Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 05 May-2025, Page No.- 4847-4854

DOI: 10.47191/etj/v10i05.14, I.F. – 8.482

Email Spam Classification Based on Logistics Regression

Iman Youssif Ibrahim[1], Omar Sedqi Kareem[2]

[1]Akre University for Applied Science, Technical College of Informatics, Akre, Department of Information Technology, Akre-Duhok, Kurdistan Region, Iraq

[2]Department of Public Health, College of Health and Medical Techniques – Shekhan , Duhok Polytechnic University, Duhok Kurdistan Region, Iraq

ABSTRACT: Email is a common communication tool used by both individuals and organizations. It involves a variety of interactions, including file sharing. In addition to the advantages it provides, there is the uninvited email sharing. This unsolicited email is referred to as "spam." Malicious content like viruses, phishing scams, and unsolicited ads can be found in spam. It is well known that communication security is crucial. In order to filter email systems for malicious tools or software, it is essential to classify them based on a variety of criteria. In these kinds of classification studies, machine learning algorithms work well. The objective of this research is to solve the problem at hand and compare the logistic regression, random forest, naive Bayes decision tree, and support vector machine (SVM) algorithms. The effects of various methods and approaches on the issue were thoroughly examined. A comparison of the various performance outcomes using the various approaches is provided. With an accuracy of 98%, logistic regression was the most accurate, followed by random forest with 97%.

KEYWORDS: Email classification, Spam detection, Machine Learning.

1. INTRODUCTION

As internet technology has advanced, more people are using email applications to share files and messages. It is widely used for a variety of reasons, including low transmission costs, quick message exchange, and ease of use (Alsuwit et al., 2024; Alshawi et al., 2024). Despite their positive effects, malicious people have started to target email applications. Malicious attachments or suspicious messages are frequent issues (Labonne & Moran, 2023). Spam, or unsolicited emails, tell the recipient that they are either irrelevant or that the sender is sending them to a large number of recipients with dubious content (Zavrak & Yilmaz, 2023; Janez-Martino et al., 2024). When recipients click on links in emails that contain such dubious content, they may be taken to malicious websites or infected computers (Qiao, 2024). The technical infrastructure is also damaged. It has become crucial to categorize spam emails and stop them from getting to users (Zhang, 2024; Occhipinti et al., 2022).

Research on this subject has been done and continues to be done. The classification problem has been studied using a variety of machine learning algorithms (Aggarwal & Zhai, 2012; Bhowmick & Hazarika, 2016). Text classification is the foundation of the email classification problem's history (Yu & Xu, 2008; Seth & Biswas, 2017). Studies on these algorithms that are tailored to the issue, like content or text classification, exist (Ohnishi & Yoshida, 2004; Youn & McLeod, 2007). Sample applications of the algorithms were analyzed within the parameters of the pertinent study (Hassan & Mtetwa, 2018; Sharma & Sahni, 2011). Six distinct machine learning techniques were modified for the spam email classification problem in this study (DeBarr & Wechsler, 2012; Labonne & Moran, 2023). This study's testing process made effective use of datasets gathered from a variety of sources. A selected dataset was used for the experiments. The research is organized as follows: Section 2 discusses the latest techniques used to classify real emails from spam. Section 3 discusses the data used and the models used in classification. Section 4 presents the results obtained by the proposed models. Section 5 discusses and compares the proposed models with previous studies to determine the effectiveness of the proposed system. Finally, Section 6 presents the main conclusions of the research.

2. RELATED WORK

The field of spam classification has witnessed advances in recent years, and various machine learning techniques have been used to improve spam detection. Zhang (2024) compared three algorithms: naive Bayes, random forest, and logistic regression. The results showed that the random forest algorithm was the most accurate It achieved an accuracy of 93%. In a paper by Occhipinti et al. (2022), the authors compared twelve machine learning models for spam classification. They used the Enron dataset and achieved an F1 score of 94% used XGBoost.. In Li (2023), the authors compared the performance of machine learning models such as SVM, KNN, and Naive Bayes in classifying spam messages. Their results showed that Naive Bayes was more

accurate than the other models. In (Ouyang et al., 2023), the authors compared two machine learning models, KNN and Naïve Bayes, for spam classification. KNN was found to perform better at classifying spam messages, while Naïve Bayes took less time. A study by (Deng, 2023) compared five spam filtering classification models, including Naive Bayes, SVM, KNN, and XGBoost. The study aimed to determine the performance of each of these techniques in classifying spam messages. In (Sutta et al., 2020), the authors studied the effect of N-Grams in preprocessing on the performance of different classification models. The results showed that using N-Grams improved the accuracy of the models in most cases.

These papers demonstrate the importance of machine learning models in spam classification. By comparing different models, we can identify the most effective algorithms and leverage them to improve spam blocking mechanisms.

Materials and Methods

In this section, information is provided about the research methodology, the methods used, and the test data. The research methodology, illustrated in Figure 1, shows the steps involved.

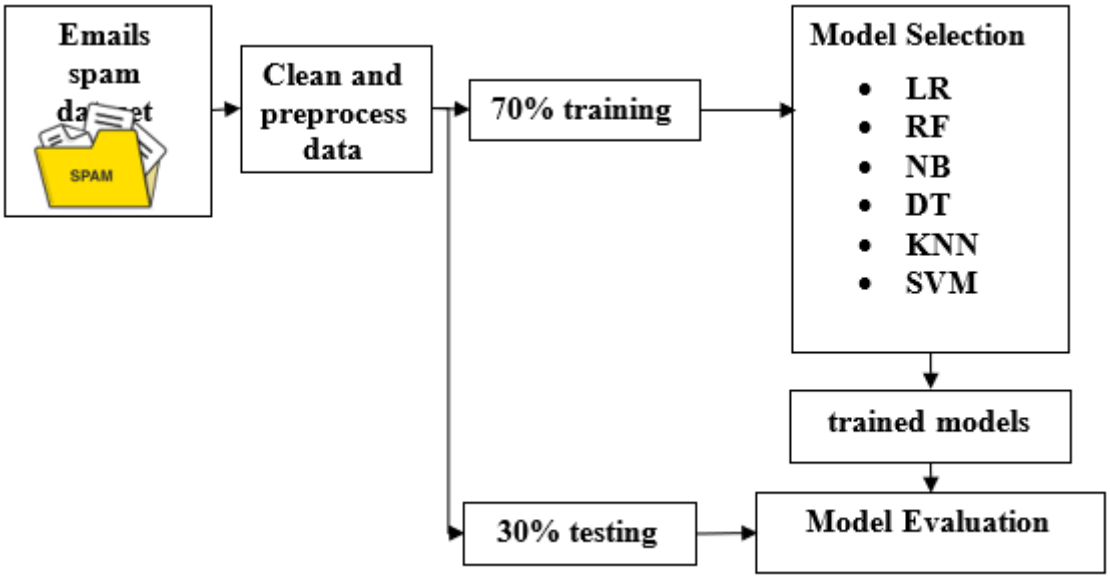


Figure 1. steps of Proposed methodology.

As shown in Figure 1, the process starts from identifying and processing the data set, then dividing the data set into a training set and a test set with a ratio of 70 to 30 percent. The selected models are trained (the logistic regression, random forest, naive Bayes decision tree, and support vector machine). After the training process, the trained models are tested using the test set and evaluated using (Accuracy, Precision, Recall, F1-Score, and ROC AUC Score).

3.1 Dataset

The dataset used in this study was a CSV file with 5,172 instances, each of which represented an email message (Biswas, B. 2020). There are 3,002 features in all. The email label appears in the first column. Personal names were replaced with numerical identifiers for identification purposes in order to maintain confidentiality. Prediction scores of 1 or 0 for spam or no spam are assigned in the dataset's final column. After deleting letters and non-alphabetical words, the remaining 3,000 includes indicate the 3,000 most common words in all emails. The dataset is tabular in structure, with each row representing a unique email message and each column representing a particular email attribute. The data is stored in a uniform and compact format, which makes it possible to scan and process the data efficiently. As a result, all 5,172 emails are consolidated into a single, compact dataframe instead of being stored separately in separate emails.

3.2 Machine Learning Models

To select the appropriate model, several models were tested and compared.

One supervised machine learning technique for categorization is the Support Vector Machine (SVM). SVM tries to find the optimal margin between the hyperplane of two classes in a training sample of two classes. The data points shouldn't be near a decent hyperplane. The hyperplane is positioned distant from each class's data points. A support vector is any point that is close to the classifier's margin. The optimal hyperplane that optimizes the separation between the two classes' decision borders is chosen via SVM (Boateng et al., 2020).

The Naive Bayes classifier is one of the most popular models based on Bayesian theory, a probabilistic classification method. This method assumes that each feature is uncorrelated and independent, meaning that no feature in a class influences the behavior of other features. This method is effective when features are uncorrelated because it relies on the conditional probabilities of each class. Because it takes

into account every possible probability, this method performs well even when the data is imbalanced (Boateng et al., 2020). The decision tree is a supervised machine learning algorithm for data classification that operates on the principle of inference. The main goal of this algorithm is to identify decision rules derived from prior information to predict the target class. Decision trees consist of a root, nodes, and terminal leaves. After generating nodes and subnodes for prediction and classification, the root node uses a set of criteria to classify the instance. The leaf nodes represent the target class, and each root node has at least two subnodes. From all the features used, the optimal node with the greatest information gain at each level is selected (Chen et al., 2023).

K-nearest neighbors (K-NN) is a supervised machine learning algorithm used to address regression and classification problems. K-NN is an instance-based learning technique that uses training data, or the nearest data, to determine an object's distance from a given location in order to classify it. K-NN uses similarity measures (distance functions) to classify new points after storing all available points. A majority vote of local nearest neighbors determines the point's classification (Boateng et al., 2020; Chen et al., 2023).

Logistic regression (LR) is another easy-to-use classification approach in machine learning. It is based on predicting the probability of a categorical output feature. This algorithm uses a logistic function to estimate probabilities to determine the relationship between the categorical output feature and one or more input features. The output feature of a logistic regression algorithm is often binary (Hassan and Amiri, 2019). Random Forest (RF) is a machine learning algorithm based on the principle of ensemble learning, specifically bagging. In this model, a set of different decision trees is trained on randomly selected parts of a dataset. The outputs of each learner are then aggregated. Majority voting is used to select the final classification/prediction, and the classification with the most "votes" wins (Chen et al., 2023; Hassan and Amiri, 2019).

Figure 2 represents the confusion matrix for each model.

4. RESULTS

To determine the most accurate model for identifying spam emails, several models are compared based on the confusion matrix and Accuracy, Precision, Recall, F1-Score, and ROC AUC Score.

Accuracy:

Represents the percentage of messages correctly classified (whether spam or not) to the total number of messages.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

Represents the percentage of messages classified as spam that were actually spam.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall or Sensitivity:

Represents the percentage of spam messages successfully detected.

$$Recall = \frac{TP}{TP + FN}$$

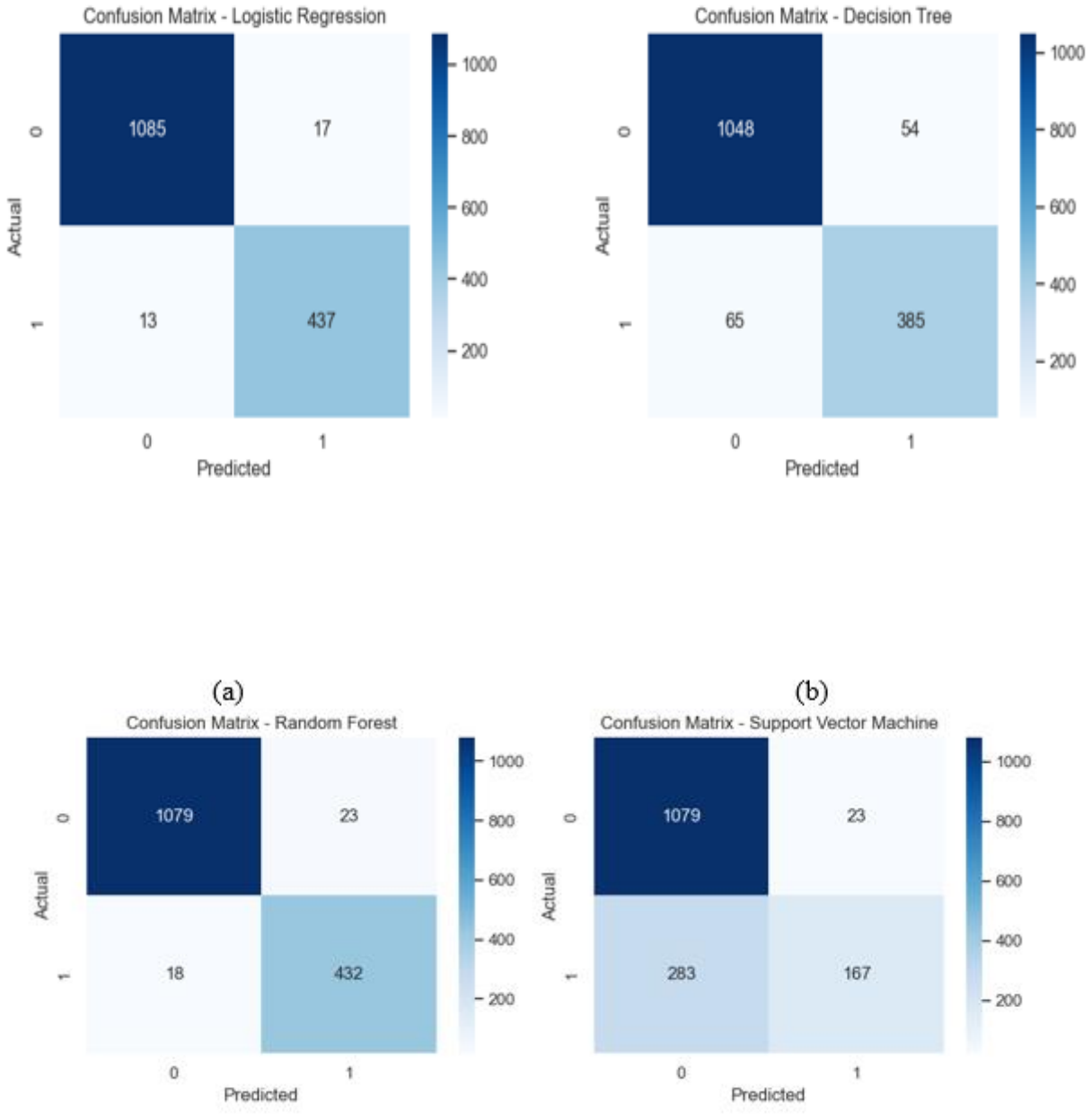
F1-Score:

The harmonic mean between precision and recall, used when a balance between the two is needed.

$$F1 - Score = 2 \cdot \frac{Precision.Recall}{Precision + Recall}$$

Area Under the ROC Curve (ROC AUC):

Measures the model's ability to discriminate between classes. The closer the value is to 1, the better the model.


Confusion Matrix - Logistic Regression
1085
17
13
437
Actual
Predicted
0
1
1000
800
600
400
200
Confusion Matrix - Decision Tree
1048
54
65
385
Actual
Predicted
0
1
1000
800
600
400
200
(a)
Confusion Matrix - Random Forest
1079
23
18
432
Actual
Predicted
0
1
1000
800
600
400
200
(b)
Confusion Matrix - Support Vector Machine
1079
23
283
167
Actual
Predicted
0
1
1000
800
600
400
200

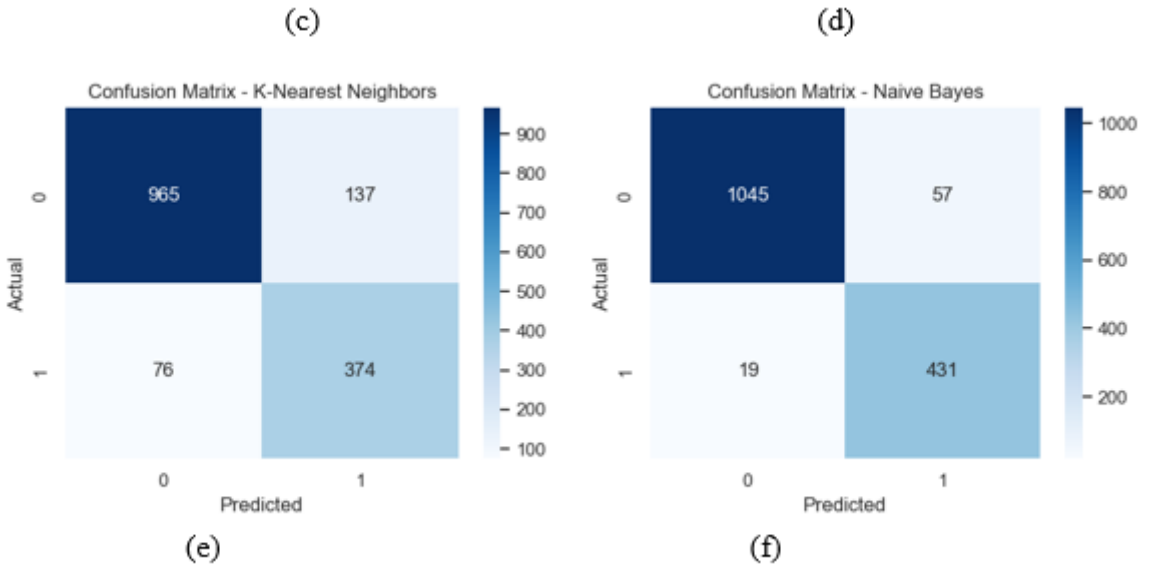
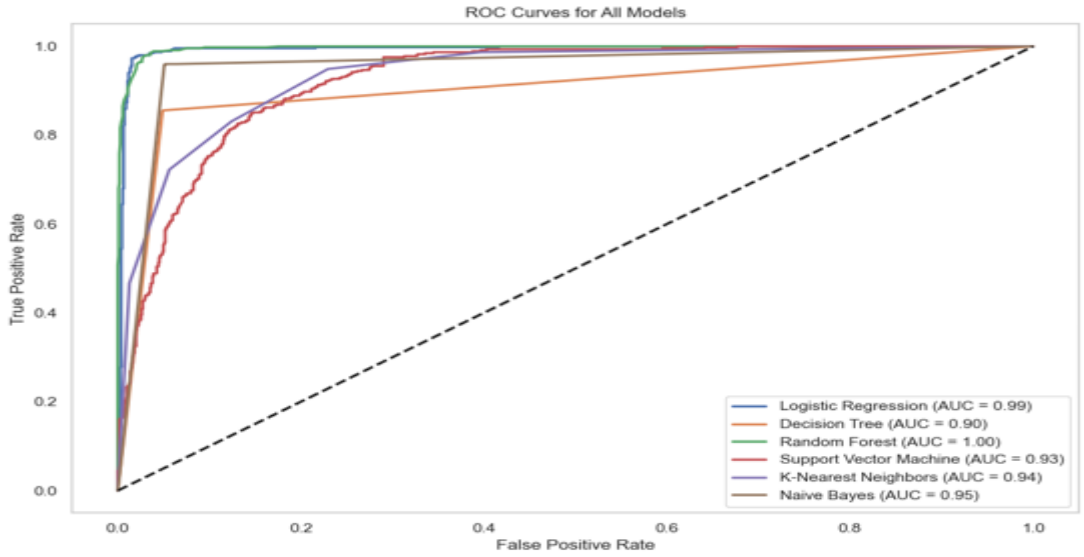


Figure 2. confusion matrix for each model: (a) Logistic Regression (LR), (b) Decision Tree (DT), (c) Random Forest (RF), (d) Support Vector Machine (SVM), (e) K-Nearest Neighbors (KNN), (f) Naïve Bayes (NB).

One of the important measures in the case of data set imbalance, as in the case of spam email classification, is to use the ROC AUC Score. Figure 3 shows the comparison between models based on the ROC AUC Score.



. **Figure 3. Comparison between models based on the ROC AUC Score.**

To compare models based on accuracy, Precision, Recall and F1-Score Table 2 shows the comparison.

Table 2 the comparison between models based on Precision, Recall and F1-Score.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.98067	0.962555	0.971111	0.966814	0.99312
Random Forest	0.973582	0.949451	0.96	0.954696	0.996243
Naive Bayes	0.951031	0.883197	0.957778	0.918977	0.954261

Decision Tree	0.923325	0.876993	0.855556	0.866142	0.903277
K-Nearest Neighbors	0.862758	0.731898	0.831111	0.778356	0.936414
Support Vector Machine	0.802835	0.878947	0.371111	0.521875	0.925154

5. DISCUSSION

To determine which machine learning models were most effective at differentiating between spam and ordinary mail, six models were tested. These models produced different outcomes.

With an accuracy of 98%, logistic regression was the most successful. This implies that it would correctly classify 98 of the 100 messages we sent it as spam or regular. This model has the great advantage of being balanced; it does not miss many real spam messages or incorrectly classify ordinary mail as spam.

Random Forest performed best on the ROC AUC metric, which gauges the model's capacity to discriminate between the two classes, but it trailed logistic regression by a small margin with an accuracy of 97%. This implies that in some circumstances, it might be a better option.

Other models, like Decision Tree and Ordinary Bayes, did well. For instance, Ordinary Bayes detected the majority of spam (about 96%) but had some false positives, meaning that some legitimate messages were mistakenly identified as spam. In contrast, Decision Tree's performance was about in the middle of all metrics. However, on this particular task, models such as SVM and KNN did not perform well. SVM in particular did a good job of not misclassifying ordinary mail as spam, but it was unable to identify the majority of spam (only 37%).

The primary finding is that, in our situation, random forest and logistic regression are the best options for the spam classification task. It's crucial to remember that these outcomes rely on the data used in the experiment; if we use different data or operate under different conditions, the outcomes could be different.

This experiment illustrates that there is no one model that works best in every scenario; rather, the decision is based on our priorities. For example, do we want to capture as much spam as possible at the expense of misclassifying some legitimate messages? Or, even if we miss some genuine spam, would we rather make sure that regular messages aren't mistakenly labeled as "spam"? Depending on the requirements of each application, the response varies. To determine the efficiency of the tested models, logistic regression was compared with related work in spam classification.

Table 3 compares the results of logistic regression and related works.

Reference	Model	Dataset	Acc	Precision	Recall	F1 Score	ROC AUC
Zhang (2024)	Random Forest	Complete Spam Assassin	0.93	0.942	0.956	0.949	N/A
Occhipinti et al. (2022)	Random Forest	Enron	0.947	0.936	0.943	0.940	N/A
Jazzar et al. (2021)	SVM	UCI	0.967	0.958	0.962	0.960	0.980
Li (2023)	Naïve Bayes	Spam Base	0.97	0.93	0.87	N/A	0.93
Ouyang et al. (2023)	KNN	Enron	0.960	0.930	0.940	0.935	N/A
This Study	**Logistic Regression**	*Email spam classification*	**0.980**	**0.962**	**0.971**	**0.966**	**0.993**

6. CONCLUSIONS

In the digital age, spam is a widespread issue that is frequently sent by scammers and advertisers who want to promote a product or service or engage in phishing. However, by examining patterns and trends in sizable email datasets, deep learning algorithms can be used to efficiently detect and weed out unwanted emails. In order to improve user experience and lower the risks associated with phishing and other malicious activities, this study applied several machine learning algorithms to classify spam emails, including logistic regression (LR), random forest (RF), support vector machine (SVM), naive Bayes (NB), and decision tree (DT). The results indicate that logistic regression achieved the best spam detection accuracy with 98%. Random forest and support vector machine models also performed well. This study provides insight into the performance of different

classification models in spam filtering. The results obtained, particularly the high performance of the logistic regression model, contribute to improving email filtering systems and making them more efficient and accurate. However, the dataset may not accurately reflect the diversity of spam messages found in the real world. More diverse and recent datasets should be included in future research to improve the model's generalizability and effectiveness against sophisticated spam tactics.

REFERENCES

1. Zavrak, S., & Yilmaz, S. (2023). Email spam detection using hierarchical attention hybrid deep learning method. Expert Systems with Applications, 233, 120977. https://doi.org/10.1016/j.eswa.2023.120977
2. Alsuwit, M. H., Haq, M. A., & Aleisa, M. A. (2024). Advancing email spam classification using machine learning and deep learning techniques. Engineering, Technology & Applied Science Research, 14(4), 14994–15001. https://doi.org/10.48084/etasr.7631 etasr.com
3. Alshawi, B., Munsh, A., Alotaibi, M., Alturki, R., & Allheeib, N. (2024). Classification of SPAM mail utilizing machine learning and deep learning techniques. International Journal on Information Technologies and Security, 16(2), 71–82. https://doi.org/10.59035/FPKO7430
4. Labonne, M., & Moran, S. (2023). Spam-T5: Benchmarking large language models for few-shot email spam detection. arXiv preprint arXiv:2304.01238. https://arxiv.org/abs/2304.01238
5. Janez-Martino, F., Alaiz-Rodriguez, R., Gonzalez-Castro, V., Fidalgo, E., & Alegre, E. (2024). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. arXiv preprint arXiv:2402.05296. https://arxiv.org/abs/2402.05296
6. Zhang, J. (2024). Machine learning-based email spam filter. Innovation in Science and Technology, 3(3), 36–51. https://www.paradigmpress.org/ist/article/view/1130
7. Qiao, Y. (2024). Spam email classification based on SVM, Transformer and Naive Bayes. Applied and Computational Engineering, 48, 161–167. https://www.ewadirect.com/proceedings/ace/article/view/10963 ewadirect.com
8. Occhipinti, A., Rogers, L., & Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. arXiv preprintarXiv:2204.06518. https://arxiv.org/abs/2204.06518
9. Bhowmick, A., & Hazarika, S. M. (2016). Machine learning for e-mail spam filtering: Review, techniques and trends. arXiv preprint arXiv:1606.01042. https://arxiv.org/abs/1606.01042
10. Zavrak, S., & Yilmaz, S. (2023). Email spam detection using hierarchical attention hybrid deep learning method. Expert Systems with Applications, 233, 120977. https://doi.org/10.1016/j.eswa.2023.120977
11. Labonne, M., & Moran, S. (2023). Spam-T5: Benchmarking large language models for few-shot email spam detection. arXiv preprint arXiv:2304.01238. https://arxiv.org/abs/2304.01238
12. Janez-Martino, F., Alaiz-Rodriguez, R., Gonzalez-Castro, V., Fidalgo, E., & Alegre, E. (2024). Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. arXiv preprint arXiv:2402.05296. https://arxiv.org/abs/2402.05296
13. Qiao, Y. (2024). Spam email classification based on SVM, Transformer and Naive Bayes. Applied and Computational Engineering, 48, 161–167. https://www.ewadirect.com/proceedings/ace/article/view/10963 ewadirect.com
14. Deng, X. (2023). Email Spam Filtering Methods: Comparison and Analysis. Humanities and Social Sciences Research, 8(4). \ https://drpress.org/ojs/index.php/HSET/article/view/5805
15. Jazzar, M. H., Aljarah, I., & Aldwairi, M. (2021). Spam Email Detection Using Machine Learning Algorithms. International Journal of Education and Management Engineering, 11(4), 35–43. https://www.mecs-press.org/ijeme/ijeme-v11-n4/v11n4-4.html
16. Li, J. (2023). A Comparison of Three Algorithms in Spam Classification. Humanities and Social Sciences Research, 8(3). https://drpress.org/ojs/index.php/HSET/article/view/5436
17. Occhipinti, G., Trovato, M., & Alessi, L. (2022). Comparative Study of Text Classification Algorithms for Spam Detection. arXiv preprint arXiv:2204.06518. https://arxiv.org/abs/2204.06518
18. Ouyang, S., Lin, H., & Zhao, Y. (2023). Spam Filtering Based on KNN and Naive Bayes Classifier. Humanities and Social Sciences Research, 8(4). https://drpress.org/ojs/index.php/HSET/article/view/5699
19. Sutta, R., Agrawal, S., & Mandal, J. K. (2020). A Study of Machine Learning Algorithms on Email Spam Classification. EasyChair Preprint. https://easychair.org/publications/paper/Jvsw
20. Zhang, M. (2024). Comparative Evaluation of Email Spam Classifiers Using Machine Learning. Information Science and Technology, 14(1).

https://www.paradigmpress.org/ist/article/view/1130Biswas, B. (n.d.). Email spam

21. Biswas, B. (2020) Email spam classification dataset CSV, Kaggle. Available at: https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv (Accessed: 16 April 2025).
22. Hassan, M. M., & Amiri, N. (2019, October 11-12). Classification of imbalanced data of diabetes disease using machine learning algorithms [Paper presentation]. IV. International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE), Istanbul, Turkey.
23. Chen, Z., Liu, Y., & Tie, N. (2023). Forest land resource information acquisition with Sentinel-2 image utilizing support vector machine, K-nearest neighbor, random forest, decision trees and multi-layer perceptron. Forests, 14(2), 254. https://doi.org/10.3390/f14020254.
24. Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. Journal of Data Analysis and Information Processing, 8(4), Article 20. https://doi.org/10.4236/jdaip.2020.84020