Assessing Machine Learning Algorithms for Detecting Email Spam: A Performance Analysis

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

Abstract

One of the alarming tactics employed by spammers is sending malicious links through spam emails. Clicking on such links can not only harm your system but also allow unauthorized access to sensitive information. Moreover, these spammers go to great lengths to create fake profiles and email accounts, deceiving recipients into thinking they are legitimate individuals. These scammers primarily target those who may not be well-versed in identifying these fraudulent schemes, making the situation even more concerning. To combat this growing menace, we must find effective ways to distinguish legitimate emails from spam. As a potential solution, we are currently working on a project that leverages the power of machine learning techniques. Our aim is to build a robust email spam detection system, capable of differentiating fraudulent messages from genuine ones. In this regard, our research will delve into various machine learning algorithms, carefully analyzing their strengths and weaknesses. By applying these algorithms to our vast datasets, we will assess their performance, with a particular focus on precision and accuracy. Ultimately, our goal is to identify the most effective algorithm that can significantly enhance the email spam detection process.

Keywords: E-mail spam, machine learning framework, logistic regression, Naive bayes classifier.

1. Introduction

Email or 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 whitelist 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: According to Wikipedia "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". Machine learning approaches are more efficient, a set of training data is used, these samples are the set of email which are pre classified. Machine learning approaches have a lot of algorithms that can be used for email filtering.

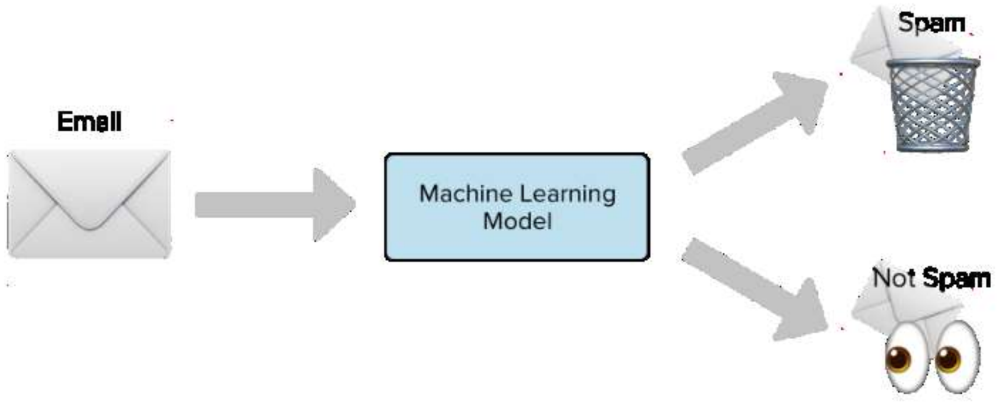


Fig. 1: Overview of spam and non-spam content.

2. Related work

There is some related work that apply machine learning methods in email spam detection, authors in [2] focused on a literature survey of Artificial Intelligence and Machine learning methods for email spam detection. In [3], authors have used the "image and textual dataset for the e-mail spam detection with the use of various methods. Harisinghaney et al. [4] have used methods of KNN algorithm, Naïve Bayes, and Reverse DBSCAN algorithm with experimentation on dataset. For the text recognition, OCR library" is employed but this OCR doesn't perform well. The feature selection hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and Rough pure mathematics are employed in [5].

2.1 Data Set

This model uses email data sets from different online websites like Kaggle, sklearn and some data sets are created by own. A spam email data set from Kaggle is used to train our model and then other email data set is used for getting result "spam.csv" data set contains 5573 lines, and 2 columns and other data sets contains 574,1001,956 lines of email data set in text format.

3. Proposed methodology

3.1 Data preprocessing

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 understands 1s and 0s.

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.”

Data Integration: In this step the addition of several databases, information files or information set is performed.

Data transformation: Aggregation and normalization is performed to scale to a specific value Data reduction: This section obtains a summary of the dataset which is very small in size but so far produces the same analytical result 1.

3.2 Stop words

“Stop words are the English words that do not add much meaning to a sentence.” They can be safely ignored without forgoing the sense of the sentence. For example, if it is tried to search a query like” How to make a veg cheese sandwich”, the search engine will try to search the web pages that contains the term “how”, “to”, “make”, “a”, “veg”, “cheese”, “sandwich”. The search engine tries to find the web pages that contains the term “how” ,”to”, ”a” than page containing the recipes of veg cheese sandwich because the terms “how”, “to”, “a” are so commonly used in English language ,If these three words are removed or stopped and actually focuses on retrieving pages that contains the keyword ” veg”, “cheese”, “sandwich” – that would give the result of interest.

3.3 Tokenization

“Tokenization is the process of splitting a stream of manuscript into phrase, symbols, words, or any expressive elements named as tokens.” The rundown of token further utilized for contribution for additional handling, for example, content mining and parsing. Tokenization is valuable in both semantics (where it is as content division), and as lexical examination in software engineering and building. It is occasionally hard to define what is intended by the term “word”. As tokenization happens at the word level. Frequently a token trusts on modest heuristics, for instance: Tokens are parted by whitespaces characters, like “line break” or “space”, or by “punctuation characters”. Every single neighboring string of alphabetic characters are a piece of one token; similarly, with numbers. White spaces and punctuation might or might not involve in the resulting lists of tokens.

3.4 Bag of words

“Bag of Words (BOW) is a method of extracting features from text documents. Further these features can be uses for training machine learning algorithms. Bag of Words creates a vocabulary of all the unique words present in all the document in the Training dataset.”

3.5 Classifiers

Classification is a form of data analysis that extracts the models describing important data classes. A classifier or a model is constructed for prediction of class labels for example:

“A loan application as risky or safe.”

Data classification is a two-step

- learning step (construction of classification model.) and

- a classification step

3.5.1 Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

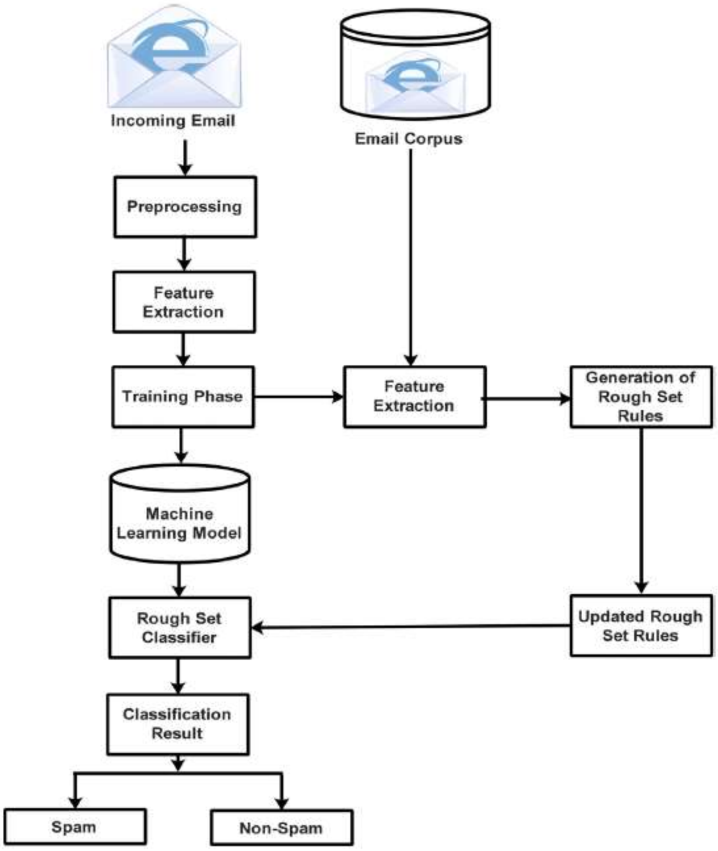


Fig. 2: Proposed email spam detection model.

- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

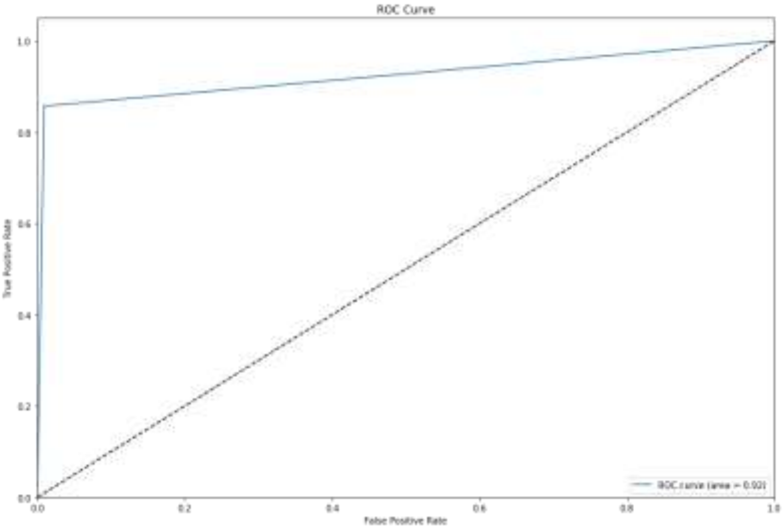
Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

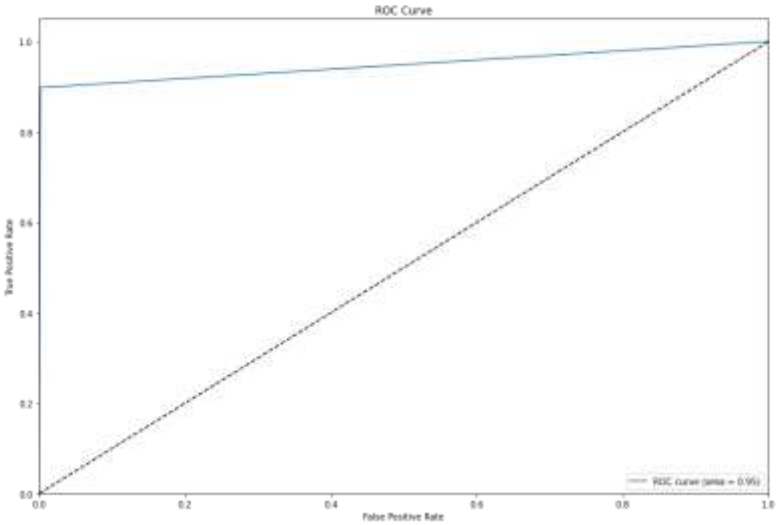
Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

4. Results



(a)



(b)

Fig. 4: ROC curve. (a) Naïve bayes classifier. (b) Logistic regression.

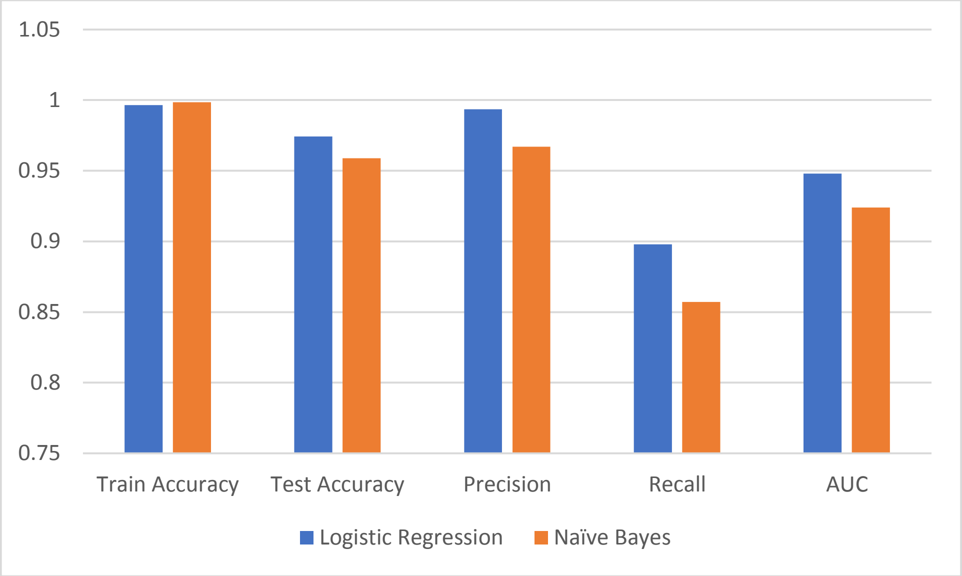


Fig. 5: Performance comparison of email spam detection algorithms.

5. Conclusion

This article proposed email spam detection and classification using machine learning algorithms such as Naïve bayes classifier and logistic regression classifier. In addition, data pre-processing, tokenization, and bag of word are also employed for extracting the features from the given data for enhanced classification accuracy. Obtained results disclosed that logistic regression performed superior in terms of training and testing accuracy, precision, recall and AUC as well. Moreover, ROC curve also demonstrated that proposed email spam detection using logistic regression classifier produced good enough results as compared to Naïve bayes classifier.

References

[1] Suryawanshi, S., Goswami, A., Patil, P. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74. 10.1109/IACC48062.2019.8971582.

[2] Karim, A., Azam, S., Shanmugam, B., Krishnan, K., Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detect ion. IEEE Access, 7, 168261-168295. https://doi.org/10.1109/ACCESS.2019.2954791

[3] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[4] Harisinghaney, A., A. Dixit, S. Gupta, A. Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliability, and Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014

[5] Mohamad, M., and Ali S., "An evaluation on the efficiency of hybrid feature selection in spam email classification." In Computer, Communications, and Control Technology (I4CT), 2015 International Conference on, pp. 227-231. IEEE, 2015.