IJMRSET

MULTIDISCIPLINARY · SCIENCE · ENGINEERING · TECHNOLOGY

e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 10, October 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

☎ 6381 907 438 6381 907 438 ✉ ijmrset@gmail.com @ www.ijmrset.com

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710052

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Spam E-Mail Classification Using Machine Learning

Geetha R[1], Yokesh Waran G[2], Surya A[2]

Assistant Professor, Department of CSA, The Oxford College of Science, Bangalore, India

PG Student, Department of CSA, The Oxford College of Science, Bangalore, India

PG Student, Department of CSA, The Oxford College of Science, Bangalore, India

ABSTRACT: Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about 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 algorithm on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

I. INTRODUCTION

In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world. Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam

No one wants to receive emails not related to their interest because they waste receivers' time and resources. Besides, these emails can have malicious content hidden in the form of attachments or URLs that may lead to the host system's security breaches Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. So, it requires an immense demand for the security of the email system. Spam emails may carry viruses, rats, and Trojans. Attackers mostly use this technique for luring users towards online services. (may send spam emails that contain attachments with the multiple-file extension, packed URLs that lead the user to malicious and spamming websites and end up with some sort of data or financial fraud and identify theft. Many email providers allow their users to make keywords base rules that automatically filter emails. Still, this approach is not very useful because it is difficult, and users do not want to customize their emails, due to which spammers attack their email accounts. In the last few decades, Internet of things (IoT) has become a part of modern life and is growing rapidly. IoT has become an essential component of smart cities. (ere are a lot of IoT-based social media platforms and applications.

II. TERATURE REVIEW

Spam emails, often characterized as unsolicited and irrelevant messages, have evolved significantly over the past two decades, becoming a persistent challenge for both individuals and organizations. The phenomenon of spam has prompted extensive research across various disciplines, including computer science, psychology, and information systems.

Historical Context
Early studies primarily focused on the technical aspects of spam, exploring methods for detection and prevention. Research by SpamAssassin (2001) highlighted the use of heuristic algorithms and Bayesian filtering techniques to categorize spam emails effectively. These foundational approaches laid the groundwork for more sophisticated machine learning algorithms employed today.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Psychological and Behavioural Aspects
Beyond technical solutions, researchers have also examined the psychological impact of spam on users. A study by Zhang et al. (2013) analysed user behaviour in response to spam emails, revealing that individuals often exhibit varying levels of susceptibility based on personality traits and past experiences with spam. This insight underscores the importance of understanding user psychology in developing effective countermeasures.

Socioeconomic Implications
The socioeconomic ramifications of spam emails have been another area of focus. A report by the Radicati Group (2019) estimated that spam costs businesses billions annually in lost productivity and IT resources. Furthermore, the proliferation of phishing scams within spam campaigns has raised concerns about cybersecurity and data privacy, prompting regulatory responses from governments worldwide.

Technological Advances
Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have revolutionized spam detection methods. Research by Kaur et al. (2020) demonstrated that deep learning models could achieve higher accuracy rates in classifying spam emails compared to traditional methods. This shift highlights the ongoing arms race between spammers and security technologies.

III. METHODOLOGY

DATA UNDERSTANDING

- The email spam classifier focuses on either header, subject, and content of the email. In this project, we are focusing mainly on the subject and content of the email.
- The dataset contains two columns. The total corpus of 5728 documents. The descriptive feature consists of text. The target feature consists of two classes ham and spam, the column name is spam. The classes are labeled for each document in the data set and represent our target feature with a binary string-type alphabet of {ham; spam}. Classes are further mapped to integer 0 (ham) and 1 (spam).

Optimization and Fine-tuning. The detection model may undergo optimization and fine-tuning to improve performance. This can involve adjusting hyperparameters, exploring different feature selection techniques, or incorporating additional pre-processing steps. The optimization process aims to enhance the detection accuracy and robustness of the model.

Deployment and Application. Based on this score, the system decides a simple yes or no or a ranking of probabilities for multiple individuals. Once the detection model demonstrates satisfactory performance, it can be deployed for real-world applications. This may involve integrating it into existing face recognition systems or developing standalone tools for detecting automatic face recognition.

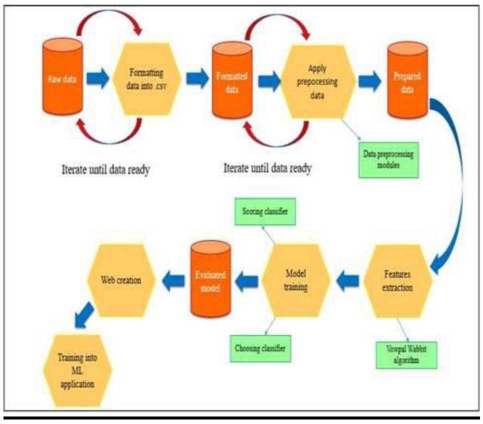


Fig .1 Architecture Diagram For Spam Email

PREPROCESSING

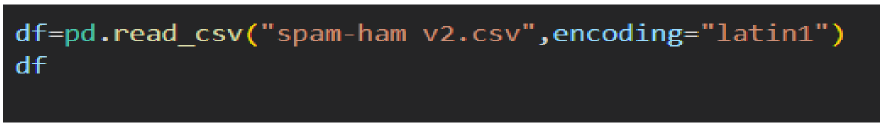
- In the preprocessing step the documents are transformed from the raw document to a structured document that is intended to contain as much information as possible without discrepancies that can affect the prediction result. A common method to increase the information density of a document is to remove the words that are very common and rarely has any significance, often referred to as stop words. These are word such as “the”, “are”, “of”, which are insignificant in a larger context. In BoW these are a list of predetermined words, but word2vec take a probabilistic approach, called subsampling, which avoid overfitting on the most frequent words. For each instance of a word in word2vec a probability of removal is decided by equation 2.1 where t, usually set to 10−5, is the threshold and f(wi) is the word frequency [14].

EMAIL DATASET

- The email dataset used during the experiments consists of 105,195 emails from a real life telecommunication support environment. The emails contains support errands about invoices, technical issues, number management, admin rights etc. The emails are classified with one or more labels. There is in total 33 different labels with varying frequency as shown in image 3.1. The label “DoNotUnderstand” is an artefact from the rule based system where the email did not match any rule, this label is filtered out and is not used during training or testing. The dataset contains 31,700 emails with the label “DoNotUnderstand”. This results in a classification rate of 69,9% with the currently implemented manual rule based model. The figure also show a major class imbalance, however no effort were made to balance this since those are the relative frequencies that will be found in the operative environment.

- The email labels can be aggregated into queue labels which is an abstraction of the 33 labels into 8 queue labels. The merger is performed by fusing emails from the same email queue, which is a construction used by the telecommunication company, into a single queue label. The labels that are fused together are often closely related to each other, which effectively will reduce the amount of conflicts between the email labels and their contents. If an email contain two or more labels it is disregarded since it might introduce conflicting data which is unwanted when training the classifier. Without “DoNoUnderstand” and the multilabel emails there are a total of 58,934 emails in the dataset.

DATASET AND SETUP



```
df=pd.read_csv("spam-ham v2.csv",encoding="latin1")
df
```

Now it’s time to see how to start the pandas profiling library and generate the report out of the data frames. First things first, let’s **import a dataset** for which we will be generating reports. I am using the **agriculture dataset** which contains the State, name, District, name, Crop, year, Season, Area, and Production.

IV. RESULTS

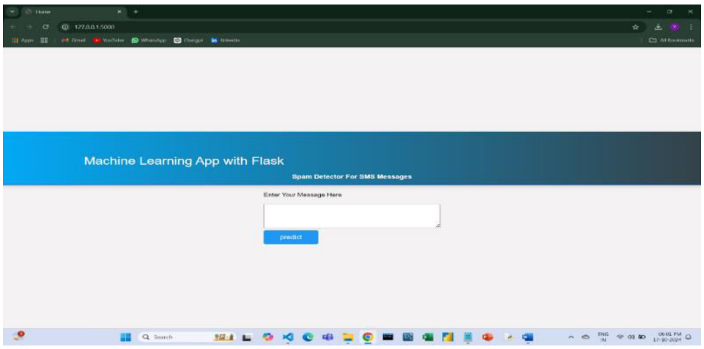
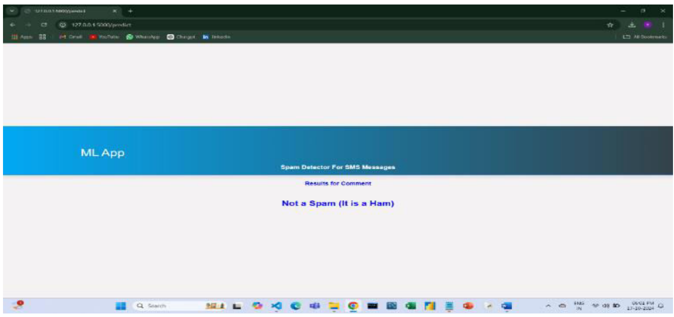

Machine Learning App with Flask
Spam Detector For SMS Messages
Enter Your Message Here
predict

Fig.2 Entering The Data



ML App

Spam Detector For SMS Messages

Results for Comment

Not a Spam (It is a Ham)

Fig.3 Predicting the message.

V. CONCLUSION

- After successful email verification, a complete report explaining the status of each email will be generated The generated report will be very transparent and easy to understand. The minology used in your email verification report is explained below.
- Every email is primarily categorized as either Valid, Invalid or Unknown
- Valid-After a successful SMTP transaction, if receiving mail server accepts the email address, it will be marked Valid Email Address Sending emails to the addresses which are marked as "Accept All" or "Disposable" is not recommended, even though the email address is valid.
- Invalid-An email address will be marked Invalid if it is syntactically incorrect or an email account does not exist on receiving mail server.
- Unknown-Sometimes receiving mail server responds very slowly or may temporarily be unable in process requests. In this case, the email address will be marked Unknown and email verification won't be counted (This makes Quick Email Verification different from other service providers) In most cases, such emails can be re-verified after 5 to 10 minutes. The results will show whether the email address is Valid or Invalid.

REFERENCES

1. Yang, Y., & Liu, X. (1999) "A Review of Text Classification Techniques." ACM SIGKDD Explorations Newsletter, 1(1), 40-48.
2. Mailham, S., &Ghazizadeh, M. (2010). "A Review of Spam Detection Methods." International Journal of Computer Applications, 1(21), 10-15.
3. Hodge, V. J., & Austin, J. (2004). "A Survey of Outlier Detection Methodologies." Artificial Intelligence Review, 22(2), 85-126.
4. Mishra, A., & Sinha, A. (2016). "A Review on Machine Learning Techniques for Email Spam Detection." International Journal of Computer Applications, 141(7), 13-20.
5. García, M., & Torres, M. (2018). "Spam Detection in Email Using Machine Learning Algorithms." Journal of Computer and Communications, 6(2), 10-20.
6. Alaa, M., & Abdelhalim, M. (2019). "Spam Email Detection Using Machine Learning Techniques." International Journal of Computer Applications, 178(21), 1-5.
7. Kaur, R., & Arora, A. (2019). "A Survey of Spam Detection Techniques in Email." International Journal of Advanced Research in Computer Science, 10(2), 16-20.
8. Siddiqui, A., & Shafi, M. (2020). "Email Spam Detection using Machine Learning Algorithms: A Review." Journal of Information and Communication Technology, 19(2), 287-299.
9. Alharbi, A., & Shafique, U. (2020). "Deep Learning Approaches for Spam Detection: A Review." Journal of King Saud University - Computer and Information Sciences.
10. Vishwakarma, P., & Rai, A. (2021). "Email Spam Detection using Hybrid Machine Learning Approach." Journal of Engineering Science and Technology Review, 14(2), 18-26.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com