iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46654

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Email Spam Detection using Naïve Bayes Algorithm

G. Revathi[1], K. Nageswara Rao[2], G. Sita Ratnam[3]
[1, 2]*Computer Science and System Engineering, Andhra University*
[3]*Computer science and System Engineering, Lendi Institute of Engineering and Technology*

Abstract: Email Spam has become a vital issue currently, with high-speed growth of internet users. Some people are using them for illegal conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and may also they will seek into our system. The need of email spam detection is to prevent spam messages from lagging into user's inbox so it'll improve user experience. This project will identify those spam emails by using machine learning approach. Machine learning is one amongst the applications of Artificial Intelligence that allow systems to read and improve from experience without being specific programmed. This paper will discuss the machine learning algorithm which is Naïve Bayes. It is a probabilistic classifier, which means it predicts on the idea of the probability of an object and it is selected for the email spam detection having best precision and accuracy.
Keywords: Machine learning, Naïve Bayes, Tokenization, Spam Detection

I. INTRODUCTION

Email spam refers to the using of email to send unsolicited emails or advertising emails to bunch of recipients. Unsolicited emails mean the receiver has not granted permission for receiving those emails. Spam has become an enormous misfortune on the web. Spam is a waste of storage, and message speed. Automatic email filtering could also be the most effective method of detecting spam mails but nowadays spammers can easily block all these spam filtering applications. Several years ago, most of the spam will be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection so Naive Bayes is one of the algorithms applied in these procedures. Naïve Bayes algorithm is supervised learning algorithm and it's used for solving classification problems which help in building the fast machine learning models that will make quick predictions.

Spam and Ham: Spam means the context of email and the use of electronic communication systems to send unsolicited bulk messages, especially advertisement; malicious links are called as spam. So, if you not known about the sender the mail can be spam. Users generally don't realize they simply signed certain those mailers after they download any free services, software or while updating the software. "Ham" is e-mail that's not Spam.

Machine learning approaches are more efficient and it focuses on developing computer programs and algorithms which will access data. So set of training data is used, and these samples are group of emails which are pre classified. Machine learning approaches have a lot of algorithms which will be used for email filtering. In this paper, the Naïve Bayes algorithm is used for detecting spam emails and it produce best accuracy.

II. LITERATURE REVIEW

There is some related work that applies machine learning methods in email spam detection.

1) They describe a focused literature survey of Artificial Intelligence Revised (AI) and Machine learning methods for email spam detection.
2) They have used the "image and textual dataset for the e-mail spam detection with the employment of various methods.
3) They have used methods of KNN algorithm, Reverse DBSCAN algorithm with experimentation on dataset. For the text recognition, OCR library" is employed but this OCR doesn't perform well.
4) They used the feature selection hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and rough math's.

III. METHODOLOGY

The methodology is used for the method of e-mail spam filtering based on Naïve Bayes algorithm.

A. Data Preprocessing

Data Preprocessing is a strategy that is used to transform the raw information into a clean data set. In other words, whenever the information is gathered from different sources it's collected in raw format which isn't feasible for the analysis. This involves the consecutive steps:

1) *Tokenization:* Tokenization is claimed to be dividing an outsized quantity of text into smaller chunks referred to as Tokens. These tokens are pretty useful to search out the patterns and that they are parted by whitespaces characters like line break, space or by punctuation characters.
2) *Dropping Values:* Dropping is the most common method to take care of the missed values. Those rows in the data set or the entire columns with missed values are dropped in order to avoid errors to occur in data analysis.
3) *Stop Words:* Stop words are English words which don't add much content to a sentence. They will safely be ignored without forgoing the meaning of the sentence.
4) *Bag of Words:* A bag-of-words is a representation of text that describes the occurrence of words within a document and it is used for extracting features from the documents.

This Algorithm contains the following steps:

a) Step 1: Consider a random email from the spam dataset for execution.
b) *Step 2:* The considered email is in basic form. To perform the feature extraction/selection and classification procedure, email is required to pre-process initially.
c) *Step 3:* Initially, tokenize the e-mail into individual keywords. Tokenization split each individual

- If the duplicate values are present within the dataset, then it'll drop the duplicate values
- Remove the stop words from the obtained tokens.
- Now we will convert the group of text into a matrix of token counts
- Splitting the dataset into training data and test data.

d) *Step 4:* By evaluating the model on the training and testing dataset it predicts the accuracy of the model.

B. Naïve Bayes Classifier

Naïve Bayes is one of the algorithms in machine learning which implies it predicts on the basis of probability of an object. It is mainly used in text classification. It can be used for classifying spam emails as word probability plays main role here. If there's any word which occurs frequently in spam but not in ham, then that email is spam. This algorithm has become a best technique for spam detection. The Naive Bayes calculates the probability of each class and maximum probability is then chosen as an output. Naïve Bayes always provide an accurate result. The Formula for Naïve Bayes algorithm is represented as follow

$P(A|B) = P(B|A) * P(A) / P(B)$

IV. IMPLEMENTATION

The platform visual studio code is used to implement this model and for this module, a dataset from "Kaggle" website is applied as a training dataset. The inserted dataset is first checked for duplicates and null values for better execution of the machine. Then the dataset is split into two datasets which is training dataset and test dataset within the portion of 80 percent and 20 percent. These datasets passed parameters for text processing. During this text process, stop words and punctuation symbols are removed and returned as clean words. Evaluate the model of training and test dataset to obtain the confusion matrix. For that we've to calculate the precision, recall, F1-score. Precision and recall help to calculate positive samples in the model. F1-score combines weighted average of precision and recall. The Confusion matrix contains the four different combinations of predicted and actual values. Accuracy is depended on the percentage of correct predictions for the test data.

Flowchart

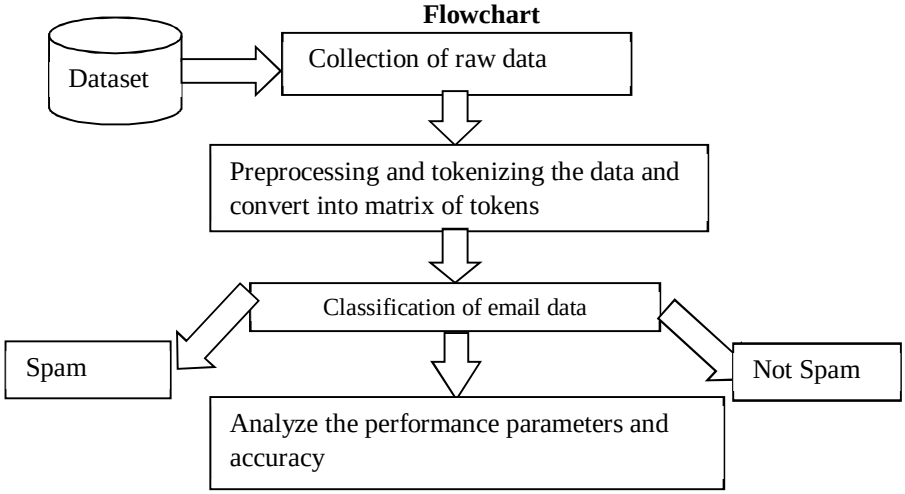


Fig1: Flowchart for Email spam detection

V. RESULT

In this project, the Naïve bayes model is used for the best accuracy and this classifier will give its estimate results to the user. The dataset is achieved from "Kaggle" website for training. The name of the dataset used is "spam.csv". The two datasets training and testing data is compared based on the percentage of correctly identified spam and non-spam. The approach of the confusion matrix is the number of occurrences of each class for the dataset being considered.

For FP, FN, TP and TN, the average of dataset as follows:

FP: Total 8 number of misclassified spam emails.

FN: Total 1 number of misclassified spam emails.

TP: Total 268 number of spam messages is correctly classified as spam.

TN: Total 862 number of non-spam e-mail that is correctly classified as non-spam.

The Accuracy that is defined by evaluating the model on the training and testing dataset is 99% and the result is shown in below figure.

	Precision	Recall	F1-Score	Support
0	1.00	0.99	0.99	870
1	0.97	1.00	0.98	269
Accuracy			0.99	1139
Macro Average	0.98	0.99	0.99	1139
Weighted Average	0.99	0.99	0.99	1139

Confusion Matrix:

[[862 8]

[1 268]]

Accuracy: 0.9920983318700615

Fig: Confusion matrix for dataset

Test dataset had 1139 text messages. Among 863 hams text messages of test dataset 862 were correctly classified as ham and remaining 1 is wrongly classified as spam. Among 276 spam messages of test data 268 were correctly classified as spam and 8 messages were wrongly classified as ham. Fig 1 shows the confusion matrix and accuracy results for test data.

VI. CONCLUSION

This project, spam detection is proficient of detecting mails giving to the content of the email. Detecting the spam emails can be done on the basis of the trusted and verified domain names. The spam email classification is incredibly significant in categorizing e-mails and distinct e-mails that are spam or non-spam. Naive Bayes could a baseline technique for managing with spam to the e-mail needs of individual users and provides low false positive spam detection rates that are generally acceptable to users. To further optimize the parameters of the Naïve Bayes approach is used, which results in increased the accuracy of the entire classification process. The accuracy of the spam detection can increase by using Naïve Bayes Classifier.

In future the other optimization algorithm can be used with Naïve Bayes algorithm. Also, the other ML approach can be used instead of NB approach. The evaluation of the experiment is done on the basis of f1-score, precision, accuracy and recall. By evaluating the results, we are able to say that the integrated concept ends up in increased accuracy and precision than the individual Naïve Bayes approach.

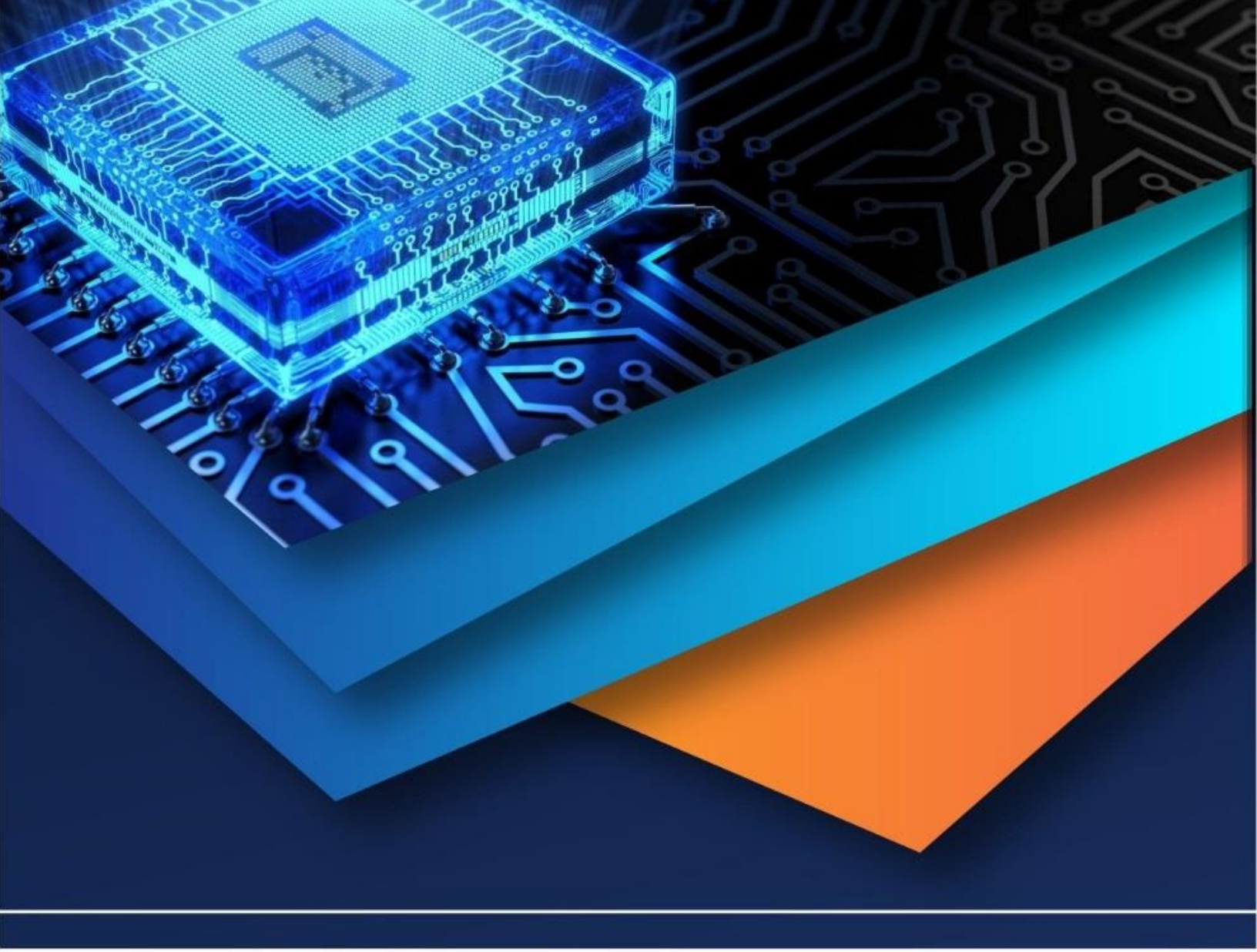
REFERENCES

[1] Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab. They describe a focused literature survey of Artificial Intelligence Revised (AI) and Machine learning methods for email spam detection.

[2] K. Agarwal and T. Kumar Harisinghaney et al. (2014) and Mohamad & Selamat (2015) have used the "image and textual dataset for e-mail spam detection with the utilization of assorted methods".

[3] Harisinghaney et al. (2014) have used methods of KNN algorithm, Reversed DBSCAN algorithm with experiments on dataset. For the text recognition, OCR library is employed but this OCR doesn't perform well.

[4] Mohamad & Selamat (2015) uses the feature selection hybrid approach of TF-IDF (Team Frequency Inverse Document Frequency) and Rough pure math.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY