offer regular email delivery modifications.*
- *Internet speed and bandwidth usage.*
- *Millions of PCs have been hacked.*
- *Billion-dollar losses all across the world.*
- *Theft identification.*
- *A rise in the number of viruses and Trojan horses.*
- *Spam can cause the mail server to fail and load up the hard drive.*

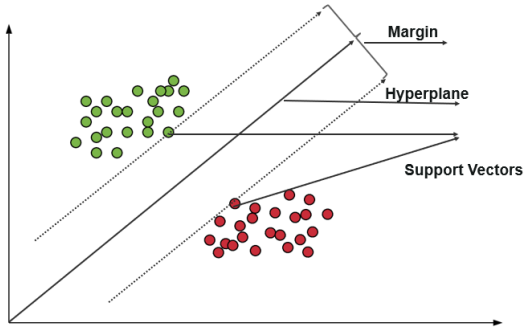

Fig. 1. Support Vectors

3. System and Methodology Proposed

When our data is completely divided into two classes, we can employ a support vector machine (SVM) (Kumari & Ghanatsala, 2020). An SVM classifies data by identifying an ideal hyperplane that separates all of a class's information purposes from those of another class (Ghanatsala et al., 2020b). For SVM, the hyperplane denotes the greatest difference between the two classes (CADe, 2020). The portion parallel to the hyperplane has a maximum width with no internal information points, as shown in the margin. One of the most essential economic kernel strategies has proven to be SVM (Ganatsala & Kumari, 2021a). SVM's excellent generalization capabilities are largely responsible for its success (Ghanatsala & Kumari, 2021b). SVM, unlike many other learning algorithms, produces sensible demonstrations without the requirement for preceding data (Ghanatsala et al., 2021b). Furthermore, the usage of positive fixed kernels in SVM can be interpreted as an associate degree embedding of the input field in a higher dimensional feature region where the classification is met, even if the exploitation does not explicitly use this feature area (Kishore et al., 2021). As a result, choosing a design for a neural network application is substituted by choosing an acceptable kernel for a support vector machine (Reddy et al., 2022). In binary classification, the support vector machine has proven to be effective (Gadde et al., 2022;, Pradeep Ghantasala et al., 2022; Ghantasala et al., 2021c). Its Theoretical Foundation is Sound, and the Learning Algorithm Rules are perfect. As a result, information classification becomes more stable. The only disadvantage is that it takes time and memory to process big amounts of data.

The proposed methodology for email spam detection will be discussed in the following section. The workflow is depicted in Figure 2.

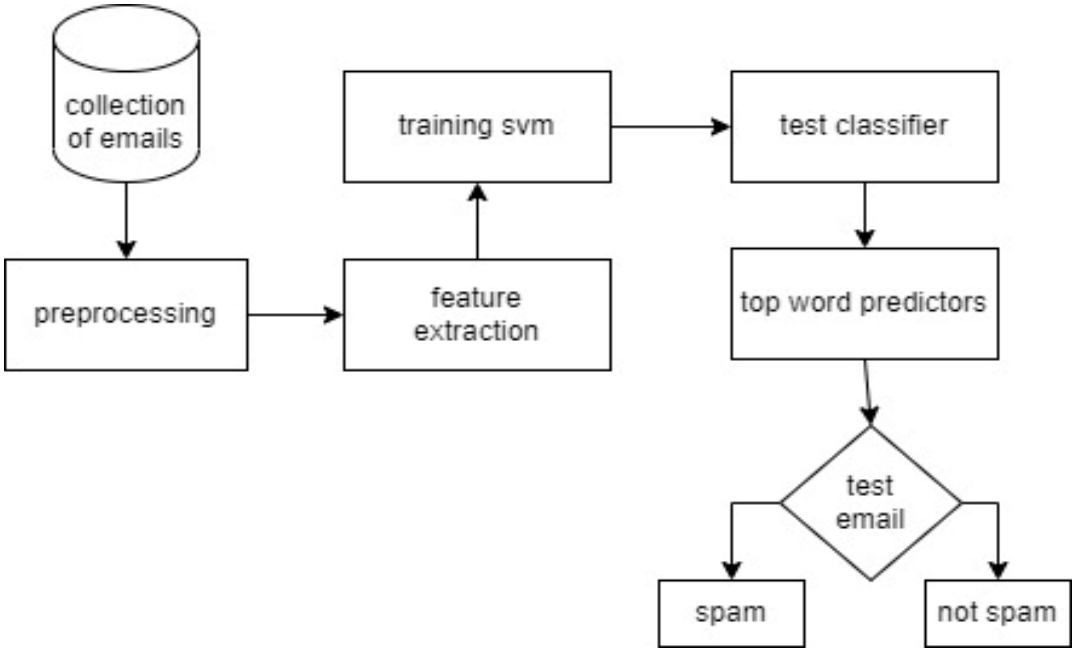

Fig. 2. Process flow of Current System

a. Pr-processing

The per-processing stage is used to filter out emails that aren't relevant and don't need to be there.

The following items are included in the per-processing phase:

- *Getting Rid of Numbers*
- *Get rid of the distinctive symbol*
- *Delete the URL*
- *HTML tags for task separation*
- *Performing Word Steaming*

b. Extraction of Characteristics

The email body is mined for crucial and relevant information using a feature extraction technique. Email 2D vector space features numbers are replaced by this feature. The dictionary list is used to map these features.

c. SVM Instruction

Spam email is utilized as a training tool. Spam content is included in the training datasets, which are used to train the classifiers. The classifier is now ready to classify spam emails after it has been trained.

d. Classifier for Tests

To ensure that the classifier is accurate, it is tested with a variety of training data. Until the proposed method, which provides 98 percent accuracy in email classification, is obtained.

e. Send a test email

Following the completion of the training phase, the classifier is given a sample email to categorize. Classifier generates output in the form of a 0 or 1, with 1 indicating spam and 0 indicating no spam.

4. Conclution

We looked at the general applicability of machine learning approaches and spam filtering in this work. To classify the communication as spam or ham, a study of the current state of the art algorithm was used. Efforts by many researchers to apply machine learning classifiers to solve the problem of spam were discussed. To get around spam filters, researchers looked into the evolution of spam messages over time. The basic structure of the email spam filter was recorded, as well as the operations involved in spam email filtering. The study looked at some of the publicly available datasets and performance measures that can be used to assess the efficacy of spam filters. A comparative assessment of machine learning techniques available in the literature was conducted to highlight the challenges of machine learning algorithms in successfully tackling the danger of spam. We have discovered several unsolved research issues with spam filters. In general, the quantity and quality of the literature we examined indicate that tremendous progress has been made and will continue to be made in this field. Following the discussion of outstanding issues in spam filtering, more study is needed to improve spam filter efficacy. It will create spam filters in order to maintain an active research area for academics and industry practitioners interested in machine learning techniques for effective spamming. We expect that this study will serve as a springboard for qualitative research in spam filtering employing machine learning, deep learning, and deep adversarial learning algorithms by research students.

5. Future Scope

However, the experiment made attempts to address the issue of spam e-mail. Legislative, behavioral, and technical measures proposed as solutions are insufficient. Spam e-mail and anti-Spam solutions are a cat-and-mouse game; spammers will come up with new ways every day. Spam e-mails are sent. This research has provided classification suggestions. Spam e-mail is a type of e-mail that is sent to Efforts will be made in the future to:

- *Accurate classification of ham E-mail as spam and spam as e-mail ham, with a zero percent (0%) error rate.*
- *Efforts will be made to prohibit phishing e-mail, which carries phishing attacks and is a source of concern these days.*
- *Work can also be expanded to make it safe from Denial of Service (DoS) attacks, which have now arisen in a distributed fashion and are known as distributed Denial of Service Attacks (DoS).*

References

- Bhowmik, C., Ghantasala, G. P., & AnuRadha, R. (2021). A Comparison of Various Data Mining Algorithms to Distinguish Mammogram Calcification Using Computer-Aided Testing Tools. In Proceedings of the Second International Conference on Information Management and Machine Intelligence (pp. 537-546). Springer, Singapore.
- CADe, M. (2020). CADx for Identifying Microcalcification Using Support Vector Machine. Journal of Communication Engineering & Systems, 10(2), 9-16p.
- Chandana, P., Ghantasala, G. P., Jeny, J. R. V., Sekaran, K., Deepika, N., Nam, Y., & Kadry, S. (2020). An effective identification of crop diseases using faster region based convolutional neural network and expert systems. International Journal of Electrical and Computer Engineering (IJECE), 10(6), 6531-6540.

- Gadde, S. S., Anand, D., Sasidhar Babu, N., Pujitha, B. V., Sai Reethi, M., & Pradeep Ghantasala, G. S. (2022). Performance Prediction of Students Using Machine Learning Algorithms. In Applications of Computational Methods in Manufacturing and Product Design (pp. 405-411). Springer, Singapore.
- Ghantasala, G. P., & Kumari, N. V. (2021a). Identification of Normal and Abnormal Mammographic Images Using Deep Neural Network. Asian Journal For Convergence In Technology (AJCT), 7(1), 71-74.
- Ghantasala, G. P., & Kumari, N. V. (2021b). Breast Cancer Treatment Using Automated Robot Support Technology For Mri Breast Biopsy. INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS, 1(2), 235-242.
- Ghantasala, G. P., Kallam, S., Kumari, N. V., & Patan, R. (2020a, March). Texture Recognization and Image Smoothing for Microcalcification and Mass Detection in Abnormal Region. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.
- Ghantasala, G. P., Kumari, N. V., & Patan, R. (2021a). Cancer prediction and diagnosis hinged on HCML in IOMT environment. In Machine Learning and the Internet of Medical Things in Healthcare (pp. 179-207). Academic Press.
- Ghantasala, G. P., Reddy, A. R., & Arvindhan, M. (2021c). Prediction of Coronavirus (COVID-19) Disease Health Monitoring with Clinical Support System and Its Objectives. In Machine Learning and Analytics in Healthcare Systems (pp. 237-260). CRC Press.
- Ghantasala, G. P., Reddy, A., Peyyala, S., & Rao, D. N. (2021b). Breast Cancer Prediction In Virtue Of Big Data Analytics. INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS, 1(1), 130-136.
- Ghantasala, G. P., Tanuja, B., Teja, G. S., & Abhilash, A. S. (2020b). Feature Extraction and Evaluation of Colon Cancer using PCA, LDA and Gene Expression. Forest, 10(98), 99.
- Kishore, D. R., Syeda, N., Suneetha, D., Kumari, C. S., & Ghantasala, G. P. (2021). Multi Scale Image Fusion through Laplacian Pyramid and Deep Learning on Thermal Images. Annals of the Romanian Society for Cell Biology, 3728-3734.
- Krishna, N. M., Sekaran, K., Vamsi, A. V. N., Ghantasala, G. P., Chandana, P., Kadry, S. & Damaševičius, R. (2019). An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using EEG signals. IEEE Access, 7, 77905-77914.
- Kumari, N. V., & Ghantasala, G. P. (2020). Support Vector Machine Based Supervised Machine Learning Algorithm for Finding ROC and LDA Region. Journal of Operating Systems Development & Trends, 7(1), 26-33.
- Mandal, K., Ghantasala, G. P., Khan, F., Sathiyaraj, R., & Balamurugan, B. (2020). Futurity of Translation Algorithms for Neural Machine Translation (NMT) and Its Vision. In Natural Language Processing in Artificial Intelligence (pp. 53-95). Apple Academic Press.
- Patan, R., Ghantasala, G. P., Sekaran, R., Gupta, D., & Ramachandran, M. (2020). Smart healthcare and quality of service in IoT using grey filter convolutional based cyber physical system. Sustainable Cities and Society, 59, 102141.
- Pradeep Ghantasala, G. S., Nageswara Rao, D., & Patan, R. (2022). Recognition of Dubious Tissue by Using Supervised Machine Learning Strategy. In Applications of Computational Methods in Manufacturing and Product Design (pp. 395-404). Springer, Singapore.
- Reddy, A. R., Ghantasala, G. S., Patan, R., Manikandan, R., & Kallam, S. (2021). Smart Assistance of Elderly Individuals in Emergency Situations at Home. In Internet of Medical Things (pp. 95-115). Springer, Cham.

- Reddy, A., Gude, V., Mamatha, K., & Rao, D. N. (2022). Smart Waste Management Systems by Using Automated Machine Learning Techniques. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 2(04), 36-45.
- Sreehari, E., & Ghantasala, P. G. (2019). Climate Changes Prediction Using Simple Linear Regression. Journal of Computational and Theoretical Nanoscience, 16(2), 655-658.