International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

Email Spam and Ham Classification Using Machine Learning Algorithm

A Monika
Assistant Professor
Department of CSE
Geethanjali College of Engineering and Technology, Hyderabad

P Naresh Kumar
Assistant Professor
Department of ECE
Geethanjali College of Engineering and Technology, Hyderabad

Abstract – ***Internet has become a major source for all kind of transactions and became a big revolution in communication as well. Users are preferring internet to send information rather than a conventional method of sending information. Email occupies major part in daily communications. This made a way for fraud spammers. Spammers may Send malicious link through emails which can harm our system and can also access the system in an unauthorized way. With the usage of fake profile, they may send spam emails. they target those peoples who are not aware about these frauds. Hence there is a great need for detecting those spam mails and save the users from the fraud. In this context machine learning techniques helps us in a Broadway to classify the received mail as spam or ham. Naïve Baye's technique has been applied in this work which resulted a good accuracy in predicting the spam mails.***

Keywords: Machine learning, Naïve Bayes, K Fold cross validation.

1. Introduction:

Electronic mail spam refers to the "using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. "The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet. Spam is a waste of storage, time and message speed. Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily. Several years ago, most of the spam can be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection. Major approaches adopted closer to junk mail filtering encompass "text analysis, white and blacklists of domain names, and community-primarily based techniques".

Text assessment of contents of mails is an extensively used method to the spams. Many answers deployable on server and purchaser aspects are available. Naive Bayes is one of the utmost well-known algorithms applied in these procedures. However, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives. Regularly clients and organizations would not need any legitimate messages to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spams. The technique is to acknowledge all the sends other than those from the area/electronic mail ids. Expressly boycotted. With more up to date areas coming into the classification of spamming space names this technique keeps an eye on no longer work so well. The white list approach is the approach of accepting the mails from the domain names/addresses openly whitelisted and place others in a much less importance queue, that is delivered most effectively after the sender responds to an affirmation request sent through the "junk mail filtering system".

Spam and Ham: the use of electronic messaging systems to send unsolicited bulk messages, especially mass advertisement, malicious links etc. are called as spam. "Unsolicited means that those things which you didn't asked for messages from the sources. So, if you do not know about the sender the mail can be spam. People generally don't realize they just signed in for those mailers when they download any free services, software or while updating the software. "Ham" this term was given by Spam Bayes around 2001 and it is defined as "Emails that are not generally desired and is not considered spam".

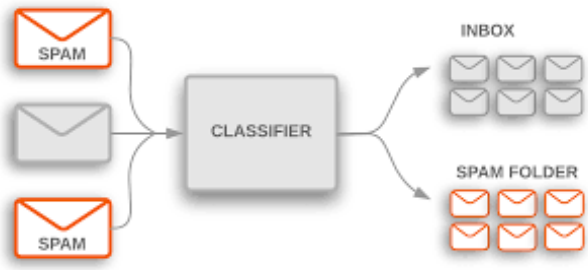

Fig.1. Classification into Spam and non-spam

2. Methodology:

When the data is considered, always a very large data sets with large no. of rows and columns will be noted. But it is not always the case the data could be in many forms such as Images, Audio and Video files Structured tables etc. Machine doesn't understand images or video, text data as it is, Machine only understand 1s and 0s.

Steps in Data Preprocessing:

i. **Data cleaning:** In this step the work like filling of "missing values", "smoothing of noisy data", "identifying or removing outliers ", and "resolving of inconsistencies is done."
ii. **Data Integration:** In this step addition of several databases, information files or information set is performed.
iii. **Data transformation:** Aggregation and normalization is performed to scale to a specific value
iv. **Data reduction:** This section obtains a summary of the dataset which is very small in size but so far produces the same analytical result

3. Machine Learning Model:

Naïve Bayes classifier was used in 1998 for spam recognition. The Naïve Bayes classifier algorithm is an algorithm which is used for supervised learning. The Bayesian classifier works on the dependent events and works on the probability of the event which is going to occur in the future that can be detected from the same event which occurred previously. Naïve Bayes was made on the Bayes theorem which assumes that features are autonomous of each other. Naïve Bayes classifier technique can be used for classifying spam emails as word probability plays main role here. If there is any word which occurs often in spam but not in ham, then that email is spam. Naive Bayes classifier algorithm has become a best technique for email filtering. For this the model is trained using the Naïve Bayes filter very well to work effectively. The Naive Bayes always calculates the probability of each class and the class having the maximum probability is then chosen as an output. Naïve Bayes always provide an accurate result. It is used in many fields like spam filtering.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = \sum_{y} P(B|A)P(A)$$

4. K Fold Cross Validation:

The method of training and testing the data will be done by using the same set of data. The given set of data is splitted into train and test parts. The model is designed based on trained part. The performance of the model and it's accuracy is measured using test data. Some hard situations may occur in this process where the train data may be completely different from test data and the accuracy lead to very low even though we have a designed a good model.

To avoid such situations, the complete data set is divided into k-folds. Later each fold is used for testing at a time and other k-1 folds are used for training. The average score is considered as the accuracy of the code.

5. Algorithm:

i. Insert the dataset or file for training or testing.
ii. Perform data cleaning process.
    a. Remove the NaN values
    b. Drop the unnecessary records from the data
iii. Perform feature engineering.
iv. Apply K Fold cross validation algorithm
v. Fit the model into Naïve Bayes Algorithm
vi. Perform Hyperparameter tuning with L1 Regularization
vii. Compare the accuracy with various training and test data combinations
viii. Show the results of the classifier

6. Results:

The K-Fold Cross Validation with k=8 is considered. The accuracy for each K value is shown in Fig 2.

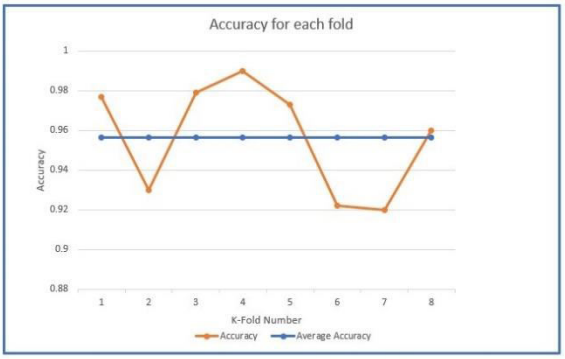

Fig.2. Accuracy for each K-Fold

After applying Naïve Bayes Algorithm, the model has classified the given message as spam or ham with an accuracy of 0.956 with the use of default parameters. After performing Hyper-parameter tuning, the accuracy increased to 0.986.

7. Conclusion:

With this result, it can be concluded that the Multinomial Naïve Bayes gives the best outcome but has limitation due to class-conditional independence which makes the machine to misclassify some tuples. Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus.

8. Future Scope:

Our project, thus spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Therefore, at this it is an only limited body of the email. There is a wide possibility of improvement in our project.

The subsequent improvements can be done:

i. "Filtering of spams can be done on the basis of the trusted andverified domain names."
ii. "The spam email classification is very significant in categorizing e-mails and to distinct e-mails that are spam or non-spam."
iii. "This method can be used by the big body to differentiate decent mails that are only the emails they wish to obtain."

References:

[1] Kaggle.com

[2] Nikhil Kumar, Sanket Sonowal, Nishanth , Email Spam Detection using Machine Learning Alogrithm, 108-113, 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] Suryawanshi, Shubhangi & Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.

[4] https://www.geeksforgeeks.org/naive-bayes-classifiers/

[5] https://scikit-learn.org/