

Hybrid Classifier for Enhancing Accuracy and Performance of Spam and Ham Email Detection

Annu Khanna Nakarmi¹, Ramesh Parajuli², Dr. Gajendra Sharma³

¹MCA, Kantipur City College, Dept. of Science and Technology, Kathmandu 44600, Nepal

²Assistant Professor, Kantipur City College, Dept. of Science and Technology, Kathmandu 44600, Nepal

³Associate Professor, Kathmandu University, Dept. of Computer Science and Engineering School of Engineering, Dhulikhel 45200, Kavre, Nepal

Abstract

In the contemporary landscape, digital communication reigns supreme, with email being a prominent channel for rapid and widespread information dissemination, also serving as evidence and a promotional tool. Yet, the issue of spam emails imperils data security. Hence, machine learning driven spam classifiers are vital for data preservation. This dissertation rigorously assesses the efficacy of diverse machine learning techniques for spam detection. In response to the exponential surge in spam, devising effective means of identification and filtration is imperative. Leveraging an email dataset, the study compares the performance of Naive Bayes, Support Vector Machines (SVM), Logistic Regression, Random Forest, and K-Nearest Neighbor (KNN) algorithms. This research examines the strengths and limitations of these methods in spam categorization. Evaluation metrics encompassing accuracy, recall, precision, F1 score, and support are considered. Naive Bayes (85%), SVM (85%), Logistic Regression Classifier (84%), Random Forest (84%), and KNN (77% at P=1 & 82% at P=2) are scrutinized, with Naive Bayes and SVM exhibiting notable accuracy. The article contrasts and scrutinizes these five techniques, seeking the most effective spam categorization approach. The findings led to the development of advanced hybrid spam detection systems, amalgamating Naive Bayes and SVM through ensemble technology, promising enhanced protection with the highest accuracy of 87%

Keywords: evaluation; accuracy; spam; machine learning; hybrid

1. INTRODUCTION

1.1 Background

According to statistics on email usage, 4.258 billion people are using email in 2022 & 300 million emails are sent every day, fifty percent of which are spam. According to statistics, by 2025 there will be around 4.6 billion email users worldwide [1]. The massive amount of spam letters streaming across computer networks has a negative impact on email servers' memory, communication bandwidth, CPU power, and user time. Additionally, it has caused untold financial loss to numerous users who have been tricked by spammers who send emails supposed to be from reputable businesses to trick people into disclosing sensitive personal information like passwords, credit card numbers, and bank verification numbers (BVNs), among other fraudulent practices [2]. Spam emails need to be detected and prevented in order to protect sensitive information, prevent the spread of viruses, malware, and hacking attempts, and restrict access to the system. Different preventative measures, such as spam filters and blacklists, were employed to identify spam emails, but they didn't prove to be any more effective. Knowledge engineering and machine learning are the two most popular methods for detecting spam emails. Knowledge engineering uses a set of rules to categorize emails as spam or ham. A set of rules must be created by the user of the filter or by the software provider who provides

the rule-based spam-filtering tool in question. The requirement to constantly update the rules means that using this strategy does not guarantee efficient outcomes. This can waste time, therefore it's not a smart option, especially for novices. The machine learning integrated spam classifier is more effective & efficient at filtering spam emails. The use of machine learning has been shown to be more effective than the use of knowledge engineering. The goal of the subfield of artificial intelligence known as machine learning is to enable robots to acquire knowledge similarly to humans [3]. The techniques that help to prevent spam emails can be sender and content-based. The email contains numerous parts such as a header, subject, body, and sender information, among which sender information such as username and writing styles can be considered as the major function in sender-based detection, and sentences and words of the body section are major features in content-based information. Context based filtering was used where the content of the subject, from, and body of the email was selected to define or categorize the words as spam, which further defines the messages as spam messages [4].

2. LITERATURE REVIEW

It focuses on importance to develop a comprehensive system for spam mail categorization utilizing NLP and URL-based filtering in order to minimize inadvertent data loss, hacking, and the promotion of immoral material. A thorough and effective spam categorization system was developed that uses a two-step process to determine whether or not a piece of mail is spam. Text classification followed by URL analysis, and filtering are used to assess whether or not any links contained in the email are dangerous. Naive Bayes and Support vector systems were highly accurate models among the several computer algorithms used for text classification used in its final model. The spam email collection came from two data sets that were designated as spam. Csv [5] available on Kaggle and Enron [6] spam data [7].

This study explains how spam detection, filtering, and classification work, how machine learning approaches aid in the spam detection process, how a general machine learning classification algorithm functions, how a specific algorithm classifies the email into constituent spam and ham messages, and the conditions under which a specific algorithm is effective. Based on three parameters provided, it is determined that, when compared to Naive Bayes, KNN, SVM, and Logistic Regression Classifier, Naive Bayes performs the best. Through the use of combinations of algorithms that excel in many areas, such as evaluation time, acquaintance cost, memory of allocation, etc., this article aids in the development of more effective and practical hybrid algorithms [3].

The classifier discovered to have the greatest accuracy is Rotation Forest, which provides 94.2%. Random Forest has demonstrated a close degree to reaching the most accurate outcome (i.e.99.72%), even though none of the algorithms achieved 100% accuracy in categorizing spam emails. The outcome demonstrates that, even with minimal feature selection, the Rotation Forest classification algorithm (0.942) outperforms certain regularly used classification algorithms, such as J48 (0.923 accuracy), Naive Bayes (0.885), and Multilayer Perceptron (0.932), in terms of email categorization. 10 folds cross validation was used because of result obtained from broad tests on various dataset with varying learning procedures, that have demonstrated that 10 is about correct number of folds to best gauge of error [8].

It states that merging algorithms or filtering the outcomes, a more intelligent spam detection classifier may be made. It includes analysis stage where data are processed, analyzed and pattern are revealed. Secondly, train stage where machine learning models are used on the obtained data and finally the last stage where

deployment of the best model is done. words. After that, the chosen algorithm will be trained, and its accuracy will be assessed using methods that are more likely to find a superior solution for spam categorization. Logistic regression and Naïve Bayes were efficient algorithm as they have the highest level of accuracy [9].

3. METHODOLOGY

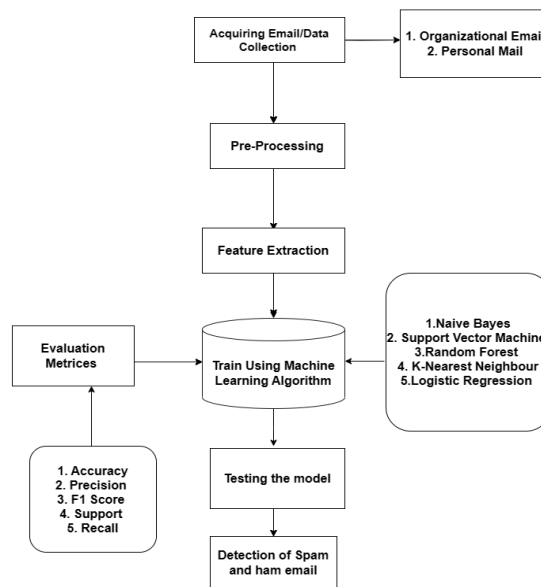


Fig 1: Research Framework

Spam emails must be distinguished from ham emails using a machine algorithm, which can only be done via automated and adaptive methods. The ML framework's methodologies had involved in extracting data from a collection of emails and using the resulting data. Fig 3.1 shows research Framework that will be used to determine the best machine algorithm among all the implemented algorithm in order to detect the spam email. Fig 1.2 shows the hybrid model [10] framework that consist of the combination of two best model obtained from the 1.1 framework using voting ensemble technique [11] and addition of the grid search algorithm that helps to combine the best features of both models to optimize the model. Obtaining the dataset is the first and most important step in developing methodologies for the detection and classification of spam emails. For this research both primary and secondary data were used to test and train the developed model using different machine learning algorithm. Firstly, models were tested using readymade data from Kaggle dataset [12]. The secondary data consist of the 5572 mails which is the set of the messages in English. And after validation of the data, the primary data were used. Primary data were extracted by using self-made Gmail API from both personal and organizational data. The combination of primary data consists of total 8047 emails which are also categorized as either ham or spam. Here there are 6059 ham emails and 1988 spam emails. The dataset consists of two columns where first column consists of category which consist result displaying given text email as ham or spam and second column consist of text which shows the text email. The dataset of emails that was extracted from the various libraries as well as personal and organizational spam email dataset has to

be preprocessed. Both primary data and secondary data requires preprocessing before implementing to model. Data can be cleaned and preprocessed in a number of ways, including lemmatization, expanding contraction, removing accented characters, eliminating tag, and removing accented characters . And for this research, preprocessing of both primary and secondary dataset was done by removing null values, duplicate values, NAN column etc. Talking about detail process, here missing values were filled with an empty string using “Fillna” method.

3.1 Feature Extraction: The preprocessed dataset requires the feature extraction before getting ready to be trained by using different machine learning algorithm. Hence the dataset involved in this study used counter vectorizer (TfidfVectorizer) method to convert textual data into 16 numerical data. The first column of the dataset categories was converted into the 0 and 1 where 0 represents ham email and 1 represent spam email.

3.2 Data Splitting: The pre-processed data was splited in two parts: Training and testing. In the training, Model training is the process of running the selected algorithm based on the training model. The quality of algorithm used will be accessed using the following measures like: precision, recall, accuracy, F1-score, support. The testing was done to check whether the given model perform the task as trained. Here the dataset was splited in the ratio of 75:25 where 75% was used for training and 25 % was used for tested.

3.3 Model Training: The model created was trained using a machine learning algorithm in this step of the methodology, which aids in categorizing spam emails. The data that was pre-processed in the preceding steps was divided into two parts, training and testing. In the training phase, the algorithm modifies the parameter for the model, evaluates thing through given parameters, and generates output, which is then further classified as spam and non-spam.

3.4 Model Evaluation/Testing: Datasets designated as testing datasets were fed into the model's training process to see how accurately spam emails are identified. This stage also produces the accuracy score, which is used to assess the model's efficacy and benchmark it against other models. This stage involved generating the accuracy of the given model, identifying overall perfection, and contrasting it to other algorithms for machine learning based

On the basis of above framework, Naïve Bayes and Support vector classifier were considered as the best classifier with the highest accuracy in order to detect the spam email. In this research, hybrid classifier was made/generated by using best classifier through ensemble technology named as voting classifier to increase the accuracy of detecting the spam.

In this research also voting ensemble Technique is used to combine the best model to increase the performance. Voting uses the principle of majority voting to determine the final prediction. Here Naïve Bayes and Support Vector machine were combined to obtain best performance using the voting Classifier.

Grid search cross validation is performed in this study to improve how well the models detect spam emails. As a result, automating the hyperparameter tuning process was helpful. Here, Param grid is utilized to investigate the ideal pairing of fit-prior and alpha. Alpha is used to manage the situation where a certain word or feature has zero frequency in the training data as well as to adjust the smoothing parameter for the Naive Bayes method. In this study, the fit prior selects whether to apply uniform prior ('False') or learn class prior probabilities from the training data ('True').

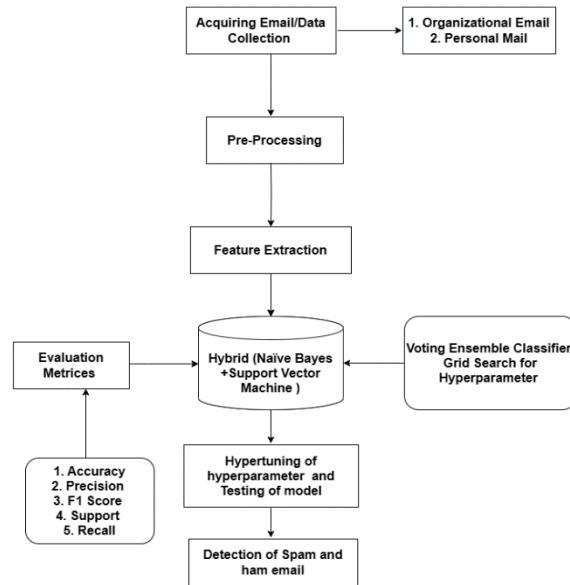


Fig 2: Hybrid Model

4. RESULT

4.1 Result of Predicted Model

Using a dataset of 8047 emails that were randomly selected from a bigger dataset of emails, various machine learning methods for spam email detection were analyzed. A training set of 6059 emails and a test set of 1988 emails were created from the dataset. Here machine learning algorithm was created and 5 best-suited algorithms determined for text classification i.e., Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Random Forest, and logistic regression were imported from the sci-kit learn library in and used for the classification and detection of spam email. They were implemented to perform the comparative analysis of their performance taking into consideration of different evaluation matrix. In order to validate the accuracy, confusion matrix for all the algorithms were also created and shown further in this documentation

Table 1: Comparative Analysis of Machine Learning Algorithm

Algorithm	0/1	Naïve Bayes	Support Vector Machine	Random Forest	Logistic Regression	KNN P=1	P=2
Precision	0	84%	87%	88%	85%	78%	85%
	1	87%	72%	68%	79%	55%	66%
Recall	0	98%	93%	91%	96%	98%	92%
	1	41%	57%	59%	43%	10%	49%
F1 Score	0	91%	90%	90%	90%	87%	89%
	1	56%	63%	63%	56%	16%	57%

Table 1 shows the comparative analysis of algorithm based on of Precision, Recall, F1 Score of all the used algorithm i.e., Naïve Bayes, Support Vector Machine, Random Forest, Logistic Regression and KNN. Naïve Bayes shows relatively higher precision, recall, and F1-score for class 0 (ham) compared to other algorithms. However, it has lower performance in identifying class 1 (spam), with lower precision, recall, and F1-score. Support Vector Machine and Random Forest have comparable performance, while Logistic Regression and KNN have varying performance depending on the value of the hyperparameter. KNN with P=2 shows better performance than P=1.

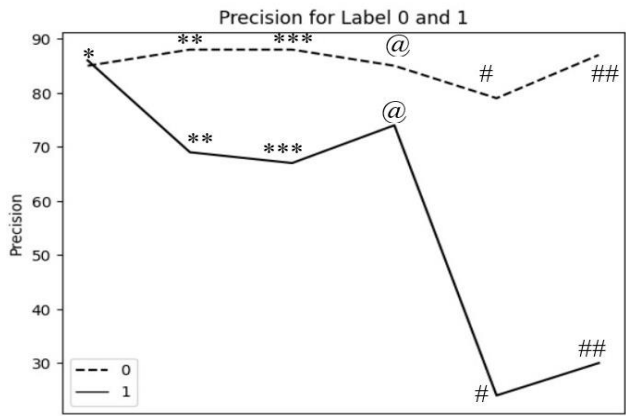


Fig. 3: Comparative Analysis on the basis of precision call

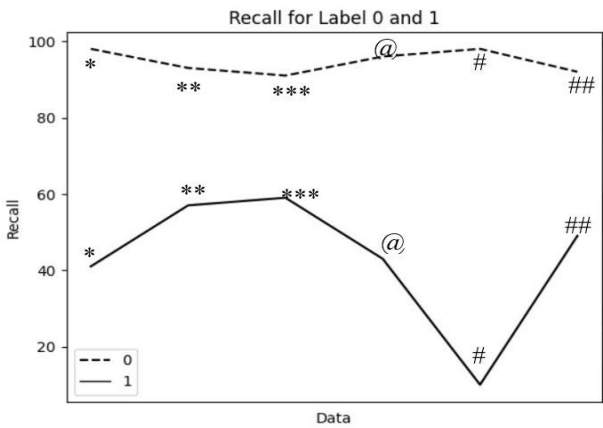


Fig. 4: Comparative Analysis based on Recall

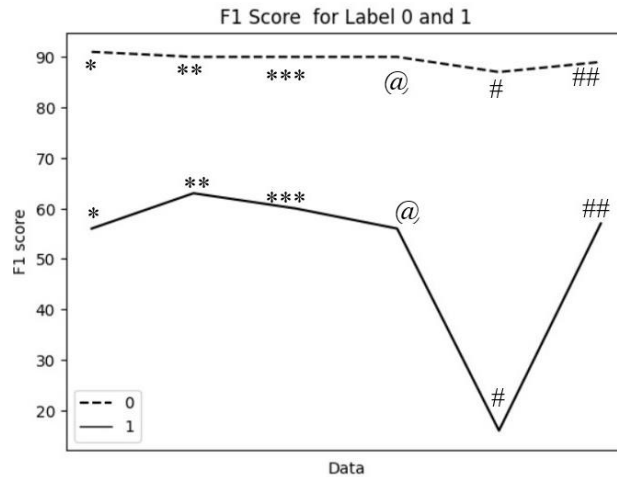


Fig. 5: Comparative Analysis based on F1 Score

Table 2: Comparative Analysis based on Accuracy

Algorithm	Naïve Bayes	Support Vector Machine	Random Forest	Logistic Regression	KNN P=1	P=2
Accuracy	85%	85%	84%	84%	77%	82%

Table 2 displays the results of numerous machine learning methods, including Naive Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN), with P set to either 1 or 2 where Naive Bayes and SVM had the anticipated predicted data and testing data greatest accuracy scores, scoring 85%, followed closely by Random Forest and Logistic Regression, scoring 84%. KNN scored 77% accuracy, the least accurate of the algorithms.

Naive Bayes and Support Vector Machine get the greatest accuracy score from the results analysis of the various machine learning algorithms shown above. Because of this, the Naive Bayes and Support Vector Machine are used to create the hybrid model using the voting ensemble classifier [13], and even different hyperparameters were used to find the better performance.

4.2 Result of Proposed Hybrid Model

Table 3: Comparative Analysis of Hybrid Classifier

Predicted Class Label	Precision	Recall	F1 Score
0	91%	92%	92%
1	67%	65%	66%

Table 3 shows the Precision, Recall and F1 Score of the Hybrid algorithm. This shows that 91% of the cases that were projected to be in class 0 were, in fact, in class 0. Furthermore, 92% of the real examples that belonged to class 0 could be accurately identified by the model. The accuracy and recall components of the

F1-score, which is 92%, are well-balanced.

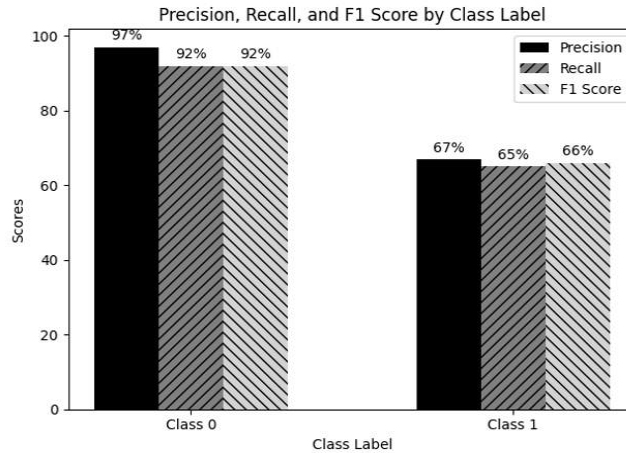


Fig. 6: Precision, Recall, and F1 Score of Proposed Hybrid Classifier

Table 4: Accuracy and Accuracy Score of the Proposed Hybrid Algorithm

Parameter	Score
Accuracy	87%
Accuracy Score	86.83%

Table 4 shows the accuracy and accuracy score of the hybrid model where accuracy is 87% and Accuracy Score is 86.83%. Overall, the results suggest that machine learning algorithms can be effectively used for spam email detection with the Naïve Bayes algorithm and Support Vector Machine being the most effective. Hence the Hybrid algorithm that combines the best features of Naïve Bayes and Support vector Machine is declared as the most efficient method of detecting spam email. However, it is important to note that the performance of algorithms may vary depending on the dataset used and the specific parameter chosen. Further research is needed to explore the effectiveness of these algorithm for different datasets and parameter settings.

5. CONCLUSION

The study may have demonstrated the viability of machine learning algorithms for spam detection by demonstrating their excellent precision and recall in properly identifying spam emails. Among the machine learning algorithms implemented, the Proposed Hybrid Classifier i.e., a combination of best features of both Naïve Bayes and Support Vector Machine seems to be the most effective machine learning algorithm for spam email detection acquiring the highest accuracy rate of 87.83%. According to the study, the quantity and quality of data used to train the algorithm have a big influence on how well it works. To obtain the best outcomes, researchers and practitioners should properly select and preprocess the dataset. The study could have shown that spam strategies change over time, necessitating ongoing algorithm updates in order to maintain effectiveness. As a result, experts in the field should think about putting an algorithm monitoring and updating system in place. A thesis dissertation on spam detection using machine learning algorithms will highlight the advantages and disadvantages of the selected approach and offer suggestions for further study and useful application.

References

- Email Marketing Statistics You Need to Know in 2022.” 10 Email Marketing Statistics You Need to Know in 2022, 25 Feb. 2022.
- Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6), e01802.
- Madhavan, M. V., Pande, S., Umekar, P., Mahore, T., & Kalyankar, D. (2021). Comparative analysis of detection of email spam with the aid of machine learning approaches. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012113). IOP Publishing.
- Onyango, L. A., Waititu, A. G., Mageto, T., & Kilai, M. (2022). A Hybrid Classification Model of Artificial Neural Network and Non-Linear Kernel Support Vector Machine. *International Journal of Data Science and Analysis*, 8(2), 47-58.
- Kaggle data set accessed on 3 November 2020, <https://www.kaggle.com/uciml/sms-spam-collection-dataset>
- Enron Spam data set accessed on 3 November <http://nlp.cs.aueb.gr/softwareanddatasets/Enron-Spam/index.html>
- A Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar and D. Karia, "E-Mail Spam Classification via Machine Learning and NaturalLanguage Processing," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 693-699, doi: 10.1109/ICICV50876.2021.9388530.
- Shafi'i, M. A., Maryam, S., Oluwafemi, O., Ismaila, I., & John, K. A. (2018). Comparative analysis of classification algorithms for email spam detection.
- Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Computer Science*, 190, 479-486.
- Onyango, L. A., Waititu, A. G., Mageto, T., & Kilai, M. (2022). A Hybrid Classification Model of Artificial Neural Network and Non Linear Kernel Support Vector Machine. *International Journal of Data Science and Analysis*, 8(2), 47-58.
- Nisar, N., Rakesh, N., & Chhabra, M. (2021, June). Voting-ensemble classification for email spam detection. In *2021 International Conference on Communication information and Computing Technology (ICCICT)* (pp. 1-6). IEEE
- Enron Spam data set accessed on 21st of November Email Spam Classifier Using Naive bayes | Kaggle Appendix
- Prince, M. S. M., Hasan, A., & Shah, F. M. (2021, April). A New Ensemble Model for Phishing Detection Based on Hybrid Cumulative Feature Selection. In *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 7-12). IEEE