International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

E-MAIL SPAM DETECTION

Jayashree Lavhare[1], Shreya Kosabe[2], Pooja Kolge[3],Pranita Wagh[4],Akshay Ghuge[5]

[1]*Jayashree Lavhare Information Technology Department & P .G. Moze College of Engineering*

[2]*Shreya Kosabe Information Technology Department & P.G.Moze College of Engineering*

[3]*Pooja Kolge Information Technology Department & P.G.Moze College of Engineering*

[4]*Pranita Wagh Information Technology Department & P.G.Moze College of Engineering*

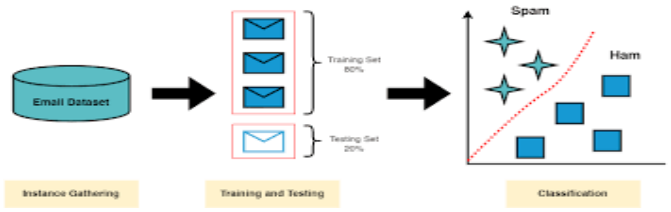
[5]*Akshay Ghuge Information Technology Department & P.G.Moze College of Engineering*

Abstract - Emails have become a ubiquitous means of personal and professional communication, often containing sensitive and confidential information. However, they are also prime targets for cybercriminals who employ techniques such as phishing to obtain private data. This paper proposes an intelligent and efficient email spam detection system that leverages data mining algorithms for classification and association. By extracting multiple features from email content, we improve classification accuracy and efficiency. The system integrates various machine learning algorithms and achieves a 30% reduction in the error rate compared to existing methods. Our approach enhances email spam detection by combining support vector machines with multi-feature extraction and classification.

Key Words: Support vector machines, Email Spam, Phishing Detection, Machine Learning, Multi-Feature Extraction.

1.INTRODUCTION

The ubiquity of emails in personal and professional communication has led to an increasing concern about the security of sensitive information shared via this medium. Cybercriminals exploit email platforms to conduct phishing attacks, attempting to trick individuals into revealing confidential data. Phishing is a lucrative type of fraud in which the criminal deceives receivers and obtains confidential information from them under pretenses. Loss of personal information if any user clicks on phishing email links. Phisher writes the program in such a way that the user's personal information is easily accessible to him/herThis paper addresses the development of an advanced email spam detection system that utilizes machine learning techniques and multi-feature extraction to enhance accuracy and efficiency in identifying phishing emails. In our model, supervised machine learning algorithms are used for classification. Supervised learning algorithms predict the nature of unknown data based on known examples.



The main bottleneck in electronic communications is the immense diffusion of unwanted, dangerous emails referred to as spam emails. A key concern is the development of acceptable filters that will sufficiently capture those emails and obtain a high-performance rate. Machine learning (ML) researchers have developed various approaches to this drawback. Inside the framework of machine learning, Fuzzy, Artificial Neural Network (ANN) and Random Forest have ready an outsized half to the event of spam email filtering. Various papers have been completed within the field of spam detection on high social networks. All of those studies have raised one or alternative options for spam detection. Some articles used for social networks have been written, and numerous contain completely different networks. Additionally, many have written on spam user accounts detection, and others were concerning spam partition post detection within social networks

2.Literature Survey

1.Yaseen et al. (2020): Introduced image analysis for spam detection, emphasizing a multidimensional approach.

2.Mohammed et al. (2019): Highlighted the importance of language diversity and visual information in spam detection.

3.Gibson et al. (2020): Used bio-inspired metaheuristic algorithms to optimize machine learning for improved spam detection.

4. Nandhini and Jeen (2020): Evaluated the performance of various machine learning algorithms and incorporated bio-inspired techniques for enhanced classifier performance.

5.Govil et al. (2020): Proposed a systematic machine learning-based spam detection mechanism focusing on feature dictionaries.

6.Chandra et al. (2019): Explored spear phishing techniques, introducing ensemble approaches for increased accuracy.

7. Bibi et al. (2020):Assessed machine learning algorithms, with a focus on naive Bayes, for reasonable accuracy and precision in spam detection.

8.Elshoush and Dinar (2019): Analyzed Adaboost and Stochastic Gradient Descent (SGD) algorithms for high accuracy and low false-positive rates.

9.Olatunji (2019): Proposed an improved spam detection model based on support vector machines, highlighting the importance of parameter tuning.

10. Hussain et al. (2019): Explored spam review detection techniques, emphasizing the significance of detection beyond traditional email domains.

Collectively, these studies contribute diverse methodologies, innovative algorithms, and valuable insights to the ongoing efforts in enhancing email spam detection.

3. Objectives

To design and develop an approach for email phishing detection from large synthetic as well as real-time data using machine learning.

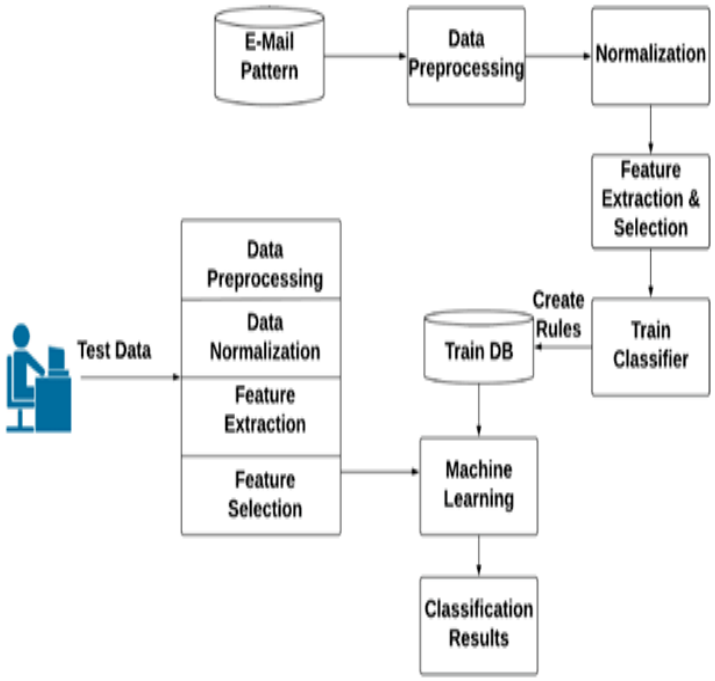
To develop an approach using various machine learning algorithms and explore the accuracy using the majority routing technique.

To develop an algorithm for extracting different kinds of features from emails to achieve better classification accuracy.

To validate and explore the system classification results with existing detection techniques.

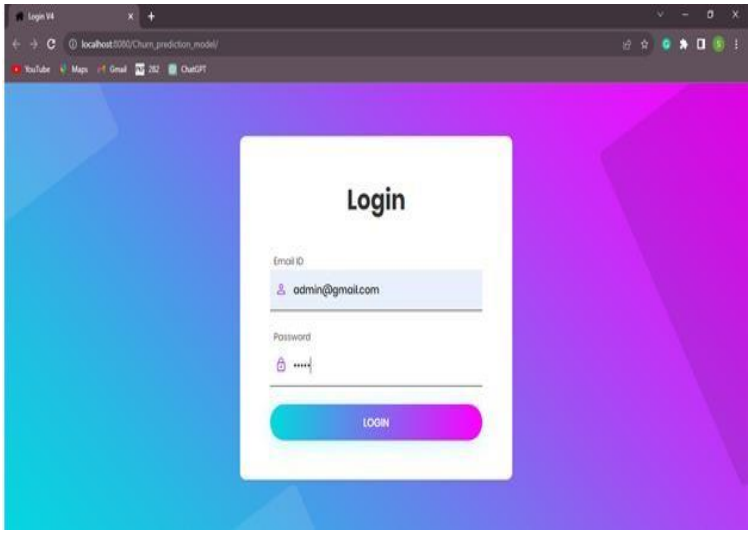
4. System Architecture

In the proposed system, detecting phished emails can be described as a classification problem with two categories i.e., ham and phished. Machine Learning is a field of artificial intelligence in which the system is given the ability to learn without being explicitly programmed. In our model, supervised machine learning algorithms are used for classification. Supervised learning algorithms predict the nature of unknown data based on known examples. These algorithms are a subset of machine learning algorithms that iteratively learn from data.

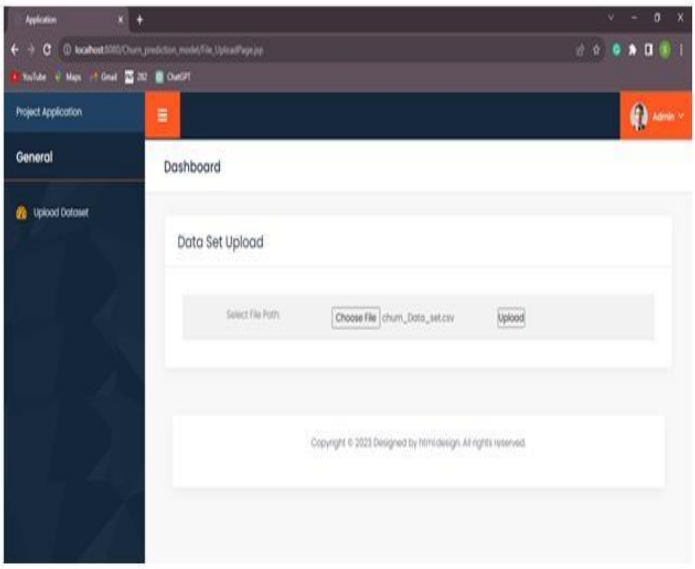


5.SNAPSHOTS

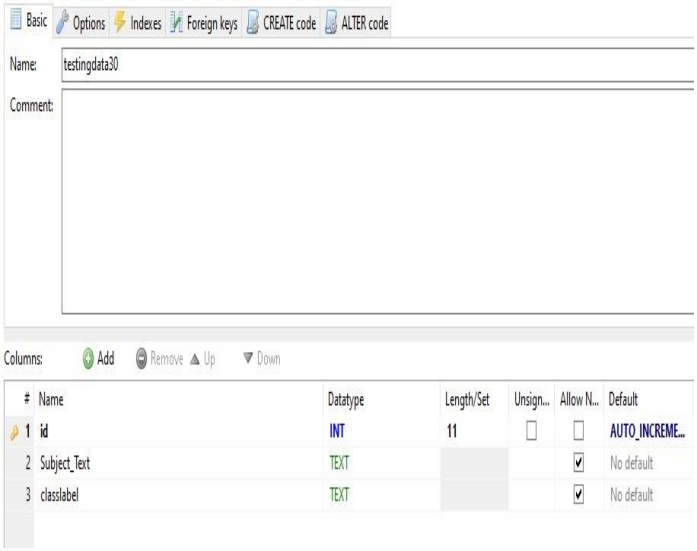
GUI login Page



File Upload Page



Database



6. CONCLUSION

In conclusion, our research project presents a comprehensive and advanced approach to email spam detection, integrating cutting-edge technologies such as supervised machine learning, support vector machines, and multi-feature extraction. By combining these methodologies, our system achieves a notable reduction in error rates, offering an efficient and accurate defense against the evolving landscape of phishing attacks.. The integration of support vector machines as a central classification tool enhances the system's ability to discern intricate patterns within emails, contributing to its robust performance .Our proposed system architecture follows a systematic process, from data collection and preprocessing to training, testing, and classification. The incorporation of hybrid machine learning algorithms and the generation of training rules result in a significant improvement over existing systems, marking a substantial stride in email security .The literature survey not only validates but enriches our approach by exploring diverse methodologies and innovations from prior studies. Insights from bio-inspired algorithms, ensemble methods, and extending spam detection beyond conventional email domains contribute to a holistic understanding of the spam landscape.

In essence, our project's efficancy, demonstrated through reduced error rates and enhanced accuracy, positions it as a valuable contribution to the ongoing efforts in mitigating email spam and phishing threats. As the digital landscape continues to evolve, our comprehensive solution stands as a robust foundation, offering insights and directions for future advancements in securing electronic communication.

6.ACKNOWLEDGEMENT

Our heartfelt gratitude extends to all those who played a pivotal role in guiding us through the successful completion of this system, an integral part of our bachelor's course syllabus.We express sincere thanks to our cooperative department for their unwavering support, providing invaluable assistance, and furnishing the necessary equipment crucial for the development of the system. Their collaborative efforts significantly contributed to the realization of this project.

A special note of appreciation goes to **Prof. Jayashree Lavhare** for being an exceptional guide. His valuable guidance, inspiration, and wholehearted involvement at every stage of the project were instrumental in bringing our concept to fruition. His professional knowledge and experience enriched the project, making it a success.

We are also thankful to **Prof. Abidali Shaikh, Head of Department – Information Technology**, for his constant enlightenment, unwavering support, and motivational guidance. His insights played a crucial role in the successful completion of our project.

Our deep appreciation is extended to **Dr. Navnath Narawade, Principal of PGMCOE, Wagholi**, for his encouragement and for providing us with the opportunity and facilities necessary to carry out this work. His support was instrumental in shaping the project's trajectory.

Finally, we express profound gratitude to our parents, friends, and well-wishers. Their continuous support, suggestions, and assistance were invaluable throughout this journey.

7. References

[1] Yaseen, Yaseen Khather, Alaa Khudhair Abbas, and Ahmed M. Sana. "Image spam detection using machine learning and natural language processing." Journal of Southwest Jiaotong University 55.2 (2020).

[2] Mohammed, Mazin Abed, et al. "An anti-spam detection model for emails of multi-natural language." Journal of Southwest Jiaotong University 54.3 (2019).

[3] Gibson, Simran, et al. "Detecting spam email with machine learning optimized with bio-inspired metaheuristic algorithms." IEEE Access 8 (2020): 187914-187932.

[4] Nandhini, S., and Jeen Marseline KS. "Performance evaluation of machine learning algorithms for email spam detection." 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). IEEE, 2020.

[5] Govil, Nikhil, et al. "A machine learning based spam detection mechanism." 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2020.

[6] Chandra, J. Vijaya, Narasimham Challa, and Sai Kiran Pasupuletti. "Machine learning framework to analyze against spear phishing." Int. J. Innov. Technol. Exploring Eng. (IJITEE) 8 (2019): 12.

[7] Bibi, Asma, et al. "Spam mail scanning using machine learning algorithm." J. Comput. 15.2 (2020): 73-84.

[8] Elshoush, Huwaida T., and Esraa A. Dinar. "Using adaboost and stochastic gradient descent (sgd) algorithms with R and orange software for filtering e-mail spam." 2019 11th Computer Science and Electronic Engineering (CEEC). IEEE, 2019.

[9] Olatunji, Sunday Olusanya. "Improved email spam detection model based on support vector machines." Neural Computing and Applications 31.3 (2019): 691-699.

[10] Hussain, Naveed, et al. "Spam review detection techniques: A systematic literature review." Applied Sciences 9.5 (2019): 987.