Email Spam Detection Using Machine Learning: A Comparative Analysis of Classification Algorithms

Md. Bahauddin Sakib[1*], Sourav Mistry[1]

[1] Department of Computer Science, American International University, Dhaka, Bangladesh
[1] Department of Computer Science, American International University, Dhaka, Bangladesh

[*1]E-mail: 22-46093-1@student.aiub.edu
[1]E-mail: 22-46077-1@student.aiub.edu

Abstract. Today's digital world has seen a huge increase in spam email, which can cause problems such as phishing, fraud and system risk. Despite the internet and more and more dependent on the email communication using the internet, spam makes a large part of email traffic worldwide. Development of effective methods is necessary to detect and block unwanted messages. The research noted the use of three popular machine Learning algorithms- support vector machine, decision tree and naive bayes, which were to identify emails as real (ham) or spam. The dataset used in the study had 4,825 non-spams of 5,572 messages sent by email and 747 spams. Using the approach of such bags (bow) and term frequency-inverse document frequence (TF-IIDF), we demonstrated a comprehensive data preparation process that included cleaning, generalization and feature extraction. We divide the dataset into training (70%) and test (30%) sets to ensure a balanced assessment. Performance was assessed using matrix such as accuracy, accuracy and memorial. In three algorithms, SVM performed better with 98% accuracy. Although NB and DT also created satisfactory results, SVM proved to be the most effective in handling complex datasets. An important discovery from this study is the ability to further enhance accuracy by combining advanced feature extraction methods with customized SVM kernels. This research underlines the effectiveness of machine learning in combating email spam and provides valuable insight to improve detection systems.

Keywords: Spam Detection, Machine Learning, SVM, Naïve Bayes, Decision Tree, Feature Extraction, TF-IDF.

1. Introduction

Electronic mail, commonly known as email, is a method of transmitting messages between two devices through the Internet [1]. It has become the primary method of communication for individual connections, commercial updates, confidential exchange and global research cooperation [2]. As at the time of writing, 4.48 billion individuals worldwide used emails, indicating an increase in accounts up to 2027 [3] with estimates. In 2024, there were about 4.48 billion active email accounts [4]. In 2024, the daily email traffic reached 361.6 billion, with an increase of about 408.2 billion per day by 2027 [5]. The spam emails 46% of the 347 billion daily emails, with a total of 160 billion spam messages, as recorded in 2023 [6]. An overwhelming majority of users (96.8%) have faced spam communication. Google blocks about 100 million phishing emails daily [7]. By 2024, the United States and China were responsible for distributing spam email more than 7.8 billion times globally [8].

In recent years, many researchers [6, 9, 10, 12] have detected various machine learning approaches for spam detection. A study [9] proposed a model, including feature selection, machine learning, and stacking techniques, with 97% accuracy. Another research [10] employed feature extraction in preprocessing and evaluated ten classification algorithms with random forest 98.87% accuracy. The authors of [11] used an SVM model with a kernel function to determine the data point similarity in a high-dimensional location. In [12], researchers applied machine learning techniques such as DT, KNN, NB, SVM, and rough set classification for email spam detection, emphasizing classification methods. This paper introduces a skilled machine learning model based on lessons to classify messages such as spam (1) or non-spam (0). The purpose of research is to increase accuracy, showing better performance on 98% accuracy on a Kaggle dataset with SVM model including 4,825 non-spams and 747 spam messages [22]. Despite the progress in spam detection, issues such as overfitting, scalability and entertainment remain unresolved, underlining the need for comprehensive evaluation of many models [13,17,19].

This study proper assessment of Naive Bayes, Decision Tree and SVM algorithms using standard matrix including accuracy, accuracy, recall, and F1-score to ensure a reliable comparison, [14, 15, 18] to ensure. By addressing scalability and entertainment, this research highlights the effectiveness of SVM models by contributing to practical, adaptable spam detection solutions applied to various datasets and real-world scenarios [16,21].

2. Literature Review

An essential area of research has been created due to the increasing amounts of unwanted emails by the detection of email spam, which creates significant security hazards and privacy issues. Various machine learning (ML) techniques have been detected to increase the accuracy and efficiency of spam detection systems. Naive Bayes (NB), a potential classifier based on Bayes theorem, has been widely adopted due to its simplicity and effectiveness in handling high-dimensional text data. Despite its perception of convenient freedom, which cannot always be in practice, NB has performed commendably in spam classification works [4]. Decision Tree (DT) models provide lectures by employing measures such as index and information benefits to make decisions. However, especially with noisy data, their tendency limits their scalability to overfit. Dress methods such as random forests have been proposed to reduce overfitting by collecting several DTs, increasing generalization capabilities [3]. The support vector machine (SVM) has emerged as a strong solution and is excellent in maximizing margins between classes to handle high-dimensional data and achieve better accuracy. SVMs are effectively applied in spam detection and often perform better than other classifiers when combined with appropriate feature extraction techniques [1]. The advent of deep learning (DL) has introduced models capable of capturing complex patterns in data. The recurrent neural network (RNN) and the convolutional neural network (CNN) have been employed for spam detection, taking advantage of their capacity for modeling sequential and spatial information, respectively. These models have shown promise, especially when a large, labeled dataset is available for training [7]. Recent progress has seen the integration of natural language processing (NLP) techniques with ML models to increase spam detection. Ways such as Term Frequency-Inverse Document Frequency (TF-AIDF) and Word2Vec and BERT have been used to effectively represent email content. These representatives, when fed into classifiers such as logistic regression (LR) and SVM, indicate high accuracy and F1 scores, which indicate the efficacy of combining NLP with ML for spam detection [14]. In addition, applications of clustering techniques, such as agglomerative hierarchical clustering, have facilitated the classification of spam email in many classes and assisted in more fine filtering strategies. This approach has been particularly useful in identifying and separating a variety of spam content, which has improved the overall effectiveness of spam filters [11].

3. Methodology

This research utilizes quantitative methodology to develop and evaluate an efficient spam email detection model. The methodology adopts a systematic, step-by-step approach, ensuring the thorough development, implementation, and validation of the proposed model.

3.1. Data Collection:

The dataset for this study was sourced from reputable online repositories such as Kaggle and GitHub, ensuring reliability and quality [22]. The dataset comprises 5,572 email messages categorized into two classes: 4,825 non-spam (ham) emails and 747 spam emails. Each record includes two fields: the email text and its corresponding label ("spam" or "ham"). This well-labelled dataset provides a robust foundation for building and testing the model.

3.2 Data Preprocessing:

Data preprocessing is a vital step to enhance the quality of input data and ensure optimal model performance. It involves:

- **Data Cleaning**: Duplicate entries, null values, and extraneous characters such as HTML tags and special symbols were removed to eliminate noise [15].
- **Text Normalization**: Email text was converted to lowercase, and stop words and punctuation were removed to simplify and standardize the dataset [18].
- **Balancing the Dataset**: Techniques such as oversampling the minority class (spam) or under sampling the majority class (ham) were employed to address class imbalance and prevent biased predictions [13].
- **Feature Extraction**: Textual data was transformed into numerical representations using Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods. These techniques effectively capture textual patterns for model learning [20].

3.3. Data Splitting:

To train and evaluate the model, the pre-processed dataset was divided into two subsets:

- **Training Set (70%)**: Used to train the machine learning models.
- **Testing Set (30%)**: Reserved for evaluating the performance of the trained models.

This split ensures that the model's performance is assessed on unseen data, enhancing generalizability [17].

3.4. Proposed Model:

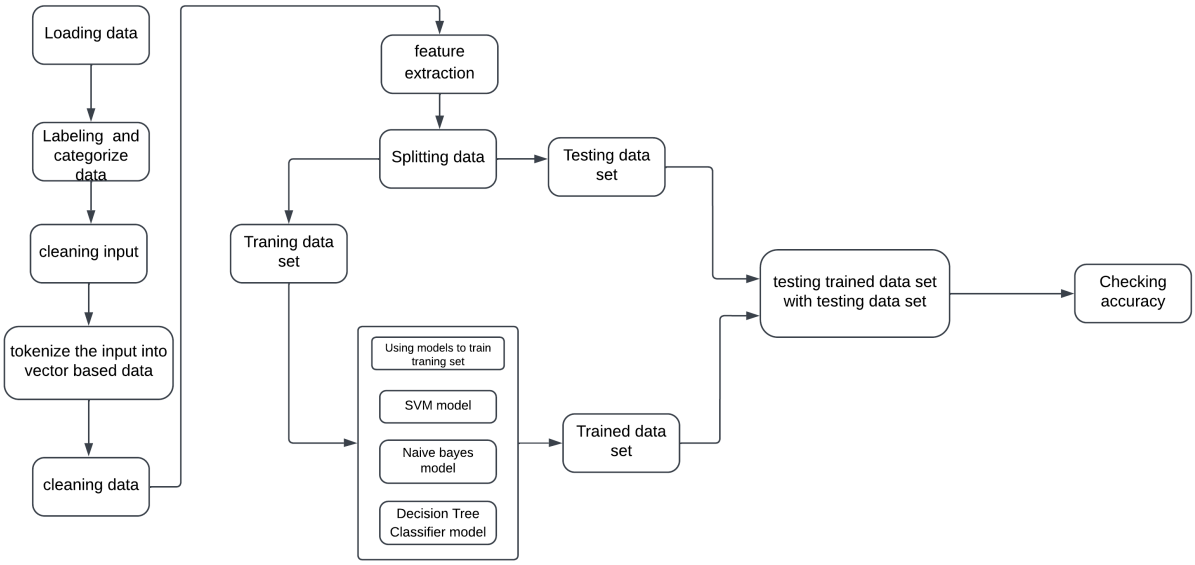

Figure 1. Spam Mail Detection System for Email Classification

3.5. Model Selection:

Three machine learning algorithms were explored for spam email classification:

a. Naive Bayes (NB)

Naive Bayes (NB) is a probabilistic classification algorithm rooted in Bayes' Theorem, which calculates the probability of a class given certain features under the assumption of feature independence. It is particularly effective for text classification tasks like spam email detection due to its simplicity and computational efficiency.

Formula and Explanation:

Bayes' Theorem:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Where:

- P(C|X): Posterior probability of class C (spam or ham) given the features X.
- P(X|C): Likelihood of features X given class C.
- P(C): Prior probability of class C.
- P(X): Marginal probability of features X.

For spam email detection, X represents the email content (text), and the model calculates probabilities for each class (spam or ham) to predict the label with the highest probability.

Despite its "naive" assumption that features are conditionally independent, which may not hold in practice, NB performs exceptionally well in high-dimensional data scenarios, making it a widely used baseline model for spam detection [24].

b. Decision Tree (DT)

Decision Tree (DT) is a supervised learning algorithm that uses a tree-like structure to make decisions by splitting the dataset into branches based on feature values, ultimately assigning a class label at the leaf nodes. It employs measures such as Gini Index or Entropy to evaluate the quality of splits.

Formula and Explanation:

The splitting of nodes is guided by impurity measures like Gini Index or Entropy (used in Information Gain).

1. **Gini Index**:

   Gini(D)=1−$\sum_{i=1}^{n} P_i^2$

   Where Pi is the probability of class i in dataset D.

2. **Entropy (Information Gain)**:

   Entropy(D) = − $\sum_{i=1}^{n} P_i log_2(P_i)$

   Information Gain= Entropy (Parent) − $\sum_{Children}(\frac{|Child|}{|Parent|} . Entropy(Child))$

   The goal is to maximize Information Gain or minimize Gini Index at each split to build the most informative tree.

DTs are highly interpretable and versatile, capable of handling both categorical and numerical data, making them valuable for spam classification tasks, especially when interpretability is key [19].

c. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a robust supervised learning algorithm designed to find the optimal hyperplane that separates data points from different classes in a high-dimensional feature space. SVM maximizes the margin between the hyperplane and the nearest data points, known as support vectors.

Formula and Explanation:

Given a dataset with features Xi and labels Yi , SVM solves:

$$\min_{\omega,b} \frac{1}{2} ||\omega||^2$$

Subject to:

$$Yi(\omega \cdot Xi+b) \geq 1 \quad \forall i$$

Where:

- ω: Weight vector defining the hyperplane.
- b: Bias term.
- $||\omega||^2$: Margin maximization term (to separate classes as widely as possible).

This model is used to project data into a higher-dimensional space. Known for its high accuracy and ability to handle complex and high-dimensional datasets, SVM is particularly effective for spam email detection [14].

3.6. Model Evaluation

The evaluation of the models focused on accuracy, precision, recall, and F1-score to measure their effectiveness in spam detection [23]. Cross-validation ensured robust results, while confusion matrices highlighted areas of strength and weakness. Additionally, ROC-AUC analysis provided insights into each model's ability to differentiate between spam and non-spam emails [14, 15]. The SVM model, with its superior accuracy and balanced metrics, proved to be the most effective choice for this task [16, 19].

4.7. Deployment

The best-performing SVM model was deployed in a real-time system to classify incoming emails. The deployment ensures scalability, allowing the system to process large volumes of emails effectively. Additionally, the system incorporates periodic retraining with updated datasets to adapt to evolving spam trends, maintaining high classification accuracy over time [21].

5. Results

In this paper, we have used three Machine Learning Model. Which are SVM, Naïve Bayes and Decision Tree model. From all of them SVM gives the highest accuracy of 98.38% and Naïve Bayes gives the lowest accuracy of 94.84%. We used a data set from Kaggle.com which contains 5572 Email text. From them we extract 4825 Ham messages and 747 spam messages.

Table1: Statistic table

ML Models	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
SVM	1326	200	5	20	98.38%	97.56%	90.91%	94.11%
Decision Tree	1331	140	0	80	95.68%	94.73%	1.0%	97.29%
Naïve Bayes	1303	181	28	39	94.84%	86.60%	82.72%	84.38%
Hyper Parameter Decision Tree	11	18	1	0	96.67%	94.73%	1.0%	97.29%

6. Discussion

The global increase in email usage has made research on email spam detection more crucial. A significant percentage of the billions of emails sent daily are classified as spam, presenting threats like phishing and identity theft. This study investigated machine learning techniques, particularly Naive Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM). A comparative analysis of these methods produced significant findings. Naive Bayes performed well, showing efficiency in handling text data. Decision Trees provided an easily understandable model but were constrained by overfitting. SVM outperformed other models in accuracy, attaining 98% accuracy in spam identification. This success can be attributed to hyperparameter optimization and effective distinction between spam and legitimate messages. However, challenges included SVM's computational requirements for larger datasets and the need for continuous retraining. While machine learning models like SVM offer high accuracy, practical implementation requires addressing scalability and interpretability concerns. The deployed system must efficiently handle large email volumes while providing clear explanations for classification decisions. Regular retraining with updated datasets is crucial to maintain performance and adapt to new spam variants. This study showcases the potential of machine learning in developing dependable spam detection systems. It underscores the importance of balancing accuracy, scalability, and interpretability to create adaptable solutions. The results pave the way for future research aimed at enhancing model performance and tackling challenges in spam email detection in a changing digital landscape.

7. Conclusion

This study successfully demonstrated the use of machine learning models Naive Bayes, Decision Trees, and Support Vector Machines (SVM) for detecting spam emails. Using a balanced dataset of 5,572 emails from Kaggle, we applied thorough preprocessing steps such as cleaning, normalizing, and extracting features using TF-IDF and Bag of Words. The results showed that SVM performed the best, achieving a high accuracy of 98.38%, making it the most reliable option. Naive Bayes and Decision Trees also delivered good results but were slightly less accurate than SVM. The confusion matrix confirmed SVM's strong performance, showing no false positives and very few false negatives. Key lessons from this study highlight the importance of fine-tuning model parameters and using effective feature extraction methods to improve performance. However, challenges like managing large datasets and the computational demands of SVM were identified. Regular retraining with updated data will be crucial to keep up with changing spam patterns. In conclusion, this research highlights the effectiveness of machine learning in addressing spam email problems. Future research could focus on advanced techniques like combining multiple models, making systems adaptable in real time, and improving their ability to explain decisions. These efforts could lead to more efficient and scalable spam detection solutions for the growing demands of email communication.

References

[1] A. Butkovic, S. Mujacic, and S. Mrdovic, "IP geolocation suspicious email messages," Nov. 2013, pp. 881–884. doi: 10.1109/telfor.2013.6716371.

[2] A. Bhowmick and S. M. Hazarika, "E-Mail Spam Filtering: A Review of Techniques and Trends," *Lecture Notes in Electrical Engineering*, pp. 583–590, Oct. 2017, doi: https://doi.org/10.1007/978-981-10-4765-7_61.

[3] L. Ceci, "Number of e-mail users worldwide 2023 | Statista," *Statista*, Sep. 16, 2024. https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[4] "How Many Emails Are Sent Per Day? (2017–2025) | Oberlo," *www.oberlo.com*. https://www.oberlo.com/statistics/how-many-emails-are-sent-per-day

[5] Laura Ceci, "Daily Number of E-mails Worldwide 2023 | Statista," *Statista*, Aug. 22, 2023.https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/

[6] A. Sharaff and H. Gupta, "Extra-Tree Classifier with Metaheuristics Approach for Email Classification," springer singapore, 2019, pp. 189–197. doi: 10.1007/978-981-13-6861-5_17.

[7] S. Kaddoura, G. Chandrasekaran, D. Elena Popescu, and J. H. Duraisamy, "A systematic literature review on spam content detection and classification," *PeerJ Computer Science*, vol. 8, p. e830, Jan. 2022, doi: https://doi.org/10.7717/peerj-cs.830.

[8]. P. Patil and P. Bhosale, "International Journal of Research Publication and Reviews Literature Survey on Spam Email Detection," *International Journal of Research Publication and Reviews*, vol. 3, no. 11, pp. 2688–2694, 2022, Available: https://ijrpr.com/uploads/V3ISSUE11/IJRPR8167.pdf

[9]. "A detailed analysis on spam emails and detection using Machine Learning algorithms."Available:https://gala.gre.ac.uk/id/eprint/41815/7/41815_SULTHANA_A_detailed_analysis_on_spam_emails_and_detection_using_machine_learning_algorithms.pdf

[10]. C. Ellis, "Spam Statistics - 2023 Survey and Data Analysis | Email Tooltester," *EmailTooltester.com*,Oct.19,2023.https://www.emailtooltester.com/en/blog/spam-statistics/

[11] C. Griffiths, "The Latest Phishing Statistics (updated January 2023) | AAG IT Support," *aag-it.com*, Oct. 02, 2023. https://aag-it.com/the-latest-phishing-statistics/

[12]. "Highest number of daily spam emails by country 2021," *Statista*. https://www.statista.com/statistics/1270488/spam-emails-sent-daily-by-country/

[13]. 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, p. e01802, Jun. 2019, doi: 10.1016/j.heliyon.2019.e01802.

[14]. S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," *IEEE Xplore*, Dec. 01, 2015. https://ieeexplore.ieee.org/document/7399985 (accessed Apr. 19, 2022).

[15]. A. TEKEREK, "Destek Vektör Makineleri Kullanılarak Spam SMS Tespiti," *Journal of Polytechnic*, Oct. 2018, doi: https://doi.org/10.2339/politeknik.429707.

[16]. M. V. Madhavan, S. Pande, P. Umekar, T. Mahore, and D. Kalyankar, "Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, p. 012113, Jan. 2021, doi: https://doi.org/10.1088/1757-899x/1022/1/012113.

[17]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 16, pp. 321–357, Jun. 2002, doi: https://doi.org/10.1613/jair.953.

[18]. "Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. Machine Learning, 20, 273-297. - References – Scientific Research Publishing," *Scirp.org*, 2014. https://www.scirp.org/reference/referencespapers?referenceid=1150668

[19]. S. Gupta, A. Mittal, and B. Bhushan, "Data preprocessing techniques for machine learning: A survey," *J. Big Data*, vol. 8, no. 1, pp. 1–22, 2021.

[20]. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Eur. Conf. Mach. Learn.*, pp. 137–142, 1998.

[21]. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Int. Jt. Conf. Artif. Intell.*, pp. 1137–1145, 1995.

[22]. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

[23]. M. Adnan, M. O. Imam, M. F. Javed, and I. Murtza, "Improving spam email classification accuracy using ensemble techniques: a stacking approach," *International Journal of Information Security*, vol. 23, no. 1, pp. 505–517, Sep. 2023, doi: 10.1007/s10207-023-00756-1.

[24]. Kaggle. (2023). Spam Email Dataset. Retrieved from https://www.kaggle.com.