International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spam E-Mail Detection Using Machine Learning Algorithms

[1]M. Hema Kalyan, [2]M. Hari Krishna

[1]Student, Department of Information Technology, GMR Institute of Technology, Rajam, A.P, India
[2]Assistant Professor, Department of Information Technology, GMR Institute of Technology, Rajam, A.P, India

ABSTRACT:

In this online world, we mostly depend on E-Mail services for faster information exchange but SPAM becomes the main drawback, an irrelevant message sent which causes cyber-attacks and misusing personal information. The ratio of spam mail is increasing day by day. The main goal of this paper is to construct a machine-learning model which detects spam emails. Machine Learning Techniques are used to automatically filter spam emails at a very successful rate. Moreover, Spam emails waste resources in terms of storage, bandwidth, and productivity. We can use machine learning algorithms like Naïve Bayes or support vector machines (SVM) or Decision trees to detect Spam and Non-spam (Ham) mails as these algorithms can achieve high accuracy. In this paper, we will discuss these algorithms and apply those algorithms to our dataset, and the algorithm which gives the best accuracy and precision will be considered.

Keywords: Machine Learning Algorithms, Naïve Bayes, Support vector machines (SVM), Datasets.

I. INTRODUCTION

In this vast world, Time complexity plays a major role in every aspect. Nowadays, we are using Electronic mail rapidly, commonly known as E-mails for a faster exchange of information. 4 Million people are using this technology daily. Therefore, it needs to be very safe to use it. It became very essential service for billions of people over the Internet. Due to its convenience and mass transfer of messages to people. Electronic mail became a trendsetter over the internet. For this technology, there is a threat known as SPAM.

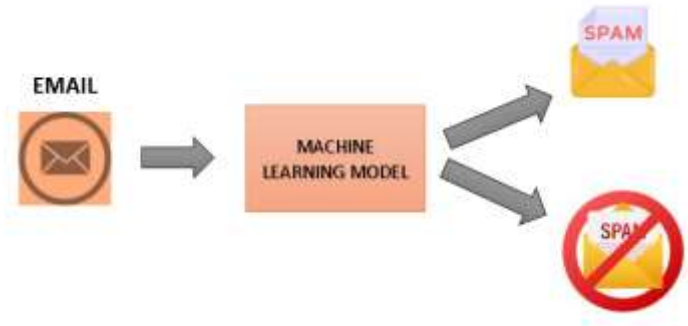

Fig 1. Spam Filtering Classification

It is a message with fraudulent content which causes misusing personal information. SPAM messages can also cause storage and bandwidth waste. To prevent this type of threat we are using some machine learning techniques. Machine learning algorithms proved that they are more efficient than the user-defined rules approaches because machine learning algorithms are more feasible, simple, and dynamic to use.

Categories	Description
Education and training	The spam of offers on seminars, courses, and online training
Personal Finance	The spam on stock marketing, loan packages
Adult content	The spam of adult content of pornography and prostitution
Health	The spam of fake medication
Promotional products	The spam on fake fashion items

Table 1. Categories of spam

Spam emails are very harmful in another way which leads to several very sensitive data breaching and some viruses like trojans, worms, unblockable ads, cryptocurrency miners, and other malware. The task of handling spam emailing is very essential because it can lead to critical situations. In other words, spam emails are quite annoying to the user.

II. LITERATURE SURVEY

Consider a situation in which by the means of the internet, you are receiving spam emails very frequently whether it may be promotions for their products to purchase them. These are negative marketing activities and fraud activities. As a receiver, we are helpless to control these spam emails. It also consumes a lot of memory and valuable time. So, there is a high need of having some mechanism to reduce these types of spam emails. In this paper, we are presenting a machine learning-based spam detection mechanism. This model consumes a dataset containing approximately 6000 emails. Using this mechanism, so much time and memory will be saved. With the help of this dataset, features can be extracted and it plays a major role in identifying the accuracy, precision, computational power, and misclassification rate of the particular algorithm.

In paper [1], they represented a machine learning model that uses a dataset of 6000 valid and invalid emails. Generate Dictionary, Generating Features, Generating a Machine learning Model and Testing of Machine Learning Model are the steps done in this paper to construct this SPAM filter model. The Model is initiated by creating a dictionary that includes a library named “Stopwords”, and removing all helping verbs. After creating Dictionary, Features Generation will take place and these extracted features will be thoroughly tested. In the Machine learning Model Generating step, they’ve used the Naïve Bayes algorithm(Probabilistic classifier), and testing will be done. For sending emails, this model has an application that provides a content spam filter. Sending Images option is disabled because it is possible that the filter got failed to detect objects in the image. A server system is activated and a spam report module is deployed which categorizes the messages into Spam / Ham. This data can be further analyzed for the determination of compromised and non-compromised machines based on a well-defined degree of fault tolerance. This server also keeps the client’s system details such as the client name and timestamp. It identifies sections of words against its spam city rather than individual words as some words are spam for one organization but not in another organization. In this model after applying the Naïve Bayes algorithm, emails are taken as inputs and the results are classified into 0 as non-spam mail and 1 as spam mail. By using this algorithm, we achieved 87.82 percent of accuracy.,

In paper [2], In this paper, they used a Logistic regression model to classify Spam mail. It is a machine-learning classification algorithm. We use this model for the prediction of the output of a categorical dependent variable. This logistic model can estimate the probability of two class responses like ham/spam. A spam mail dataset is taken which contains 5572 emails including both ham and spam emails. The dataset is divided into two sets, one for training and another for testing. They have taken 80 percent of emails for Training and 20 percent for testing and the output of the model will be shown with the help of 0 and 1. By using this logistic regression algorithm to achieve an accuracy of 96.59 percent.

In paper [3], A dataset is taken from spam assassin which contains nearly 5000 emails that will be read by a python package “Pyzmail”. during this text pre-processing phase, email structures are going to be extracted and the contents will be converted to plain text for analysis. during this model, two feature sets are prepared i.e., stopwords with N-gram and tf-IDF (term frequency-inverse document frequency) and therefore the most frequent word count with count vectorization. N-gram and tf-IDF are created by exploring the text structure to take advantage of the contextual features. the foremost frequent word count with count vectorization is based on counting the most frequently occurring words from email content. A pipeline is made to feed data with the feature set and it is also easier to compare results. This pipeline consists of three algorithms i.e., Naïve Bayes, Logistic Regression, and Support vector machines. The evaluation criteria are supported by Accuracy, Precision, Recall, and F1 Score. Among the models which are using feature set 1 i.e., N-gram and tf-IDF, Logistic Regression got the very best precision i.e.,98.33%. Among the models which are using feature set 2 i.e., the most frequent word count with count vectorization, Logistic Regression got the very best precision i.e., 99.33%

In paper [4], a dataset is taken consisting of 5674 mail with columns showing the category of the mail. This dataset is categorized into training data and testing data. during this dataset, spam email is taken into account as tokens and a unique number is assigned to each word i.e., “Tokenisation”. Countvectorizer.fit() is the method used in machine learning to read the complete email, identify all available features, and return the count. Term frequency(tf) gives the number of times the word is repeated in the given dataset and inverse document frequency(IDF) minimizes the importance of the feature. Useful keywords aren't removed by TFIDF. Stemming are going to be done to convert the words into similar structures. Classifier evaluations are often done based on the following metrics i.e., Accuracy, precision, recall, and F1 score. After comparing the obtained results, the ensemble classifier got promising results and therefore the speed of testing is also better with an accuracy of 98 percent.

III. METHODOLOGY

A. DATA COLLECTION

In the initial step, we are visiting download the dataset from Kaggle (https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification) which contains 5572 records of two columns i.e., Message and Category. we'd wish to import the required python libraries to perform operations such as NumPy, Pandas, and sklearn then the downloaded dataset has got to be uploaded by the method “read_ csv” in pandas’ library. Data Pre-processing must be done by checking null values within the data.

B. LABEL ENCODING

Later, Label encoding must be done and it's defined as the process of converting labels into numerical values to machine-readable form. Spam is labeled as “0” and Ham Mail is labeled as “1”.

```
mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```

C. FEATURE EXTRACTION

Now the info in the spam dataset is categorized into Training data and Testing data in the ratio of 80:20 and then feature extraction is done using tf-idf vectorizer which id Term frequency-inverse document frequency. this is often a very common algorithm to transform the text into a meaningful representation of numbers which is used to fit machine algorithms for prediction. TF-IDF Vectorizer may be a measure of the originality of a word by comparing the number of times a word appears in the document with the number of documents the word appears in.

D. MODEL TRAINING

In this model, we are employing a Logistic Regression Classifier for predicting spam mail. Logistic regression is one among the most popular machine learning algorithms, which comes under supervised machine learning algorithms. it's used for predicting the specific dependent variable using a given set of independent variables. the outcome must be a discrete or categorical value. It is frequently either Yes or No, 0 or 1, true or false, etc., but rather than providing the precise values of 0 and 1, it provides the probability values that fall between 0 and 1.

E. MODEL EVALUATING

Model evaluation is the process of using different evaluation metrics to understand a machine learning model’s performance, also as its strengths and weaknesses. Model evaluation is vital to assess the efficacy of a model during initial research phases, and it also plays a task in model monitoring.

F. RESULTS

```
prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

```
print("Accuracy_on_training_data:", accuracy_on_training_data)
```

```
Accuracy_on_training_data: 0.9661207089970832
```

Fig 1. ACCURACY IN TRAINING DATA

```
prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
```

```
print("Accuracy_on_test_data:", accuracy_on_test_data)
```

```
Accuracy_on_test_data: 0.9623318385650225
```

Fig 2. ACCURACY OF TEST DATA

IV. PSEUDO CODE OF THE METODOLOGY

1. Import required python libraries
2. Dataset collection
3. Data pre-processing
4. Label encoding
5. Feature extraction
6. Model training
7. Evaluating model

8. Results

V. CONCLUSION

By the above results, we will conclude that the Naïve Bayes classifier outperforms all other classifiers. In present scenarios, spam emails are increasing rapidly. We'd like a better model to identify spam emails to handle that scenario. Our proposed model witnesses the naïve Bayes classifier, which provides the probabilistic statistics that identify whether the email is spam. Our proposed model achieves a mean of 95 percent accuracy.

VI. REFERENCES

[1]. Nikhil Govil, Kunal Agarwal, Ashi Bansal, Astha Varshney. "A Machine Learning based Spam Detection Mechanism", IEEE Xplore Part Number: CFP20K25-ART; ISBN:978-1-7281-4889-2.

[2]. Chode Abhinav, K Jayachandra, Kommu Pranith Kumar, V Sowmya "Spam Mail Detection using Machine Learning", International journal for applied science & engineering technology research.

[3]. Manoj Sethi, Sumesha Chandra, Vinayak Chaudhary, Yash, "Email Spam Detection using Machine Learning", International Research Journal of Engineering and Technology (IRJET).

[4]. Nikhil Kumar, Sanket Sonowal, Nishant, "Email Spam Detection Using Machine Learning Algorithms", IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.

[5]. Jyoti Dake, Gunjan Memane, Prerana Katake, Samina Mulani "Email Spam Detection and Prevention using Machine Learning", International Journal of Advanced Research in Computer and Communication Engineering.

[6]. "An Efficient Spam Filtering using Supervised Machine Learning Techniques" in IJSRCSE, Vol.6, Issue.2, pp.33-37, April (2018).

[7]. Vinodhini. M, Prithvi. D, Balaji. S "Spam Detection Framework using ML Algorithm" in IJRTE ISSN: 2277- 3878, Vol.8 Issue.6, March 2020.

[8]. Deepika Mallampati, Nagaratna P. Hegde "A Machine Learning Based Email Spam Classification Framework Model" in IJITEE, ISSN: 2278-3075, Vol.9 Issue.4, February 2020.

[9]. Linda Huang, Julia Jia, Emma Ingram, Wuxu Peng, "Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection", 2018 17th IEEE International Conference on Trust, Security, and Privacy in Computing and Communications.

[10]. http://www.securelist.com/en/threats/spam?chapter=97.

[11]. Deepika Mallampati, K.Chandra Shekar and K.Ravikanth "Supervised Machine Learning Classifier for Email Spam Filtering", © Springer Nature Singapore Pte Ltd. 2019 and Engineering, https://doi.org/10.1007/978- 981-13-7082-341.

[12]. Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.

[13]. Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. https://doi.org/10.1109/ACCESS.2019.2954791.

[14]. W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology. 3. 10.5121/ijcsit.2011.3112.

[15]. K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690