WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Email Phishing Attacks Detection Using Random Forest Algorithm

Dr. D. J Samatha Naidu[1], M. Gowri[2]

[1]Principal, Department of MCA Annamacharya PG College of Computer Studies, Rajampet. Email: samramana44@gmail.com
[2]MCA Student, Dept. of MCA Annamacharya PG college of Computer Studies, Rajampet. Email: Gowrireddymooli79@gmail.com

A B S T R A CT

The prevalence of email spam has increased significantly in recent years, along with the exponential rise in internet users. They are being utilized for fraud, phishing, and other unethical and criminal activities. sending malicious links through spam emails that could damage our system and gain access to yours. By posing as a real person in their spam emails, spammers may quickly construct a fake profile and email account. People who are not aware of these scams are the targets of these spammers. Therefore, it is important to recognize spam emails that are bogus. These spam emails will be identified by this initiative using machine learning methods. The algorithms used in machine learning will all be covered in this article and applied to our data.

Keywords: Emails, email spam messages, machine Learning.

I. Introduction

Sending unsolicited or commercial emails to a list of subscribers is known as email spam, also known as electronic mail spam. Unsolicited emails are those that have not been requested to be sent to the receiver. Utilizing spam emails has gained popularity over the past ten years. Online spam has become a serious issue. Spam is a message delivery, space, and time waster. Even if automatic email filtering may be the most effective method of preventing spam, contemporary spammers may easily circumvent all of these tools.

EXISTING WORK

Spam emails have increased in volume over the past few years, along with the growth in email subscribers. Handling a large variety of emails for data mining and machine learning has recently become much more difficult. In order to compare how well different classification algorithms work and how accurately they classify emails using a variety of performance criteria, numerous academics have conducted comparison studies. Therefore, it's critical to identify an algorithm that produces the best results for each given parameter for the accurate classification of emails as either spam or ham.

II. Related work

The increasing reliance on email for communication and business transactions, the issue of email spam has become a persistent problem. Unsolicited emails not only clutter our inboxes but also pose significant security risks. To combat this challenge, researchers and experts have developed various techniques to detect and filter out spam emails. This article explores the related work in the field of email spam detection using machine learning algorithms.

AUTHORS

Suryawanshi, Shubhangi& Goswami, Anurag & Patil, Pramod

It is preferred to use recent hardware and software developments for email communications. However, the unwanted emails have a negative impact on communication. The demand for spam email detection and classification is present. Models are created for the current research on email spam detection and classification. We have employed a variety of machine learning classifiers, including ensemble classifiers with a voting mechanism, Naive Bayes, SVM, KNN, Bagging and Boosting (Adaboost), and SVM. On the email spam dataset from the UCI Machine Learning repository and the Kaggle website, classifiers are evaluated and tested. Various accuracy measures are employed, including Accuracy Score, F measure, Recall, Precision, Support, and ROC. The initial findings indicate that an ensemble classifier with a voting mechanism is the most effective. It generates the fewest false positives.

PROPOSED WORK

The data is always thought of as very huge data sets with several rows and columns. However, this is not always the case because the data could also be in the form of image, audio, or video files. Detailed tables, etc. Machines simply comprehend 1s and 0s; they are incapable of understanding photos, video, or text data. Data preprocessing steps: cleansing data The tasks of "filling in missing values," "smoothing noisy data," "identifying or removing outliers," and "resolving inconsistencies" are completed in this step. Integration of data: Several databases, information files, or information sets are added in this step. Transformation of data Scaling up to a particular value is accomplished by aggregation and normalization.

MODULES

USER

ADMIN

ALGORITHM

Random Forest Algorithm

Step1:-Insert the dataset or file for training or testing.

Step2:- Check the dataset for supported encoding.

Step3:- If one of the supported encodings, then go to step

Step4:- If not one of the supported encoding, then go to step

Step5:- Change the encoding of the inserted file into one of the supported encodings. Then try again for reading.

Step6:-Select whether you want to “Train”, “Test’ or “Compare” the models using the dataset.

Step7:- If “Train” is selected, then go to step

Step8:- If “Test” is selected, then go to step

Step9:- If “Compare” is selected, then go to step

Step10:- “Train” selected:

Step11:- Select which classifier to train using the inserted dataset. Step12:-Check for duplicates and NAN values.

Step13:- Find the values from Hyper parameter Tuning.

Step14:- Process the text for feature transform.

Step15:- Train the model

Step16:- Save the model and features. Show the results.

Step17:- Select which classifier to test using the inserted dataset.

Step18:- Check for duplicates and NAN values.

Step19:- Load the model and features saved in the training phase of the model.

Step20:- Using the loaded values for testing the dataset.

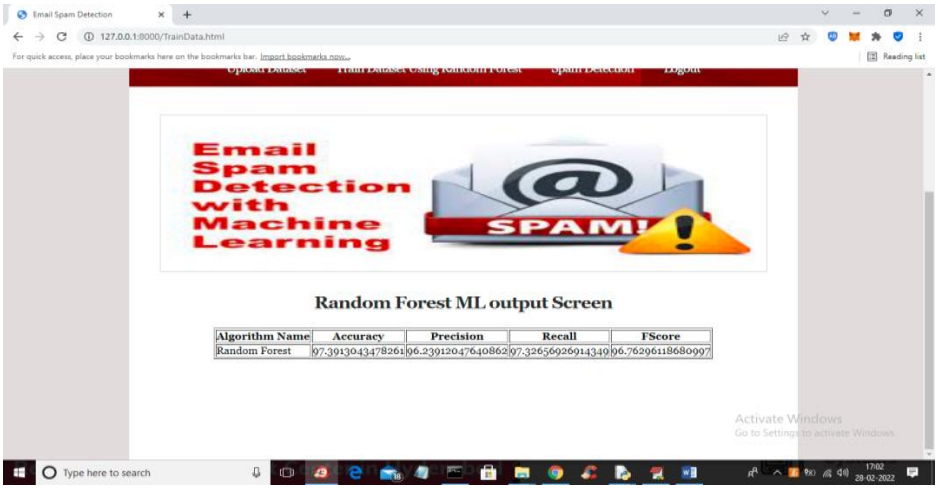
Step21:-Show the results 1.6. “Compare” selected:

Step22:- Compare all the classifiers using the inserted dataset.

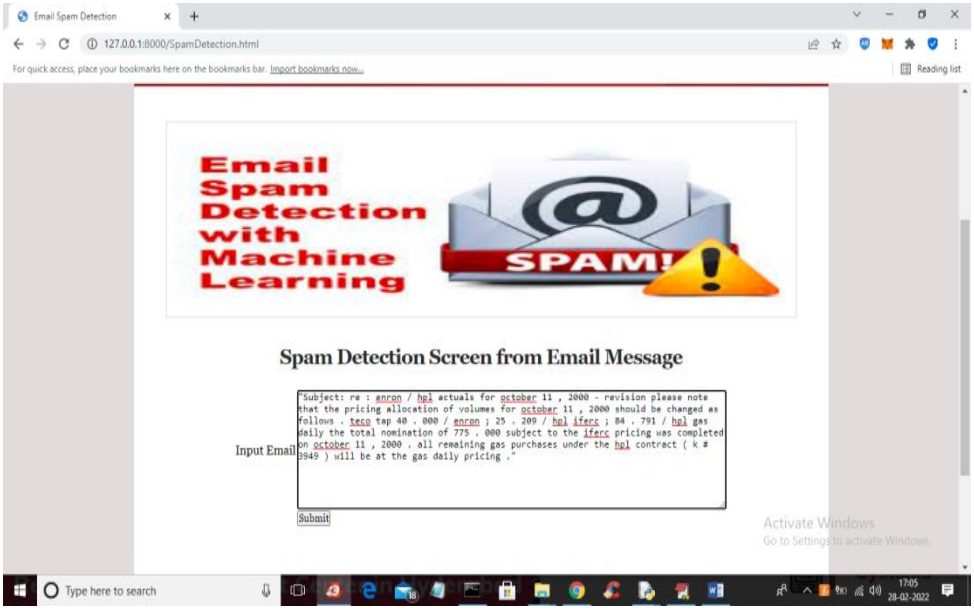
Step23:- Show the results of the classifiers.

Sample Screens:-

Screen1:-User Login Screen



Screen2:-Random Forest Algorithm



Screen3:-Spam detection screen from email message

Screen4:-View Dtection Screen results

Conclusion

With this output, it is clear that multinomial naive Bayes produces the greatest results, but it has limitations caused by class-conditional independence, which causes the machine to misclassify some tuples. On the other hand, ensemble approaches have been shown to be effective because they combine several classifiers to predict classes. Nowadays, a large number of emails are sent and received, which makes it challenging for our project because it can only test emails using a small corpus. Our project's spam detection is capable of filtering emails based solely on the content of the message, rather than the domain names or any other factors.

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. Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. 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, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014

5. Mohamad, Masurah, and Ali Selamat. "An evaluation on t he efficiency of hybrid feature selection in spam email classification." In Computer, Communications, and Control T echnology (I4CT), 2015 International Conference on, pp. 227 -231. IEEE, 2015

6. Shradhanjali, Prof. T oran Verma “E-Mail Spam Detection and Classification Using SVM and Feature Extraction”in International JouranlOf Advance Reasearch, Ideas and Innovation In T echnology,2017 ISSN: 2454-132X Impact factor: 4.295

7. W.A, Awad& S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology. 3. 10.5121/ijcsit.2011.3112.

8. A. K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, T urkey, 2018, pp. 1-4.

9. Diren, D.D., Boran, S., Selvi, I.H., &Hatipoglu, T . (2019). Root Cause Detection with an Ensemble Machine Learning Approach in the Multivariate Manufacturing Process.

10. T asnim Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman. "Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach”, 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018