Spam Email Classification by Hybrid Feature Selection with Advanced Machine learning Algorithm – Future Perspective

B. Vivekanandam[1], Balaganesh[2]

[1,2]Associate Professor, Faculty of computer science and Multimedia, Lincoln University College, Malaysia

E-mail: [1]vivekresearch2014@gmail.com, [2]balaganesh@lincoln.edu.my

Abstract

Recently, email has become a common way for people to communicate and share information both officially and personally. Email may be used by spammers to transmit harmful materials to Internet users. The data must be protected from unauthorized access, which necessitates the development of a reliable method for identifying spam emails. As a result, a variety of solutions have been devised. An innovative hybrid machine learning strategy for effectively detecting spam emails has been discussed in this study. This means that identifying spam and non-spam email is a difficult process. Spam email categorization has undergone a significant evolution in recent years, as shown by the research given below. For locating spam, this study uses a mixed approach. Different email categorization algorithms are used to rank them for future perspective.

Keywords: Hybrid Neural Networks, email spam classification, Feature selection; artificial bee colony; ant colony optimization

1. Introduction

Email has become an essential element of many people's daily routines. They utilize email for a variety of reasons, including work, school, and so on. In part, this is due to the fact that it has become the most popular, most affordable, and quickest method of contacting one another. On a daily basis, it is not uncommon for a user to get hundreds of emails. Spam makes up around 92% of all emails [1, 2]. Recently, the advertising through mail for a wide range of goods and services, such as medications and electronic goods, jewellery, stocks, gambling, loans etc. and other potentially harmful programs (such as phishing) have increased.

DOI: https://doi.org/10.36548/jscp.2022.2.002

Received: 17.05.2022, received in revised form: 12.06.2022, accepted: 26.06.2022, published: 06.07.2022

An internet user's inbox is referred to as "spam" when they receive undesired trash mail. Spammers are able to send millions of emails throughout the globe at no expense because of this convenience. Filtering and detection methods are employed most of the time. They use simply the message's text and a few additional factors to determine whether a message is spam or not [3-6].

Spam is ubiquitous in today's online dialogues and is well-known for degrading the effectiveness of the medium in which it appears. Anti-spamming tactics have been used on a variety of additional online venues to protect them against spam attacks. In spite of the fact that spam detection has been studied extensively, the present notion of spam detection traction strategies is still unsatisfactory. Bulk spam is mostly caused by the web graph, an ever-changing and intractable social media network [8]. Spam is made possible by the fact that user-generated material on social media is unrestricted and unfettered by any standards or controls [7-9].

Using a search engine to find information online is essential. Search engine spam is becoming a major factor in most search engines. The content and links of a website are manipulated to generate revenue for a spammy search engine page. Search engine spammers is the terms used to describe those who send spam to search engines. Spamdexing is another name for search engine spam. It is a combination of the terms "Spam" and "Indexing" that is known as Spamdexing. Convey invented the word "Spamdexing" in 1996 to describe the purposeful manipulation of search engine indexes, which he has been doing since the early 1990s.

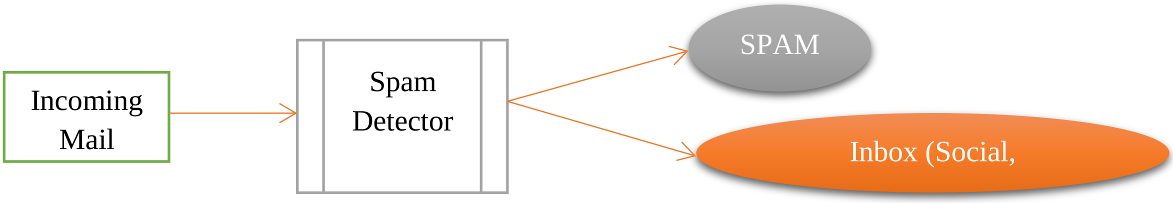

Figure 1. Block diagram of Conventional Detector

Spam has become a major societal concern because of the settings that encourage it to circulate online. Our inboxes are plagued with spam mail. Most email users spend a significant amount of time each day deleting spam emails, which takes up storage space on servers and consumes bandwidth on networks. As a result, being able to tell the difference between real and spam email is critical. There has been a rise of photo spam as an email spam, new phenomena despite the effectiveness of text-based anti-spam solutions [5].

Randomization and a variety of picture-taking methods make photo spam more difficult to detect than other types of spam [10]. Figure 1 shows simplified blocks of conventional SPAM detector.

The interaction between the colleagues and friends is commonly through any communication medium especially email and chatting about the discussion of technology development. The job advertisements and recruiting, healthcare communications, transactional financial information and inter- and intra-organizational contact are only a few examples of the vast range of human activities in which an email has been widely used [11]. However, spam detection technologies have made it more difficult for consumers to use email. In order to avoid being tricked by spam, you should never open an email from a person you don't know.

Spam must be prevented from reaching users' inboxes by employing software approaches that identify between spam and non-spam email. It is countable and measures details throughout the globe; the reported accuracy demonstrates that more effort is needed in this area.

2. Literature Survey

Various methods for recognizing spam photos have been examined in this investigation. One method uses Principal Component Analysis (PCA) to evaluate spam photos, and the scores are calculated by projecting the frames to the proprietary spaces that arise from this study. Getting a wide variety of picture attributes and picking a suitable subset of Support Vector Machine (SVM) devices is the focus of the second method to image classification. The machines are to be thanked for the excellent accuracy and low difficulty of these two ways of identification. In actuality, using PCA or SVM, a fresh spam photo dataset may be created. Image spam detection should benefit from this additional data gathering. When it comes to spotting spam emails on mobile social networks, email categorization is an effective and extensively utilized strategy.

2.1 Supervised learning algorithms

There are many methods included in supervised learning approach, as follows;

1. Naive Bayes (NB),
2. K-Nearest Neighbour (k-NN),

3. Support Vector Machines,
4. Ensemble Learning,
5. Decision Trees.

According to Gebali et al., the proposed neural network classifier was used in the email categorization for junk mail control unit. For example, the authors developed a Naive Bayes classifier that employs a look-up table to minimize complexity of the process for the Logarithmic Number System (LNS). It seems that their system can manage a high volume data [12].

Meizhen et al., presented a fuzzy based detection of spam control and its filter system that can compute information gathered to study and select out behavioural elements of emails. Using a decision tree-based technique, spam emails may be identified, and ensemble learning is shown in [13]. According to public dataset assessments and benchmark algorithms like SVM, KNN and Naive Bayes, the recommended strategy was superior [14].

According to Firte et al., spam filters may be generated using KNN and SVM. For the learning step, they developed neural network based on kNN algorithm to detect the junk mail in an email reception domain. Datasets and most frequently used terms in texts may be updated while this model is being evaluated [15].

There are many advanced classification algorithms that Drucker et al., compared SVM with, when determining if an email message is real or spam. In the case of binary features, Support Vector Machines outperformed the other three approaches on both datasets [16].

2.2 Semi-Supervised learning procedure

Unlabelled and labelled data were classified using semi-supervised learning because of the vast quantity of the dataset that is providing good results during validation and testing section, and the authors of [17] showed how classification accuracy might be improved. Semi-supervised classification employs labelled public domain emails as a training set to categorize user's emails, during the neural network gets trained with large standard dataset for better classification procedure.

The authors of [18] presented advanced ensemble classifier to categorize the junk mail detection in an easier way. SVM to identify email messages with high accuracy, was developed by Gao et al. to identify image spam emails. For completely managed learning, the cost of obtaining enough labelled data to train was too high as compared to fully supervised learning approaches. A huge quantity of unlabelled data as well as some labelled data may be used to detect erroneous emails and train a classification algorithm [19].

Using semi-supervised spam filtering, the authors in [20] take advantage of a specific circumstance that can provide the better accurate detection rate for email spam classification. The authors presented two spam filtering strategies for this case. Their methodology outperformed novel approach through spam filtering process in the assessment.

3. Neural Networks

3.1 Hybrid Neural Network

Many of the black circles indicate the pre-processed word embedding layer, which is the initial layer of the proposed technique. Convolution layers with varying window widths and feature maps make up the next layer in the CNN stack. Several experiments using dilated convolutions in temporal convolution layers have also been conducted, with encouraging results. Regularization may also be accomplished by the practice of dropping out of school. An encoder of the data focusing on a single context vector is sequential. Besides, it uses soft max activation function of many neuron setup in the trained neural network at tail part. More generalizable, significant, and abstract characteristics are detected using CNN's hierarchical representation. The suggested technique detects additional characteristics and enhances generalization in email spam classification. The Fast Text (FT) may be used to start word embedding by using sub-word information to obtain representations for less frequent terms [21-24]. Figure 2 shows some advanced feature selection in hybrid neural network in Artificial Intelligence (AI) model training.

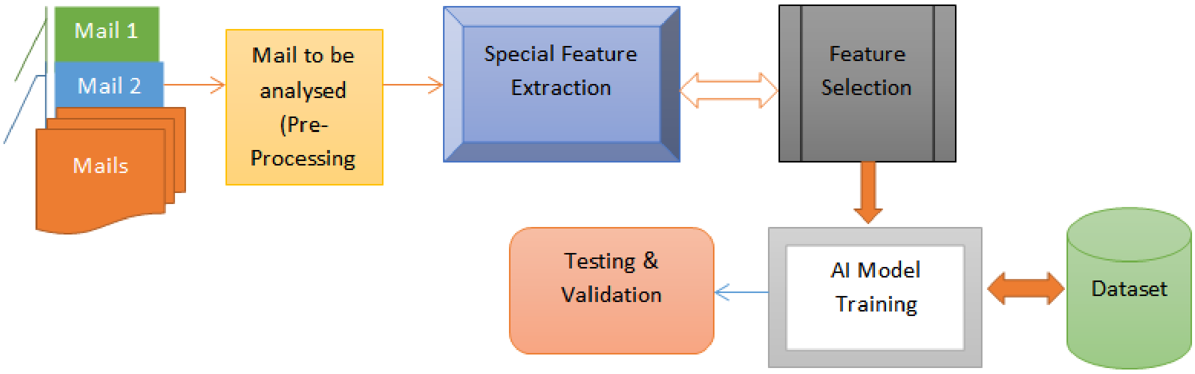

Figure 2. Recent advanced SPAM detector

3.2 Hybrid Genetic Algorithm based Decision Tree

Using metaheuristic search methods, the Genetic Algorithm attempts to replicate the process of natural selection. In the Evolutionary Algorithm, "survival of the fittest" is a guiding principle. The feature selection issue has a possible solution in every chromosome in the population. The chromosomes are a collection of genes that make up a person's DNA. In the same way, the potential solution is a collection of characteristics. Random selection is used to populate the initial population. Then, each chromosome is generated using a random binary vector. In order to identify search engine spam, the suggested method employs a number of different components. There are two options for subgroup selection. N-bit binary vectors are often used to illustrate possible solutions. The functionality is not activated if a bit is set to zero. The quality is decided if the bit is set to 1.

In the Genetic Algorithm, there are selection, crossover, and mutation operators for evaluating potential solutions. In order to pick the best feature, these evolutionary operators are adjusted as mean proportional selection, child occurrence-based crossover, and adaptive mutation. The fitness function is a way to gauge a person's overall worth. A high fitness rating is used to identify the person, and their information is then saved separately. The precise procedure is done over and over again until the best possible option is discovered. It is possible to get stuck in a local optimum with the genetic algorithm's quick convergence. Because of this, it uses tabular searching to avoid the local optimum. The tabular search maintains a list of features to prevent an unending loop of feature selection.. Each feature's classification rate is analyzed. A decision tree is built with the greatest overall quality possible from the characteristics selected. A 10-fold cross-validation test is used to verify the results. Finally, the user interface receives the categorization results. Different populations and web page bank sizes are used for each iteration [25 -27].

3.3 Ensemble learning

It's a method of improving the student under the guidance of an instructor. Ensemble classification is based on the premise that many models may be combined by creating "experts" and allowing them to vote. There are several areas where it has shown to be more effective than other machine learning paradigms. There are a number of different techniques to analyse the data, and some of them try to build a single intelligible framework. Boosting is one of the most common strategies. Instead of using voting and averaging, this strategy weighs the models according to their performance. On the new model, this promotes to become a "expert" for the cases that were misclassified by previous models. As long as the learning method is compatible with weighted resampling, it is possible to boost without penalties. Adaptive boosting is the approach employed in the ensembles.

3.4 Artificial Bee Colony (ABC)

This approach was introduced and described by Karaboga in 2005 as an optimization method. Swarm intelligence-based systems like this one replicate the behavior of a honeybee hive. Employee bees, observers, and scouts make up the ABC model's three groupings. Only one artificially hired bee is used for each food source. Bees that are employed proceed to the beginning of the input data in various section of the process. The ABC approach may be used in a variety of ways to find the best collection of attributes for spam detection. Figure 3 shows some advancement blocks of ABC.

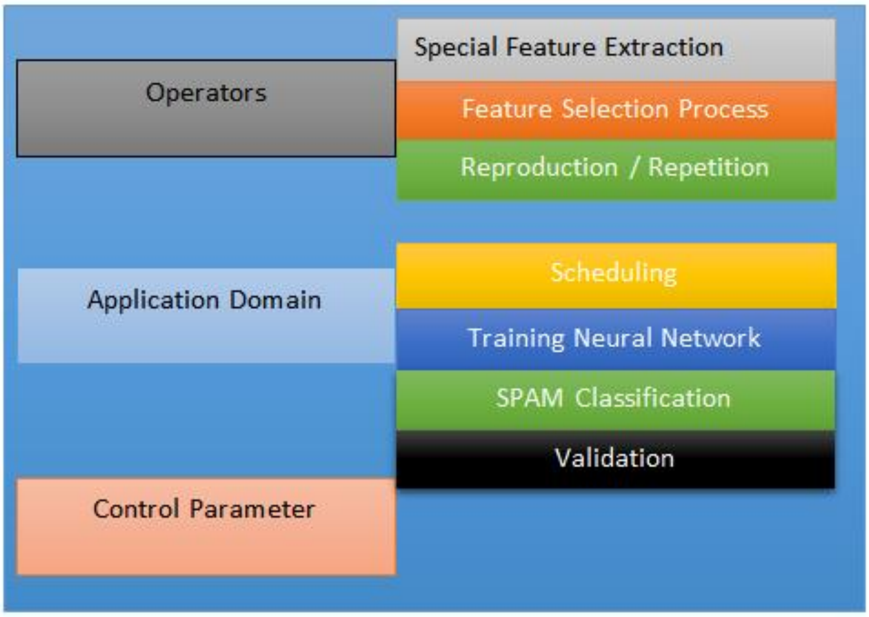

Figure 3. Advancement of Artificial Bee Colony

3.5 Ant Colony Optimization (ACO)

Many computational problems may be solved using Ant Colony Optimization (ACO) techniques, which are probabilistic. This approach is based on ant behavior, which seeks the most direct route between the colony and the food supply. It was Marco Dorigo's Ph.D. thesis in 1992 that first presented the ACO approach. There are three phases to this method: First, ants randomly build a solution; next, the pathways established by the ants are compared; and lastly, the pheromone levels on each edge are updated. Several sorts of study employ ACO in spam detection with acceptable results.

3.6 Random Forest Classifier (RF)

Decision trees are used in a Random Forest (RF) classification technique. Many of the researchers are trying to make it one of the most important and accurate algorithms ever devised. To boost the tree's diversity, RF uses a variety of bootstrap samples taken from the data. Since there are only a limited number of available features, the number of features picked is less than the total number of possible features. Thus, RF's speed and efficiency in dealing with huge datasets is a considerable benefit [28].

4. Conclusion

Every internet user has to deal with spam at some point. Two spam detection and prevention methods are proposed in this study. The email spam classification dataset is used to test model one for spam detection. It's probable that the future study may concentrate on further enhancement in order to achieve classification accuracy that is very precise and interpretable. Extending the current approach to scenarios when there are several sorts of spam emails, may be considered as another potential path. New technologies and more powerful hybrid swarm intelligence approaches may be used by researchers. When compared to other techniques of classification, the accuracy of an integrated hybrid genetic algorithm-based decision tree approach is higher. However, the current process requires more time to get the best possible characteristics. The time it takes to form a decision tree is reduced after picking an optimized element. This search engine spam is detected by the present system. As a consequence, the search engine does not have to deal with an increase in site traffic or extra crawling, indexing, or query processing. Users may also avoid being bombarded with useless results from spammers when browsing the web. More tools for identifying spam in search engines may be added in the future. By using additional classification algorithms, optimization-based methods, and local search algorithms, it is also capable of generating novel approaches to the problem area.

References

[1] Li Z, Shen H. ”Soap: A social network aided personalized and effective spam filter to clean your e-mail box.” in Proceedings of INFOCOM, 2011, pp. 1835-1843

[2] BIGGIO, B., FUMERA, G., PILLAI, I., and ROLI, F. (2007) Image spam filtering using visual information. In: Proceedings of the 14th International Conference on Image

Analysis and Processing, Modena, September 2007. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 105-110.

[3] BOUZERDOUM, A., HAVSTAD, A., and BEGHDADI, A. (2004) Image quality assessment using a neural network approach. In: Proceedings of the 4th IEEE International Symposium on Signal Processing and Information Technology, Rome, December 2004. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, pp. 330-333.

[4] Zhang,Y.,Wang,Y.,Gong,D., Sun, X. (2021). Clustering-guided particle swarmfeature selection algorithm for high-dimensional imbalanced data with missing values. IEEE Transactions on EvolutionaryComputation. DOI 10.1109/TEVC.2021.3106975.

[5] Song, X., Zhang, Y., Gong, D., Gao, X. (2021). A Fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for High-Dimensional Data. IEEE Transactions on Cybernetics. pp. 1–14. DOI 10.1109/TCYB.2021.3061152.

[6] Song, X., Zhang, Y., Guo, Y., Sun, X., Wang, Y. (2020). Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. IEEE Transactions on Evolutionary Computation, 24(5), 882–895. DOI 10.1109/TEVC.2020.2968743.

[7] Hu, Y., Zhang, Y., Gong, D. (2021). Multiobjective particle swarm optimization for feature selection with fuzzy cost. IEEE Transactions on Cybernetics, 51(2), 874–888. DOI 10.1109/TCYB.2020.3015756.

[8] Bilge, D., Bahriye, A. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 91, 1–18. DOI 10.1016/j.asoc.2020.106229.

[9] Faris, H., Aljarah, I., Al-Shboul, B. (2016). A Hybrid approach based on particle swarm optimization and random forests for email spam filtering. 8th International Conference on Computational Collective Intelligence. Greece.

[10] Alqatawna, J., Faris, H., Jaradat, K., Al-Zewairi, M., Adwan, O. (2015). Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution. International Journal of Communications, Network and System Sciences, 8(5), 118–129. DOI 10.4236/ijcns.2015.85014.

[11] Khoi-Nguyen, T., Alazab, M. (2013). Towards a feature rich model for predicting spam emails containing malicious attachments and URLs. Eleventh Australasian Data Mining Conference, pp. 161–171. Canberra, Australia.

[12] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, “Binary LNS-based naïve Bayes hardware classifier for spam control,” Proceedings - IEEE International Symposium on Circuits and Systems. IEEE, pp. 3674–3677, 2006.

[13] W. Meizhen, L. Zhitang, and Z. Sheng, “A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree,” 2009 Ninth IEEE International Conference on Computer and Information Technology. IEEE, 2009 [Online]. Available: http://dx.doi.org/10.1109/cit.2009.136

[14] L. Shi, Q. Wang, X. Ma, M. Weng, and H. Qiao, “Spam email classification using decision tree ensemble,” Journal of Computational Information Systems, vol. 8, no. 3, pp. 949–956, Mar. 2012.

[15] L. Firte, C. Lemnaru, and R. Potolea, “Spam detection filter using KNN algorithm and resampling,” Proceedings - 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing, ICCP10. IEEE, pp. 27–33, 2010 [Online]. Available: http://dx.doi.org/10.1109/iccp.2010.5606466

[16] H. Drucker, D. Wu, and V. N. Vapnik, “Support vector machines for spam categorization,” IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1048–1054, 1999, doi: 10.1109/72.788645. [Online]. Available: http://dx.doi.org/10.1109/72.788645.

[17] V. Cheng and C. H. Li, “Combining supervised and semi-supervised classifier for personalized spam filtering,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4426 LNAI. Springer Berlin Heidelberg, pp. 449–456, 2007

[18] V. Cheng and C. h. Li, “Personalized Spam Filtering with Semi-supervised Classifier Ensemble,” 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI’06). IEEE, 2006 [Online]. Available: http://dx.doi.org/10.1109/wi.2006.132

[19] Y. Gao, M. Yang, and A. Choudhary, “Semi Supervised Image Spam Hunter: A Regularized Discriminant EM Approach,” Advanced Data Mining and Applications. Springer Berlin Heidelberg, pp. 152–164, 2009.

[20] J. S. Whissell and C. L. A. Clarke, “Clustering for semi-supervised spam filtering,” ACM International Conference Proceeding Series. ACM Press, pp. 125–134, 2011 [Online]. Available: http://dx.doi.org/10.1145/2030376.2030391.

[21] Olatunji, S.O.: ‘Improved email spam detection model based on support vector machines’, Neural Comput. Appl., 2019, 31, (3), pp. 691–699.

[22] Jain, G., Sharma, M., Agarwal, B.: ‘Optimizing semantic LSTM for spam detection’, Int. J. Inf. Technol., 2019, 11, (2), pp. 239–250.

[23] Yang, H., Liu, Q., Zhou, S., et al.: ‘A spam filtering method based on multimodal fusion’, Appl. Sci., 2019, 9, (6), p. 1152.

[24] I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, “A combined negative selection algorithm–particle swarm optimization for an email spam detection system”, Engineering Applications of Artificial Intelligence, vol. 39, pp.33-44, 2015.

[25] A. Adeleke, et al., “A two-step feature selection method for quranic text classification,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 16, no. 2, pp. 730-736, 2019.

[26] Maneet Singh. “Classification of spam email using intelligent water drops algorithm with naive bayes classifier” In Progress in Advanced Computing and Intelligent Engineering, pages 133–138. Springer, (2019).

[27] Surender Singh and Ashutosh Kumar Singh. Web-spam features selection using cfs-pso. Procedia computer science, 125:568–575, (2018).

[28] Sudeep D Thepade, Deepa Abin, Rik Das, and Tanuja Sarode. Human face gender identification using thepade’s sorted n-ary block truncation coding and machine learning classifiers. International Journal of Intelligent Engineering Informatics, 8(2):77–94, (2020).

Author’s biography

B. Vivekanadam is an Associate Professor in the Department of Computer Science and Multimedia at Lincoln University College in Malaysia. His major area of research are machine learning, neural network algorithms, image processing, video and signal processing, cloud computing, deep learning, artificial intelligence, object recognition, complex feature extraction and vision graphics.

Balaganesh is currently working as an Associate Professor in the Department of Computer Science and Multimedia at Lincoln University College in Malaysia. His area of research includes web mining, IoT, data science, machine learning, blockchain, signal processing and data mining.