EMAIL SPAM DETECTION USING HYBRID ALGORITHM

RAGAVI R [1], JANANI G[2],

[1]Student, Dept. of Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
[2] Student, Dept. of Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India

---***---

Abstract - *Email spam is still a continuous and changing issue that threatens user experience and calls into question the reliability of digital communication. By automating the detection of spam emails, machine learning techniques have become a potent weapon in the fight against this problem. In this study, a hybrid algorithm that combines both Logistic Regression and Neural Networks for email spam detection is introduced. The simplicity of Logistic Regression and pattern recognition capabilities of neural networks have been combined. Also, to prepare the data for the model, feature extraction and data transformation is involved. At first, Logistic Regression identifies the spam indicators in the mail which is followed by deeper analysis by Neural Networks to enhance classification. The hybrid algorithm overtakes individual models, according to experimental findings, and achieves a high level of accuracy in differentiating spam from authentic emails leading to have a more secure digital communication environment.*

Key Words: *machine learning, Neural Networks, Secure Digital Communication, Logistic Regression.*

1. INTRODUCTION

In the present technological world, the widespread use of email communication has completely changed how we interact and share our information both official and personal. The actual problem here is, as the usage of email grows immensely, Spam, phishing and all other malicious activities also grow along with it. The immense increase in the number of unwanted emails called spam and the person sending them are spammers. Online spam has grown to be a serious issue. It blocks the user from making good usage of storage and time. Also, the massive amount of spam mails flourishing through the network have devastative effects on the memory space of mail servers, CPU power and communication bandwidth. The level of danger of spam emails is on the increase in daily basis and is responsible for over 77% of global email traffic. These spam emails have resulted in financial loss to many users who were the victim of fraudulent practices of spammers who send emails pretending to be from reputed companies with the aim to gain sensitive information like passwords, credit card numbers etc., These has resulted in the need for development of the model "spam detection and filtering". This proposed hybrid machine learning algorithms provide a proactive and reliable way to successfully combat these risks.

Email spam is still a problem, and fraudsters' strategies are getting more and more complex. Regular spam filters frequently find it difficult to keep up with these developing dangers. In order to achieve greater accuracy in identifying and blocking harmful emails, our solution acknowledges the need for a more flexible and robust approach, integrating the strengths of several machine learning algorithms. The leading email providers such as Gmail and Yahoo have employed different machine learning models such as Naïve Bayes, Neural networks, Support vector machine and K Nearest neighbor. We are thrilled to present our cutting-edge method for email spam identification, utilizing the strength of hybrid machine learning algorithms, to battle this ever-evolving threat. In this project, the model has been made with the use of hybrid algorithms, which combine the strengths of different machine learning models like neural networks, decision trees, and support vector machines. These algorithms can be combined to more effectively capture the many patterns and attributes found in email content, headers, sender behavior, and attachment properties.

2. LITERATURE SURVEY

Kriti Agarwal, Tarun Kumar [2020] makes an effort to apply machine learning techniques to recognize a pattern of recurring keywords that are categorized as spam. The system also suggests categorizing emails based on additional variables found in their structure, such as Cc/Bcc, domain, and header. When a parameter is applied to the machine learning algorithm, it is regarded as a feature. The pre- trained machine learning model will have a feedback mechanism to discriminate between a correct output and an ambiguous output. This approach offers an alternate architecture for the implementation of a spam filter.

Aakash Atul Alurkar [2019] examined the usage of string matching algorithms for identifying spam emails. The performance of six well-known string matching algorithms, including the Longest Common Subsequence (LCS), Levenshtein Distance (LD), Jaro, Jaro-Winkler, Bi-gram, and TFIDF, is specifically examined and compared in this paper using the Enron corpus and the CSDMC2010 spam dataset as two different datasets. They found that in both datasets, the Bi-gram algorithm performs the best at detecting spam.

T.Verma [2018] proposed a technique for filtering emails using the SVM algorithm and feature extraction. This method includes a number of processes, including Email Collection, where data is taken from the dataset. It is then routed via preprocessing, wherein extraneous content is eliminated and only desired content is sent on to the next step. Then comes feature extraction followed by SVM model training. The dataset from the Apache Public Corpus was used by the author. Special symbols, HTML tags, URLs, and extraneous alphabets were deleted from the suggested solution. All of the dictionary words were mapped by the author using the vocabulary file. A 98% accuracy was achieved using the SVM method on a pre-processed dataset.

Abdulhamid Muhammad Shafi [2018] conducted performance analyses for various machine learning classification methods, including Random Forest, Lazy Bayesian Rule, and Radial Basic Function (RBF) Network. Bayesian Logistic Regression, J48, and Tree. There was a based on a comparison of all of these provided algorithms the F-measure, precision, recall, root mean squared error, and Accuracy. They made use of the UCI Machine dataset. Learning Archive. For determining the recall and precision They used the F-measure approach to determine value. Rotation forest algorithm was used to acquire the maximum measure, and Naive Bayes algorithm has the lowest. The best result was reached for the Rotation Forest Algorithm with 87.9 using the Kappa Statistics for the statistical results. Rotation Forest technique produced the highest accuracy (94%), while REP Tree algorithm produced the lowest accuracy (89%). Other algorithms, such Naive Bayes and J48, provided accuracy of 88.5% and 92.3%, respectively.

3. OBJECTIVE AND METHODOLOGY

This project mainly focuses on developing an efficient email spam classifier which can identify and filter emails more accurately than ever. It prevents spam messages from getting loaded into the user' inbox, thereby improving user experience. It is known that nowadays emails are just flourishing in the inbox which consists of both spam and ham. These ham emails are very dangerous because of the fraudulent practices of spammers who sent these kinds of illegitimate emails. Because of this, users have even suffered from financial loss. This is considered to be one of the reasons behind our objectives. Along with this, spam emails also occupy most of the storage space. It also wastes the time because without spam detection and filtering, the user has to do it on their own.

Compared to the existing algorithms, this model is designed to be more scalable and can handle large volumes of email data. The hybrid approach which combines both neural networks and logistic regression can add the strengths of both the techniques. While logistic regression offers interpretability and can act as a useful final classifier, neural networks are powerful for feature learning and catching complicated patterns. This model supports false positives reduction.

Even these days, many spam emails are incorrectly classified as ham and ham emails are incorrectly classified as spam which creates a lot of confusion for the user. This issue is known as false positives. This model overcomes the above issue as the algorithm used here is neural networks. The model can generalize good to new spam patterns and those unseen ones without getting the training data overfitted. It increases the overall user experience by reducing its exposure to spam emails in the inbox.

4. CONTRIBUTION

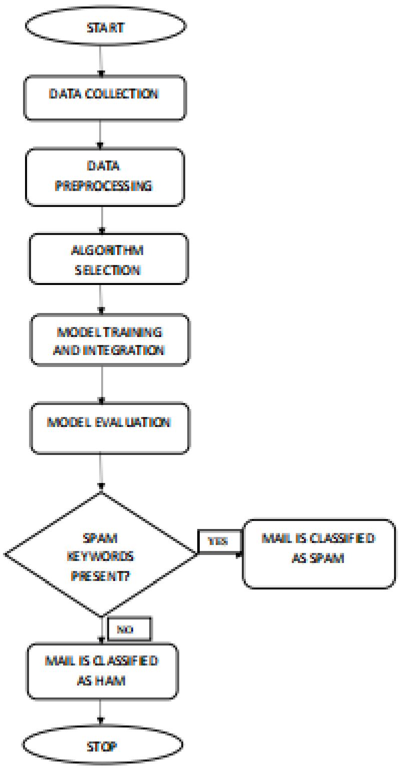
The contribution for the successful completion of the model has been broken into two parts where my part is completely focused into training and testing stages. In the context of machine learning, these two stages are considered very crucial for making sure of the model's effectiveness and accuracy in serving its calculated functions. In order to begin with I have started to thoroughly research all of the different algorithms that were relevant to this project. Because, only with proper research, it is possible to choose the most suitable algorithm.

After understanding all the algorithms, each was put into implementation for getting the accuracy rate for further proceedings in the project which was considered to be a baseline for our endeavour. At first the model was individually trained with each algorithm. After getting results, it soon became clear that using just one algorithm wouldn't be enough to produce the necessary results. As a result, we made the decision to research the idea of hybrid algorithms. In an effort to achieve greater accuracy, we tried mixing various algorithms.

During the individual evaluation of each algorithm, it is known that logistic regression and k nearest neighbour algorithm has shown greater accuracy as compared with the others. But the thing is, all just work upon the frequency of occurrence of each word. We were aware of the limits of these conventional algorithms as our research developed. They had some effectiveness, but they were unable to understand the larger context of text messages.

Instead of using these algorithms, it is good to switch to the concept of neural networks where the model will be reading the whole text message, understands the context and semantics and then proceed with further jobs. This capacity was expected to considerably improve the precision and potency of our model. From all the above findings, it was decided that it is better to use the combination of both neural networks and logistic regression.

5. FLOW DIAGRAM



EXPLANATION OF FLOW DIAGRAM:

In this step, the email data will be prepared so that the machine learning models may be trained and tested.

• Lowercasing: we have converted the uppercase letter to lowercase letters to make sure that the model should not treat them different as both are same.

• Removal of special characters and punctuations: As punctions and special characters like '$','&' etc., don't carry any necessary information, we have removed them in order to make the dataset compact.

• Handling newlines and blank spaces(whitespaces): In addition to the above process, we have also removed extra newlines to support text formatting.

• Tokenization: we had split the text document into a list of words that is known to be tokens. Each token in the text serves as a unique unit of meaning. We consider this step as very important because it enables you to treat each word as a separate object, enabling fine-grained analysis and processing of the text. The machine learning model uses each token as a feature in order to make predictions.

• Removal of Stop Words: In this step, we have removed the words that occur very frequently in text and also doesn't carry any necessary information in it. Performing this step help in dimensionality reduction which improves the model performance further. In English, the words that frequently occur are 'the', 'and', 'or', 'in', 'is' etc.,

•Stemming and Lemmatization: Lemmatization and stemming are methods for breaking down words to their root or basic forms. The main objective is to combine words with comparable meanings, which can aid in feature engineering and feature reduction. When you wish to minimize dimensionality but don't necessarily need proper words, stemming is a good approach. But in the case of lemmatization, it is a more complex method that reduces words to their dictionary form. The outcome is always a meaningful word. When you wish to keep the words' grammatical correctness and interpretability, lemmatization is preferred.

• Feature Extraction: we have created numerical features by converting the preprocessed text data. These numerical features are considered to be the inputs for our machine learning model. This technique includes term TF- IDF which we have implemented. It's a numerical statistic that is frequently used in information retrieval and natural language processing to assess the significance of a word inside a document. For text-based machine learning applications which includes text categorization and information retrieval, TF-IDF is especially helpful. The frequency of a particular word in a document is quantified by the TF.

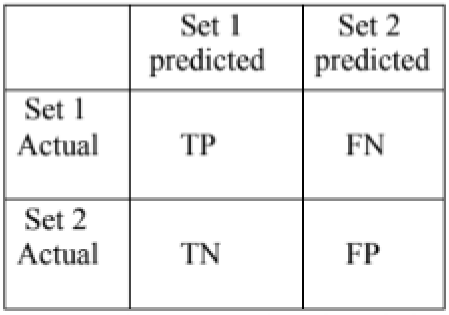
• Data Preparation for Neural Networks: As we used neural networks, we made all the email texts to be in same length by truncating larger texts or padding shorter ones with zeros. This is known as padding sequences. Apart from this, we replaced the spam and legitimate labels into numerical values that is, replacing '0' for legitimate or ham labels and '1' for spam labels. This process is known as encoding labels.

• Data Preparation for Logistic Regression: As we have used logistic regression for training our model, we used the preprocessed features like the tokens, TF-IDF vectors from which it classifies accordingly. When a new mail arrives after all the preprocessing features have been generated, it calculates the frequency of the words and then decides whether these words occurred mostly in spam or ham emails from which it has been trained. And then, it classifies as spam or ham.

• Data Splitting: We have splitted the preprocessed dataset into training and testing data in a ratio of 80:20. The machine learning model is trained using the data from the subset 'training set'. This part of the data is used by the model to learn patterns and relationships. The performance of the final model is assessed using the testing set. It is a chunk of the data that has been withheld and has not been viewed by the model during training or validation. Apart from training and testing set, we have validation set. This set helps us in selecting the good model among others. It evaluates the performance of the model at the time of training. Also, it prevents overfitting. It is good to randomize the rank of the samples to make sure that the sets reflect a vast number of examples. Also, in order to maintain the class distribution in each split, we have utilized stratified sampling. This made sure that each set accurately represents the proportions of the entire class.

6.RESULT AND DISCUSSION

A confusion matrix is a summary of the results of a categorization challenge. The number of right and bad predictions are totalled with count values and dispersed by each class. The confusion matrix's key is this. The confusion matrix shows how your classification model becomes perplexed while making predictions. It gives us insight into the types of errors that are being made, as well as into the mistakes that specific people have made when using a classifier. Only after the true values of the test data are known can it be determined.



	Set 1 predicted	Set 2 predicted
Set 1 Actual	TP	FN
Set 2 Actual	TN	FP

TABLE 1: General Representation of Confusion matrix

Here,

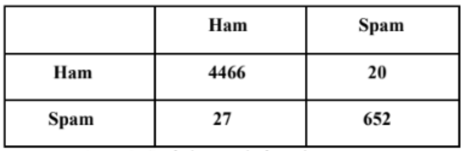
Set 1: Positive

Set 2: Negative

- TP: Examination will be positive, also it's predicted value will be positive.
- FN: Examination will be positive, but it's predicted value will be negative.
- TN: Examination will be negative, but it's predicted value will be negative.
- FP: Examination is negative, but it's predicted value will be positive

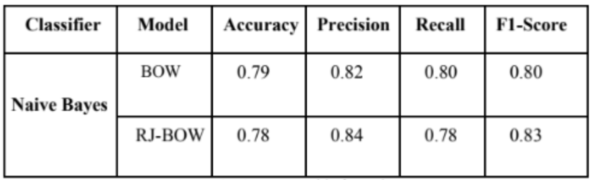
6.1. NAIVE BAYES

Naive Bayes is a popular machine learning method for email spam identification. Since Naive Bayes is a simple understandable algorithm, it has a short implementation time. When you just have a tiny amount of labelled data, Naive Bayes is advantageous since it performs well even with small training datasets. Using Bayes' theorem, we have calculated the posterior probabilities for both the spam and non-spam classes in order to categorize a new email as spam or not. In conclusion, Naive Bayes also have drawbacks, its naive independence assumption and constrained expressiveness. It can be a good option for simple spam detection jobs, but other algorithms, such as logistic regression, neural networks may produce much more good results.



	Ham	Spam
Ham	4466	20
Spam	27	652

TABLE 2: Confusion Matrix for Naive Bayes

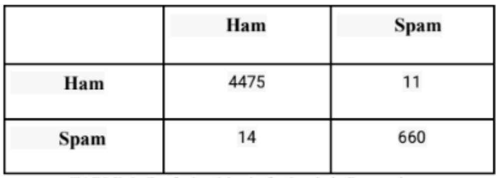


Classifier	Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	BOW	0.79	0.82	0.80	0.80
	RJ-BOW	0.78	0.84	0.78	0.83

TABLE 3: Accuracy table for Naive Bayes

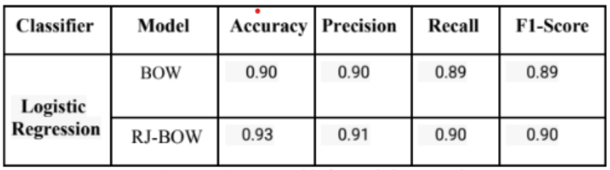
6.2. LOGISTIC REGRESSION

Logistic Regression is another algorithm that can be used for email spam detection in machine learning. Using the logistic (sigmoid) function, logistic regression models the probability that an email belongs to the spam class. In order to maximize the likelihood of the observed data, the logistic regression algorithm learns the coefficients for each feature during training. By penalizing large coefficients, regularization approaches like L1 or L2 regularization can be used to prevent overfitting. The probabilities generated by logistic regression fall between [0, 1]. To choose the class label, we selected a threshold almost like 0.5. The precision and recall trade-off can be adjusted by changing the threshold. Results from logistic regression are simple to interpret. It is a reliable option when working with noisy data or when you have a limited amount of labelled data because it has a decreased risk of overfitting compared to more complex algorithms. It also works on the frequency of the words present.



	Ham	Spam
Ham	4475	11
Spam	14	660

TABLE 6: Confusion Matrix for Logistic Regression

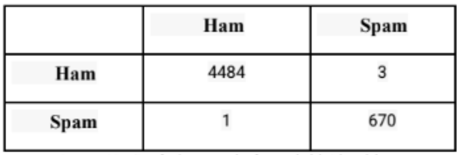


Classifier	Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	BOW	0.90	0.90	0.89	0.89
	RJ-BOW	0.93	0.91	0.90	0.90

TABLE 7: Accuracy table for Logistic Regression

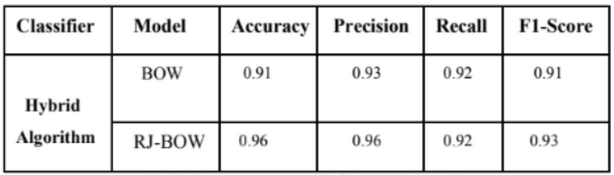
6.3. HYBRID ALGORITHM

Combining Logistic Regression and Neural Networks in email spam detection can offer several advantages, often referred to as a hybrid approach. Performance can be enhanced by combining various techniques, such as neural networks and logistic regression. This hybrid technique efficiently captures both simple and complicated patterns by combining the advantages of both models. We fed the neural network's newly learned features into a logistic regression model after it has been trained. A set of input features are used in a simple linear model called logistic regression to provide an output that may be understood as the likelihood that an email is spam. For binary classification problems like email spam detection, it is a widely used model. you use the neural network's output as a set of logistic regression's input features. Logistic regression uses the features that the neural network has extracted as its input features. The following is the accuracy we obtained from our hybrid model. Hybrid algorithms play a prominent role in improving the search capability of algorithms. Hybrid algorithms exploit the good properties of different methods by applying them to problems they can efficiently solve. The fusion of more than one algorithm makes the model to effectively classify the email thus making it the best one.



	Ham	Spam
Ham	4484	3
Spam	1	670

TABLE 12: Confusion Matrix for Hybrid Algorithm



Classifier	Model	Accuracy	Precision	Recall	F1-Score
Hybrid Algorithm	BOW	0.91	0.93	0.92	0.91
	RJ-BOW	0.96	0.96	0.92	0.93

TABLE 13: Accuracy table for Hybrid Algorithm

7. CONCLUSION

In conclusion, the hybrid strategy that combines logistic regression and neural networks for email spam detection offers a complete and efficient solution. The ability of neural networks to recognize complex patterns in email content and the clarity and simplicity of logistic regression are two advantages of this combination. The resulting hybrid model is perfect for real-time email identification because it not only boasts improved accuracy but also feature extraction abilities and scalability. Its robustness is highlighted by its capacity to adjust to various spam forms and shifting strategies. However, to ensure user privacy and uphold legal requirements, the successful implementation of such a system calls for rigorous model tuning, continual monitoring, and ethical considerations. By achieving these goals, the hybrid model guarantees customers a dependable and seamless email experience, effectively eliminating spam while upholding openness and credibility throughout.

8. SUGGESTIONS FOR FUTURE WORK

To further improve the efficiency of email spam detection system, there are some suggestions which provide directions for achieving the above. Using advanced deep learning architectures such as BERT model for email spam detection can be said as a good approach. It is a transformer-based model. Other than its much higher computational resources, it can be considered as a very good approach.

Other than this, we can also include integration of user feedback. It integrates feedback on the emails that are mis classified in order to improve the email spam detection and filtering system performance.

REFERENCES

[1] T. A. Almeida and A. Yamakamil, “Facing the spammers: A very effective approach to avoid junk Emails,” Expert Systems with Applications, vol. 39, Issue 7, 1 June 2012, Pages 6557–6561.

[2] T. Fawcett. “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, issue 8, pp. 861-874, June 2006.

[3] T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to spam filtering,” Expert Systems with Applications, vol. 36, pp. 10206–10222, 2009.

[4] T. Guzella and W. Caminhas, “A review of machine learning approaches to spam filtering,” Expert Syst. Appl., vol. 36, no. 7, pp. 10206–10222, 2009.

[5] Q. Le and T. Mikolov, “Distributed representation of sentences and documents,” presented at the 31 the

International Conference on Machine Learning, Beijing, China, 2014.

[6] M. Woitaszek, M. Shaaban, and Czernikowski. "Identifying junk electronic mail in Microsoft outlook with a support vector machine," in Proc. Symposium on Applications and the Internet, 2003, pp. 66–169.

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," arXiv:1405.4053.

[8] J. Hovold, "Naive Bayes spam filtering using word-position-based attributes," presented at 2nd Conference on Email and Anti-Spam, Stanford, CA, 2005.

[9] I. Kanaris, K. Houvardas, and E. I. Stamatatos, "Words vs. Character n-grams for anti-spam filtering," International Journal on Artificial Intelligence Tools, vol. 16, no. 6, pp. 1047–1067, 2007.

[10] G. Salton, Introduction to Modern Information Retrieval, A. G. K. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker, "On the relationship between feature selection and classification accuracy," in Proc. the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery, 2008, p. 16.Auckland, McGraw-Hill, 1983.

[11] J. S. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," Pattern Recognition, vol. 48, issue 9, pp. 2761-2771, 2015.

[12] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based Framework for Text Categorization," Procedia Engineering, vol. 69, pp. 1356-1364, 2014.

[13] M. Sahami, S. Dumaisis, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," AAAI Technical Report WS-98-05, 1998.

[14] G. E. Hinton, "Learning distributed representations of concepts," inProc. the Eighth Annual Conference of the Cognitive Science Society,"

[15] D. Talbot, "Where SPAM is born," MIT Technol. Rev., vol. 111, no. 3, p. 28, 2008.

[16] C. Li, L. Ji, and J. Yan. "Acronym disambiguation using word embedding," in Proc. the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[17] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in Proc. 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014.

[18] A. Barushka et al., "Spam filtering using regularized neural networks with rectified linear uni.," presented at XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, 2016.

[19] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch "Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis" IEEE GLOBECOM, 2008

[20] Steve Webb, James Caverlee, Calton Pu, Introducing the Webb Spam Corpus: using Email spam to identify web spam automatically, CEAS, 2016.