INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 919-922

SPAM EMAIL DETECTION

P. Jaya Sri Valli[1], P. Hima Sai Naga Deepthi[2], P. S. Renuka Varma[3], S. Kavya Kranthi Rekha[4]

[1,2,3,4]Information Technology Department Jawaharlal Nehru Technological University, Kakinada Shri Vishnu engineering college for women Bhimavaram, India.

ABSTRACT

Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conduct, phishing and fraud. Sending malicious links through spam emails can harm our system and can also seek in into your system. Creating a fake profile and email account is much easier for the spammers, they pretend to be a genuine person in their spam emails, these spammers target those people who are not aware of these frauds. So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithms on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

Keywords: Ham/Spam, Email, Machine learning, Naïve Bayes, Online Webpage.

1. INTRODUCTION

Email 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 to receive those emails. "The popularity of using spam emails has been increasing since the 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 could be blocked manually coming from certain email addresses. A 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 message to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spam. 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 working 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 spam. "Unsolicited means those things which you didn't ask 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 are not considered spam".

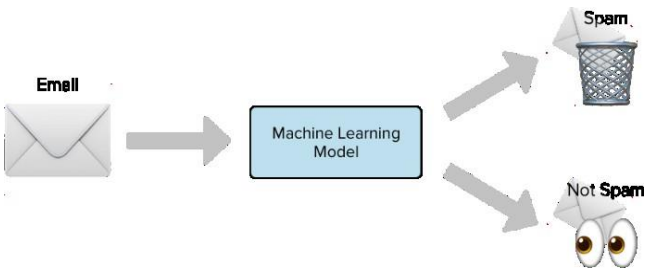

Fig.1. Classification into Spam and non-spam.

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. These algorithms include Naïve Bayes.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 919-922

2. LITERATURE REVIEW

In the field of email filtering, machine learning techniques have emerged as powerful tools for automatically categorizing emails as spam or non-spam. The seminal work by Sahami et al. [2] introduced a machine learning approach to email filtering based on message classification, laying the foundation for subsequent research in this area. Building upon this foundation, Androutsopoulos et al. [3] investigated the effectiveness of different variants of the Naïve Bayes algorithm for spam filtering, providing insights into the optimal choice of classifier for this task. Subsequent studies, such as Cormack and Lynam [4], have addressed real-world challenges in spam detection and highlighted the importance of robust machine learning models and data preprocessing techniques. More recently, Saeed et al. [5] applied supervised machine learning algorithms, including Naïve Bayes, to classify emails from the Enron-Spam dataset, demonstrating the continued relevance and efficacy of machine learning in email classification tasks.

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 “emails.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. METHODOLOGY

A. Data preprocessing:

When the data is considered, always a very large data set 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.

Steps in Data Preprocessing:

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.

Stop words:

“Stop words are the English words that do not add much meaning to a sentence.” They can be safely ignored without forgetting 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.

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 tokens is 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 be involved in the resulting lists of tokens.

Bag of words

“Bag of Words (BOW) is a method of extracting features from text documents. Further these features can be used 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.”

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 919-922

B. CLASSIC 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 steps

C. NAÏVE BAYES:

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 the 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 provides an accurate result. It is used in many fields like spam filtering.

This Algorithm contains the following steps:

a. Step 1: Consider a random email from the spam dataset for execution.

b. Step 2: The considered email is in basic form. To perform the feature extraction/selection and classification procedure, email is required to pre-process initially.

c. Step 3: Initially, tokenize the e-mail into individual keywords. Tokenization split each individual If the duplicate values are present within the dataset, then it'll drop the duplicate values Remove the stop words from the obtained tokens.

Now we will convert the group of text into a matrix of token counts Splitting the dataset into training data and test data.

d. Step 4: By evaluating the model on the training and testing dataset it predicts the accuracy of the model

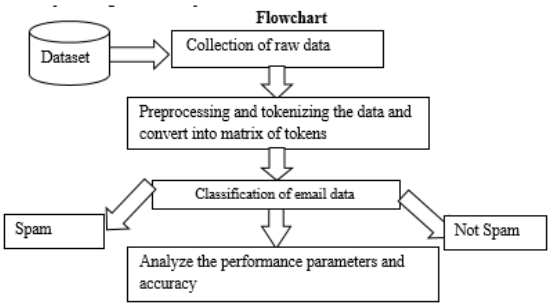

Fig.2. Flowchart of the model

D. IMPLEMENTATION

The implementation of spam email detection system involved a systematic approach leveraging Python programming language and various libraries and frameworks. We acquired the dataset, "emails.csv," which served as the foundation for our training and testing procedures. Preprocessing of the dataset was paramount, encompassing tasks such as data cleaning, text normalization, and feature extraction. Feature engineering played a crucial role in transforming raw text data into a format suitable for machine learning algorithms. Subsequently, we employed the Naïve Bayes classifier due to its simplicity and effectiveness in text classification tasks. Training of the classifier involved feeding it with labeled examples from the training dataset and optimizing its parameters using techniques such as cross-validation. We rigorously evaluated the performance of the classifier using metrics such as accuracy, precision, recall, and F1-score to ensure its effectiveness in distinguishing between spam and non-spam emails. Furthermore, we developed a user-friendly web interface using the Flask framework, allowing end-users to interact with the system seamlessly. The deployment of the system emphasized scalability, robustness, and real-time response, catering to the practical needs of email users. Overall, our implementation combined machine learning techniques with web development to create an efficient and reliable solution for combating email spam.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 919-922

4. RESULT

In this project, the Naïve bayes model is used for the best accuracy and this classifier will give its estimate results to the user. The dataset is achieved from "Kaggle" website for training. The name of the dataset used is "spam.csv". The two datasets training and testing data are compared based on the percentage of correctly identified spam and non-spam. The approach of the confusion matrix is the number of occurrences of each class for the dataset being considered.

For FP, FN, TP and TN, the average of dataset as follows:

FP: Total 8 number of misclassified spam emails.

FN: Total 1 number of misclassified spam emails.

TP: Total 268 spam messages are correctly classified as spam.

TN: Total 862 number of non-spam e-mail that is correctly classified as non-spam.

The Accuracy that is defined by evaluating the model on the training and testing dataset is 99% and the result is shown in the figure below.

TABLE 1. classification report

	precision	recall	f1-score	support
Ham	0.99	0.99	0.99	871
Spam	0.98	0.98	0.98	268
accuracy			0.99	1139
macro avg	0.99	0.99	0.99	1139
weighted avg	0.99	0.99	0.99	1139

confusion matrix

[[865 6] [5 263]]

5. CONCLUSION

Our research has presented a comprehensive approach to tackling the pervasive issue of email spam through the development of a machine learning-based detection system. Leveraging techniques such as preprocessing, feature engineering, and supervised learning with the Naïve Bayes classifier, we have demonstrated the effectiveness of our system in accurately distinguishing between spam and legitimate emails. Our evaluation metrics, including accuracy, precision, recall, and F1-score, indicate promising results, showcasing the system's robustness and reliability. Furthermore, the deployment of the system in a user-friendly web interface underscores its practical usability and potential for real-world application. While our research has made significant strides in combating email spam, there remain opportunities for future enhancements, including the integration of advanced machine learning techniques and the adaptation of the system to evolving spam tactics. Overall, our work contributes to the ongoing efforts in enhancing email security and improving user experience in managing email communications in an ever-evolving digital landscape.

6. REFERENCES

[1] 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.

[2] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Machine Learning Approach to Email Filtering Based on Message Classification. Proceedings of the 1998 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4, 23-30.

[3] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., & Stamatopoulos, P. (2000). Spam Filtering with Naive Bayes – Which Naive Bayes? CEAS 2000.

[4] Cormack, G. V., & Lynam, T. R. (2007). Learning to Detect Spam in the Real World. CEAS 2007.

[5] Saeed, A., Tafreshi, S., & Abadeh, M. S. (2019). Enron-Spam Email Classification using Supervised Machine Learning Algorithms. Journal of Information Science, 45(1), 22-32.