Detection of Social Network Spam by using Naïve Bayes Classification

Mrs.V.Gayathri[1] P.Subha Sree[2] M.Shesha Vardhini[3] S.Srimathi[4]
[1]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]K.L.N. College of Engineering, Tamil Nadu, India

Abstract — To know the classification of email spam using different machine learning algorithms. To develop machine learning-based methods for detecting spammers on E-mail. Spam refers to all emails of unsolicited content that arrive in a user's email box. Spam can often lead to network congestion and blocking or even damage to the system for receiving and sending electronic messages. To implement different machine learning algorithms such as Naive Bayes classification Algorithm.

Keywords: Social Network Spam, Naïve Bayes Classification

I. INTRODUCTION

This study purpose to know the classification of email spam with ham using different machine learning algorithms. Now we have to develop machine learning based methods for detecting spammers on E-mail. Spam refers to all emails of unsolicited content that arrive in a user's email box. Spam can often lead to network congestion and blocking or even damage to the system for receiving and sending electronic messages.

II. RELATED WORKS

A. D.M.Ablel-Rheem, ''Hybrid feature selection and ensemble learning 899 method for spam email classification,'' Int. J. Adv. Trends Comput. Sci. 900 Eng., vol. 9, no. 1.4, pp. 217–223, Sep. 2020.

The data mining techniques produce good work in many domains. The spam emails are becoming a serious dilemma and an important matter to have different solutions, and enhanced methods and algorithms. Using Ensemble methods which are well-established classifiers. In this paper data mining techniques used to classify spam email using the UCI spam base dataset. The results achieved by the machine learning tools and techniques, and the Ensemble learning methods, after applying feature selection methods on the data set; which gave better result, and better classification accuracy. For the evaluation method used the cross-validation for testing and training option, and the confusion matrix to show the accuracy and the performance result of the chosen classifiers; which are Naïve Bayes, decision tree, ensemble boosting and ensemble hybrid boosting classifiers.

B. S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli, and W. A. H. M. Ghanem, 908 ''E-mail spam classification using grasshopper optimization algorithm 909 and neural networks,'' Comput., Mater. Continua, vol. 71, no. 3, 910 pp. 4749–4766, 2022.

Spam has turned into a big predicament these days, due to the increase in the number of spam emails, as the recipient regularly receives piles of emails. Not only is spam wasting users' time and bandwidth. In addition, it limits the storage space of the email box as well as the disk space. Thus, spam detection is a challenge for individuals and organizations alike. To advance spam email detection, this work proposes a new spam detection approach, using the grasshopper optimization algorithm (GOA) in training a multilayer perceptron (MLP) classifier for categorizing emails as ham and spam. Hence, MLP and GOA produce an artificial neural network (ANN) model, referred to (GOAMLP). Two corpora are applied Spam Base and UK-2011 Web spam for this approach. Finally, the finding represents evidence that the proposed spam detection approach has achieved a better level in spam detection than the status of the art.

C. S.A.A.Ghaleb, M.Mohamad, S. A. Fadzli, and W. A. H. M. Ghanem, 912''Spam classification based on supervised learning using grasshopper optimization algorithm and artificial neural network,'' Commun. Comput. Inf. 914 Sci., vol. 1347, pp. 420–434, Dec. 2021.

The electronic mailing system has in recent years become a timely and convenient way for the exchange of multimedia messages across the cyberspace and computer networks in the global sphere. This proliferation has prompted most (if not all) inboxes receiving junk email messages on numerous occasions every day. Due to these surges in spam attacks, a number of approaches have been proposed to lessen the attacks across the globe significantly. The effect of previous detection techniques has been weakened due to the adaptive nature of unsolicited email spam. Hence, resolving spam detection (SD) problem is a challenging task. A regular class of the Artificial Neural Network (ANN) called Multi-Layer Perceptron (MLP) was proposed in this study for email SD. The main idea of this research is to train a neural network by leveraging a new nature-inspired metaheuristic algorithm referred to as a Grasshopper Optimization Algorithm (GOA) to categorize emails as ham and spam. Evaluation of its performance was performed on an often-used standard dataset. The results showed that the proposed MLP model trained by GOA achieves high accuracy of up to 94.25% performance compared to other optimization.

III. PROBLEM STATEMENT

Spam is a waste of time to the user since they have to sort the unwanted junk mail and it consumed storage space and communication bandwidth. Spammers expect only a small number of recipients to respond or interact with their message, but they can still swindle their way to a big payday because they can easily send their shady message to so many emails addresses in a single stroke. That is why spam continues to be a big problem in the modern digital economy.

IV. METHODOLOGY

The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. In our project, we can use

the naïve bayes algorithm such as Multinomial model for classifying the email message is spam or non-spam.

V. IMPLEMENTATION

A. Modules:

1) Data Selection:
The email dataset was collected from dataset repository. The data selection is the process of detecting the mail (i.e.) spam or ham. In python, we have to read the dataset by using the pandas packages.Our dataset, is in the form of '.csv' file extension.

2) Data Preprocessing:
Data pre-processing is the process of removing unwanted data from the dataset. Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning. Missing data removal,Encoding Categorical data, Missing data removal is the process, the null values such as missing values and Nan values are replaced by 0.

3) NLP:
NLP is a field in machine learning with the ability of a computer to understand, analyse, manipulate, and potentially generate human language.Pre-processing the data typically consists of a number of steps that are Remove Punctuation, Tokenization, Stemming, Padding.

4) Data Splitting:
Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.

5) Classification
In our process, we have to implement the machine learning algorithm such as NB. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

6) Result Generation:
The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like Accuracy, Precision, Recall,F-measure.

VI. RESULT AND CONCLUSION

We conclude that, the email dataset was collected from dataset repository as input. The input dataset was mentioned in our research paper. We are implemented the NLP techniques and classification algorithms (i.e.) machine learning algorithm. Then, machine learning algorithms such as naïve bayes. Finally, the result shows that the accuracy for above mentioned algorithm and visualize the output in the form of graph. Then, analyse the mail is ham or spam.

VII. FUTURE ENHANCEMENT

In the future, we should like to hybrid the two different machine learning. In future, it is possible to provide extensions or modifications to the proposed classification algorithms to achieve further increased performance. Apart from the experimented combination of data mining techniques machine algorithms can be used to improve the detection accuracy. Finally, the sentiment analysis detection system can be extended as a prevention system to enhance the performance of the system.

REFERENCES

[1] D. M. Ablel-Rheem,''Hybrid feature selection and ensemble learning 899method for spam email classification,''Int. J. Adv. Trends Comput. Sci. 900 Eng., vol. 9, no. 1.4, pp. 217–223, Sep. 2020. 901
[2] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, ''Email 902 classification research trends: Review and open issues,'' IEEE Access, 903 vol. 5, pp. 9044–9064, 2017. 904
[3] A. Kumari, N. Agrawal, and U. Lilhore, ''Clustering malicious spam in 905 email systems using mass mailing,'' in Proc. 2nd Int. Conf. Inventive Syst. 906 Control (ICISC), Jan. 2018, pp. 870–875. 907
[4] S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli, and W. A. H. M. Ghanem, 908''E-mail spam classification using grasshopper optimization algorithm 909 and neural networks,'' Comput., Mater. Continua, vol. 71, no. 3, 910 pp. 4749–4766, 2022. 911
[5] S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli, and W. A. H. M. Ghanem, 912''Spam classification based on supervised learning using grasshopper opti- 913mization algorithm and artificial neural network,'' Commun. Comput. Inf. 914 Sci., vol. 1347, pp. 420–434, Dec. 2021. 915
[6] M. Shuaib, S. M. Abdulhamid, O. S. Adebayo, O. Osho, I. Idris, 916 J. K. Alhassan, and N. Rana, ''Whale optimization algorithm-based email 917 spam feature selection method using rotation forest algorithm for classifi- 918 cation,'' Social Netw. Appl. Sci., vol. 1, no. 5, p. 390, May 2019. 919
[7] S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli, and W. A. H. M. Ghanem, 920 ''An integrated model to email spam classification using an enhanced 921 grasshopper optimization algorithm to train a multilayer perceptron neural 922 network,'' Commun. Comput. Inf. Sci., vol. 1347, pp. 402–419, Dec. 2020. 923
[8] I. Idris, A. Selamat, N. T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and 924 M. Penhaker, ''A combined negative selection algorithm-particle swarm 925 optimization for an email spam detection system,'' Eng. Appl. Artif. Intell., 926 vol. 39, pp. 33–44, Nov. 2015. 927
[9] O. M. E. Ebadati and F. Ahmadzadeh, ''Classification spam email with 928 elimination of unsuitable features with hybrid of GA-naive Bayes,'' J. Inf. 929 Knowl. Manage., vol. 18, no. 1, Mar. 2019, Art. no. 1950008. 930.
[10] A. Karim, S. Azam, B. Shanmugam, and K. Kannoorpatti, ''An unsu- 931 pervised approach for content-based clustering of emails into spam and 932 ham through multiangular feature formulation,'' IEEE Access, vol. 9, 933 pp. 135186–135209, 2021.

