Email spam filtering approaches

Author: Kothakota Viswanadh
Co-Author/Corresponding Author: Jimoh Abdulganiyu

1.1 INTRODUCTION:

In recent times, undesirable business mass messages called spam has become an immense issue on the web. The person sending the spam messages is referred to as the spammer. Such an individual assembles email addresses from various sites, chatrooms, and infections. Spam keeps the client from making full and great utilization of time, stockpiling limit, and system data transfer capacity. The gigantic volume of spam sent moving through the PC systems effectively affects the memory space of email workers, correspondence transfer speed, CPU force, and client time. The threat of spam email is on the expansion on yearly premise and is answerable for over 77% of the entire worldwide email traffic. Clients who get spam messages that they didn't demand think that it's extremely bothersome. It has additionally come about to untold money related misfortune to numerous clients who have fallen casualty of web tricks and other deceitful acts of spammers who send messages claiming to be from trustworthy organizations with the aim to convince people to uncover touchy individual data like passwords, Bank Verification Number (BVN) and Mastercard numbers.

To viably deal with the danger presented by email spams, driving email suppliers, for example, Gmail, Yahoo mail, and Outlook have utilized the mix of various AI (ML) procedures, for example, Neural Networks in its spam channels. These ML procedures have the ability to learn and recognize spam sends and phishing messages by dissecting heaps of such messages all through an immense assortment of PCs. Since AI has the ability to adjust to differing conditions, Gmail and Yahoo mail spam channels accomplish something beyond browsing garbage messages utilizing prior standards. They produce new principles themselves

dependent on what they have realized as they proceed in their spam separating activity. The AI model utilized by Google has now progressed to the point that it can recognize and sift through spam and phishing messages with about 99.9 percent precision. The ramifications of this is one out of a thousand messages prevail with regards to avoiding their email spam channel. Measurements from Google uncovered that between 50-70 percent of messages that Gmail gets are spontaneous mail. Google's identification models have additionally consolidated devices called Google Safe Browsing for recognizing sites that have noxious URLs. The phishing-location execution of Google has been upgraded by the presentation of a framework that postpones the conveyance of some Gmail messages for some time to complete extra exhaustive investigation of the phishing messages since they are simpler to distinguish when they are examined by and large. The reason for deferring the conveyance of a portion of these dubious messages is to lead a more profound assessment while more messages show up at the appropriate time of time and the calculations are refreshed progressively. Just about 0.05 percent of messages are influenced by this conscious deferral

1.2 BACKGROUND OF STUDY:

There is a fast increment in the enthusiasm being appeared by the worldwide examination network on email spam separating. It is assessed that 70 percent of all email sent comprehensively is spam, and the volume of spam keeps on developing since spam stays a worthwhile business. Spammers get perpetually modern and inventive in their strategies to get their messages into your inboxes and unleash their devastation. Spam filtering solutions should consistently be refreshed to address this advancing danger.

Numerous scientists and academicians have proposed distinctive email spam classification methods which have been effectively used to classify data into groups. These strategies incorporate probabilistic, decision tree, support vector machine (SVM), artificial neural systems (ANN), and case-based procedure. It has appeared in writing that it is conceivable to utilize these grouping strategies for spam

mail separating by utilizing a content-based filtering strategy that will distinguish certain highlights (ordinarily watchwords as often as possible used in spam messages). The rate at which these highlights show up in messages determine the probabilities for every trademark in the email, after which it is estimated against the threshold value. Email messages that exceed the limit classified as spam.

1.3 STATEMENT OF PROBLEM:

My aim here is to review some of the algorithms being applied for classification of messages as either spam or ham is provided.

1.4 MOTIVATION FOR STUDY:

The Internet is viewed as a useful asset. Email is a proficient method to trade data. Thinking about the development of the Internet and wide utilization of email, the pace of increment of spam is of incredible concern. Spam may start from anyplace in the World Wide Web. Despite tools to prevent spam, it has been expanding every day. One approach to evaluate the current circumstance is that associations analyze accessible implies that can be utilized to try and tally the measure of spam. These methods incorporate corporate email frameworks, gateways, spam separating and end client training. Web clients can't dismiss this significant issue of the cutting edge Internet world. Absence of motorized frameworks to forestall spam will bring about a spam-soaked World Wide Web, destruction of Internet items and extreme loss of transmission capacity.

1.5 AIM AND OBJECTIVES OF THE STUDY:

1.5.1 AIM:

This work aims to review machine learning approaches and their application to the field of spam filtering.

1.5.2 OBJECTIVES OF THE STUDY:

To realize the stated aim, the subsequent objectives shall be morally followed: The methods used must be simple and fast enough to detect the spam and classify it as a spam or ham as quickly as possible.

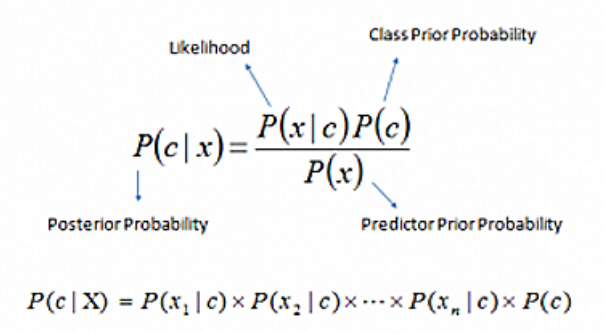
1.6 MATERIALS AND METHOD:

Spam mail classification is regularly dealt with by machine learning (ML) algorithms intended to separate spam and non-spam messages. Machine learning algorithms accomplish this by using an automatic and adaptive technique. Rather than depending on hand-coded rules that are susceptible to the perpetually varying characteristics of spam messages, ML methods can obtain data from a set of messages provided, and then use the obtained data to classify new messages that it just received. In this segment, we will review some of the most popular machine learning methods that have been applied to spam detection.

1.6.1 Naive Bayes classifier:

The Bayesian classification exemplifies a supervised learning technique and at the same time a statistical technique for the classification. It acts as a fundamental probabilistic model and lets us seize ambiguity about the model in an ethical way by influencing the probabilities of results. It is used to provide solutions to analytical and predictive problems. The classification offers practical learning algorithms and previous knowledge and experimental data can be merged.

Mathematical concept of Naive Bayes classifier:



Here,

- P(c|x): posterior probability of class(c,target) given predictor(x,attributes). This represents the probability of c being true, provided x is the true.
- P(c): is the prior probability of class. This is the observed probability of class out of all the observations.
- P(x|c): is the likelihood which is the probability of the predictor-given class. This represents the probability of the x being true, provided x is the true.
- P(x): is the prior probability of the predictor. This is the observed probability of the predictor out of all observations.

1.6.2 Decision Tree Model:

A Decision tree is a specific type of probability tree that enables you to make a decision about some kind of process. Decision trees are useful both in case of classification and in regression analysis. Tree's accuracy majorly depends on the decision of strategic splits. Decision criteria may vary for classification and for regression. Information theory is a typical measure to define the degree of disorganization in a system known as the Entropy. Entropy will be zero if the sample is completely homogeneous and it will be one if the sample is divided equally (50-50). The lesser the entropy, the better its way to decide. It chooses the split which has the lowest entropy compared to the parent node and other splits.

Mathematical concept of Decision Tree Model:

Entropy is defined as:

$$H = -\sum p(x) \log p(x)$$

The information gain Gain(S,A):

$$Gain(S,A) = \underbrace{Entropy(S)}_{\text{original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)}_{\text{relative entropy of S}}$$

- **S** = Each value v of all possible values of attribute A
- **Sv** = Subset of S for which attribute A has value v
- **|Sv|** = Number of elements in Sv
- **|S|** = Number of elements in S

Information Gain (n) =

Entropy(x) — ([weighted average] * entropy(children for feature))

1.6.3 Random Forest:

Numerous Decision trees are involved in Random Forest algorithm, each decision tree has the same nodes, but data used is differ it leads to different leaves. It merges the decision of multiple decision trees in order to provide an optimal solution, which denotes the average of all these decision trees.

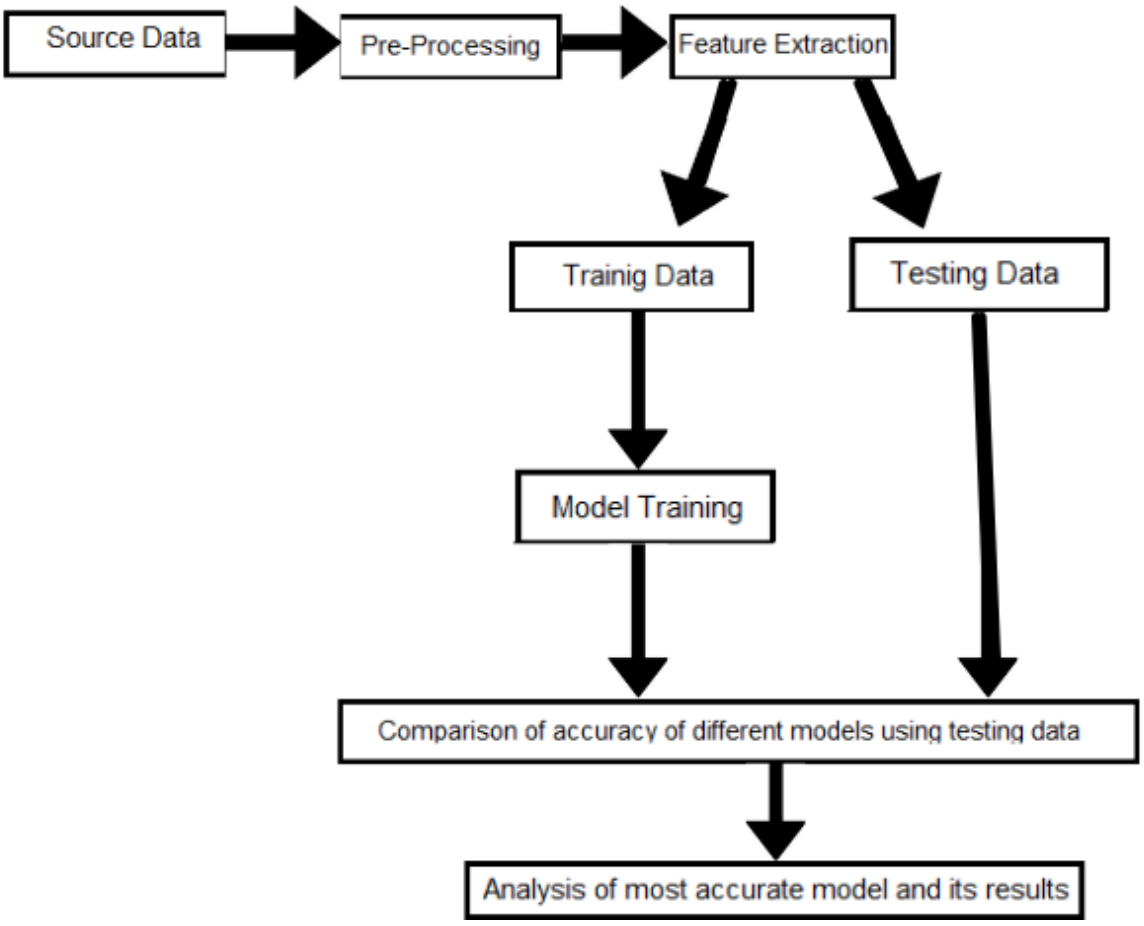
Mathematical concept of Random forest:

The Gini formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- *pi* represents the relative frequency of the class we are observing in our dataset.
- *C* represents the number of classes

1.7 PROPOSED MODEL:



1.8 SCOPE OF THE STUDY:

This study focuses on the problem of Spam and tries to give an overview of Spam characteristics and various models useful to detect spam. Spam is one of the problems that is continuing to grow from day to day, costing corporations billions of dollars in lost productivity. Fortunately though, there are different spam blocking techniques to help counter the various types of the spam. Because spammers are always trying to bypass anti-spam techniques by changing the methods they use to send spam, it's best for corporations to protect themselves with spam blocking solutions that use more than one spam blocking technique. Each one of these techniques has disadvantages, advantages as well as limitations. To minimize the

amount of spam that enters in an organization, a spam blocking solution that includes a combination of the most effective techniques should be implemented.

1.9 SIGNIFICANCE OF THE STUDY:

While new PC security threats may come and go, spam remains a constant irritation for nonprofits. At a minimum, spam can interrupt your day to day life, forcing you to spend time opening and deleting emails hawking herbal remedies or once in a lifetime investment opportunities. In a more serious scenario, spam could unleash a nasty virus on your association's network, crippling your servers and desktop machines. Anti-spam services tend to fix the pace of spam at somewhere in the range of 50 to 90 percent of all messages on the Internet. In spite of the fact that keeping industrious spammers from sending garbage mail may never be conceivable, installing an anti-spam application on your organization's mail server or individual computers can majorly reduce the amount of spam your staffers have to deal with. Anti-spam applications typically use one or more filtering methods to identify spam and stop it from reaching a client's inbox. Yet, in light of the fact that anti-spam programs are intended to do a similar occupation doesn't mean they all go about it similarly. For example, some spam-filtering techniques run a progression of checks on each message to decide the probability that it is spam. Other spam-filtering procedures simply block all email transmissions from known spammers or only allow email from specific senders. And while some spam-filtering methods are completely transparent to both the sender and recipient, others require some level of client connection.

1.10 DEFINITION OF TERM:

The word "Spam" as applied to Email means "Unsolicited Bulk Email". Unsolicited means that the Recipient has not granted verifiable permission for the message to be sent. Bulk means that the message is sent as part of a larger collection of messages, all having substantively identical content. A message is Spam only if it is both Unsolicited and Bulk.

Technical Definition of Spam :

An electronic message is "spam" if the beneficiary's very own personality and setting are immaterial on the grounds that the message is similarly pertinent to numerous other possible beneficiaries; AND the beneficiary has not unquestionably allowed purposeful, express, and still-revocable consent for it to be sent.

2.1 RELATED WORKS:

Many algorithms have been proposed for classifying spam and legitimate emails. J. Provost evaluated the rule-learning RIPPER algorithm compared to the NB algorithm. RIPPER generates keyword-spotting rules to set and bag valued attributes. It performs its classification according to the predefined rules that determine the impact of having certain words in the header fields or content of the emails. The experiments performed on junk email provided by several users and on legitimate ones from the inbox of the author achieved 90% accuracy after training it with 400 emails whereas NB reached 95% after only 50 training emails.

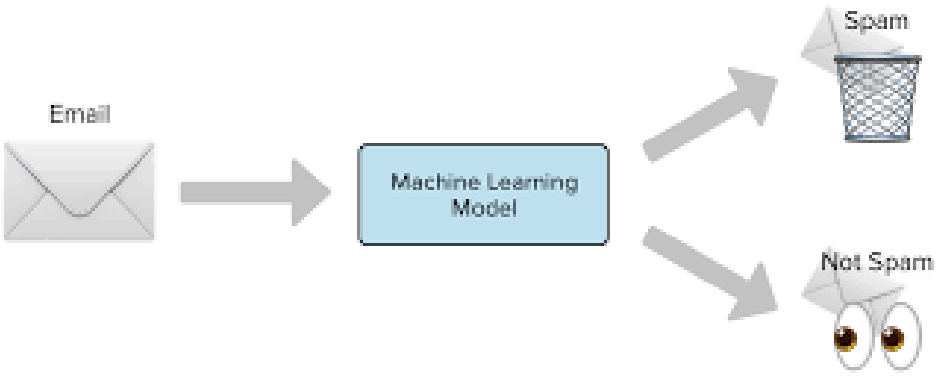
B. Medlock proposed the smoothed n-gram language interpolation and modeling which assumes that the probability of a specific word in a sequence is solely dependent on the previous n-1 words. Separate language models were built for legitimate and spam email followed by computing the probability that the model generated this text message. Bayes rule was applied later to find the class with the highest probability for the provided message. An accuracy of 98.84% and 97.48% was obtained for the adaptive bigram and unigram language model classifier after applying it to the GenSpam corpus of 9072 legitimate and 32332 spam emails.

3.1 METHODOLOGY:

Spamming activities can be appropriately modelled by this solid theory and, subsequently, a young Internet Security Industry has recently emerged to fight against spam. However, the enormous intensification of spam conveyances during last years has prompted to the need of achieving a significant improvement in filter accuracy. In this specific circumstances, current research efforts are mainly focused on providing a wide variety of content-based techniques able to overcome common

spam filtering inconveniences. Although theoretical filtering evaluation is generally taken into the consideration in scientific works, last portion of the evaluation protocols are not appropriate to correctly assess the performance of models during filter activity in real environments. To cover the gap between the basic research and the applied deployment of notable spam filtering techniques, this work proposes a novel clear evaluation methodology able to rank the available models using four unique but the integral perspectives: static, dynamic, adaptive and internationalization.

3.2 SYSTEM ARCHITECT OF THE MODEL:



Data Collection:

A shuffled sample of emails with its subjects and bodies belonging to the spam and ham in different proportions is considered as a dataset to train the model. Downloaded the spam and ham emails through Google's takeout service as an ".mbox" file, which contains the subjects and bodies of emails.

Data Preprocessing:

By using the 'mailbox' package we are going to read the mbox files into lists. Each mail denoted by each element in the list. In the first iteration. As the emails are in packed format we are going to unpack each email and concatenate their subject

and body. As subjects of emails also play a major role in indication of spam or ham we are including the subjects too. Then convert the lists into a dataframe, join the ham and spam dataframes and shuffle the dataframe.

Data Splitting:

Split the dataframe into train and test dataframe in a ratio of 85:15 . The train data was 85%, test data was 15% of the original dataset.

Algorithm/Approach:

Inorder to achieve the most accurate results, different machine learning algorithms like Naive Bayes, Decision Tree, Random Forest Classifier were used.

Prediction:

The data was thoroughly fitted and the spam - ham detection could be accurately tracked by the accuracy, precision, recall, F1 score and AUC(Area Under Curve).

Model Evaluate:

The models are run on the dataset and the accuracy of the model is evaluated with the help of accuracy, precision, recall values obtained by the models prediction on test data.

- **Accuracy** = (TP + TN) / All
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F-Score** is a statistical method for determining the accuracy accounting for both precision and recall. It is essentially the harmonic mean of precision and recall.
- **AUC** denotes the area under the ROC curve (receiver operating characteristic curve). The closer the AUC value to 1 denotes the model performance.

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

	Accuracy	precision	recall	f_score	AUC
Naive Bayes	0.909091	0.937500	0.9375	0.937500	0.885417
Descision Tree	0.772727	0.866667	0.8125	0.838710	0.739583
Random Forest	0.863636	0.933333	0.8750	0.903226	0.854167

Discussion

- Among the three models Naive Bayes achieved the highest accuracy, suggesting that the model is good at predicting or classifying the mail as ham or spam.
- Precision value is also good at 0.9375, denoting that model has a low false positive rate compared to Decision Tree(0.866) and Random Forest (0,93).
- A high Recall denotes the low false negative rate. Naive Bayes model obtained the greater value of 0.9375 while the other models obtained are less.
- When comparing F-Score & AUC for 3 models , the Naive Bayes performed much better than the Decision Tree and Random Forest Classifier.
- Overall the Naive Bayes model was best in determining the spam.

Limitation

The highest accuracy obtained among all the three was from Naive Bayes(0.90) but there might be a scope to get faulty classification of ham-spam when the test data contained a higher percentage of ham than spam.

Recommendation/Further Studies

We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense. We can also use complex models to get better accuracy in classifying ham - spam.

Conclusion

After a comprehensive analysis and research about different algorithms to classify ham and spam , I conclude that the Naive Bayes Algorithms proved to be most efficient.

REFERENCES:

M. Awad, M. Foqaha

Email spam classification using hybrid approach of RBF neural network and particle swarm optimization

Int. J. Netw. Secur. Appl., 8 (4) (2016)

Google Scholar

Kaspersky Lab Spam Report (2017)

spam_report_April_2012

D.M. Fonseca, O.H. Fazzion, E. Cunha, I. Las-Casas, P.D. Guedes

Measuring characterizing, and avoiding spam traffic costs

IEEE Int. Comp., 99 (2016) Google Scholar

J.S Whissell, C.L.A Clarke

Clustering for semi-supervised spam filtering

Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-abuse And Spam Conference (CEAS '11) (2011)

Decision Tree, Random Forest mathematical concepts source:

https://medium.com/meta-design-ideas/decision-tree-a-light-intro-to-theory-math-code-10dbb3472ec4