Webology (ISSN: 1735-188X)
Volume 18, Number 4, 2021

Email Spam Detection & Prediction Using Deep Clustering And Multi-Regression Models

I V S Venugopal[1] , D Lalitha Bhaskari[2] , M N Seetaramanath[3]

[1,3]Department of IT, G V P College of Engineering (A), Andhra Pradesh, 530048, India

[2]Department of CS&SE, AUCE (A), Andhra Pradesh, Visakhapatnam, India

---

Abstract— The growth in the digital form of information exchange have also cultivated the demand for higher security of the information and the information have become more sensitive for exchanges. The primary and most preferred form of the digital communication is email and majority of the attackers attempt to sabotage this communication method in the form of spam emails. Recognition of email spam is been the focal point of examination for more than 10 years at this point and numerous autonomous analysts and associations are attempting to assemble the most powerful type of spam filters in email servers. Throughout the investigation, this work revealed that the current spam separating, or recognition techniques are exceptionally time complex and a greater part of the cases are over fitted, hence conventional intentions is profoundly unimaginable. Additionally, the spam email location techniques are far away from the prescient recognition of the emails into ham or spam classes. Consequently, this work proposes two novel techniques for email location and expectation. The first deep clustering method demonstrates the email categorization process with 99.6% accuracy and the second method demonstrates nearly 99% accuracy for the prediction of the email spams. The proposed methods together are one of the benchmarked research outcomes in this domain of the research for making the email based communications a more desirable method.

Keywords— email spam, ham, email categorization, deep clustering, multi-regression

I. INTRODUCTION

The use of machine learning-driven strategies for identification or categorizations of spam or ham emails are widely popular. Few early demonstrations of such methods can showcase the use of support vector mechanism by W. Feng et al. [1]. The support vector machine is a penalty-driven method for enforcing the guided learning process for categorizations and detections. Such processes are always criticized for the over fitting of the model to the dataset and often found irrelevant for the newer datasets with newer parameters. The criticism about such methods is rightly highlighted in the work by E. G. Dada et al. [2]

with the overall applicability of the machine learning methods for spam detections over email datasets. Nevertheless, many of the early research attempts have showcased the use of hybrid models and highlighted the benefits of using regression methods as demonstrated in the work by A. Wijaya et al. [3].

Thus, proposing a machine learning-based method for generic purposes is the demand of the current research. Henceforth, in this work, a novel method is proposed to detect and predict the emails into two different classes as spam and ham.

The remainder of the work is organised as follows: Section – II presents fundamental mathematical models; Section – III analyses parallel research outcomes; Section – IV presents proposed solutions using the mathematical models; Section – V presents proposed algorithms; Section – VI discusses the obtained results; Section – VII presents a comparative analysis; and Section – VIII concludes the research.

II. FUNDAMENTALS OF DEEP CLUSTERING & REGRESSION MODEL

After setting the context of the research in the previous section, in this section of the work, the detailed elaboration on the deep clustering and the regression models as the foundation of the research are furnished.

Firstly, the mathematical model for the deep clustering is analysed. Assuming that, any dataset, D is a collection of multiple attributes or properties, A[], and each attribute can be represented as A_i. Thus, for n number of attributes in the dataset, the relation can be formulated as,

$$D = \sum_{i=1}^{n} A_i \qquad \text{(Eq. 1)}$$

Also, each attribute is part of the attribute collections for the same dataset and the following relation can be furnished,

$$A[] = \sum_{i=1}^{n} A_i \qquad \text{(Eq. 2)}$$

Thus, $D = A[]$ (Eq. 3)

Further, as the clustering method must be applied on the complete dataset based on a specified class variable, A_C, thus for a total number of X clusters and possibly Y number of class variations, the relation can be formulated as,

$$C[] \leftarrow \sum_{i=1}^{X} \sum_{j=1}^{Z} D_{i,j}[A_C] \qquad \text{(Eq. 4)}$$

Thus, the new collection C[] defines the clustered group and can be visualized here [Fig – 1] with the initial 3 cluster groups.

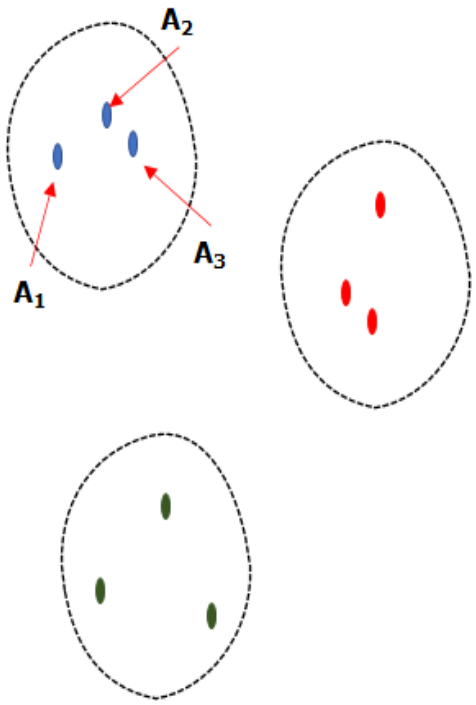

Fig. 1 Illusion of three cluster groups as initial clustering

Further, as per the illustration above, three elements A_1, A_2, and, A_3 are clustered based on the class variable in the same cluster. Nevertheless, looking into the detailed characteristics of the same cluster, it is natural to realize that the first element A_1 is not having the same Euclidian distance as the other two elements in the same cluster.

Thus, one further clustering will eliminate the overlapping of two or more clusters for better realizing of the data categorization.

$$C'[] \leftarrow \sum_{i=1}^{X'} \sum_{j=1}^{Z'} C[] \qquad \text{(Eq. 5)}$$

This can be visualized here [Fig – 2].

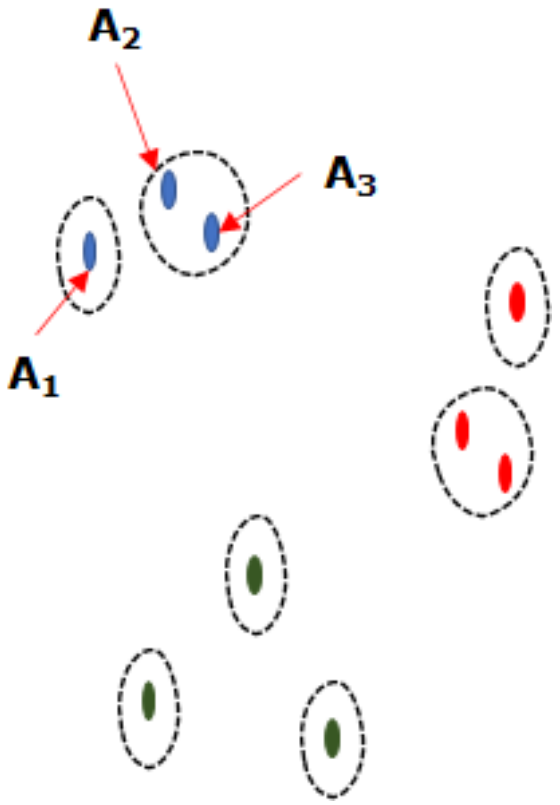

Fig. 2 Illusion of seven cluster groups as final clustering

Henceforth, this can be stated conclusively that deep clustering can generate more efficient data categorizations compared with initially generated clusters. Thus, this concept is adopted in this work.

Secondly, the regression method is analyzed here. The primary purpose of the regression method is to analyze the relationship between multiple attributes in the dataset and finally generatea prediction based on the independent attributes to the class attribute. Assuming that, any dataset DS, is a collection of two attributes as X and Y, where X is the independent attribute and Y is the dependent or class variable. Thus, the following relation can be established with β_0 and β_1 as intercept or correction factor and slope coefficient for the independent variable, X,

$$Y = \beta_0 + \beta_1 X \qquad \text{(Eq. 6)}$$

Further for n number of observations in the total dataset, the intercept and the coefficient can be calculated as,

$$\beta_0 = \frac{\sum Y.\sum X^2 - \sum X.\sum XY}{n.\sum X^2 - (\sum X)^2} \qquad \text{(Eq. 7)}$$

And,

$$\beta_1 = \frac{n.\sum XY - \sum X.\sum Y}{n.\sum X^2 - (\sum X)^2} \qquad \text{(Eq. 8)}$$

Thus, the dependent variable, Y, can be utilized for any prediction such as email spam detection. Thus, this analogy is used in this work for email spam or ham detection.

Further, in the next section of this work, the parallel research advancements are analysed to identify the gap in the current research.

III. PARALLEL RESEARCH OUTCOMES

After setting the context of the research and with the fundamental understanding of the research proposal, in this section of the work, the parallel research outcomes are analysed with significant discussions.

The process of text-based detection or logical analysis of the email is always been a time complex process and multiple research attempts have recommended incorporating optimization methods to reduce the time complexity. One such outcome from the work of K. Agarwal et al. [4] has showcased the use of machine learning methods with optimization to detect email spam. With these methods gaining popularity, yet another outcome from the work of A. I. Taloba et al. [5] can be seen by demonstrating the use of genetic optimizations for email classifications. These methods have significantly reduced the time complexity.However, the complexity of the algorithm makes it difficult to adapt.

The classification or identification of the email is also challenging with the inclusion of local linguistics and special symbols. The work by S. L. Marie-Sainte et al. [6] is one such attempt. Nevertheless, such models have higher dependencies on language libraries and highly irrelevant for the use of generic problems.

In recent times, a good number of tools can be observed to detect email spam. The work by A. Géron et al. [7] and S. Zhu et al. [8] have significantly demonstrated the drawbacks and the benefits of using such tools. Nonetheless, given the perimeters of the restrictions of these tools, the scope for the improvements is highly complex to achieve. Thus, many of the parallel research attempts have criticized the use of tool-based analogies as deprecated in the work by S. Sawla et al. [9], which is again justified by G. Bonaccorso et al. [10]. The overall comparisons of multiple approaches can be observed in the survey conducted by G. Singh et al. [11].

Yet another direction of the research also directs that, the content mentioned in the emails play a major role in detecting the type of the email. As proposed in the work by A. Karim et al. [12] have successfully deployed clustering approach on the email contents to categorize the emails into spam or ham or normal.

Nonetheless, the complexity of the proposed algorithms, in such cases, have increased to a greater extend. Thus, the work by S. A. A. Ghaleb et al. [13] must be taken under consideration, which showcases inclusion of optimization methods. The final recommendation, in this direction of the research, comes from the work by S. Gibson et al. [14], which significantly motivates the researchers to adapt to machine learning and bio genetic optimization methods. The benefits of these methods with the research bottlenecks are elaborated in the study conducted by A. Karim et al. [15].

Yet another direction of this research is to adapt to the contemporary methods for email spam detections as showcased in the work by G. Al-Rawashdeh et al. [16].

Henceforth, it is clear to realize that the use of machine learning methods independent of the tools and with or without the optimization methods can be highly efficient to detect spam or ham emails from any given dataset in real-time. Thus, this work proposed the solution, based on this understanding, in the next section of this work.

IV. PROPOSED SOLUTIONS – MATHEMATICAL MODELS

After the detailed understanding of the parallel research outcomes, in this section of the work, the proposed solutions are furnished using the mathematical modelling technique.

Lemma – 1: The knowledge-based dependent spam and ham categorization using the deep clustering method can improve the time complexity and accuracy.

Proof: The categorization of the emails into spam and ham can be achieved using the classification method, however the deep clustering method can also categorize the emails into these two categories with severity. Thus, the use of deep clustering shall be adopted. Also, the time complexity can be highly reduced with the use of a predefined knowledgebase.

Assuming that, the initial knowledgebase, KB[] is a collection of individual words as W_i. Thus, for n number of words, this can be presented as,

$$KB[] = \prod_{i=1}^{n} \langle W_i \rangle \qquad \text{(Eq. 9)}$$

Also, any email text data, T[], is a collection of individual text items, T_i, and can be represented from number items as,

$$T[] = \sum_{i=1}^{m} T_i \qquad \text{(Eq. 10)}$$

Further, the probability set, P[] for all the words or terms in the knowledge base must be calculated against the collected text samples as,

$$P[] = \frac{P(W_i)}{P(W_i) \cup P(KB[])} \cdot \frac{P(W_i)}{P(W_i) \cup P(T[])} \quad \text{(Eq. 11)}$$

Thus, the final categorization data, DS[], can be formulated as,

$$DS[] = \langle KB[], P[] \rangle \quad \text{(Eq. 12)}$$

Here, in DS[] dataset, the P[], probability distribution sets will be acting as the class variable.

Further, the Euclidian distance, $\partial[]$, among the P[] set elements must be calculated as,

$$\partial[] = \prod_{i=1}^{n} |P[i] - P[i+1]| \quad \text{(Eq. 13)}$$

Henceforth, based on the elements of $\partial[]$, the initial clustering must be done as,

$$C[] = \frac{\prod_{i=1}^{n-1} \partial[]}{\Delta\partial} \quad \text{(Eq. 14)}$$

Further, the highest two densities from the cluster sets, C[] shall be identified for spam and ham detection. During this process, if multiple clusters contain a similar number of elements, then, the clustering process shall be repeated to obtain two genuine highest number of elements from the existing clusters.

Finally, as a result,two specified clusters for spam and ham detection as C_1 and C_2 must be reported.

Furthermore, from Eq. 5 and Eq. 10, the time complexity, τ_1, must be calculated and compared.

$$\tau_1 = \frac{C'[]}{T[]} \quad \text{(Eq. 15)}$$

And, from Eq. 14 and Eq. 10, the time complexity, τ_2, shall also be calculated as,

$$\tau_2 = \frac{C[]}{T[]} \quad \text{(Eq. 16)}$$

As the number of clusters in the generic method will be much higher compared with the deep clustering and the following can be stated,

$$\phi(C'[]) >> \phi(C[]) \qquad \text{(Eq. 17)}$$

Thus, it is natural to realize that, the time complexity will also reduce during the deep clustering process and the following can be stated,

$$\tau_2 << \tau_1 \qquad \text{(Eq. 18)}$$

Further, it is also natural to realize that, due to the deep clustering process, the removal of the outliers will be maximum, thus, the result of the deep clustering process will also generate more accuracy.

Henceforth, the deep clustering process will significantly reduce the time complexity and increase the accuracy of the email categorization process.

Lemma – 2: The prediction of spam or ham emails can be realized using the multi-regression model.

Proof: The predictive analysis for detection of spam or ham email is realized using the classification method. Nevertheless, the classification method is highly dependent on the correctness of the data. The correctness relies on the removal of the outliers and the missing values. For text data, such as email it is obvious to have some missing value or some outliers, or some noise in the actual data. Hence detection or prediction using the classification is not an optimal solution.

Thus, the regression method based on multiple classes as a combined detection method can be realized. From Eq. 14, it is natural to realize that finally two classes can be realized as C_1 and C_2. Thus, for a detection class, C_k, the regression formula can be realized as,

$$C_k = \beta_0 + \beta_1 C_X + \beta_2 C_Y + \varepsilon \qquad \text{(Eq. 19)}$$

Where $\beta_{0..2}$ sets are the regression coefficients and ε are the correction factor. Further, the error correction factor and the coefficients can be calculated as,

$$C_k(t+1) = C_k(t) \pm \varepsilon \qquad \text{(Eq. 20)}$$

Or,

$$\varepsilon = \left| C_k(t) \pm C_k(t+1) \right| \qquad \text{(Eq. 21)}$$

And,

$$\beta_0 = |C_k(t) - C_k(t+1)| \pm \varepsilon \qquad \text{(Eq. 22)}$$

Along with,

$$\beta_1 = \frac{C_X}{\Delta C_Y}, \beta_2 = \frac{C_Y}{\Delta C_X} \qquad \text{(Eq. 23)}$$

Thus, it is conclusive that, the regression method can be deployed to predict the class of the email where the detected class can be a combination of spam or ham, or both.

Henceforth, based on the proposed mathematical models, in the next section of this work, the proposed algorithms are furnished.

V. PROPOSED ALGORITHMS

In the previous section of the work, the proposed mathematical models have proven the benefits of the proposed methods over existing methods. Henceforth, in this section of the work, based on the mathematical models, the proposed algorithms are furnished.

Algorithm - I: Spam and Ham Email Detection using Knowledgebase Deep Clustering (**SHED-KDC**) Algorithm
Input: Knowledgebase, KB[] and Email text set, T[] **Output**: Final clusters as C[]
Process: Step - 1. Build the pre-existing spam keywords into KB[] Step - 2. For each element in KB[] a. Build the probability distribution as P[] Step - 3. Accept the text data as T[] Step - 4. For each element in T[] a. Identify the probability distribution as T[i] == KB[i], Assign TP[i] = P[i] b. Build the dataset, DS[] as <<T[],TP[i]>> c. Calculate the Euclidian distance, ED[j] = |TP[i] - TP[i+1]| Step - 5. Build the initial cluster set, C[] as Step - 6. For each element in ED[] a. C[k] = ED[j]+ED[j+1]/Mean(ED[]) Step - 7. Repeat Step –4.C to 6.A until only length of C[] == 2 Step - 8. Report the final clusters as C[]

Profound grouping structures join highlight extraction, dimensionality decrease, and bunching into a start to finish model, permitting the profound neural organizations to learn appropriate portrayals to adjust to the presumptions and rules of the grouping module that is utilized in the model. This mitigates the need to perform complex learning or dimensionality decrease on huge datasets independently, rather than joining it into the model preparation.

Algorithm - II: Spam and Ham Prediction using Multi-Regression (**SHP-MR**) Algorithm
Input: Spam & Ham Data Clusters as C[] **Output**: Detected Class as CK
Process: Step - 1. Import the Cluster, C[] Step - 2. For each entity of C[] as C[i] as elements as C[X] and C[Y] with E = 0 a. Calculate the B1 = C[X] + C[X+1] / Mean (C[X]) b. Calculate the B2 = C[Y] + C[Y+1] / Mean (C[Y]) c. Calculate the CK = B0 + B1.C[X]+ B2.C[Y] d. Calculate the Error, E = |CK - C[i]| e. Repeat Step - 2, until E = 0 Step - 3. Report the predicted email class as CK

Direct relapse is a capability that enables an investigator or analyst to predict one variable based on data about another one. Direct relapse must be used when two constant variables exist—an autonomous variable and a dependent variable. The free factor is the boundary around which the dependent variable or outcome is determined. A model with several relapses may be reduced to a few illustrative components.

Further, the obtained results from the proposed algorithm implementations are discussed in the next section of this work.

VI. RESULTS & DISCUSSIONS

After the detailed understanding of the proposed mathematical models and the algorithms to solve the intended research problem, in this section of the work, the obtained results are discussed.The algorithms are tested on the benchmarked dataset provided by Kaggle [17].

Firstly, the summary of the research outcomes is furnished here [Table – 1].

TABLE I SUMMARY OF THE DETECTION PROCESS

Parameter Name	Value	Percentage
"Correctly Classified Instances"	270	99.631
"Incorrectly Classified Instances"	1	0.369
"Kappa statistic"	0.8553	-
"Mean absolute error"	0.0065	-
"Root mean squared error"	0.0665	-
"Relative absolute error"	-	19.9835
"Root relative squared error"	-	55.1179
"Total Number of Instances"	271	100

The result is visualized graphically here [Fig – 3].

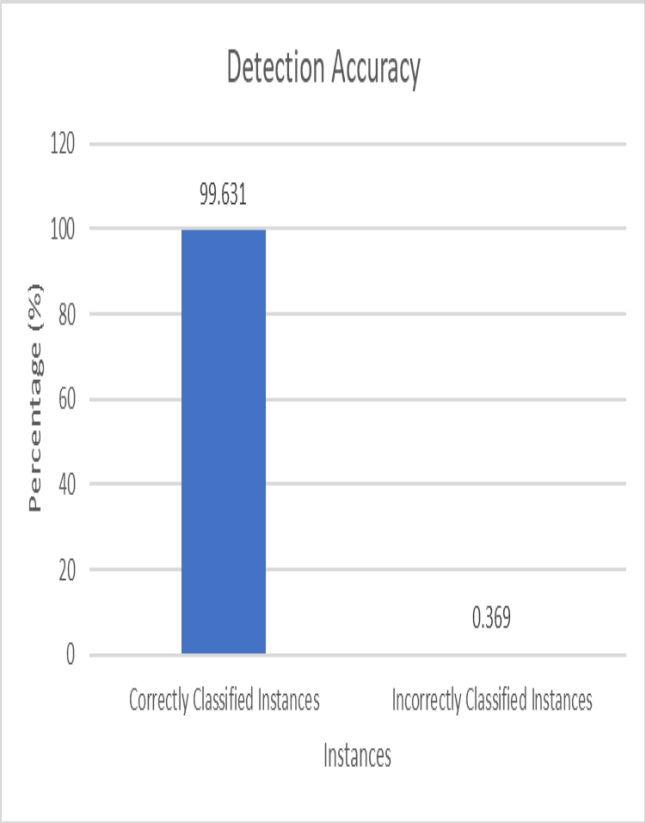

Fig. 3 Detection Summary

The proposed method demonstrates a significant 99.6% accuracy.

Secondly, the detailed accuracy analysis is furnished here [Table – 2].

TABLE II DETAILED DETECTION ANALYSIS

CLASS	SPAM	HAM	WEIGHTED AVG.
TP Rate	0.75	1	0.996
FP Rate	0	0.25	0.246
Precision	1	0.996	0.996
Recall	0.75	1	0.996
F-Measure	0.857	0.998	0.996
MCC	0.864	0.864	0.864
ROC Area	1	1	1
PRC Area	1	1	1

It is natural to realize that, the precision is nearly 1 for ham emails and 1 for spam emails, which significantly denotes the higher-performing measures of the proposed algorithm.

Further, the deep clustering method analysis is furnished here to demonstrate the benefits of the proposed algorithm and the improvement of the accuracy [Table – 3].

TABLE III DEEP CLUSTERING RESULTS

Accuracy (Scale of 1 = 100%)	Time Complexity (sec)
-1	0
-0.916	0.369
0.072	0.738
0.488	1.107
0.916	1.476
0.952	1.845
0.976	2.214
0.98	2.583
0.996	14.022
1	100

The result is visualized graphically here [Fig – 4].

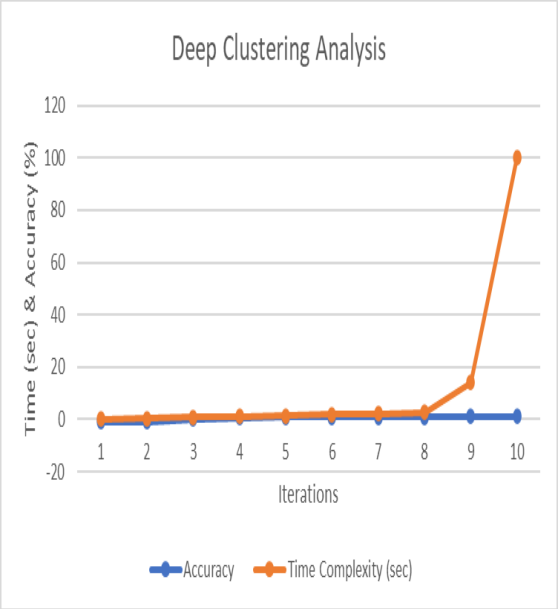

Fig. 4 Deep Clustering Results

Regardless of to mention, that the proposed deep clustering method defines the divergence point at iteration – 9 and after iteration – 9, the time complexity increases astronomically. Thus, the final accuracy is considered is 0.996 or 99.6%.

Henceforth, after obtaining highly satisfactory results, in the next section of this work, the proposed methods are compared with the parallel research outcomes.

VII. COMPARATIVE ANALYSIS

After the detailed discussion on the obtained results, in this section of the work, the proposed method is compared with the parallel research outcomes [Table – 4].

TABLE IV COMPARATIVE ANALYSIS

Author, Year	Research Method	Algorithm Complexity	Time Complexity (Sec)	Accuracy (%)
W. Feng et al. [1], 2016	SVM	$O(n^2)$	29.33	91%
K. Agarwal et al. [4], 2018	Naïve Bayes	$O(n^2)$	25.01	92.97%
A. I. Taloba et al. [5], 2019	Genetic Algorithm	O(n log n)	23.22	93%
G. Singh et al. [11], 2019	Multinomial	$O(n^2)$	22.05	97%
Proposed SHED-KDC and SHP-MR Algorithm	Deep Clustering and Multi – Regression	$O(n^2)$	14.02	99.6

Henceforth, it is natural to realize that the proposed method has outperformed the other parallel research contributions.

Further, post the comparative analysis, in the next section of this work, the final research conclusion is presented.

VIII. CONCLUSION

With the growth in communication in the digital form, the sensitivity of information protection is also increasing.The primary type of attack during the digital information exchange is email spam. Detection of email spam is been the focus of research for a decade now and multiple independent researchers and organizations are working to build the most robust form of spam filters in email services. During the course ofthe study, this work reported that the existing spam filtering or detection methods are highly time complex and a majority of the cases are overfitted, thus applicability to generic purposes is highly impossible. Also, the spam email detection methods are far away from the predictive detection of the emails into safe or unsafe categories. Hence, this work proposes two novel methods for email detection and prediction. The first deep clustering method is designed to categorize the emails into two different classes as spam and ham. The designed algorithm has showcased great benefits due to the deep clustering process and demonstrated 99.6% accuracy. Also, the second multi-regression method has demonstrated nearly 99% accuracy for the prediction of spam emails. Henceforth, this work must be treated as one of the benchmark research in this domain for making email-based communication a safer world.

REFERENCES

[1] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, "A support vector machine-based Naive Bayes algorithm for spam filtering", Proc. IEEE 35th Int. Perform. Comput. Commun. Conf. (IPCCC), pp. 1-8, Dec. 2016.

[2] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: Review approaches and open research problems", Heliyon, vol. 5, no. 6, Jun. 2019.

[3] A. Wijaya and A. Bisri, "Hybrid decision tree and logistic regression classifier for email spam detection", Proc. 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), pp. 1-4, Oct. 2016.

[4] K. Agarwal and T. Kumar, "Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization", Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), pp. 685-690, Jun. 2018.

[5] A. I. Taloba and S. S. I. Ismail, "An intelligent hybrid technique of decision tree and genetic algorithm for E-Mail spam detection", Proc. 9th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS), pp. 99-104, Dec. 2019.

[6] S. L. Marie-Sainte and N. Alalyani, "Firefly algorithm based feature selection for arabic text classification", J. King Saud Univ.-Comput. Inf. Sci., vol. 32, no. 3, pp. 320-328, Mar. 2020, [online] Available: https://www.sciencedirect.com/science/article/pii/S131915781830106X.

[7] A. Géron, Hands-On Machine Learning With Scikit-Learn Keras and TensorFlow, Newton, MA, USA:O'Reilly Media, 2019.

[8] S. Zhu and F. Chollet, Working With RNNs, Nov. 2019, [online] Available: https://keras.io/guides/working_with_rnns/.

[9] S. Sawla, Introduction to Naive Bayes for Classification, Oct. 2018, [online] Available: https://medium.com/@srishtisawla/introduction-to-naive-bayes-for-classification-baefefb43a2d.

[10] G. Bonaccorso, Machine Learning Algorithms, Birmingham, U.K.:Packt Publishing, 2018.

[11] G. Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between multinomial and Bernoulli Naïve Bayes for text classification", Proc. Int. Conf. Autom. Comput. Technol. Manage. (ICACTM), pp. 593-596, Apr. 2019.

[12] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "An Unsupervised Approach for Content-Based Clustering of Emails Into Spam and Ham Through Multiangular Feature Formulation," in IEEE Access, vol. 9, pp. 135186-135209, 2021.

[13] S. A. A. Ghaleb, M. Mohamad, S. A. Fadzli and W. A. H. M. Ghanem, "Training Neural Networks by Enhance Grasshopper Optimization Algorithm for Spam Detection System," in IEEE Access, vol. 9, pp. 116768-116813, 2021.

[14] S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in IEEE Access, vol. 8, pp. 187914-187932, 2020.

[15] A. Karim, S. Azam, B. Shanmugam and K. Kannoorpatti, "Efficient Clustering of Emails Into Spam and Ham: The Foundational Study of a Comprehensive Unsupervised Framework," in IEEE Access, vol. 8, pp. 154759-154788, 2020.

[16] G. Al-Rawashdeh, R. Mamat and N. Hafhizah Binti Abd Rahim, "Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection," in IEEE Access, vol. 7, pp. 143721-143734, 2019.

[17] Wessel van Lit, Ham and Spam emails from SpamAssasin, https://www.kaggle.com/veleon/ham-and-spam-dataset.