

Detecting spam mail using machine learning algorithm

Muddala Bhavani¹, Machetti Bala Santoshi², Ummidi Anu Pravallika³, Nuni Madhav⁴

¹²³⁴Final Year B.Tech, CSE, Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P, India

Guided by: Mrs. Dr. K.N.S. Lakshmi, Professor, SVPEC, Visakhapatnam, A.P, India

Abstract - Spam emails defined as unrequested and unwanted commercialized emails or deceptive emails received by a specific person or a company. Some of the Spam identified through natural language processing and machine learning methodologies. ML (machine learning) methods are used to render spam classifying emails to either valid messages or unwanted messages by applying of Machine Learning classifiers. The proposed work useful for differentiating features of the content of documents. Huge work that has been applied in the area of spam filtering which is restricted 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 model-based approaches is developed to review the machine learning methodologies used for automatic spam detection.

Key Words: NLP, Feature Selection, spam detection.

1. INTRODUCTION

E-mails square measure used everybody, they additionally keep company with unessential, undesirable bulk mails, that are referred to as Spam Mails [15]. Anyone with access to the net will receive spam on their devices. Email system is one amongst the foremost effective and usually used sources of communication. The rationale of the recognition of email system lies in its value effective and quicker communication nature. Sadly, email system is obtaining vulnerable by spam emails. Spam emails square measure the uninvited emails sent by some unwanted users additionally referred to as spammers [1] with the motive of creating cash. The e-mail users pay most of their valuable time in sorting these spam mails [2]. Multiple copies of same message square measure sent associate degree again over and over however additionally irritates the receiving user. Spam emails don't seem to be solely intrusive the user's emails however they're additionally manufacturing great amount of unwanted knowledge and therefore touching the network's capability and usage. In this paper, a Spam Mail Detection (SMD) system is planned which can classify email knowledge into spam and ham emails.

The method of spam filtering focuses on 3 main levels: the e-mail address, subject and content of the message [4]. All mails have a standard structure i.e. subject of the e-mail and therefore the body of the e-mail. A typical spam mail may be classified by filtering its content. The

method of spam mail detection relies on the belief that the content of the spam mail is totally different than the legitimate or ham mail. As an example words associated with the packaging of any product, endorsement of services, qualitative analysis connected content etc. The method of spam email detection may be broadly speaking categorized into 2 approaches: information engineering and machine learning approach [5].

Knowledge engineering is a network-based approach in which IP (internet protocol) address, network address along with some set of defined rules are considered for the email classification. The approach has shown promising results however it's terribly overwhelming. The upkeep and task of change rules isn't convenient for all users. On the opposite hand, machine learning approach doesn't involve any set of rules and is economical than information engineering approach [6]. The classification algorithmic rule classifies the e-mail supported the content and alternative attributes. For many of the classification issues the method of feature extraction and choice is extremely necessary. Options play an important role within the method of classification. During this paper, a correlation primarily based feature choice (CFS) [7] method is employed for feature extraction. The CFS approach extracts the simplest options from the pool of options for economical classification results. The planned spam mail detection system is impressed from the effectiveness of machine learning approach. In spam mail detection system, at the start email information is collected. The e-mail information collected is raw and unstructured in nature. So as to scale back the computations and get correct results, email information must be pre-processed. The info is pre-processed by removing stop words, stemming and word tokenization is additionally performed to acquire valuable info. Then, CFS primarily based i.e. correlation primarily feature choice is performed to induce the simple selected options from the pool of options. The pre-processing step reduces the spatial property of knowledge and options within the sort of bag of words area unit then extracted. For the classification a bagged hybrid approach (which is combination of Naïve scientist classifier and J48) is used therefore on produce the classification stronger and extra correct. Spam emails unit capable of filling up inboxes or storage capacities, deteriorating the speed of the net to a wonderful extent. These emails have the ability of corrupting one's system by importing viruses into it, or steal useful data and scam gullible of us. The identification of spam emails

could also be a very tedious task and may get frustrating typically. Where as spam detection is finished manually, filtering out an outsized vary of spam emails can take very long and waste some time. Hence, the necessity of spam detection software package has become the necessity of the hour. To solve this disadvantage, varied spam detection techniques unit used presently. The foremost common technique for spam detection is that the utilization of Naive Bayesian[5] methodology and have sets that assess the presence of spam keywords. The foremost purpose is to demonstrate an alternate theme, with the use of Decision Tree[4] arrangement that utilizes a group of emails sent by several users, is one in each of the objectives of this analysis. One different purpose is that the event of spam detection with the help of Artificial Decision Trees, resulting in nearly ninety eight 98.8% accuracy

EXISTINGSYSTEM:

Due to the rise within the range of email users, the number of spam emails have also increases in the past years. It's currently become even more difficult to handle a good vary of emails for data processing and machine learning. Therefore, several researchers have dead comparative studies to ascertain varied classification algorithms performances and their ends up in classifying emails accurately with the assistance of variety of performance metrics. Hence, it's vital to seek out associate algorithmic program that provides the most effective potential outcome for any specific metric for proper classification of emails and spam or ham. This systems of spam detection area unit dependent on 3 major ways:-

Linguistic Based Methods

Unlike humans, agency will grasp linguistic constructs at the side of their exposition, machines cannot and thus it's necessary to show machines some languages to assist them perceive these constructs.

This is often the technique that's utilized in places like search engines so as to determine succeeding terms for suggestions to the user whereas they're typewriting their search. Sentences are divided into 2 Unigrams (words taken are one by one) and 2 Bigrams (words that are taken 2 at a time). Since this method needs that each expression be remembered, this technique isn't possible and conjointly time-intensive[9].

Behavior-Based Methods

This technique is Metadata-based. This approach needs that users generate a group of rules, and therefore the users should have a radical understanding of those rules. Since the attributes of spam amendment over time that the rules additionally ought to be reformed from time to time. As a result, it still needs somebody's to scrutinize[9].

Graph-Based Methods

This technique uses one graphical illustration by incorporating various, heterogeneous particulars. Graph-based anomaly recognition algorithms are dead that discover abnormal forms within the knowledge showing behaviors of spammers. This methodology isn't dependable, therefore it's onerous to recognize faulty opinions[9].

Feature Engineering principally depends on the industrial charm of terms and is totally content-oriented, and doesn't rely on statistics of these attributes result in an interesting decline of this structure.

PROPOSED SYSTEM

The dataset is taken from Spam Assassin, non-spam messages belong to easy-ham and that they ought to be simply differentiated from spam.

rather than victimization refined and hybrid models, this study depends on comparatively easy classification algorithms to unravel this downside like provision Regression, Naive mathematician, and Support Vector Machine.

The idea of call Trees is additionally accustomed choose the simplest activation operate for spam detection. The dataset is within the kind of TEXT files that are reborn into plaintext throughout text pre-process.

This paper has used 2 feature sets to seek out the foremost optimum feature set and individual models. so as to perform economical operations, compressed distributed Row (CSR) is employed to feed knowledge to models.

Hence, the info is reborn into a compressed distributed row matrix format for modeling. an ideal (or best) model ought to be the one that reduces under-fitting or over-fitting. There are 3 practices for identification. They're datasets ripping, cross-validation, and bootstrap.

In projected work to forestall under-fitting and over-fitting, the modeling results are evaluated initial through a 10-fold cross-validation score, then evaluated by analysis metrics of classification.

METHODOLOGY

The Python program can load all the required Python libraries which will assist the mil modules to classify the emails and observe the spam emails.

A. ADDING CORPUS

This section will load all the email datasets within the program and distribute into training and testing data. This process will be accepting the datasets in '*.txt' format for individual email (genuine and Spam). This is to assist

perceive the real-world problems and the way will they be tackled..

B. TOKENIZATION

Tokenization is the method where the sentences within an email are broken into individual words (tokens). These tokens are saved into an array and used towards the testing data to identify the occurrence of every word in an email. This will help the algorithms in predicting whether the email should be considered as spam or genuine.

C. FEATURE EXTRACTION AND STOP WORDS

This was used to remove the unnecessary words and characters within each email, and creates a bag of words for the algorithms to compare against .The module 'Count Vectors' from Sklit-learn assigns numbers to each word/token while counting and provides its occurrence within an email. The instance is invoked to exclude the English stop words, and these are the words such as: A, In, The Are, As, Is etc., as they are not very useful to classify whether the email is spam or not. This instance is then fitted for the program to learn the vocabulary Once tokenized, the program applies 'Tfidf-Transformer' module to compute the Inverse Document Frequency (IDF). The most occurring words within the documents will be assigned values between 0-1, and lower the value of the

word means that they are not unique. This allows the algorithms/modules to browse the info. The TF-IDF can be calculated by the Equation-11 where (t, d) is the term frequency (t) within a document (d):

$$tf - idf(t, d) = tf(t, d) \times idf(t)$$

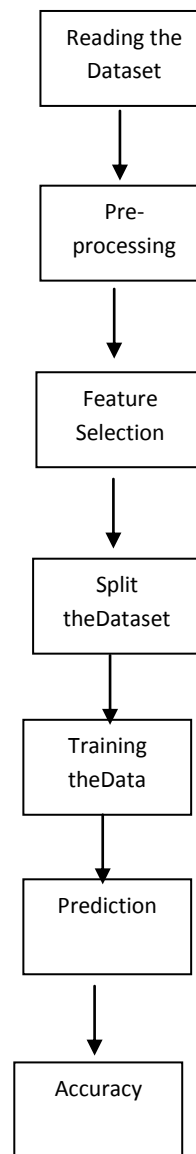
where IDF is calculated by the Equation given n is the number of documents:

$$idf(t) = \log(n/idf(t)) + 1$$

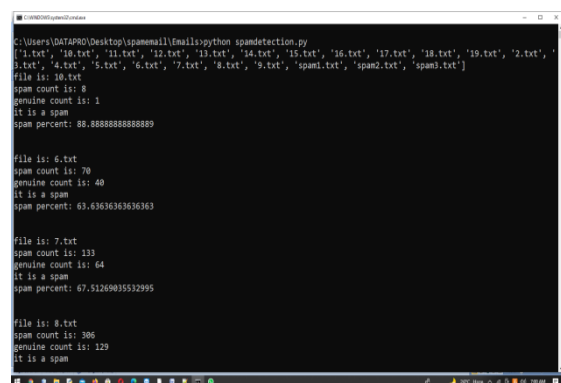
D. MODEL TRAINING AND TESTING PHASE

As mentioned through the analysis, supervised learning strategies were used and therefore the model was trained with notable knowledge and tested with unknown knowledge to predict the accuracy and different performance measures. To acquire the reliable results K-Fold cross validation was applied. This method does have its disadvantages such as, there is a chance that the testing data could be all spam emails, or the training set could include the majority of spam emails. This was resolved by Stratified K-fold cross validation, which separates the data while making sure to have a good range of Spam and genuine into the distributed set. Lastly, parameter tuning was conducted with the Sklit-Learn and bio-inspired algorithms approach to try and improve the accuracy of ML models. This provides a

platform to compare the Scikit-learn library with the bio-inspired algorithms



Results and analysis

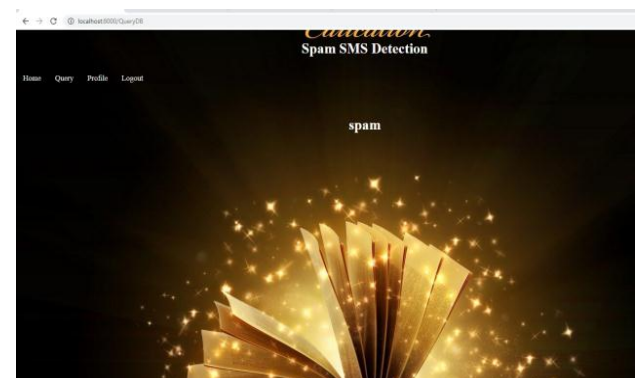
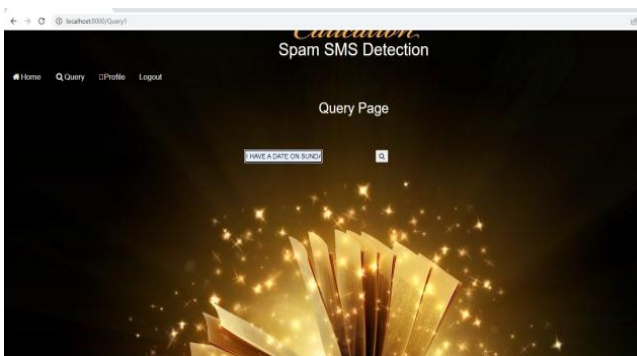


```
C:\Users\DATAPRO\Desktop\spammemail\mail\python spamdetection.py
['1.txt', '10.txt', '11.txt', '12.txt', '13.txt', '14.txt', '15.txt', '16.txt', '17.txt', '18.txt', '19.txt', '2.txt', '3.txt', '4.txt', '5.txt', '6.txt', '7.txt', '8.txt', '9.txt', 'spam1.txt', 'spam2.txt', 'spam3.txt']
file is: 0.txt
spam count is: 8
genuine count is: 1
it is a spam
spam percent: 88.88888888888889

file is: 6.txt
spam count is: 70
genuine count is: 40
it is a spam
spam percent: 63.63636363636363

file is: 7.txt
spam count is: 233
genuine count is: 64
it is a spam
spam percent: 67.5126983532995

file is: 8.txt
spam count is: 200
genuine count is: 129
it is a spam
```



Conclusion

All the models supported the feature set a pair of most-frequent-word-count have higher accuracy and F1 score than those supported the feature set one stop words + n-gram + tf-IDF. If the utilization case is to introduce a beta version of associate email spam detector like no-spam within the inbox.

During this case, Tree with activation operate and also the feature set one stop words + n-gram + tf-IDF serves this purpose in line with the graphs in Figure four, if the utilization case is to introduce associate email spam detector to scale back dangerous user expertise in finding out vital emails from junk mailboxes and filtering spam from the inbox.

During this case, call Tree with a feature set a pair of - 'most frequent word count' provides an improved user expertise generally. The longer-term work includes testing the model with numerous normal datasets.

This analysis proposes that the result that's obtained ought to be compared with further spam datasets from numerous sources. Also, a lot of classification and have algorithms ought to be analyzed with email spam datasets

Future Scope:

In the future, we have a tendency to attempt to wear down tougher issues like the analysis and management of report in spam SMS filters storing. we are going to focus Transformer model for higher accuracy.

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 THIS CORPUS", 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/>

[11] Wikipedia, "History of email spam," Internet Free Encyclopedia, 2001.
https://en.wikipedia.org/wiki/History_of_email_spam

[12] Rohith Gandhi, "Support Vector Machine" The Machine Learning Mastery, June 7, 2018.
<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms934a444fca47>

[13] Jason Brownlee, "Logistic Regression for Machine Learning" The Machine Learning Mastery, April 1, 2016.
<https://machinelearningmastery.com/logisticregression-for-machine-learning/>

[14] Jason Brownlee, "How to Encode Text Data for Machine Learning with scikit-learn" The Machine Learning Mastery, September 29, 2017.
<https://machinelearningmastery.com/prepare-text-datamachine-learning-scikit-learn/>

[15] I. Androutsopoulos, J. Koutsias, K. Chandrinou and C. D. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal email messages," Computation and Language, pp. 160-167, 2000.



U.Anu Pravallika
Persuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College



N.Madhav
Persuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

BIOGRAPHIES



Dr.K.N.S.Lakshmi
Currently working as associate professor from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering college



M.Bhavani
Persuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College



M.B.Santoshi
Persuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College