Email Spam Detection Using Machine Learning

Prof. Prachi Nilekar, Tamboli Abdul Salam, Manish Kumar Gupta, Krishna Sharma, Safwan Attar

ALARD COLLEGE OF ENGINEERING & MANAGEMENT
(ALARD Knowledge Park, Survey No. 50, Marunje, Near Rajiv Gandhi IT Park, Hinjewadi, Pune-411057)
Approved by AICTE. Recognized by DTE. NAAC Accredited. Affiliated to SPPU (Pune University).

Abstract – *Nowadays, Email spam has become a big problem, with the fast growth of internet users, email spams are also increasing. People are using them for phishing, illegal and unethical practices and frauds. Sending malicious links through spam emails that can harm for our system and may also get into your system. It is very simple for spammers to create a fake profile and email account, they show like a real person in their spam emails, these spammers simply target people who are not aware of these frauds. then there is a need to identify those spam mails which are frauds, this project will identifies those spams using techniques of machine learning, this paper will discuss machine learning algorithm's and apply all these algorithm's to our dataset. it select the best algorithm, for this project algorithm will be chosen based on the best accuracy and precision in email spam detecting.*

Key Words*: *(Machine Learning, Naive Bayes, Support Vector Machine, DTS, Random Forest, Bagging, Boosting)

1. INTRODUCTION

Machine learning approaches are more efficient, a set of training data is used, these samples are the set of email which are pre classified. Machine learning approaches have a lot of algorithms that can be used for email filtering, these algorithms are "Naive Bayes, support vector machines, Neural Networks, K-nearest neighbor, Random Forests, etc."

Why Machine Learning: Machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data.

What is DATASET: Dataset is a collection of data or related information that is composed for separate elements. A collection of datasets for e-mail spam contains spam and non-spam messages.

What is Train and Test datasets: The main difference between training data and test data is that training data is the subset of original data that is used to train a machine learning model, whereas test data is used to check the accuracy of the model. The training dataset is usually larger in size than the test dataset. Train and test dataset are two key concepts in machine learning, where the training dataset is used to fit the model, and the test dataset is used to evaluate the model.

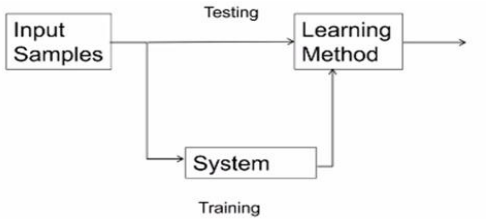

Fig -1: Train and Test Model

Machine learning algorithms used to classify the text into two different categories, spam and ham. The algorithm will predict the score more accurately. The objective of developing this model is to detect and score word faster and accurately.

2. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Naive Bayes: Naive Bayes is a classification algorithm suitable for both binary and multiclass classification. Naive Bayes performs better for categorical input variables than for numerical variables. It is useful for making predictions based on historical results and forecast data.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

P(A) is Prior Probability: The possibility of a hypothesis before seeing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Support Vector Machine: SVMs are used in intrusion detection, face detection, email classification, gene classification, web pages, etc. It can handle classification and regression on linear and non-linear data.

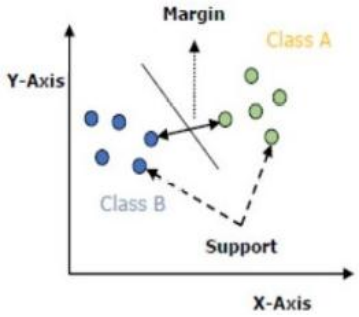

Fig -2: Support Vector Machine

Decision tree: Decision trees are extremely useful for data analytics and machine learning because they break down complex data into more manageable chunks. They are often used in these fields for predictive analysis, data classification, and regression.

Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

KNN: The KNN algorithm can compete with highly accurate models because it makes highly accurate predictions. The KNN algorithm use for applications that require high accuracy but do not require a human readable model. The quality of the predictions is depends on the distance measurement. Formula:

$$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

Random forest classifiers: Random forest classifiers can be used to solve regression or classification problems. The random forest algorithm is composed of a collection of decision trees, and each tree in the ensemble consists of data samples drawn from the training set with replacement, called bootstrap samples.

3. OBJECTIVES OF THE STUDY

Machine learning algorithms used to classify the text into two different categories, spam and ham, the algorithm will predict the score more accurately. The purpose of developing this model is to recognize and score the word rapidly and accurately.

4. SCOPE OF THE STUDY

The proposed system of the project will effectively detect spam mails and the system will extract spam mails using some machine learning algorithms and it gives results with more accuracy and good performance. This project required a coordinated scope of work. These project scopes will help focus the project. The scopes are:

- Modifying existing machine learning algorithm.
- Use and classify data sets, including data preparation, classification, and visualization.
- Score the data to determine the accuracy of spam detection.
- This proposed system will detect the credibility of the mail and it will filter spam messages.
- This proposed system will save the time of the user and it will eliminate the risk of spam mails.

Use case diagrams describe the high-level functions and scope of the system, these diagrams also identify the interactions between the system and its actors. A Use case diagram outlines how external entities user interact with an internal software system.

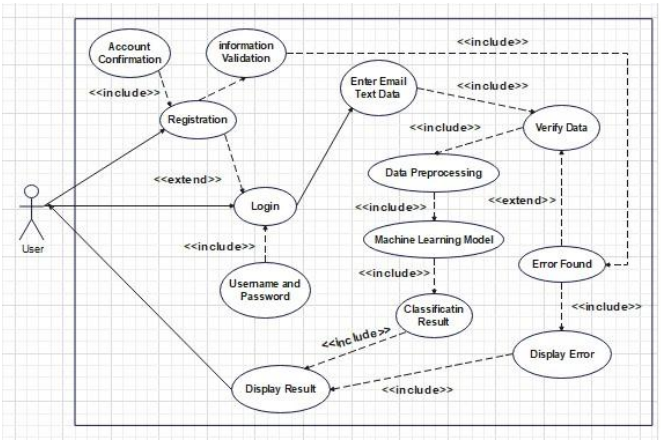

Fig -3: Use Case Diagram

A state diagram consists of states, transitions, activities, and events. It describes the different states that an object moves through or provide an abstract description of the behavior of a system.

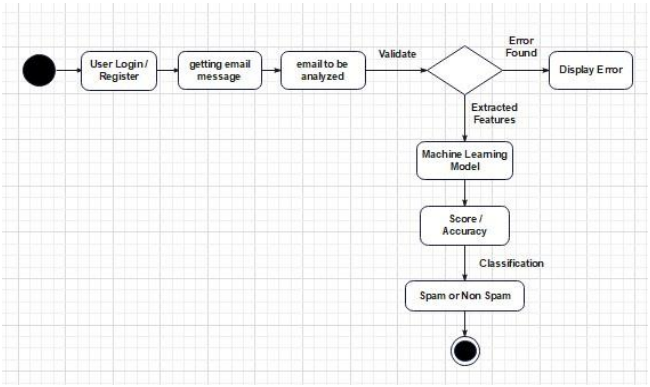

Fig -4: State Diagram

Activity diagrams are graphical representations of workflows with support for selection, repetition, and concurrency of step-by-step activities and tasks.

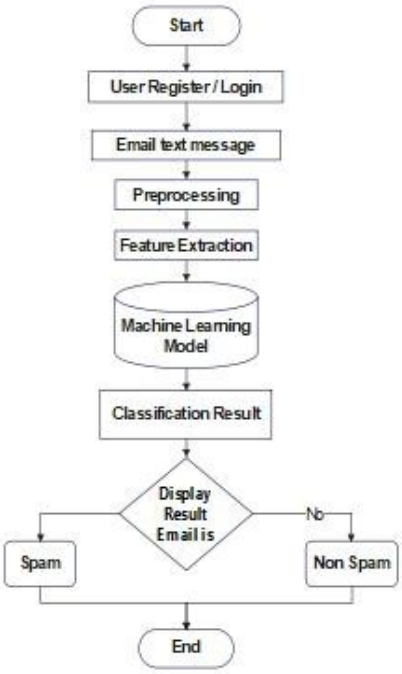

Fig -5: Activity Diagram

5. PROJECT ARCHITECRURE DIAGRAM

An architectural diagram is a visual representation that shows the physical implementation of the components of a software system. It shows the general structure of the software system and the associations, boundaries and limits between each element.

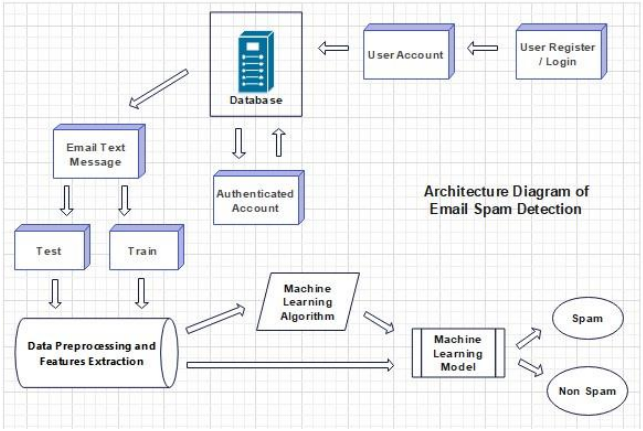

Fig -6: Architecture Diagram of Email Spam Detection

6. CONCLUSIONS

This system, in addition to lessening the work load, it also fixes any false data about the users that they may have. It is a benefit for the users' who's important time and data is preserved, for the Affected users or authority whose data is immensely important, whose data will secured from misuse.

We are able to classify email as spam or non spam. With huge number of emails if people are using the system it will be difficult to handle all the possible mails as our project deals with only limited amount.

The website use for end user, it is user friendliness. Because of the end user it uses without any other help and without any conflicts. The website goal is "Email Spam or Non Spam" using machine learning, related to its use for free and maintenance (coding, updates, uploading data, datasets, etc) cost is less. The many goals was successfully completed and achieved by us.

ACKNOWLEDGEMENT

This paper was supported by Alard College of Engineering & Management, Pune 411057. We are very thankful to all those who have provided us valuable guidance towards the completion of this Seminar Report on "Email Spam Detection Using Machine Learning" as part of the syllabus of our course. We express our sincere gratitude towards the cooperative department who has provided us with valuable assistance and requirements for the system development. We are very grateful and Prof. Prachi Nikelar for guiding us in the right manner, correcting our doubts by giving us their time whenever we required, and providing their knowledge and experience in making this project.

REFERENCES

[1] A Sharaff and Srinivasarao U (2020), "Towards classification of email through selection of informative features," First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, pp. 316-320, DOI: 10.1109/ICPC2T48082.2020.9071488.

[2] Adebayo A. Alli, Modupe Odusami, Olusola A. Alli and Sanjay Misra (2019), A review of soft techniques for SMS classification: methods, approaches and applications, Engineering Applications of Artificial Intelligence, DOI: 10.1016/j.engappai.2019.08.024.

[3] A. Sharma & H. Kaur, Improved email spam classification method using integrated particle swarm optimization and decision tree. In Next Generation Computing Technologies 2nd International Conference on pp. 516-521, DOI: 10.1109/NGCT.2016.7877470.

[4] A. Sharaff, A. Dhadse and Naresh K. Nagwani (2016), Comparative study of classification algorithms for spam email detection, in Emerging Research in Computing, communication and applications, Information, pp. 237-244, Springer, Berlin, Germany, DOI: 10.1007/978-81-322-2553-9_23.

[5] Alfandi O., Dahmani N. and Kaddoura S., "A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach", IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, France, Bayonne, pp. 193-198, DOI: 10.1109/WETICE49692.2020.00045.

[6] Amin, Hossain N. & Rahman M. M., "A Bangla Spam Email Detection and Datasets Creation Approach based on Machine Learning Algorithms," 2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering, Bangladesh, Rajshahi, 2019, pp. 169-172, DOI: 10.1109/ICECTE48615.2019.9303525.