Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 22 (2024) 560-566

Original

Efficient clustering of e-mails by applying supervised machine learning algorithms

D. Quirumbay Yagual[a]* • B. Soria Méndez[b] • V. Cruz Ruiz[b]

[a]*Universidad Estatal Península de Santa Elena, UPSE*
[b]*Universidad Estatal de Milagro, UNEMI*

Received 11 30 2023; accepted 06 05 2024
Available 08 31 2024

Abstract: In today's digital age, effective detection of unwanted e-mails, commonly known as "spam", has become a priority for individuals and organizations. As e-mail inboxes fill up with un-solicited messages, it has become evident that the predefined rules and heuristics used by traditional spam filters have lost their effectiveness. This persistent problem poses challenges at both the personal and business level.

Despite efforts to protect e-mail accounts with anti-virus, which in many cases come at a cost, spam remains a growing concern. For businesses, implementing costly firewalls can be an unnecessary burden. The problem of spam persists, and its impact on the efficiency and security of e-mail communication is indisputable.

The primary objective of this paper is to investigate and evaluate machine learning algorithms specifically designed to address the challenge of automatic spam detection. This is achieved by using text classification techniques applied to mail servers and personal computers. Three key algorithms are examined: Random Forest, decision tree and Naive Bayes, with the intention of determining their applicability in both environments.

This study relies on two essential research methodologies. First, feature selection, a crucial process that identifies the most relevant variables in mail classification, including keywords and word frequencies, is conducted. In addition, performance evaluation, which uses metrics such as accuracy, recall and F1-score, is employed to understand the performance of machine learning models in detecting spam and legitimate e-mails.

The results of this study are presented in the form of comparative tables showing the hit and miss rates of the three models evaluated. Notably, it is determined that the Random Forest model, when applied in conjunction with tokenization techniques, exhibits superior efficiency compared to the other two models.

The choice of the right machine learning model is critical to ensure efficiency in e-mail classification, and this study provides a solid basis for making informed decisions in the implementation of e-mail security systems in real-world business environments. Spam detection, supported by machine learning algo-rhythms, remains an evolving field and offers a promising solution to address a persistent problem in the digital world.

Keywords: Machine learning, text classification, artificial intelligence, spam

*Corresponding author.
E-mail address: dquirumbay@upse.edu.ec (D. Quirumbay Yagual).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

E-mail services are one of the most widely used means of personal and professional communication worldwide, in 2021 it is estimated that there were 319.4 billion users, in 2022 there are 333.2 billion users, and it is estimated that for the years 2023, 2024 and 2025 the percentage of users will grow by 4.1% (THE RADICATI GROUP, 2021).

Nowadays, the internet connects the world in a faster way, the amount of e-mails transmitted worldwide has increased considerably, in 2021, it is estimated that globally 300 billion e-mails have been sent and received daily, creating expectations that this figure will grow by more than 17% until 2025 (Fernández, 2023). The possibilities offered by this form of communication have been reflected in the use by businesses, government institutions, education, banks, and personal use. The management of this service is especially important, as companies consider it as a fast, cheap, and accessible method to implement in institutions, increasing the efficiency and productivity of employees (NIBUSINESS, n.d.)

Information sent via e-mail is personal, confidential, and sensitive data that companies manage internally, and it is a risk to a company's reputation if this data gets into the hands of third parties.

Cybercriminals try to gain access to this information by compromising security flaws in networks, services that a server has or by applying social engineering on people, with financial gain being one of the main objectives for this type of cybercrime. In 2021, economic losses in companies caused by cybercriminals are estimated to exceed 5 billion dollars worldwide, and it is predicted that by 2025 this figure will exceed 10 billion dollars (Morgan, 2020). Phishing is one of the techniques most used by cybercriminals to steal information from companies through e-mails; in 2020, around 241,000 phishing attacks and economic losses of up to $54,241,075 have been recorded worldwide (Bischoff, 2022).

To provide a solution to this problem, the concept of cybersecurity arises, which consists of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks, and one of the most common categories is information security. Phishing is an ever-present threat, and while staff training remains crucial, effectively addressing this challenge requires advanced solutions. In this context, solutions utilizing machine learning and deep learning techniques have been developed to detect spam, phishing, and spoofing. Algorithms such as Naive Bayes, KNN, decision tree, Random Forest, and support vector machine have emerged as valuable tools in the fight against cyber threats.

A notable example is the system developed by Junnarkar et al. (2021), which employs Naive Bayes and support vector machine algorithms to enhance defenses against these attacks.

The significant contribution of this work concentrates on a comprehensive solution, employing decision tree, Naive Bayes, and Random Forest algorithms.

The proposed architecture lies in section two, where the research and development methodology are discussed, the latter being based on KDD and OMSTD. Section three presents the results, highlighting the superior performance of certain algorithms using visual tables. This article concludes in section four, offering valuable insights derived from extensive testing of key spam detection algorithms.

2. Development

For the development of this research, we took some alternatives of machine learning algorithms that identify and prevent unwanted e-mails of which only three alternatives were taken to perform the training, identify the efficiency, and error rate.

For this, we trained and analyzed each of the algorithms to detect the best performance for the automatic detection of spam in e-mails, training, and analysis of some supervised machine learning algorithms of which many were discarded considering only three alternatives that are detailed below:

Decision tree: The decision tree is a machine learning technique commonly used for both classification and regression, with a structure like a flow chart. It uses a recursive subdivision process to evaluate features or attributes present in the data, based on purity indices. The most popular indices are the Gini index and entropy, which are used to determine which features should be placed at the root or internal nodes of the tree. The decision tree is capable of handling both continuous and categorical variables (Junnarkar et al., 2021).

Naive Bayes classifier: A classification algorithm that uses a probabilistic approach based on Bayes' theorem to determine the conditional probability, with two assumptions: All features are equally important, and all features are independent of each other. In the case of text classification, the features can be modelled by a multinomial distribution or, if continuous, by a Gaussian distribution. This algorithm is effective for dealing with high-dimensional data points in feature space and has a fast training speed, requires few training examples and is widely used for text classification, with performance comparable to advanced methods if adequate data preprocessing is used (Junnarkar et al., 2021)

Random Forest: A machine learning algorithm that uses many uncorrelated decision trees to create a model that performs with high accuracy on new datasets due to better generalization to new, unseen datasets. It reduces variation and always avoids overfitting of the model by using Bootstrap Aggregation or Bagging technique from where several subsets of data are created for training and randomly chosen with replacement (Junnarkar et al., 2021).

2.1. Suggested architecture

Figure 1 illustrates the architecture of the spam detection agent developed in Python. This agent is installed on the e-mail server and performs continuous monitoring of the institutional inbox, specifically on the Zimbra platform. Its main function is to constantly identify and categories incoming e-mails, classifying them as spam or non-spam (Nichols et al., 2018). This approach, implemented server-side, is positioned as an effective method for initiative-taking detection of unwanted content in the institutional e-mail environment.

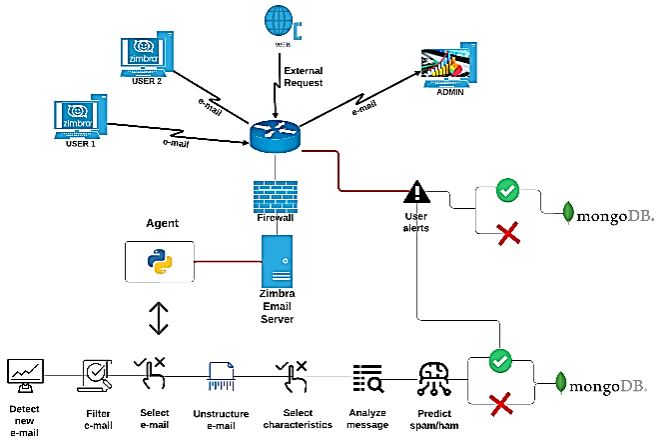

Figure 1. SPAM detection architecture.

2.2. Evaluation metrics

In this research, several metrics are used to assess the performance of the models. The selected metrics are detailed below:

1.Accuracy: Accuracy is a crucial metric that assesses the validity of the estimated results. By focusing on cases classified as positive, it determines what percentage of the positive predictions made by the model are truly positive. High accuracy indicates the model's ability to avoid misclassifying negative examples as positive (Messina Valverde, 2021).

2.True positives (recall): Recall, also known as sensitivity, is essential in machine learning. It represents the proportion of true positive cases correctly identified by the model. It measures the model's ability to capture most existing positive cases, i.e., how many of the relevant cases are recovered by the model (Powers, 2020).

3.F1-score: The F1-score is a comprehensive measure that considers both precision and recall calculating the overall score. It can be interpreted as a weighted average of the precision and recall values, where it reaches its maximum value at 1 and its mini-mum value at 0. It provides a balanced assessment of a classification model's ability to maintain a balance between precision and recall (The State of Phishing in the US: Report and Statistics, 2022).

3. Methods

3.1. Research

This study employs two essential methodologies. First, "feature selection" is applied to identify relevant variables, including keywords and word frequency, fundamental to e-mail classification. Then, "performance evaluation" is conducted, using metrics such as precision, recall and F1-score. These metrics are used to understand the performance of three machine learning algorithms: Random Forest, decision tree and Naive Bayes. The results are presented in comparative tables, allowing the selection of an optimal model for spam detection. This methodology provides a sound basis for decision making in the implementation of e-mail security systems in e-mail environments.

In addition, data was collected from tests conducted by different authors referenced in this article, and in relation to the results obtained, it is proposed to recommend the best applicable algorithm for this problem.

3.2. Algorithm training

The methodological approach concentrates on the exploration of various machine learning algorithms to refine the automatic detection of spam e-mails, focusing on the context of the Zimbra service.

During this stage, different practice platforms offering specific datasets in csv format were evaluated, with a particular focus on identifying e-mails classified as spam. Platforms used included Kaggle and GitHub. A representative example of such datasets is the fraud e-mail dataset, which contains two columns: The first one corresponds to the content of the e-mail, i.e. the text, and the second one, known as the label, specifies whether the content is spam or not (Radev, 2008).

The detection of spam will be based on predictive analysis, using the KDD methodology. This methodology provides a structured and systematic framework for discovering valuable knowledge from large volumes of data, merging discovery and analysis (Timarán Pereira, 2016). The key phases of this methodology are described in figure 2.

By applying machine learning techniques and data analysis, the use of KDD (knowledge discovery in databases) facilitates the identification of distinctive spam features, enhancing filtering to differentiate between legitimate and malicious messages. The integration of KDD with AI provides a solid foundation for the continuous evolution of cybersecurity defense mechanisms. This methodology not only improves the effectiveness of spam filters but also provides a robust foundation for the development of more advanced solutions in the field of cybersecurity.

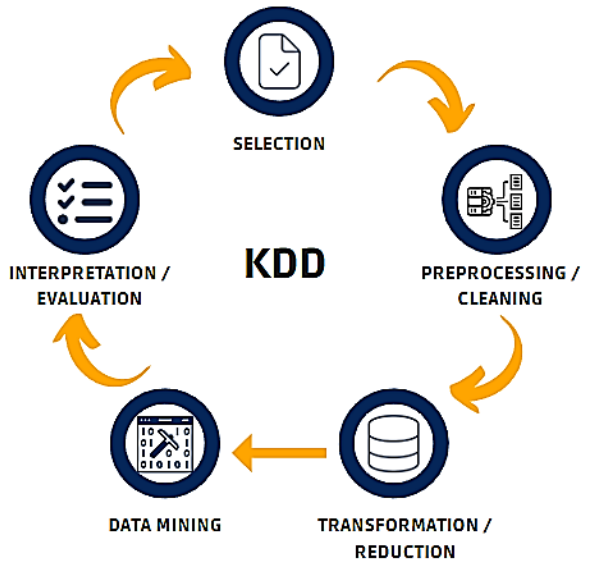

Figure 2. Life cycle KDD.

Systems based on this approach achieve significantly high detection rates, reducing false positives and improving the end-user experience. Furthermore, the ability of these systems to adapt and learn from new data allows them to remain effective against the changing tactics of spammers. In summary, the combination of KDD and AI not only optimizes the accuracy of spam filters but also ensures the sustainability and scalability of cybersecurity solutions over time.

Another methodology used in this research is OMSTD, which is a methodology and set of best practices to achieve the development of well-built security tools, mainly based on hacking tools written in Python, although it is not especially limited to this language (OMSTD Project, 2014). Its phases are visualized in figure 3.

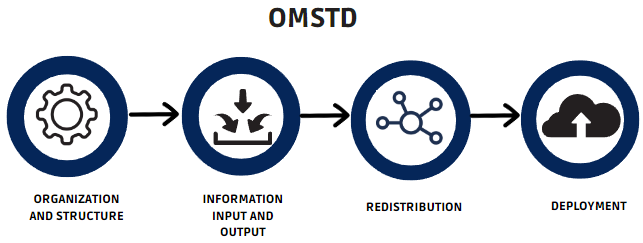

Figure 3. Life cycle OMSTD.

The methodology offers a robust and adaptable approach for spam detection, focusing on the selection and cleaning of relevant data from large volumes of e-mails, optimizing key parameters and features to improve model accuracy. In the modeling phase, AI algorithms such as neural networks and decision trees are developed and trained to identify complex and subtle patterns in e-mail data that distinguish between legitimate messages and spam. Through simulations, these models are rigorously evaluated and refined to maximize their effectiveness.

The testing phase involves validating the models with realistic and diverse datasets to ensure their reliability and robustness. Finally, in the deployment stage, spam detection systems are integrated into production environments, where they continue to learn and adapt to new spam tactics through online machine learning techniques. The results of applying OMSTD with AI in spam detection have shown significant improvements in detection rates, reducing false positives, and maintaining consistent performance as spam threats evolve. This methodological approach not only enhances the effectiveness of spam filtering systems but also ensures their long-term adaptability and sustainability in the fight against spam.

4. Discussion and results

In this section, we summarize and discuss the results obtained, highlighting potential avenues for future research on optimization and automated spam detection.

The results shown below demonstrate the evaluations performed on different supervised machine learning algorithms specific to spam detection where each of them was trained and different techniques were applied to evaluate their performance.

The algorithms selected in this phase for subsequent training are:

- Random forest
- Decision tree
- Naïve Bayes

The amount of data for training the algorithm amounted to 60,000 records. To ana-lyse the e-mails, tokenization techniques were implemented, which involves the conversion of phrases and words into an array. Non tokenization, where words are examined directly, was also exploited. Additional parameters, such as test_size = 0.002, random_state =42, and Chunk_size = 41000, were employed to fine-tune the process.

For Random Forest training, a non-tokenized dataset was chosen. This approach resulted in a process that took about 3 minutes to analyses 8006 training data, leaving the remaining data for testing. This strategic tuning and selection of parameters contributed to the efficiency of the model and optimization of analysis time.

In the next stage, training of the Random Forest algorithm was conducted using the tokenized dataset. This process involved splitting a whole text into words, which generated a more complex and robust dataset. However, this approach had a longer analysis time of around 10 minutes for both the training and testing phases. Detailed results are presented in table 1.

This processing delay is attributed to the greater complexity of the tokenized dataset. Although this method provided a more detailed representation of the content, the results obtained were inferior compared to training using the non-tokenized dataset. This finding highlights the importance of selecting the right approach according to the specific characteristics and objectives of the analysis.

Table 1. Particle size and power consumption.

Without tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.97	0.98	0.98	4384
Spam	0.97	0.98	0.98	3662
Accuracy			0.9731	8006
With tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.95	0.99	0.97	4415
Spam	0.99	0.94	0.96	3591
Accuracy			0.9665	8006

During the running time of the decision tree algorithm with the untokenized dataset, it was found that the running time was approximately 5 minutes between training and testing data, with the following data, in which a total of 8006 of the original datasets was used, equivalent to 20% of both spam and ham data. Detailed results are presented in table 2.

The training of the algorithm with the tokenized dataset took approximately 10 minutes of waiting, where the results improved for the evaluation metrics of Recall and F1-score compared to ham, but decreased for the accuracy metrics, Recall and F1-score, resulting in an overall decrease of approximately 1%.

Table 2. Decision tree results without and with tokenization.

Without tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.94	0.93	0.94	4384
Spam	0.92	0.93	0.93	3622
Accuracy			0.9323	8006
With tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.92	0.94	0.93	4415
Spam	0.93	0.90	0.91	3591
Accuracy			0.9245	8006

The results obtained in the training of the Naive Bayes algorithm with the untokenized data showed results in most metrics of 0.94, equivalent to 94% and even reaching up to 99%, giving a first result of 0.96, equivalent to 96%. Detailed results are presented in table 3.

When training was established with the tokenized dataset, performance was reduced with respect to the other dataset, where the most affected metrics were spam accuracy, Recall in the ham category and a reduction in F1-score in both categories.

Table 3. Naive Bayes results without and with tokenization.

Without tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.99	0.95	0.97	4384
Spam	0.94	0.99	0.96	3622
Accuracy			0.9660	8006
With tokenization				
Type	Precision	Recall	F1-score	Support
Ham	0.99	0.91	0.95	4384
Spam	0.90	0.99	0.94	3622
Accuracy			0.9454	8006

During the process of the previous stages, data sets were selected to be processed and prepared according to established parameters, such as eliminating those characteristics that would not be relevant at the time of training the algorithms and those that do not have the appropriate format.

To clean and process the data, several techniques were implemented such as translation into Spanish, elimination of duplicate and null values, as well as counting the total values of the Enron dataset, determining that the amount was relatively large, so the technique of separating the dataset into small segments and applying processing threads to perform the translation more quickly was applied, and then these were merged into the original dataset.

After processing and data cleaning, the technique of tokenization was applied to the dataset for more efficient training, where three automatic learning algorithms were selected for further analysis.

During the data mining phase, the effectiveness of each algorithm was evaluated with the dataset containing the tokenized and un-tokenized data, giving different results as shown in that phase, giving the Random Forest algorithm with the best training statis-tics on the same dataset, but with different techniques.

As expected, the algorithm to be recommended in this research will be the Random Forest due to the statistics obtained during the various phases applied in the KDD and OMSTD methodology, obtaining superior results compared to algorithms such as Naïve Bayes and decision tree. This integration of methodologies ensured optimal performance, highlighting the importance of a structured approach to enhancing spam detection capabilities through AI, as visualized in Figure 4.

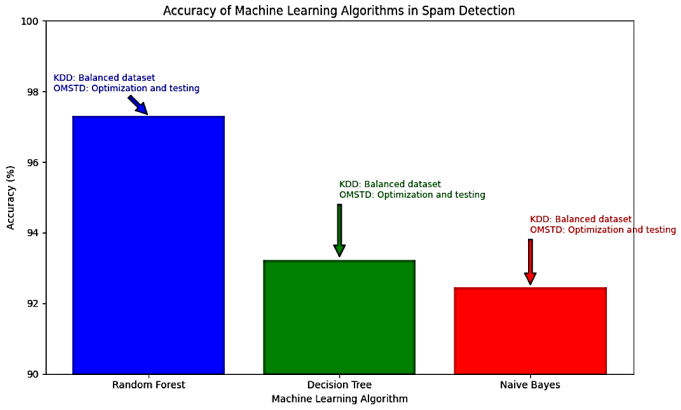

Figure 4. Accuracy of machine learning algorithms in spam detection.

4.1. Figures (science style)

In addition, a table summarizing the alerts sent to e-mail and the e-mails reported by users during a week of work activity in the institutional environment is presented. This information provides a detailed view of the efficiency of the system in a real-world scenario.

Table 4. Results of application of the algorithm.

Day	Analyzed e-mails	Ham	Spam	Alerts sent	Re-ported mail
1	45	34	11	18	0
2	153	136	17	17	1
3	218	202	16	18	2
4	444	417	27	27	0
5	303	292	11	11	0
Total	1163	1081	82	91	3

The analysis of this table 4 reveals patterns and trends in the alerts generated and in the responsiveness of users to potential threats. This operational insight is essential to continuously adjust and improve the spam detection system, ensuring an initiative-taking and efficient defense against potential cyber risks.

5. Conclusions

This study highlights the need to address human vulnerability in information security, underlining that, despite robust policies implemented by antivirus and firewalls, the user remains the weakest link when interacting on the network.

Comparing various machine learning algorithms, Random Forest leads with an accuracy of 97.31%, standing out as the preferred choice for spam detection. Decision tree, with 93.23%, performs solidly, and Naive Bayes, with 92.45%, proves to be a respectable choice in the fight against spam.

The success of these algorithms lies in a balanced dataset, classified into mail types, and trained appropriately. The KDD methodology provided a robust dataset, while OMSTD provided the technological basis for agent development.

Test results in a Zimbra environment show the effectiveness of the agent, detecting 70 out of 75 normal mails and 70 out of 75 spam mails, with low false positives. This approach has potential applications in intrusion detection, fraud, and other fields.

In conclusion, the successful implementation of machine learning algorithms depends crucially on the quality of the data extracted from the e-mails. This study lays the groundwork for future research, exploring challenges and improvements in deep learning algorithms for spam detection.

Conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

The authors would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

Funding

The authors received no specific funding for this work.

References

Bischoff, P. (2022, August 28). The State of Phishing in the US: Report and Statistics 2021. Comparitech.

Fernández, R. (2023). E-mails: correos electrónicos recibidos diariamente en el mundo hasta 2025 | Statista.

Junnarkar, A., Adhikari, S., Fagania, J., Chimurkar, P., & Karia, D. (2021). E-mail spam classification via machine learning and natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 693-699). IEEE. https://doi.org/10.1109/ICICV50876.2021.9388530

Messina Valverde, A. (2021). *Diseño e implementación de una extensión de Chrome para la detección de sitios web de Phishing utilizando aprendizaje automático* (Bachelor's thesis). http://hdl.handle.net/10486/700048

Morgan, S. (2020). *Cybercrime To Cost The World $10.5 Trillion Annually By 2025.* Cybercrime Magazine. (2020, November 15) https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

NIBUSINESS. (n.d.) *Advantages and disadvantages of using email for business, nibusinessinfo.co.uk*. Business. https://www.nibusinessinfo.co.uk/content/advantages-and-disadvantages-using-email-business

Nichols, J. A., HW, H. C., & Baker, M. A. B. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, *11*(1), 111-118. https://doi.org/10.1007/s12551-018-0449-9

OMSTD Project. (2014). Conceptos de desarrollo — OMSTD - Open Methodology for Security Tool Developers :Documentation.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies.

Radev, D. (2008). CLAIR collection of fraud e-mail. *ACL Data and Code Repository*. ACL Wiki (aclweb.org)

THE RADICATI GROUP, INC. (2021). A Technology Market Research Firm. Statistics Report, 2021-2025.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. *Universidad Cooperativa de Colombia, Facultad de Ciencias Sociales, Programa de Maestría en Educació*n, Bogotá, Colombia, 00000. https://doi.org/10.16925/9789587600490

The State of Phishing in the US: Report and Statistics 2021. (2022, March 14) *Comparitech*. https://www.comparitech.com/blog/information-security/state-of-phishing/#:~:text=In