ISSN No:-2456-2165

Analysis of Spam Email Filtering Through Naive Bayes Algorithm Across Different Datasets

Ritik Singh
Galgotia College of Engineering and Technology
Greater Noida, INDIA

Rahul Tyagi
Galgotia College of Engineering and Technology
Greater Noida, INDIA

Tushar Chaudhary
Galgotia College of Engineering and Technology
Greater Noida, INDIA

Abstract:- Today, if you are living a life in this world, email plays an important role in your life for communication. And due to its advantages like low cost, speed etc, a lot of people use it for advertisement of thier products . And that type of emails are called as spam email. And there are many method in market to differentiate spam email from ham email. Machine Learning algorithm used for some of the effective method.In this exploration, we will test Naive Bayes calculation for email spam classification based on two datasets and test its exhibition. First we are going to analyze by using a spam image dataset . And after that test the performance on the basis of SPAMBASE datasets.

Keywords:- *Component, Formatting, Style, Styling, Insert.*

I. INTRODUCTION

Emails are use consistently by numerous individuals for correspondence and for mingling. Recently spam messages, become a major difficulty over the web. Spam is waste of memory storage, time and correspondence data transmission. The issue of spam email has been expanding for years. Together with the headway of innovation and email, spam has filled significantly as of late. To distinguish approaching mail is spam or ham is a major issue for clients. Each client attempts to perceive spam mail and to erase it. However, the spam mail is coming in colossal sum, it contains notice of some business sites for buying their items, likewise contains extortion messages to open free saving record and apply for Mastercard and to certification data of client causes phishing assaults. There are many spam classification approaches and machine learning calculations to distinguish spam messages.

Spam Classification is the technique for separating spam mail from inbox and moved to the spam folder. For spam arrangement, different spam classification strategies are utilized. The capacity of a spam channel is to recognize spam email and keeps it from going to the letter box. A few AI calculations have been utilized in spam email classification, however Naive Bayes calculation is especially well known in business and open-source spam channels. The Naive Bayes algorithm is very easy to use, that is why it is most popular compare to other Machine Learning algorithm. This is its straightforwardness that make them simple to actualize and simply need short preparing time or quick assessment to channel email spam.

The way to deal with AI is by all accounts more fruitful than the way to deal with programming on the grounds that no rules are required. Then again, an assortment of testing models are an assortment of email messages pre-grouped. A specific calculation is then utilized to know from these email messages the grouping rules. AI methods have been concentrated broadly and numerous calculations can be utilized to handle messages.

The training to the classifier can be given by a past arrangement of spam and non-spam messages and by the arrangement of past spam pictures. So it monitors each word that happens just in spam, in non-spam messages, and in both.. We can use Naive Bayes in different datasets and that is what we are going to do in this project.

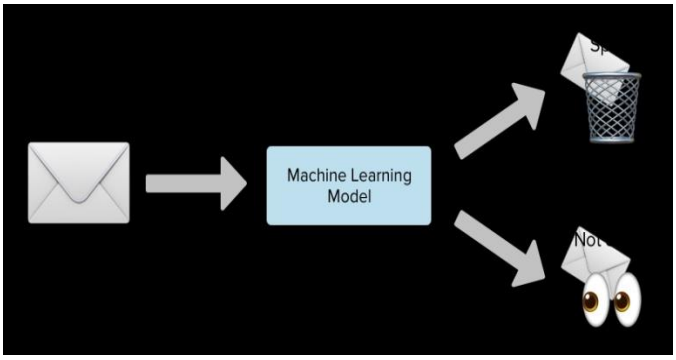

Fig 1. Basic working of spam email filter

II. RELATED WORK

Firstly we do a brief servay on previous massage filtering or machine algorithm. There is a quick expansion in the interest being appeared by the worldwide exploration local area on email spam classification. In this part, we present comparative surveys that have been introduced in the literature in this area. Spam classification has become a provoking zone to lead research as spammers are discovering approaches to change the data identifying with spam words through the expansion of intricacies. Different AI classifiers have been utilized in the examination to handle such issues like Boosting algorithm, Support Vector Machines (SVM) algorithm and Naive Bayes algorithm. Bayesian methods are very usefull in the field of word classification and its popularity in text and spam classification applications are increasing .

In [1] the author perceive the distinctive highlights of the material of reports. Their analysis gets the positives strategies just as certain deficiencies of current ML techniques and open spam channels study difficulties. They use Deep Neural Network, Support Vector Machine and Naive Bayes algorithm for comparison. And in [2] they do comparison between some famous Machine Learning algorithms like Support Vector Machine-nearest neighbor, Random Forest and Naive Bayes algorithm by using their own dataset and compare their accuracy.

III. METHODOLOGY

In this section we are going to discuss the methodology that is used in this research. For email classification, Machine Learning algorithms is used.

Machine Learning **:** Machine Learning field is a subfield from the expansive field of Artificial Intelligence, this expects to make machines ready to like human. Learning here methods comprehended, notice and address data about some factual wonder. In solo learning one attempts to reveal shrouded normalities (bunches) or to distinguish oddities in the information like spam messages or organization interruption. In email sifting task a few highlights could be the sack of words or the title examination. The system is utilized for the interaction of email spam sifting dependent on Naive Bayes calculation.

Naive Bayes algorithm **:** In the past days this method named naive bayes classifier was used . It was used around in 1998 for the purpose of classification of spam and hams. Naive Bayes classifier was used for spam acknowledgment. It's calculation is a calculation which is utilized for regulated learning. This classifier works upon the reliant occasions and deals with the likelihood of the occasion which will happen later on that can be distinguished from a similar occasion which happened already and we are going to use the this bayes theorem twice in our work. Here we taking bags of the most repeated words in any of the spam mail , for example as we all know the words like free are more likely to come in the spam mails so we will add that word in our bag. In the similar manner we are going to take the bag of the images . There also we will take only those images which have the good possibilities to be in a spam mail. Now firstly we will apply the Bayes theorem to the image part and if we find a good frequency of those spam images then we proved the mail spam mail and if the mail don't have any image or the images present in the mails doesn't matched to our spam mails bag . Then we move to our second step authentication where we will be applying the Bayes theorem on the text database. In this way we are applying our Bayes theorem and this Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. Conditional probability is the main concept here and this algorithm too mainly depends on it.

The formula for Bayes' theorem is given as:

$P(X/Y) = (P(X/Y)P(X))/P(X)$

$P(X/Y)$ = probability of occurrence of X when given Y is true

$P(Y/X)$ = probability of occurrence of Y when given X is true

$P(X)$ = the independent probability occurrence of X

$P(Y)$ = the independent probability of occurrence of Y

For getting the results for our research, Email Spam Classification is divided into three subtasks and that is :

1. Pre-processing
2. Tokenization
3. Feature Selection

1**.** Pre-processing **:** This is the initial step when a mail is sent. In AI, information pre-handling is main step. Information assortment strategies may prompt out-of-range esteems or incomplete values, etc. False outcomes can be produced if analyzing technique for input information isn't chosen appropriately. This stage requires tokenization.

2**.** Tokenization **:** Tokenization, when applied to information security, is the way toward subbing a delicate information component with a non-touchy same, alluded to as a token, that has no outward or exploitable importance or worth. The token is a reference that guides back to the touchy information through a tokenization framework.

When it is applied on email it is the way toward parting text into more modest pieces, called tokens. Every token is a contribution to the machine learning calculation. Tokenization is a utility capacity that tokenizes a text into tokens while keeping just the words that happen the most in the content corpus.

3**.** Feature Selection **:** Feature Selection is the interaction where you automatically or manually select those highlights which contribute most to your prediction variable or yield in which you are keen on. Having unimportant highlights in your information can diminish the precision of the models and cause your model to learn dependent on insignificant highlights.

And in other words we can say feature selection is the strategy of picking a subgroup of related highlights for development of learning models and it likewise lessens the dimensionality of highlight space.

ISSN No:-2456-2165

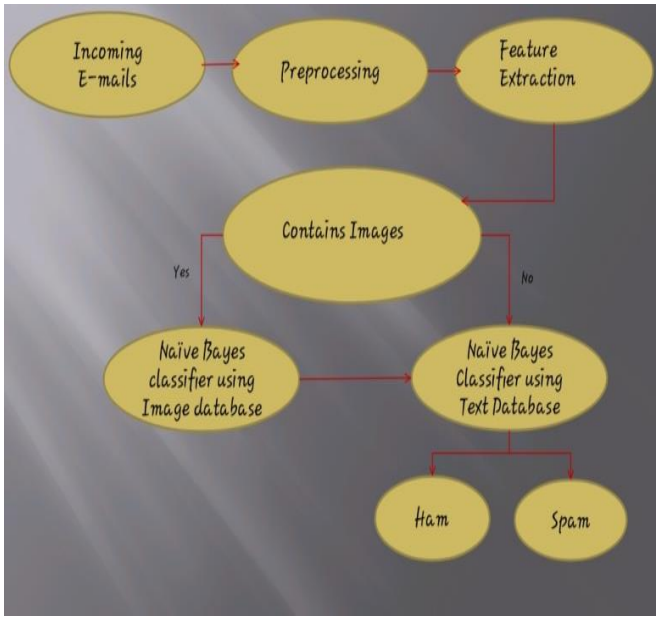

Fig 2. Proper analyzed flow chart for spam email filter.

IV. RESULT

In this section we are going to discuss about the experimental result using Naive Bayes algorithm. It is significant that spam sends don't arrive at the inbox of the clients as this diminishes effectiveness of tasks. Yet, more significantly it is essential that no ham mail goes to the spam organizer as those lead difficult issues to the users.

In this research we used two datasets. First we are using a dataset of different images that are taking from different spam emails in which spammers used different types of images. And if the mail don't have any images then it is not able to classify the spam email. After that we used the second dataset name SPAMBASE. In this dataset there are different words that are used in different spam emails. But if the mail contains spam images then the results is not good.

Fig 3. A basic output of spam email filter.

V. CONCLUSION

Email spam classification is a significant issue in the organization security and machine learning strategies; It is significant that spam sends don't arrive at the inbox of the clients as this lessens proficiency of tasks. And again, spam messages present even more a risk to the security of your own data. These massages or email look to take your own data for vindictive methods. When individual data, for example, account logins or charge card data is taken, your own records can be hacked or you can turn into a victim of financial fraud. Naive Bayes classifier that utilized has a vital part in this interaction of filtering email spam. In any case, more critically it is important that no ham mail goes to the spam envelope as those lead major issues to the client. With this outcome, it tends to be inferred that the Multinomial Naïve Bayes gives the best result yet has constraint because of class-restrictive autonomy which makes the machine to misclassify some tuples. , we took a gander at proficient strategies towards spam sifting of messages utilizing AI draws near. Likewise, the investigation of these improved techniques was completed to characterize messages as spam or ham. Nowadays, heaps of messages are sent and gotten and it is troublesome as our task is simply ready to test messages utilizing a restricted measure of corpus.

REFERENCES

[1]. Deepika Mallampati, Nagaratna P. Hegde, "A Machine Learning Based Email Spam Classification Framework Model: Related Challenges and Issues." International Journal of Innovative Technology and Exploring Engineering (IJITEE),ISSN: 2278-3075, Volume-9 Issue-4, February 2020

[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 Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2

[3]. SIMRAN GIBSON, BIJU ISSAC, LI ZHANG, SEIBU MARY JACOB, "DETECTING SPAM EMAIL WITH MACHINE LEARNING OPTIMIZED WITH BIO-INSPIRED METAHEURISTIC ALGORITHMS" Published in IEEE Access (Volume: 8)Page(s) 187914 - 187932 Electronic ISSN: 2169-3536

[4]. SHRAWAN KUMAR TRIVEDI, "A STUDY OF MACHINE LEARNING CLASSIFIERS FOR SPAM DETECTION." 2016 4th INTERNATIONAL SYMPOSIUM ON COMPUTATIONAL and BUSINESS INTELLIGENCE,IEEE INSPEC Accession Number 16466970.

[5]. ADITYA SHRIVASTAVA , Dr. RACHANA DUBEY, "CLASSIFICATION of SPAM MAIL USING DIFFERENT MACHINE LEARNING ALGORITHM" 2018 INTERNATIONAL CONFERENCE on ADVANCED COMPUTATION and TELECOMMUNICATION (ICACAT), INSPEC Accession Number 19257292

[6]. W. A. Awad, S. M. EL seuofi, "Machines Learning Methods For Spam E-mail Classifications" INTERNATIONAL JOURNAL of COMPUTER SCIENCE & INFORMATION TECHNOLOGY (IJCSIT), Vol 3, No 1, Feb 2011
[7]. JIA-NING LUO, MING HOUR YANG, "USING E-MAIL AUTHENTICATION and DISPOSABLE E-MAIL ADDRESSING for FILTERING SPAM" 2009 10th INTERNATIONAL SYMPOSIUM on PERVASIVE SYSTEMS, ALGORITHMS, and NETWORKS,INSPEC Accession Number 11085213
[8]. NURUL FITRIAH RUSLAND, NORFARADILLA WAHID, SHAHREEN KASIM, HANAYANTI HAFIT, "ANALYSIS of NAIVE BAYES ALGORITHM for EMAIL SPAM FILTERING ACROSS MULTIPLE DATASETS" IOP CONFERENCE SERIES: MATERIALS SCIENCE and ENGINEERING, VOLUME-226, INTERNATIONAL RESEARCH and INNOVATION SUMMIT (IRIS2017), 6-7 May 2017, MELAKA, MALAYSIA.