SPAM IDENTITY IN E-MAIL USING ML

Mrs. Palakollu Divya [1], Mr. Vemula Pranay [2], Dr. Pulime Satyanarayana [3]
Department of Computer Science Engineering, Samskruti College of Engineering and Technology
.

Abstract**:** Spam emails are known as unrequested commercialized emails or deceptive emails sent to a specific person or a company Spams can be detected through natural language processing and machine learning methodologies. Machine learning methods are commonly used in spam filtering. These methods are used to render spam classifying emails to either ham (valid messages) or spam (unwanted messages) with the use of Machine Learning classifiers. The proposed work showcases differentiating features of the content of documents There has been a lot of work that has been performed in the area of spam filtering which is limited to some domains. Research on spam email detection either focuses on natural language processing methodologies on single machine learning algorithms or one natural language processing technique on multiple machine learning algorithms In this Project, a modeling pipeline is developed to review the machine learning methodologies.

Keyword: Email Spam Detection, Spam Detection, Machine Learning, Neural Networks, Naive Bayes, Support Vector Classifier, Logistic Regression, Spam, Social Media.

I. INTRODUCTION

Technology has become a vital part of life in today’s time. With each passing day, the use of the internet increases exponentially, and with it, the use of email for the purpose of exchanging information and communicating has also increased, it has become second nature to most people. While e-mails are necessary for everyone, they also come with unnecessary, undesirable bulk mails, which are also called Spam Mails. Anyone with access to the internet can receive spam on their devices. Most spam emails divertpeople’s attention away from genuine and important emails and direct them towards detrimental situations. Spam emails are capable of filling up inboxes or storage capacities, deteriorating the speed of the internet to a great extent.These emails have the capability of corrupting one’s system by smuggling viruses into it, or steal useful information and scam gullible people.

The identification of spam emails is a very tedious task and can get frustrating sometimes.

While spam detection can be done manually, filtering outa large number of spam emails can take very long and wastea lot of time. Hence, the need for spam detection soft wares has become the need of the hour. To solve this problem, various spam detection techniques are used now. The most common technique for spam detection is the utilization of Naive Bayesian method and feature sets that assess the presence of spam keywords. The main purpose is to demonstrate an alternative scheme, with the use of Neural Network (NN) classification system that utilises a collection of emails sent by several users, is one of the objectives of this research. One other purpose is the development of spam detection with the help of Artificial Neural Networks, resulting in almost 98.8% accuracy.

II.LITERATURE SURVEY

Email:

Electronic mail (email) is a messaging system that electronically transmits messages across computer networks. Anyone is free to use email services through Gmail, Yahoo or people can even register with an Internet Service Provider (ISPs) and be provided with an email account. Only an internet connection is required, otherwise being a free service.

Spam:

Bulk mails that are unnecessary and undesirable can be classified as Spam Mails. These spam emails hold the powerto corrupt one's system by filling up inboxes, degrading the speed of their internet connection.

Spam Detection:

Many spam detection techniques are being used now-adays. The methods use filters which can prevent emails from causing any harm to the user. The contributions and their weakness have been identification.

Categories	Description
Health	The spam of fake medication
Promotional products	The spam of fake fashion items like clothes, watches and bags
Adult content	The spam of adult content
Finance and marketing	The spam of stock kiting, tax solutions and loanpackages
Phishing	The spam of Phishing offraud

Table: Spam caterogies

There are several methods that are accessible to spam, for example location of sender, it's contents, checking IP address or space names. Spammers use refined variations to avoid spam identification. Few measures connected with spam identification are; Blacklist and white-list, Machine learning approaches, Naïve Bayes, Support Vector Machine,Neural Network Classification. A mobile system was proposed by Mahmoud ei at. with the motive of blocking and identifying spam SMS. In their work, they attempted to protect smartphones by filtering SMS spam that contains abbreviations and idioms. The system was based on the Artificial Immune System (AIS) and Naïve Bayesian (NB) algorithm. By the use of the Naive Bayes algorithm, the messages are classified based on their features. It used an SMS dataset with 1324 messages. Results from this system gave detection rate 82%, 6% positive rate and 91% accuracy.

III.EXISTING SYSTEMS

Due to the increase in the number of email users, the amount of spam emails have also risen in number in the pastyears. It has now become even more challenging to handle awide range of emails for data mining and machine learning. Therefore, many researchers have executed comparative studies to see various classification algorithms performancesand their results in classifying emails accurately with the help of a number of performance metrics. Hence, it is important to find an algorithm that gives the best possible outcome for any particular metric for correct classification of emails and spam or ham. The present systems of spam detection are reliant on three major methods:-

A. Linguistic Based Methods Unlike humans, who can grasp linguistic constructs along with their exposition, machines cannot and hence it is necessary to teach machinessome languages to help them understand these constructs. This is the technique that is used in places like search engines in order to ascertain the next terms for suggestions to the user while they are typing their search. Sentences are divided into two Unigrams (words taken are one by one) and two Bigrams (words that are taken two at a time). Since this technique requires that every expression beremembered, this method is not feasible and also time- intensive.

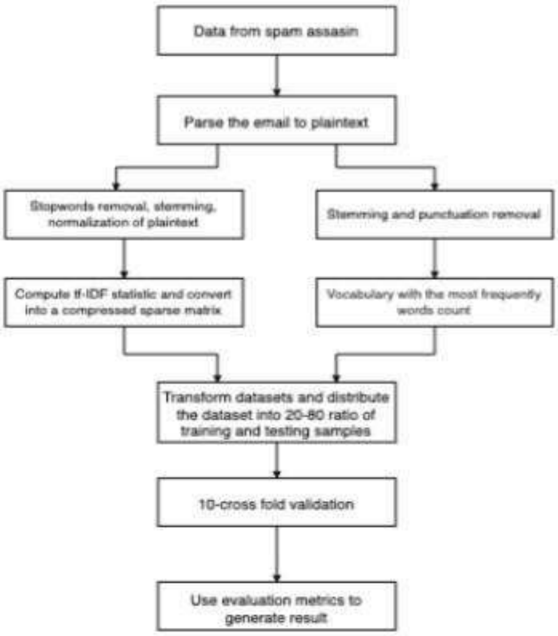
Data from spam assasin
Parse the email to plaintext
Stopwords removal, stemming, normalization of plaintext
Stemming and punctuation removal
Compute tf-IDF statistic and convert into a compressed sparse matrix
Vocabulary with the most frequently words count
Transform datasets and distribute the dataset into 20-80 ratio of training and testing samples
10-cross fold validation
Use evaluation metrics to generate result

Fig 1: Flow Chart of Method

Behavior-Based Methods This technique is Metadata-based. This approach requires that users generate a set of rules, and the users must have a thorough understanding of these rules. Since the attributes of spam change over time sothe rules also need to be reformed from time to time.

As a result, it still requires a human to scrutinise the details and ismajorly user-dependent.

B. Graph-Based Methods This technique uses a single graphical representation by incorporating numerous, heterogeneous particulars. Graph-based anomaly recognition algorithms are executed which detect abnormal forms in the data showing behaviours of spammers. This method is not dependable, so it is taxing to recognise faulty opinions.

IV.PROPOSED SYSTEM

The dataset is taken from SpamAssassin , 2500 nonspam messages belong to easy_ham and they should be easily differentiated from spam. Instead of using sophisticated and hybrid models, this study relies on relatively simple classification algorithms to solve this problem like Logistic Regression, Naive Bayes, and Support Vector Machine. Theconcept of Neural Networks is also used to select the best activation function for spam detection. The dataset is in the form of HTML files which are converted into plaintext during text preprocessing. This paper has used two feature sets to find the most optimal feature set and respective models. In order to perform efficient operations, Compressed Sparse Row (CSR) is used to feed data to models. Hence, the data is converted into a compressed sparse row matrix format for modeling.

A perfect (or best) model should bethe one that reduces underfitting or overfitting. There are three practices for identification. They are datasets splitting, cross-validation, and bootstrap. In proposed work to prevent underfitting and overfitting, the modeling results will be evaluated first through a 10-fold cross-validation score, and then evaluated by evaluation metrics of classification.

V.EMAIL SPAM FILERING METHOD

At present, number of spam email has increased for several criteria such as an advertisement, multi-level marketing, chain letter, political email, stock market advice and so forth. For restricting spam email, several methods or spam filtering system has been constructed by using various concept and algorithms. This section concluded by describing few of spam filtering methods to understand the process of spam filtering and its effectiveness.

A) Standard Spam Filtering Method

Email Spam filtering process works through a set of protocols to determine either the message is spam or not. At present, a large number of spam filtering process have existed. Among them, Standard spam filtering process follows some rules and acts as a classifier with sets of protocols. Figure.1 shows that, a standard spam filtering process performed the analysis by following some steps. First one is content filters which determine the spam message by applying several Machines learning techniques. Second, header filters act by extracting information from email header. Then, backlist filters determine the spam message and stop all emails which come from backlist file. Afterward, “Rules-based filters” recognize sender through subject line by using user defined criteria. Next, “Permission filters” send the message by getting recipients pre-approvement. Finally, “Challenge response filter” performed by applying an algorithm for getting the permission from the sender to send themail.

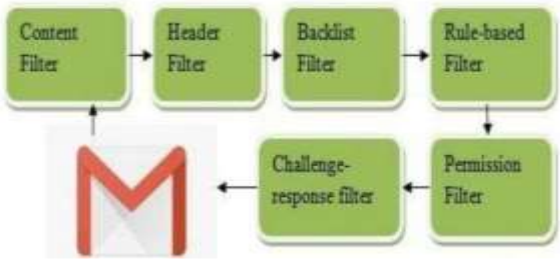

Fig 2: A standard process of Email spam filtering system

B)Client Side and Enterprise Level Spam Filtering Methods

A client can send or receive an email by just one clicking through an ISP. Client level spam filtering providessome frameworks for the individual client to secure mail transmission. A client can easily filter spam through these several existing frameworks by installing on PC. This framework can interact with MUA (Mail user agent) and filtering the client inbox by composing, accepting and managing the messages .

Enterprise level spam filtering is a process where provided frameworks are installing on mail server which interacts with the MTA for classifying the received messages or mail in order to categorize the spam message onmore efficiently. By far most; current spam filtering frameworks use principle based scoring procedures. An arrangement of guidelines is connected to a message and calculates a score based principles that are valid for the message. The message will consider as spam message when it exceeds the threshold value.

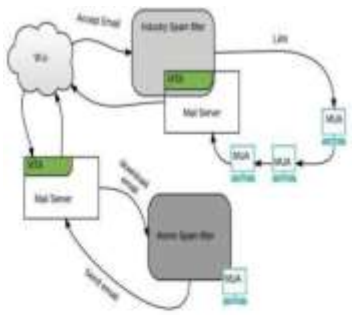

Fig3: Client Side and Enterprise level Email spam filtering system

B) Case Base Spam Filtering Method

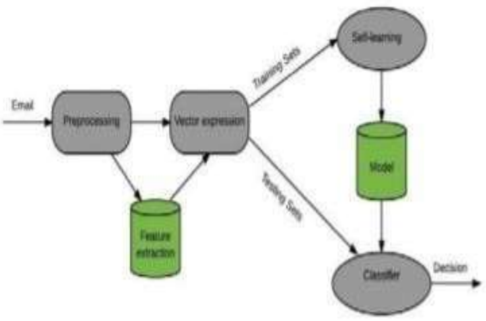

Fig 4: Case Base Spam Filtering System

Among several spam filtering methods; case base or samplebase filtering is one of the prominent method for machine learning methods

Here, describes a sample of case base spam filtering architecture by applying Machine learning techniques in detail. The full process perform through several steps which followed by the figure 3. At the first step, extracted all email (spam email and legitimate email) from individual users email through collection model. Then, the initial transformation starts with the pre-processing steps through client interface, highlight extraction and choice, email data classification, analyzing the process and by using vector expression classifies the data into two sets. Finally, machine learning technique is applied on training sets and testing sets to determine email whether it is spam or legitimate. The final decision makes through two steps; through self observation and classifier's result to make decision whether the email is spam or legitimate.

VI. SPAM DETECTION TECHNIQUES

There are lots of existing techniques which try to preventor reduce the expansion of huge amount of spam or junkemail. The available techniques usually move around usingof spam filters. Generally, spam detection techniques or Spam filters inspect different sections of an email messageto determine if it is spam or not. On the basis of differentsections of email messages; Spam detection techniques canbe classified as Origin based spam detection techniques andContent based spam detection techniques . Generally, mostof the techniques applied to the issue of spam detection areeffective but the important role in minimizing spam email isthe content based filtering. Its positive outcome has forcedspammers to regularly change their methods, behaviors, andto scam their messages, in order to avoid these kinds offilters. Spam detection techniques are discussed below: Origin-Based Technique:

Origin or address based filters are techniques which based on using network information to detect whether a emailmessage is spam or not. The email address and the IP address are the most important parts of network information used. There are few main categories of origin-Based filters like Blacklists; Whitelists based systems

1) Blacklists:

Blacklists are records of email addresses or IP addresses thathave been earlier used to send spam . In creating a filter; if the sender of mail has its entry in the black list then that mail is undesirable and will be considered as spam . For example those websites can be put in blacklist which have a past record of fraudulent or which exploits browser's vulnerabilities. The main problem of a blacklist is maintaining its content to be accurate and up-to-date.

2) WhiteLists:

It is opposite to the black list concept. It consists of the list of entries which can penetrate through and are authorized. These mails are considered as ham mails and can be accepted by the user. It has a set of URLs and domain names that are legitimate . Spam is blocked by a white list with a system which is exactly opposite to existing blacklist.Rather than define which senders to block mail from, a white list define which senders to permit mail from; these addresses are placed on a trustedusers list . The main difficulty of white listings is the assumption that trustworthy contacts do not send junk, for a while this theory could be invalid. Great number of spammers uses PCs that have been harmed using viruses and Trojans for sending spam, to every single one contacts of address book, thus we could receive a spam message from a recognized sender if a virus has infected his computer. Seeing as these contacts are present in the white list, all messages arriving from them arelabeled as secure .

3) Realtime Blackhole List (RBL):

This spam-filtering method acts something like the same toa accepted blacklist on the converse less hands-on maintenance is required, and the Mail Abuse Prevention System Control Theory . and System administrators (third- party) operate it using spam detection tools li. This filter basically needs to connect to the third-party system whenever an email comes in, to authenticate the sender's IP address against the list. As the list is probably to be preserved by a third party, we don't have as much of controlon what addresses are there on the
List.

Content Based Spam Detection Techniques:

Content based filters are based on examining the content of emails. These content based filters are based on manually made rules, also called as heuristic filters, or these filters are learned by machine learning algorithms .These filters try to interpret the text in respect of examine its content and make decisions on that basis have spread among the Internet users, ranging from individual users at their personal computers, to big commercial networks. The success of content-based filters for spam detection is so large that spammers have performed more and more complex attacks intended to avoid them and to reach the users mailbox. There are various popular content based filters such as: RuleBased Filters, Bayesian filters, Support Vector Machines (SVM) and Artificial Neural Network (ANN).

Rule- Based Filters: The Rule-Based Filters use a set of rules on the words incorporated in the whole message to find out whether the message is spam or not. In this approach, a comparison is done between each email message and a set of rules to find out whether a message is spam or ham. A set of rules contains rules with a variety of weights assigned to each rule. In the beginning, each received email message has a zero score. Then email is parsed to detect the existence of any rule, if it exists. If the rule is found in the message, then the weight of the rule is added to the final score of the email. At the end, if the final score is found to be exceeding some threshold value, then the email is declared as spam .

VII. RESULTS

The evaluation criteria is simply based on the followingevaluation metrics:

- Accuracy
- Precision
- Recall
- F1 score

These four factors comprehend the performance of a model with the feature set. In the figure 4 above, it is shown how different models perform with these respective metrics. As shown by the accuracy graphs it can be seen that the artificial
neural network has the highest detection rate of whether afile is spam or ham. Also as shown by recall and F-Score itcan be seen that the Neural Network out performs everyother model.
However results can also be seen in terms of precision logistic regression is the better,
however it's notthe best model as its poor performance compared to others. The output of the results of feature set 1 and feature set 2with the models respectively. cv_score_mean refers to cross validation score, and is used to verify accuracy results.

VIII. CONCLUSION

As shown in all the models based on the feature set 2 most-frequent-word-count have higher accuracy and F1 score than those based on the feature set 1 stop words + n-gram + tf-IDF. If the use case is to introduce a beta version of an email spam detector like no-spam in the inbox.In this case, the model: Neural Network with tanh activation function and the feature set 1 stop words + n-gram + tf-IDF serves this purpose. According to the graphs in Figure 4, if the use case is to introduce an email spam detector to reducebad user experience in searching for important emails from junk mailboxes and filtering spam from the inbox. In this case, Neural Network with a feature set 2 - ‘most frequent word count’ gives a better user experience in general. The future work includes testing the model with various standarddatasets. This research proposes that the outcome that is obtained should be compared with additional spam datasets from various sources. Also, more classification and feature algorithms should be analyzed with email spam datasets.

IX. REFERENCES

[1] AKINYELU, A. A., & ADEWUMI, A. O. (2014). “Classification of phishing email using random forest machine learning technique”. Journal of Applied Mathematics.

[2] Vinodhini. M, Prithvi. D, Balaji. S “Spam Detection Framework using ML Algorithm” in IJRTE ISSN: 2277- 3878, Vol.8 Issue.6, March 2020. [3] YUsKSEL, A. S., CANKAYA, S. F., &UsNCUs, It. S. (2017). “Design of a Machine Learning Based Predictive Analytics System for Spam Problem.” Acta Physica Polonica, A., 132 (3).[26] GOODMAN, J. (2004, July). “IP Addresses in
Email Clients.” In CEAS.

[4] Deepika Mallampati, Nagaratna P. Hegde “A Machine Learning Based Email Spam Classification Framework Model” in IJITEE, ISSN: 2278-3075, Vol.9 Issue.4, February 2020.

[5] Javatpoint, “Machine Learning Tutorial” 2017
https://www.javatpoint.com/machine- learning

[6] SpamAssassin, “Spam and Ham Dataset", Kaggle, 2018.
https://www.kaggle.com/veleon/ham-and-spam-dataset

[7] Apache, “open-source Apache SpamAssassin Dataset”,2019
https://spamassassin.apache.org/old/publiccorpus/

[8] SpamAssassin, “Spam Classification Kernel”, 2018
https://www.kaggle.com/veleon/spam-classification

[9] SpamAssassin, “REVISION HISTORY OF THISCORPUS”, 2016
https://spamassassin.apache.org/old/publiccorpus/readme.html

[10] Jason Brownlee, “Naive Bayes for Machine Learning”The Machine Learning Mastery, April 11, 2015. https://machinelearningmastery.com/naive-bayes- formachine- learning/

.