**Efficient machine learning approach for spamming electronic mail detection**

**Abstract:**

Electronic mail is a standard way of communication for all of us but an increasing number of email users have increased different unethical activities like irrelevant or unsolicited messages sending malwares, etc in email transferring procedure. Such emails are known as spam emails. For the Detection of such spam emails, features of email are used as parameters in machine learning algorithms like logistics regression, decision tree, random forest, SVM etc. In our model, we have collected data from different email users related to features of email like text, sender reputation, phishing website, special character, multiple language, attachment present in it and used this feature as a parameter for modeling our machine learning algorithm. Finally, we got accuracy …. From ……algorithm and …. Form …. algorithm

**Introduction:**

Every person on earth is using email for sending official letters and contacting people all over the world. Such expansion of this technology has benefited many people by saving time and money but along with its development, it has created major problems in this digitalized world like fraud text, transferring virus through file sharing, irrelevant or unsolicited messages etc. Such a problem as a combine is called spam email.

In this research work, we have used different machine learning algorithms for email spam detection based on the different features of email like email text, sender reputation, phishing websites and many more. The main focus is on data preprocessing in which we have considered all the features of email in our dataset for finding whether the corresponding email is spam or ham. We have vectorized all the additional features using TF-IDF and used different machine learning algorithms like logistic regression, SVM, Naive Bayes, Decision tree, Random Forest, for determining whether the given email is a spam or ham mail.

**Literature survey/related work**

According to Statista, in September 2021, the estimated daily spam volume was 88.88 billion spam emails a day. On average, that’s 21 spam emails per person per day. This proves that a strong email spam detection algorithm is needed to counter this problem.

There are mainly three types of spam filter which are used for spam detection:

**Third Party spam filter**

The most popular third-party spam filter is spam titan which uses a machine learning algorithm to separate spam from ham emails.

**Cloud Based spam filter**

This is an online spam filtering service which is hosted in the cloud and offered as subscription-based service by a third-party provider.

**Gateway spam filter**

Gateway spam filter is a type of spam filter that is installed at the gateway or the point of entry between a company’s email system and the internet.

Although there are many other spam filters, the spam traps will still be there. So, many organizations and researchers are aiming to build robust and efficient filters to stop this problem. Many models based on machine learning algorithms show excellent performance to detect whether an email is spam or ham. Many researchers have got above 90% accuracy in their related algorithms along with that they have included many features of email like email text, header information and many more. They have used popular machine learning algorithms like logistic regression, SVM, Naive Bayes, Decision tree for training their model but for data collection most of them have collected their data from Kaggle or any other database.

Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath [1] has conducted email spam detection using IP address and email text as parameter and Naive Bayes, Naive Bayes’ Classifier as machine learning algorithm. Similarly, focusing on data pre-processing procedure, Nikhil Kumar, Sanket Sonowal, Nishant [2] have trained their pre-processed data to their model.

**METHODOLOGY** Data collection:

We have collected data from different people based on different parameter like email text, sender reputation, links in email, attachment, number of special characters, phishing websites.  
 For training our model, data from Kaggle website is used and for testing purpose, practical collection of data is made.

**Data pre-processing:**

Data preprocessing is the procedure of converting the raw data into a vectorized dataset which can be used for training our models. The first step for data preprocessing is data cleaning which corrects the missing values, removes inconvenient data entry and modifies the data set as per the condition of the dataset.

The main focus of this research paper is vectorizing the dataset. We have taken various parameters like sender reputation, website present in the email, email text, special character, multiple language, attachment and many more. For vectorization of email text, we used feature extraction in which TF-TDF (term frequency- inverse document frequency) is calculated for each word present in the email text.

**TF IDF:**

w(i,j)=f(i,j) x log( )

where,

w(i,j): word i in document j

f(i,j): frequency of i in document

jN: total number of document

f(i): total frequency of word i in all document

Using this TF-IDF, a sparse matrix is calculated which will have n row (distinct types of word in the document) and m column (total number of text). In addition to this sparse matrix, we add more columns which give the other information about email like sender reputation, phishing website etc. For additional column phishing websites, we range the level of the phishing website for example, the most popular phishing website: the value is high and safe website: the value is almost zero.

**Modeling:**

For spam detection, a model should be made which can separate ham from spam emails based on the information about that email. Making a model for a test is done by using different supervised machine learning algorithms like logistic regression, random forest, SVM, Naive Bayes etc.

**Logistic Regression:**

In logistic regression, we find the optimum value of parameter θ using cost function and gradient descent. This overall method of finding the parameter is known as modelling. Further, calculating the value of the sigmoid (logistic) function using the determined parameter θ and given data, provides us the value between 0 to 1. Then using the probability concept, we can determine whether a given email is spam or ham.

Mathematical expression for logistic regression:

For modelling, θ value is calculated using gradient descent:

θ=θ-Σ[h(x)-y].x

where, m is the total number of features

h(x)= : sigmoid function

Y : 0: spam and 1:ham

X: features of email

**Naive Bayes:**

This algorithm is based on Bayes Theorem which gives the conditional probability of an event A given another event B had occurred. Using this theorem, we can find the probability of spam email with the help of data present on it.

P(A/B) =

where,

P(A/B): conditional probability of A given B

P(B/A): conditional probability of B given A

P(A): probability of event A

P(B): probability of event B

**SVM:**

SVM stands for support vector machine. In this algorithm, a margin line is determined to distinguish data points between two different classes. In addition, two extreme hyperplanes (positive and negative) are also determined to separate the classes. Moreover, kernel function is applied in a given data set to manipulate it which makes it easier to find the margin line for that given set.  
 **Random Forest:**

Random forests (RF) construct many individual decision trees at training. The result of random forest is calculated by averaging all the predicted values of all decision trees. Following equation are used to find the result:

RFfii=

Normfii=

Fii=

Nij=WjCj - Wleft(j).Cleft(j) - Wright(j).Cright(j)

**Decision tree:**

Decision tree split a dataset into two groups based on features. The importance of that feature to distinguish two different classes is determined by calculating gini impurity, reducing the variance and calculating the chi-square of the child nodes

.

**Testing:**

Testing a model based on a given training set will help to determine whether the algorithm is working efficiently or not. Confusion matrix is a tool used in the field of machine learning and statistics to visualize and evaluate the performance of a classification model. A confusion matrix is usually a square matrix with rows and columns corresponding to the true classes and predicted classes, respectively.

Here’s a breakdown of the confusion matrix of a binary classification problem:

| True positive | False positive |
| --- | --- |
| False Negative | True Negative |

Here's a quick explanation of the terms in the confusion matrix:

True Positive (TP): Instances that are correctly predicted as belonging to the positive class.

True Negative (TN): Instances that are correctly predicted as belonging to the negative class.

False Positive (FP): Instances that are incorrectly predicted as belonging to the positive class when they actually belong to the negative class (Type I error).

False Negative (FN): Instances that are incorrectly predicted as belonging to the negative class when they actually belong to the positive class (Type II error).

There are some other parameters which help to declare the efficiency of machine learning algorithms:

**Accuracy**

In machine learning, accuracy refers to the measure of how well a classification model correctly predicts the class labels of the input data points. It is a commonly used metric to evaluate the performance of classification algorithms.

**Accuracy:**

**Recall**

Recall, also known as sensitivity or true positive rate, is a metric that quantifies the ability of a classification model to correctly identify all positive instances from the total actual positive instances in a dataset.

**Recall:**

**Precision**

Precision is a metric that measures the accuracy of positive predictions made by a classification model.

**Precision:**

**F1-Score**

F1 score is a metric that combines both precision and recall into a single value, providing a balanced measure of a model's performance, particularly in situations where class imbalance exists.

F1- Score:

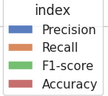
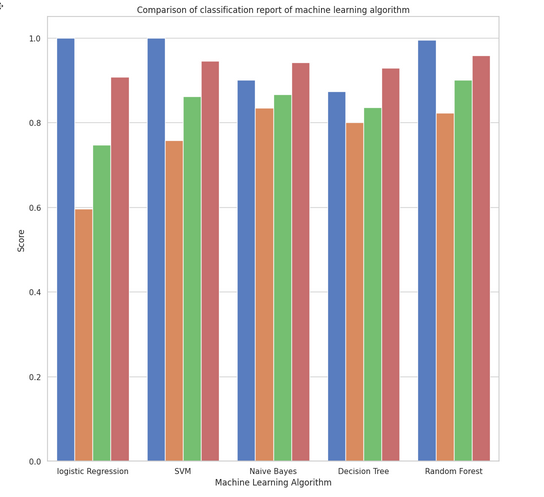
**Experiment result and Analysis**

In this section, the result obtained after testing all the machine learning algorithms is presented. The model is trained with different machine learning algorithms like Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, Random Forest and K mean clustering. We have collected data for testing our model and used four different parameters (accuracy, recall, precision, f1-score) for comparative analysis.

In this research work, we have modeled our algorithms based on various characteristics of emails in which we have used sender reputation, phishing websites etc. Comparing two models, one with additional features of email and another without additional features of email, we have found that using additional features of email like sender reputation, special symbol, phishing website etc increases the accuracy of the model.

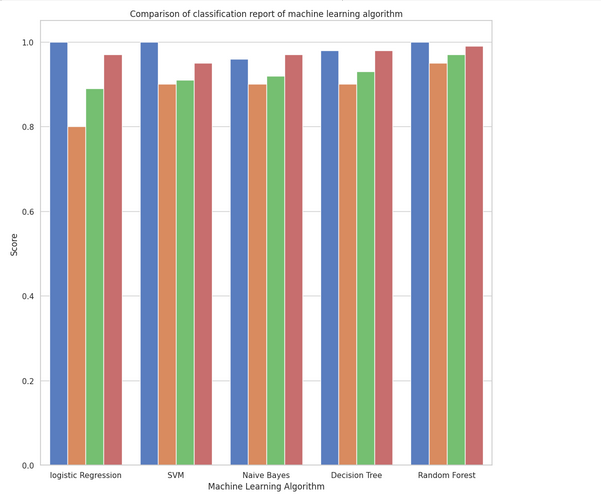
We got highest of 0.96 accuracy from random forest algorithm whereas lowset of 0.91 from logistic regression model where we have not used additional features. The tabulation form of that research is given below:

|  | | Precision | Recall | F1-Measure | Accuracy |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | Ham | 0.89 | 1.00 | 0.94 | 0.91 |
| Spam | 1.00 | 0.60 | 0.75 |
| SVM | Ham | 0.93 | 1.00 | 0.97 | 0.94 |
| Spam | 1.00 | 0.76 | 0.86 |  |
| Decision Tree | Ham | 0.95 | 0.97 | 0.96 | 0.94 |
|  | Spam | 0.91 | 0.82 | 0.86 |  |
| Random Forest | Ham | 0.95 | 1.00 | 0.97 | 0.96 |
|  | spam | 1.00 | 0.82 | 0.90 |  |
| Naive Bayes | Ham | 0.95 | 0.97 | 0.97 | 0.94 |
|  | Spam | 0.90 | 0.83 | 0.87 |  |

The graphical representation of above tabulated data is also given below:

In the experiment in which additional features of email is used, we got highest accuracy of 0.99 from random forest algorithm whereas lowest accuracy of 0.95 from SVM. The detailed value of accuracy of each machine learning algorithm is tabulated below:

|  | | Precision | Recall | F1-Measure | Accuracy |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | Ham | 1.00 | 0.9 | 0.94 | 0.97 |
| Spam | 1.00 | 0.7 | 0.82 |
| SVM | Ham | 1.00 | 1.00 | 1.00 | 0.95 |
| Spam | 1.00 | 0.8 | 0.88 |  |
| Decision Tree | Ham | 0.98 | 0.95 | 0.96 | 0.98 |
|  | Spam | 0.98 | 0.85 | 0.91 |  |
| Random Forest | Ham | 1.00 | 0.95 | 0.97 | 0.99 |
|  | spam | 1.00 | 0.95 | 0.97 |  |
| Naive Bayes | Ham | 0.97 | 0.9 | 0.93 | 0.97 |
|  | Spam | 0.95 | 0.89 | 0.91 |  |



This experiment suggest that the random forest model is better than other remaining algorithms.

**Conclusion**

Email has been the most popular electronic way of communication with people. Moreover, billions of data are transferred through emails daily. Such a vast increasing population of email users has increased the number of spam emails. Spam emails are those emails which contain malware in the form of files,unsolicited messages and many unethical works. This problem has caused email users to gain wrong information and billions of memory are filled with spam emails only.Spam emails generated from various websites or botnets can be detected using the different features of coming emails at every gateway using machine learning algorithms. In this research work we have compared two different approaches for modeling. On one hand, we are using additional features of emails, on another we are using only text information of email. Comparing both of these, we got a better result in the first one. Overall, machine learning algorithm will run with better efficient when proper parameter are included in data set.References

**Refrences**

[1] Sultana, Thashina, et al. "Email based Spam Detection." *International Journal of Engineering Research & Technology (IJERT)* (2020).

[2] Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020.

[3] Bhowmick, Alexy, and Shyamanta M. Hazarika. "E-mail spam filtering: a review of techniques and trends." *Advances in Electronics, Communication and Computing: ETAEERE-2016* (2018): 583-590.

[4] Shafi’i, Muhammad Abdulhamid, et al. "Comparative analysis of classification algorithms for email spam detection." (2018).