Comparative Analysis for Email Spam Detection Using Machine Learning Algorithms

Gayatri Gattani, Shamla Mantri, and Seema Nayak

Abstract The prominence of the use of communication over the Internet is increasing progressively. Being economical, faster, and easy user interface, the number of email users is increasing tremendously. These led to the gradually increasing activity of spam. Spam emails are unrequested and unimportant emails in bulk. Due to this, there arise major Internet and email security issues that also include a problem of electronic storing space and waste of time. Thus, the identification of spam emails is very necessary. In this paper, four supervised machine learning algorithms, which are Naïve Bayes, support vector machine (SVM), logistic regression, and random forest classifier, are proposed for spam and ham emails classification. Experiments using these four algorithms are performed on prepared feature sets on two different datasets to select the best model with the highest accuracy and less overfitting or underfitting for spam detection. To automate the workflow of building the model and its evaluation, a machine learning pipeline is used in this project. Experimental results show that the overall accuracy of the random forest classifier model is the highest and also has less complexity.

Keywords Emails · Spam email · Spam email identification · Machine learning · Naïve Bayes · Support vector machine · Logistic regression · Random forest classifier

G. Gattani (✉) · S. Mantri
School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT WPU, Pune, India
e-mail: gayatrigattani2001@gmail.com

S. Mantri
e-mail: shamla.mantri@mitcoe.edu.in

S. Nayak
Department of Electronics and Communication Engineering, IIMT College of Engineering, Greater Noida, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

R. Agrawal et al. (eds.), *Modern Electronics Devices and Communication Systems*,
Lecture Notes in Electrical Engineering 948,
https://doi.org/10.1007/978-981-19-6383-4_2

1 Introduction

In recent decades, the use of technology and the Internet has reached to peak. Being fast, cheap, and accessible, the extension of the use of email has increased tremendously. This resulted in a dramatic increase in spam emails [1]. These emails are junk emails that are almost identical and sent to multiple recipients randomly [2]. The changing way of communication by Internet on a very large scale has led to the expansion of new communication services, such as email [3]. According to a recent study, over 4 billion of the population use email. Due to its simplicity and accessibility, the mark of people using email is increasing day by day. It is extremely fast and cost-effective. With the escalation in the broadening of emails, there is also a rise in spam emails, and the unnecessary and undesirable bulk mails sent to several users haphazardly. Spam mails not only cause the problem of electronic storing space but also are the carrier of malware and hoard the network bandwidth, space, and computational power [4]. A study estimates that approximate measure of spam emails is 85%.

While the number of spam emails increases, the certainty of a user not reading a non-spam email increases. Due to the loss of network bandwidth and time consumed by users to demarcate between normal and spam [5], various spam filtering techniques have been introduced. These techniques can be categorized based on the use and non-use of machine learning algorithms. The use of ML algorithms provides an automated approach where the model trains itself based on features extracted from the dataset. As easy to implement and short training time, Naïve Bayes is a popular spam filter [6]. The main objective is to collate the accuracy of four major classification systems that include SVM, random forest classifier, Naïve Bayes, and logistic regression and select the best model for spam detection.

2 Literature Survey

Spam: Unnecessary emails sent by unknown people randomly in bulk are spam mails. These spam mails are vulnerable to major user security and also cause the problem for electronic storing space. The following are the major spam categories (Table 1).

Table 1 Frequency of major scam categories and danger level caused by them

Categories	Frequency of receiving	Danger
Ads	High	Moderate
Chain letter	Low	High
Email spoofing	Low	High
Hoaxes	Moderate	Moderate
Money scams	Moderate	High

Table 2 Some previously used techniques in spam filtering and their accuracy [9–12]

Year	Author	Classification technique	Dataset	Highest accuracy
2008	Bo Yu, Zong-ben Xu	Naive Bayes, NN, SVM, RVM	SpamAssassin & Babble Text	SVM: 95.2% and 96.0%
2011	W. A. Awad & S. M. ELseuofi	Naïve Bayes, KNN, ANN, SVM, artificial immune	SpamAssassin spam corpus	NB: 99.46%
2013	Sumant Sharma & Amit Arora	ML techniques provided by WEKA tool	SPAMBASE	94.28%
2014	Andronicus A. Akinyelu & Aderemi O. Adewumi	Classification of phishing email Using random forest ML technique	2000 phishing and ham emails set	99.7%
2017	A. S. Yuksel, S. F.Cankya, & I. S. Uncu	Cloud-based approach combining predictive analysis and ML techniques (SVM and decision tree)	SpamAssassin	SVM: 97.6%
2018	Deepika Mallampati	Naïve Bayes, J48, MLP	–	MLP: 99.3%
2021	Manoj Sethi, Sumesha Chandra, Vinayak Chaudhary & Yash	NB-multinomial, logistic regression, SVM, NN	SpamAssassin	NN: 99.02%

Spam classification: Email systems without spam classification techniques are highly open to risks. The dangers open to email systems without spam filtering are spyware, phishing, ransomware [7]. Thus, the classification of such messages can be seen as another defense mechanism against such dangers. In the previous years, various techniques of spam identification have been developed. Domain name server blacklist (DNSBL) and white list, high-volume spammers (HVSs) and low-volume spammers (LVSs) classification, machine learning-based Web spam classification, support vector machine classifier model, TruSMS systems, cloud-based approach [3], and ML algorithms like Naïve Bayes, random forest classifier, neural networks [8] are some of the classification techniques developed by researchers earlier (Table 2).

2.1 Existing Approaches

Global email users are increasing day by day. In 2024, it is set to grow up to 4.48 billion [13]. As the use of email increases, spam increases too. This causes a decrement in productivity since manually spam filtering is time-consuming, and also the electronic

storing space is reduced. Spam also increases the cyber threat to users through various phishing and malware attacks. Not only this, it has been discovered that on yearly basis, spam is accountable for over 77% of whole global email traffic [13].

In today, two common approaches, namely knowledge engineering and machine learning, are used for spam filtering. A collection of rules in knowledge engineering are used to identify mails as ham or spam. This method can lead to large time wastage and also does not guarantee the results as there is a continued need for an update in the specified set of rules. Thus, it is mainly used by naïve users [14].

Machine learning is completely based on the datasets. It just needs the training datasets, and the algorithm used itself learns the classification rules from the set of training samples from datasets. Thus, machine learning is proved to be more effective than knowledge engineering [14]. Examples of machine algorithms used for the classification of spam include Naïve Bayes, support vector machine, artificial immune systems, neural networks, logistic regression, deep learning, and many more.

The best possible outcome for any algorithm can be checked using various evaluation techniques in machine learning. This evaluation technique also helps in recognizing the overfitting and underfitting of the model. Cross-validation score, F1-score, confusion matrix, precision, recall, accuracy, regression metrics, and mean squared error can be used for evaluating the model. The three major metrics to weigh up a classification model are accuracy, precision, and recall [15].

Three major methods that are reliable for present spam detection systems are linguistic-based (used in places like a search engine), behavior-based (user-dependent since the need of change in rules from time to time), and graph-based (detect abnormal forms in data showing the behavior of spammers) [16].

3 Proposed Method

3.1 Proposed Algorithm and Workflow

Two datasets are used for this experiment to select the best algorithm with the highest accuracy. Dataset 1 is taken from Kaggle SMS Spam Collection [17]. This dataset contains 5574 messages tagged ham or spam. Dataset 2 is taken from the collection of emails from_Apache SpamAssassin's public datasets_ available on Kaggle as spam or not spam dataset [18]. There are 2500 non-spam and 500 spam emails in this dataset. The experiment is performed using four simple machine learning classification algorithms that are logistic regression, support vector machine (SVM), random forest classifier, and logistic regression on a prepared feature set of two datasets.

Through evaluation using confusion matrix, evaluation metrics, k cross-validation score, and accuracy, the perfect model with the highest accuracy and reduced underfitting or overfitting is selected. Selection of parameter k in k cross-validation score and splitting ratio of datasets play an efficient contribution in assessing the accuracy

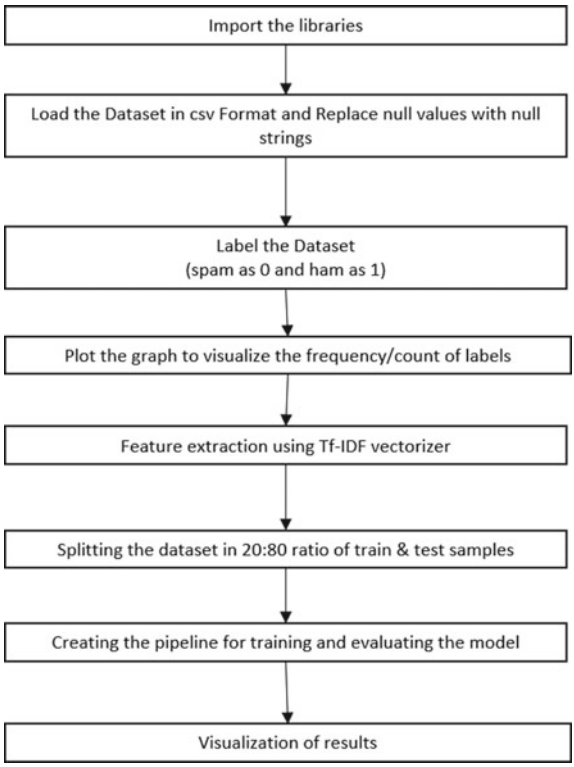

Fig. 1 Flowchart of model

of the model. The accuracy and overfitting/underfitting results are visualized using a heat map (Fig. 1).

Data Preprocessing

Both the datasets are taken from Kaggle [17, 18]. Dataset 1 comprises 5574 messages tagged according to being ham or spam. Here, we need to label the spam messages as 0 and ham messages as 1 for further simplicity. Dataset 2 comprises a collection of 3000 emails taken from "_Apache SpamAssassin's public datasets_". There are 2500 non-spam emails and 500 spam emails in this dataset. Here, the dataset initially contains the labeled data, that is, spam mails as 1 and ham emails as 0. All the null values in both datasets are converted to null strings for the normalization of plain text (Fig. 2).

Feature Extraction

The feature set will be prepared using term frequency-inverse document frequency (TF-IDF) vectorizer by transforming the feature text into feature vectors and converting it to lowercase. Parameter min_df is set to 1 that means to ignore the terms that appear in less than one document. The terms that appear irregular, min_df is used to remove them [19]. The next parameter, stopwords, is set to English to return

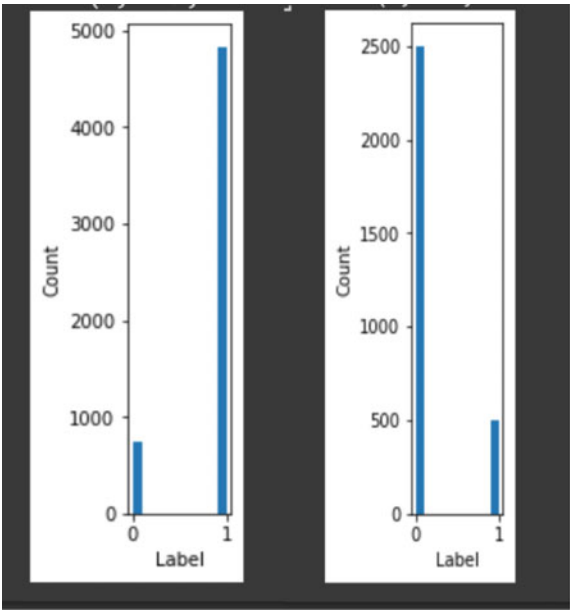

Fig. 2 Visualization of labeled dataset 1 (left) and dataset 2 (right)

the relevant stop list. The parameter lowercase is set to true to convert all characters to lowercase.

Pipeline

To automate the workflow of producing a machine learning model and evaluation of spam detection using different algorithms, a pipeline is created. The different algorithms used in this experiment are as follows:

Logistic Regression

A supervised machine learning algorithm is used for solving classification problems. It is a simple yet very effective algorithm for binary classification. The basis of this algorithm is the logistic function (sigmoid function), which takes any real-valued number and maps it in the value between 0 and 1 [20].

$$\text{Logistic Function: } y = 1/(1 + e^{-x})$$
$$\text{i.e., } 1 + e^{-x} = 1/y$$
$$e^{-x} = (1-y)/y$$
$$e^{x} = y/(1-y)$$
$$x = \log(y/(1-y))$$

Naïve Bayes

A simple probabilistic classifier uses the Bayes theorem that calculates a set of probabilities by counting the frequency and combination of values in the dataset [21].

$$P(A|B) = P(B|A)\, P(A)/P(B)$$

Using Bayesian probability terminology, the above equation can be written as [22]

$$\text{Posterior} = \text{Prior} * \text{Likelihood/Evidence}$$

Random Forest Classifier

It uses ensemble learning and regression technique to solve data classification problems [23]. It is a supervised machine learning algorithm that gets a prediction from each decision tree created.

Support Vector Machine

SVM is a supervised machine learning algorithm that classifies the data points by finding an optimal hyperplane. There are support vectors that help to maximize the classifier margin.

3.2 Performance Evaluation Criteria for Algorithm

This section is to measure and analyze the accuracy of different algorithms used in the model to estimate the results that fit best between the model and testing dataset. In this experiment, we have computed confusion matrix, evaluation metrics, and k cross-validation to assess our model and different algorithms.

Confusion Matrix

The table having four outcomes computed by the binary classifier is called confusion matrix. Measures, such as error rate, accuracy, specificity, sensitivity, and precision, are derived from the confusion matrix [24]. The four outcomes mentioned above are true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Accuracy, recall, precision, and F-score are calculated using these four outcomes. Here, in this experiment, we have considered accuracy, recall, precision, F-score, and error rate to evaluate the models.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{Total no. of dataset samples})$$

Sensitivity ***is*** also known as recall or true positive rate. It is used to measure the ability of a test to be positive when the condition is present [25].

$$\begin{aligned}\text{Sensitivity or Recall} &= \text{TP}/(\text{TP} + \text{FN}) \\ &= \text{TP}/(\text{Total positive})\end{aligned}$$

Precision is also known as positive predictive value [25]. The value ranges from 0 to 1.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

F-score is calculated with precision and recall, as follows:

$$\text{F-score} = (2 * \text{precision} * \text{recall})/(\text{precision} + \text{recall})$$

K Cross-validation Score

Cross-validation is a data resampling method to assess the generalization ability of predictive models, and cross-validation is a resampling (in such a way that no two samples overlap) method to assess the abstraction ability of models to predict the outcomes and stave off the overfitting [25]. The parameter k is the number of sets in which the sample is to be split, such that no set contains element in common. In this experiment, value of k is 4. cv_score_mean and cv_score_std are calculated to verify the accuracy results and find deviation in cv_score, respectively.

4 Results

Accuracy and precision are the important parameters in the above experiment to evaluate the different algorithms used. Other functions such as F1-score, error rate, and recall are also calculated to compare the performance of four algorithms on the above-mentioned two datasets. As shown in Fig. 3, the accuracy for random forest classifier is highest followed by the logistic regression classification algorithm. F1-score and precision of random forest classifier outperform all the other algorithms in the experiment. Figures 5 and 6 depict the evaluation results of all four algorithms on both datasets, respectively (Fig. 4).

The error rate in identifying whether the mail is spam or ham is lowest for random forest classifiers and highest for Naïve Bayes. Cv_score_std for Naïve Bayes in dataset 1 (0.0027) and logistic regression in dataset 2 (0.0066) is the lowest out of the four algorithms. Lower the cv_score_std, lower is the overfitting. However, cv_score_mean that verifies the accuracy results is highest for random forest, 0.9757 and 0.9673 in dataset 1 and 2, respectively.

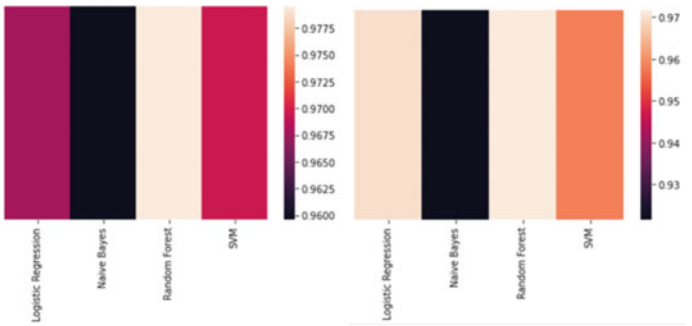

Fig. 3 Accuracy of dataset 1 (left) and dataset 2 (right)

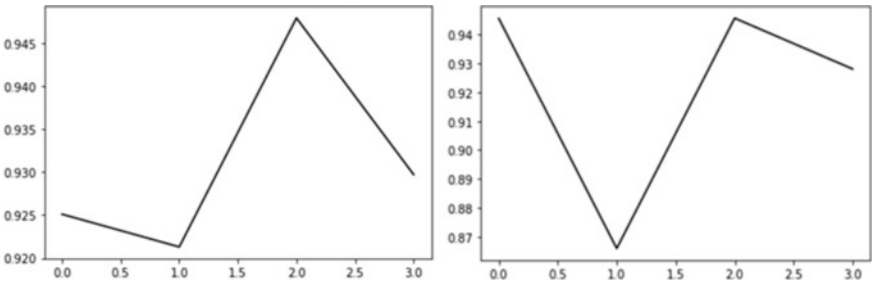

Fig. 4 F1-score of dataset 1 (left) and dataset 2 (right)

	Model	Accuracy	f1-score	precision	recall	error_rate
0	Logistic Regression	0.9677	0.9251	0.9819	0.8839	0.0322
1	Naive Bayes	0.9596	0.9213	0.8963	0.9522	0.0403
2	Random Forest	0.9767	0.9479	0.9835	0.9188	0.0233
3	SVM	0.9695	0.9297	0.9829	0.8903	0.0304

==========

	Model	cv_mean	cv_std
0	Logistic Regression	0.9669	0.0034
1	Naive Bayes	0.9549	0.0027
2	Random Forest	0.9757	0.0041
3	SVM	0.9691	0.0036

Fig. 5 Evaluation of dataset 1

Among all the models used, random forest has the highest accuracy, that is, 0.9767 and 0.9717 in both datasets, respectively.

	Model	Accuracy	f1-score	precision	recall	error_rate
0	Logistic Regression	0.9700	0.9456	0.9456	0.9456	0.0300
1	Naive Bayes	0.9216	0.8659	0.8476	0.8883	0.0783
2	Random Forest	0.9717	0.9457	0.9738	0.9222	0.0283
3	SVM	0.9583	0.9281	0.9090	0.9507	0.0416

==========

	Model	cv_mean	cv_std
0	Logistic Regression	0.9616	0.0066
1	Naive Bayes	0.8793	0.0300
2	Random Forest	0.9673	0.0075
3	SVM	0.9093	0.0304

Fig. 6 Evaluation of dataset 2

5 Conclusion and Future Work

In the comparative analysis of machine learning algorithms to classify emails as spam or ham using two different datasets, the random forest classifier is the best binary classifier out of all the four supervised algorithms. Feature extraction is done using the TF-IDF vectorizer, and the application of pipeline automates the workflow of training and evaluating the model using four different classification algorithms and different evaluation methods. Here, two different datasets are used to analyze the results on different data to select the model with high accuracy and less error rate.

The future work includes assessing the model with various effective algorithms to automate the task of filtering spam and non-spam emails using different features. This research proposes to test the model using different feature sets on different types of datasets to analyze and increase the efficiency of the prototype to identify the email as spam or non-spam.

References

1. Yu B, Xu Z-b (2008) A comparative study for content-based dynamic spam classification using four machine learning algorithms. Knowledge-based systems, vol 21. Elsevier, pp 355–362
2. Abdulhamid SM, Shuaib M, Osho O, Ismaila I, Alhassan JK (2018) Comparative analysis of classification algorithms for email spam detection. MECS, I J Comput Netw Inf Secur, pp 60–67
3. Yuksel AS, Cankya SF, Uncu IS (2016) Design of a machine learning based predictive analytics system for spam problem. In: ICCESEN, pp 500–504
4. Huang L, Jia J, Ingram E, Peng W (2018) Enhancing the Naïve Bayes spam filter through intelligent text modification detection. IEEE, pp 849–854
5. Guzella TS, Caminhas WM (2009) A review of machine learning approaches to Spam Filtering. Elsevier, pp 10206–10222

6. Rusland NF, Wahid N, Kasim S, Hafit H (2017) Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In: IRIS
7. Lloyed-Davis F (2017) The dangers of Spam email. Acronyms, November
8. Sethi M, Chandra S, Chaudhary V, Yash (2021) Email Spam detection using machine learning and neural networks. IRJET 8(4):349–355
9. Awad WA, ELseuofi SM (2011) Machine learning methods for spam e-mail classification. IJCSIT 3:173–184
10. Sharma S, Arora A (2018) Adaptive approach for spam detection. IISTE 7:14–21
11. Mallampati D (2018) An efficient spam filtering using supervised machine learning techniques. IJSRCSE, pp 14–21
12. Akinyelu AA, Adewumi AO (2014) Classification of phishing email using random forest machine learning technique. Hindawi, Article ID 425731
13. Spam Mail Detection using Machine Learning. PerfecteLearning, April 2021
14. Dada EG, Bassi JS, Chiroma H, Abdulhamid SM, Adetunmbi AO, Ajibuwa OE (2019) Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5(6)
15. Evaluating a Machine Learning Model. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/
16. Mohammed M, Mostafa S, Ibrahim Obaid O, Zeebaree S, Abd Ghani MK, Mustapha A, Jubair M, Hassan M, Ibrahim D, Al-Dhief F (2019) An anti-spam detection model for emails of multi-natural language
17. SMS spam collection dataset. https://www.kaggle.com/uciml/sms-spam-collection-dataset
18. Apache, Spam or not spam dataset. A collection of emails from spamassassin.apache.org
19. TF-IDF vectorizer. Scikit Learn Library (2021)
20. Yeldirim S (2020) Logistic regression as a classification algorithm. Towards Data Science
21. Yang S (2019) Introduction to Naïve Bayes classifier. Towards Data Science
22. Venkatraman A (2019) Naïve Bayes for machine learning—from zero to hero. FloydHub
23. Dada E, Joseph S (2018) Random forests machine learning technique for email spam filtering, vol 9. University of Maiduguri, pp 29–36
24. Classeval (2017) Basic evaluation measures from the confusion matrix. Classevalblogs
25. Berrar D (2018) Cross-validation, vol 1. Elsevier, pp 542–545