www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May-2022, pp : 44-48

e-ISSN : 2583-1062

Impact Factor : 2.205

MACHINE LEARNING FOR EMAIL SPAM MESSAGES DETECTION

Hardik N Patel[1], Shilpa Serasiya[2]

[1]PG Student Computer Engineering, KITRC, Kalol, Gujarat, India,

[2]Head Of the Department, Computer Engineering, KITRC, Kalol, Gujarat, India

ABSTRACT

In recent years, email spam has become a major problem with a big economic impact on society. An overgrowing increase in popularity the number of unsolicited data has also increased rapidly. Email spam known as unwanted email messages Other related forms of spam are increasingly appearing as a problem of importance.

To filtering data different approaches, exist which automatically detect and remove these untenable messages. There are different methods to filter your data that automatically detect and delete these untenable messages We present a popular machine learning based email spam filtering approaches.

In our work we machine learning technique to email spam messages filter. Therefore, it is necessary to identify these spam e-mail which are fraud, this project can identify spam using machine learning techniques, this paper will discuss the machine learning technique and apply Artificial Neural Networks classifier, Recurrent Neural Network, Logistic Regression, SVM, Unsupervised learning algorithm on our data sets and best algorithm is selected for the email spam messages detection having best precision and accuracy.

Keywords: - Deep Learning, Artificial Neural Networks classifier, Logistic Regression, Unsupervised learning, SVM

1. INRODUCTION

This year email spam increases to unwanted email messages in all device. Email spam creates unnecessary behaviour that like hacking unsolicited data and fraud.

This rapid growth of spam mails has led to issues like over-flow of user's mailboxes, consumption of It takes a lot of time for the user to clear and sort these mails as spam and these issues have increased the need for efficient and effective email filters that filter emails into spam. Or ham spam filters prevent spam emails from getting into a user's inbox.

Machine learning is a study of computer algorithms that can automatically improve through experience and data using machine learning algorithms to build models based on the sample data provided.

It is also known as training data. The use of training data is explicitly programmed to make predictions and make decisions. Spam and ham emails are trained to classify spam filtering models based on the system. further research needs to be done to increase the effectiveness of spam filters.[11]

This makes the system intelligent enough to classify spam emails on the datasets described in this. The success ratio of these machine learning algorithms varies. Machine learning techniques apply Artificial Neural Networks and Unsupervised learning algorithm, SVM on our data sets and best algorithm is selected for the email spam messages detection having best precision and accuracy.

1. Artificial Neural Networks classifier

The Artificial Neural Network (ANN), also called the "Neural Network" (NN), is a computational model based on biological neural networks. It consists of an interconnected collection of artificial neurons. An artificial neural network is an adaptive system that changes its structure based on information that flows through the artificial network during the learning period. ANN is based on the principle of learning by example.

There are two types of training for neural networks.

i. **Supervised: -** Here the network is equipped with a set of input and corresponding output patterns, known as training data sets, for network training.

ii. **Unsupervised: -** In this case, the network is trained by creating pattern groups. The system does not provide any previous set of training data.

2. Recurrent Neural Network

A recurrent neural network (RNN) is an ideal result to overcome the progressive problem of learning traditional neural networks. It has a unique character to store information when reading the input sequence at each step so that the "" status "' Assignment is an important property of RNN, which it generalizes services to the model on the input arrangement of different matches. The basic structure of RNN is the same as for feed forward neural network with differential operator contacts between neurons. As a one-way replacement interconnection, where data flows into layer neurons in another, neurons may have controlled phases on the net. They have their own loops or links.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May-2022, pp : 44-48

e-ISSN : 2583-1062
Impact Factor : 2.205

3. Logistic Regression

Logistic regression is a classification algorithm used to determine the probability of success and failure of an event. Used when the dependent variable is binary (0/1, True / False, Yes / No). It supports the categorization of data into discrete classes by studying the relationship of a given set of highlighted data. They learn the linear relationship of a given data set and then identify the nonlinearity in the form of a sigmoid function.

Logistic regression is also known as binomial logistic regression. It is based on a sigmoid function, where the output is random and the input can be from -infinity to + infinity.

4. Support Vector Machine

Support Vector Machine as SVM is one of the best-known algorithms for Supervised Learning, used for classification as regression problems. However, it is primarily used for machine learning classification problems.

The goal of the SVM algorithm is to make the best decision about a line or boundary that can segregate n-dimensional space into classes so that we can easily place a new data point in the correct category in the future. This optimal decision limit is called superficial.

SVM selects extreme points / vectors that help create the hyperplane. These extreme cases are called support vectors, and therefore this algorithm is called a support vector machine. Consider the diagram below, where two different categories are classified with a decision boundary or hyperplane:

SVM can be of two types

1 Linear SVM: - Linear SVM is used for linearly separable data, which means that if a data set can be classified into two classes using a straight line, then such data is termed linearly separable data. and the classifier used is called the SVM linear classification.

2 Nonlinear SVM: - Nonlinear SVM is used for nonlinearly separated data, which means that if a data set cannot be classified as a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

2. METHODOLOGY

a. Dataset

Email spam detection dataset https://drive.google.com/drive/folders/1tzFtW4qGA3nyYErD-zjvmSppTikIYyEy?usp=sharing the name of the dataset “spam,csv” .This data set contains 5572 rows and 2 columns.

Here is a project based on machine learning e-mail spam detection, we use various machine learning algorithms to detect spam e-mails. Based on feature extraction, we use various studies to find out which method is more suitable for feature extraction and to find a more accurate result compared to other methods.

1.Data Cleaning

2.Detection of spam/ ham emails

A. Artificial Neural Networks classifier

B. support vector machine

C. Logistic Regression

3. Feature Extraction

A. spam emails

B. Ham emails

B: - Support Vector Machine

The SVM model is a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner way in SVM in order to minimize the error. The goal of SVM is to divide the datasets into classes in to find the maximum marginal hyperplane (MMH).

her followings are important concepts in SVM -

1. Support vectors: - Data points that are closest to the hyperplane are called support vectors The separating line will be defined using these data points.

2. Hyperplane: - As we can see in the above diagram, it is a decision plane or space divided between a set of objects with different classes.

3. Margin: - it can be defined as a space between two lines in a data point box of different classes. It can be calculated as the perpendicular distance from the line to the supported vectors. A large margin is considered a good margin and a small margin is considered as a bad margin.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May-2022, pp : 44-48

e-ISSN : 2583-1062
Impact Factor : 2.205

SVM Kernels

In simple words kernel converts non-separable problems into separate problems by adding more dimensions to it. it makes SVM more powerful, flexible and accurate.

1. **Linear kernel: -** that the product between two vectors says x & xi is the sum of the multiplication of each pair of input values.

   K(x,xi)=sum(x∗xi)

2. **Polynomial kernel: -** This is a more general form of a linear kernel and distinguish curved a curved and a nonlinear input space.

   k(X,Xi)=1+sum(X∗Xi)^d

   Here d is the degree of the polynomial that we need to specify manually in the learning algorithm.

3. **Radial Basis Function (RBF) Kernel: -** Here's gamma ranges from 0 to 1. We need to manually in the learning algorithm. The ideal default gamma value is 0.1.

   K(x,xi)=exp(−gamma∗sum(x−xi^2)),

3. MODELING AND ANALYSIS

A. PROPOSED SYSTEM

The pre-processing step is used to spam email detection check out a spam and hum.

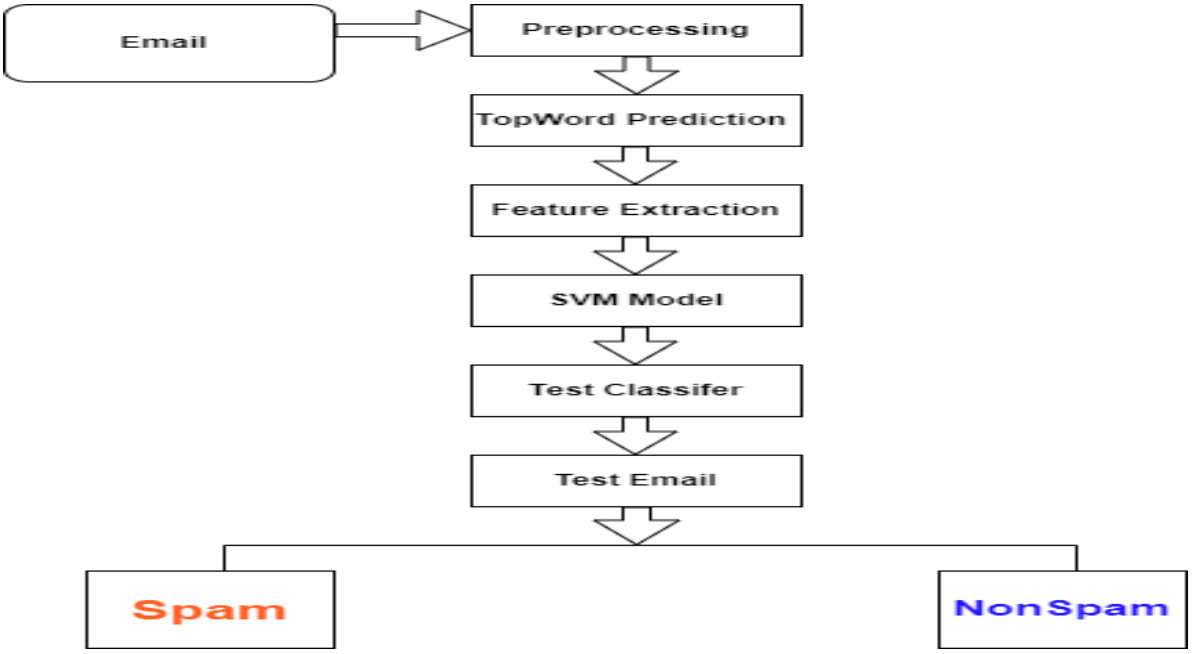

Figure 1: - Flow of process

These feature extraction/encoding approaches convert the words/text of spam emails to a numeric vector that can be used for classification

Supporting vector machines are used for classification as well as for regression problems, where datasets are used to train SVM to classify any new data it receives. A controlled algorithm that teaches the machine to work.

This feature is used for e-mail spam training in this area. The training dataset contains spam content and the classification is trained to use it. After training, the classifier allows them to classify spam emails.

To test the accuracy of the test classifier, the classifier is tested. A set of test data is combined with several training data. In this process the classification of emails with good accuracy of dataset is achieved by the proposed solution.

After complete the training phase, a new sample email will be provided as a classification input for the email classification. Classification produces output as 0 or 1, 1 means it is spam and 0 means it is not spam

B. PROPOSED WORK

In recent days spam email is increasing day by day and it is causing problem to user by spam [1] so we are going to stop spam by using SVM. Different techniques have been examined to email spam filtering have used different algorithms which are based on machine learning Algorithm. But some algorithms give better results to email spam filtering and some algorithms decreases which do not give better results. SVM achieved 94.06% testing accuracy [10] So, in this field its required that algorithms can give better results and improved in such a way email spam filtering.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May-2022, pp : 44-48

e-ISSN : 2583-1062

Impact Factor : 2.205

C. IMPLEMENTATION

A Jupyter Notebook used for the Implementation

a. Logistic Regression

The following steps were performed for logistic regression. In this, the data is loaded into the data frame using pandas. Pandas methods info () and describe were used to display statistics and general information about dataset. Data pre-processing was performed as described above with library form sklearn. The data were split into training and tests as described above. The Logistics Regression module is imported from sklearn.linear_models. The training data was fit to the classifier and then a prediction is made in the testing set. Performance: This model provides an accuracy of 98.9% in invisible data.

b. Support Vector Machine

Vector Machine Support is imported first. The SVM classifier is fitted to the training set. The prediction is performed on the SVM test set. The accuracy in the first prediction is just over 98.9% percent, which is logistic Regression, but only marginally after that Parameter tuning was done on it, after which the accuracy is slightly improved and 99.0 % above. SVM Linear kernel and RBF Kernel accuracy better than Poly kernel.

4. RESULTS AND DISCUSSION

a. RESULT

SVM is used to detect e-mail spam. The final SVM parameters for machine learning have been selected because they are more than 99.0% provides the best percentage accuracy. This was achieved by tuning each parameter and using a grid search for each parameter.

However, when a new model is created, its accuracy is a. Logistic reflection is only slightly better than the accuracy it provides. SVM accuracy is better than other models.

Table 1: - Accuracy Results

Model	**Accuracy**
Logistic Regression	98.9%
SVM	99.0%
SVM Linear kernel	99.5%
SVM Polynomial kernel	94.6%
SVM RDF kernel	99.0%

b. ANALYSIS

SVM algorithm is used to check ham and spam emails. We see 86.59% ham and 13.41% spam. With SVM it is easy for us to detect email spam and SVM gives a good result. We have used Logistic Regression and SVM to stop spam emails. SVM has given better results in email spam detection. After that we used linear kernel of SVM to detect email spam and linear kernel gave us 99.5% accuracy.

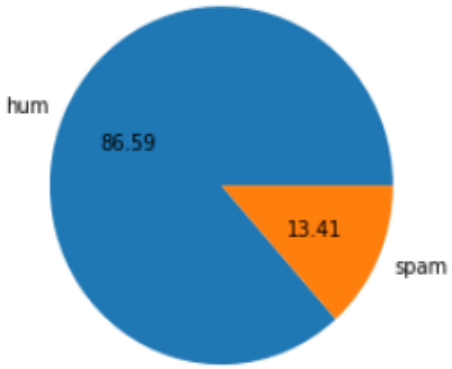

Figure 2: - ham/spam

Through the graph we can see that a total of 976 times the mail was ham (0) and the model predicts it well and 134 times it is spam and the model predicts spam (1), so in total - and we have a created a good model. however, you can experiment with the parameters, layers, and network architecture to increase it.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May-2022, pp : 44-48

e-ISSN : 2583-1062

Impact Factor : 2.205

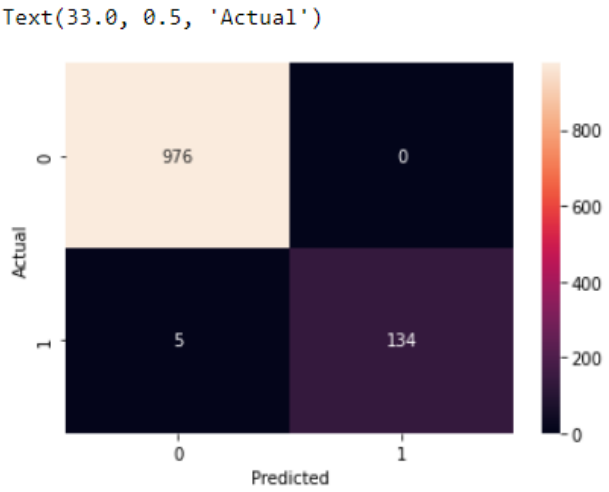

Figure 3: - ham/spam

5. CONCLUSION

We have used machine learning algorithm to detect spam email. We have used these two best techniques for email spam detection, SVM and Logistic Regression. By using this SVM, we have got Email spam accuracy 99.0% and Logistic Regression accuracy 98.9%. After that we used linear kernel of SVM to detect email spam and linear kernel gave us 99.5% accuracy. SVM better than result provide to detect ham/spam emails. Future Work SVM takes less time to detect spam emails and delivers better results.

6. REFERENCES

[1]. Nikhil Govil, Kunal Agarwal, Ashi Bansal, Astha Varshney "A Machine Learning based Spam Detection Mechanism" Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020) IEEE

[2]. Nikhil kumar, sanket sonowal, nishant "email spam detection using machine learning algorithms" Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE

[3]. Simran Gibson, Biju Issac, Li Zhang, Seibu Mary Jacob "Detecting Spam email with machine Learning Optimized with bio-inspired metaheuristic algorithms" IEEE VOLUME 8, 2020

[4]. Ganiev Salim Karimovich, Khamidov Sherzod Jaloddin ugli, Olimov Iskandar Salimbayevich "Analysis of machine leaning method for filtering spam messages in email services " 2020 International Conference on Information Science and Communications Technologies (ICISCT) | IEEE nov 2020

[5]. Mansoor RAZA and Nathali Dilshani Jayasinghe, Muhana Magboul Ali Muslam "A Comprehensive Review on email spam classification using Machine Learning Algorithms" 2021 International Conference on Information Networking (ICOIN) IEEE 02 February 2021

[6]. Nandhini.S, DR.Jeen Marseline.K.S "Performance Evaluation of machine algorithms for email spam detection" 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) IEEE 2020

[7]. Mahammad Abdullahi, Abdulmalik D. Mohammed, Opeyemi O. Abisoye "A review on Machine Learning Techniques for images-based spam emails detection" Proceedings of the 2020 IEEE 2nd International Conference on Cyberspace (Cyber Nigeria) IEEE 2020

[8]. Priya.S, Annie Uthra.R "An Effective Concept Drift Detection Technique with Kernel Extreme Learning Machine for Email Spam Filtering" Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020] IEEE 2020

[9]. Fahima Hossain, Mohammed Nasir Uddin, Rajib Kumar Halder "Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection" International IOT, Electronics and Mechatronics Conference (IEMTRONICS) IEEE 2021

[10]. Sunday Olusanya Olatunji "Extreme Learning Machines and Support Vector Machines Models for Email spam detection" Canadian Conference on Electrical and Computer Engineering (CCECE) IEEE 2017

[11]. Shivam Pandey, Ashish Taralekar, Ruchi Yadav, Shreyas Deshmukh and Prof. Shubhangi Suryavanshi "E-mail Spam Detection and Classification using SVM" International Journal of Computer Science and Information Technologies, Vol. 11 (1), 2020, 6-8