Spam analysis and classification of the dynamic message using a vectorizing technique using NLP

Tushar Gupta
SRM Institute of Science and Technology
Networking and Communications
Tamil Nadu, India
tg1340@srmist.edu.in

SahilPandey
SRM Institute of Science and Technology
Networking and Communications
Tamil Nadu. India
sp9774@srmist.edu.in

Dr. Thenmalar S
SRM Institute of Science and Technology
Networking and Communications
Tamil Nadu, India
thenmals@srmist.edu.in

Abstract - Spam analysis and classification of dynamic messages is an essential task in order to combat the ever increasing volume of unsolicited and malicious emails. One effective approach is to employ a vectorizing technique along with a multi-model machine learning algorithm. This approach involves representing email messages as high-dimensional vectors, capturing various features such as word frequencies, presence of specific keywords, and structural characteristics. By transforming the text into numerical representations, the machine learning algorithm can then learn patterns and make predictions based on these representations. The use of a multi-model algorithm allows for the integration of different classification models, each with its own strengths and weaknesses, to enhance the overall performance. This approach can achieve high accuracy by leveraging diverse learning methods and combining their predictions. Furthermore, the approach is dynamic in nature, meaning that it can adapt to new forms of spam and evolving attack strategies. The key challenge lies in selecting appropriate features and tuning the parameters of the algorithm to ensure optimal performance. The results of this study can contribute to the development of more effective and efficient spam detection systems, helping users to filter out unwanted and potentially harmful messages. Vectorizing and multi-model machine learning methods for spam message dynamic analysis and categorization are novel features of this study. This research vectorizes spam communications into high-dimensional representations to better grasp their content and context than static feature extraction methods. The system can adapt to different spam messages and increase its classification accuracy by using a variety of machine learning models, such as neural networks, decision trees, and ensemble approaches. This dynamic and comprehensive method will improve spam identification and categorization, improving email filtering and cybersecurity. Despite the advancements in spam analysis, existing systems face challenges in accurately distinguishing between legitimate and spam messages. The sheer volume and evolving nature of spam necessitate sophisticated methods. Traditional approaches often struggle to adapt to new tactics employed by spammers, leading to false positives or negatives.

I. INTRODUCTION

Spam analysis and classification of dynamic messages have become critical in the realms of information security and data analysis. With the escalating volume of unsolicited and malicious messages, developing efficient methods for spam detection is imperative. Advanced techniques such as vectorizing and multi-model machine learning algorithms are crucial in addressing this challenge.

The concept of vectorizing involves converting text-based data into numerical representations for processing by machine learning algorithms. This technique, proven effective in natural language processing tasks, enables the extraction of important patterns and characteristics. Vectorizing allows the inclusion of both message content and metadata, providing a comprehensive analysis of dynamic messages. Metadata, encompassing sender details, time of sending, and message length, enhances the ability to distinguish between legitimate and spam messages. By incorporating content and metadata, a more reliable classification model can be constructed.

Despite the advancements in spam analysis, existing systems face challenges in accurately distinguishing between legitimate and spam messages. The sheer volume and evolving nature of spam necessitate sophisticated methods. Traditional approaches often struggle to adapt to new tactics employed by spammers, leading to false positives or negatives. Additionally, the reliance on single-model algorithms may limit the overall effectiveness of the system. Addressing these challenges is paramount for ensuring robust spam detection capabilities.

The workflow of the proposed system involves the initial step of vectorizing dynamic messages. This includes converting text data into numerical representations, incorporating both message content and metadata. The vectorized data is then utilized in training multi-model machine learning algorithms. These algorithms, leveraging diverse features such as statistical measures, keyword frequencies, and linguistic patterns, collectively contribute to accurate spam classification. Evaluation of the system will involve assessing its performance on a dataset containing a mix of legitimate and spam messages. Metrics such as precision, recall, and F1 score will be employed to quantify the system's effectiveness.

II. Literature Review

"Intelligent detection: a classification-based apporach to e-mail (text) filtering." done by Reguig, Sara Azzouz defines spam analysis and classification of dynamic messages using vectorizing techniques and multi-model machine learning algorithms. Through this research, a comprehensive framework is developed to enhance the accuracy and efficiency of email filtering systems by incorporating advanced methods for detection and classification. The study contributes valuable insights into the field of email security and showcases the potential for improving spam detection mechanisms in real-time settings. Reguig, Sara Azzouz's work highlights the significance of employing cutting-edge technologies for effective email filtering and emphasizes the importance of continuous innovation in combating spam.

"Artificial Intelligence Methods in Email Marketing-A Survey Check for updates." done by Jach, Anna proposed that by utilizing a vectorizing technique in conjunction with multi-model machine learning algorithms, the study delves into enhancing email marketing strategies. This gexploration was presented at the Eighteenth International Conference on Dependability of Computer Systems DepCoS-RELCOMEX in Brunów, Poland in July 2023. The work published in Springer Nature's Proceedings addresses the evolving landscape of email marketing through innovative AI approaches.

"Deep learning for phishing detection: Taxonomy, current challenges and future directions." proposed by Do, Nguyet Quang, Ali Selamat, Ondrej Krejcar, Enrique Herrera-Viedma, and Hamido Fujita. explains the comprehensive study ""Deep Learning for Phishing Detection: Taxonomy, Current Challenges, and Future Directions"" published in IEEE Access, Do, Nguyet Quang, Ali Selamat, Ondrej Krejcar, Enrique Herrera-Viedma, and Hamido Fujita share valuable insights on spam analysis and classification of dynamic messages. Their research delves into the utilization of vectorizing techniques coupled with multi-model machine learning algorithms for more effective spam detection. This scholarly work addresses pressing issues in the field, proposing innovative solutions and paving the way for future advancements in combating phishing activities.

"Comparative analysis on deep neural network models for detection of cyberbullying on Social Media." done by Balakrishna, Sivadi, Yerrakula Gopi, and Vijender Kumar Solanki co-authored a research paper titled ""Comparative analysis on deep neural network models for detection of cyberbullying on Social Media"" published in the journal Ingeniería Solidaria. Their study focused on spam analysis and classification of dynamic messages using a vectorizing technique with multi-model machine learning algorithms. The research aimed to explore the efficacy of deep neural network models for detecting cyberbullying on social media platforms, emphasizing the importance of advanced machine learning approaches in addressing online harassment and promoting a safer digital environment.

"Twenty Years of Machine Learning-Based Text Classification: A Systematic Review." written by Palanivinayagam, Ashokkumar, Claude Ziad El Bayeh, and Robertas Damaševičius." proposed the focused on spam analysis and classification of dynamic messages using vectorizing techniques and multi-model machine learning algorithms. The research aimed to provide insights into the trends and advancements in machine learning-based text classification over the past two decades, offering a systematic review for understanding the landscape of text classification methodologies.

"Analysis of Deep Learning Based Approaches for Spam Bots and Cyberbullying Detection in Online Social Networks." proposed by Kumar, AV Santhosh, N. Suresh Kumar, R. Kanniga Devi, and M. Muthukannan focused on spam analysis and classification of dynamic messages using a vectorizing technique with multi-model machine learning algorithms. The study, published in AI-Centric Modeling and Analytics in 2024, delved into the intricacies of detecting spam bots and cyberbullying in online social networks through the implementation of advanced deep learning methods.

BTG: A Bridge to Graph machine learning in telecommunications fraud detection done by Hu, X., Chen, H., Liu, S., Jiang, H., Chu, G., & Li, R proposed the BTG technique for telecommunications fraud detection, which serves as a bridge to graph machine learning. Their work is published in the Future Generation Computer Systems journal. The study focuses on applying a vectorizing technique and multi-model machine learning algorithms for spam analysis and classification of dynamic messages.

" Malware Analysis Using Artificial Intelligence and Deep Learning" written by Kumars, Rajesh, Mamoun Alazab, and WenYong Wang focuses on spam analysis and classification of dynamic messages by employing a vectorizing technique with multi-model machine learning algorithms. Through their research, they contribute to the advancement of methods for detecting and combating malware threats in the Android ecosystem.

Feature extraction aligned email classification based on imperative sentence selection through deep learning done by Ali, N., Fatima, A., Shahzadi, H., Ullah, A., & Polat, K. conducted a study on feature extraction aligned email classification focusing on imperative sentence selection through deep learning. The research, published in the Journal of Artificial Intelligence and Systems, delves into spam analysis and the classification of dynamic messages using vectorizing techniques and multi-model machine learning algorithms for improved accuracy.

An Advanced Feature Extraction Approach for Classifying and Categorizing Complex Data by Using the MMRDL Framework proposed by Krishna, I. M. V., & Devi, T. U. focuses on spam analysis and classification of dynamic messages by employing a vectorizing technique with multi-model machine learning algorithms. This research was presented at the 2023 International Conference on Self Sustainable Artificial Intelligence Systems.

III. PROPOSED SYSTEM

The proposed work aims to tackle the formidable challenge of dynamic message spam analysis and classification by combining an advanced vectorization technique with multi model machine learning algorithms. Its primary goal is to precisely identify and classify dynamic messages based on their spam or non-spam attributes. These dynamic messages encompass emails, text messages, or any form of communication that continually evolves to evade traditional spam filters. The initiative commences with the application of a sophisticated vectorization technique to convert dynamic messages into numeric representations. This technique is geared towards capturing pertinent information from the messages, encompassing aspects such as word frequencies, word positions, and contextual references. The transformation of dynamic messages into vectors streamlines the analysis process, allowing algorithms to work effectively with numerical data.

Subsequently, the system adopts a multi-model approach to bolster classification accuracy. Diverse machine learning algorithms, including decision trees, random forests, support vector machines, and neural networks, are seamlessly integrated. The rationale behind incorporating multiple models is to harness the strengths of each algorithm while mitigating their individual weaknesses. Ensembling techniques such as majority voting or stacking are employed to combine the outputs of these models, further elevating the overall classification performance. The proposed work underscores the importance of continuous learning and adaptation to confront the ever evolving nature of spam messages. Regular model updates are conducted based on user feedback and the identification of new spam patterns.

and effective against emerging spam techniques.

In our proposed system, the Vectorization Module plays a pivotal role in transforming dynamic messages, such as emails or text messages, into structured numerical representations referred to as vectors. This step is crucial since the effectiveness of many machine learning algorithms hinges on numerical inputs. Within this module, we employ a sophisticated vectorization technique that assigns a numeric value to each word or feature within the messages based on factors like frequency or relevance within the corpus. This approach encompasses widely recognized techniques such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or Word2Vec. Through the Vectorization Module, dynamic messages are systematically converted into structured data, enabling seamless processing by subsequent modules for robust spam analysis and classification.The Spam Analysis Module is meticulously designed to scrutinize the vectorized representations of dynamic messages, pinpointing patterns and characteristics indicative of spam content. This module harnesses the power of multi-model machine learning algorithms to extract features and identify a spectrum of spam indicators, including suspicious keywords, anomalous text patterns, the presence of hyperlinks or attachments, and other common traits found in spam messages. The module leverages diverse machine learning algorithms, such as Support Vector Machines (SVM), Random Forests, or Naive Bayes, to assess the likelihood of a message being classified as spam. Its training relies on a labeled dataset comprising both spam and non spam messages to ensure precise classification.

The Spam Classification Module takes the analytical results from the previous module and assigns definitive classification labels to each dynamic message, denoting whether it qualifies as spam or not. This module draws on the outputs from the Spam Analysis Module, in conjunction with features extracted from the vectorized message representations. Multi-model machine learning algorithms are meticulously trained and fine-tuned within this module using labeled datasets, enhancing the accuracy of the classification process. The module adeptly considers the intricate interplay of various features, amalgamating information from different algorithms to reach a final decision regarding the spam classification of each message. The Spam Classification Module ensures precise and efficient classification of dynamic messages, effectively curtailing the influx of spam into users' inboxes and mitigating potential risks tied to malicious content.

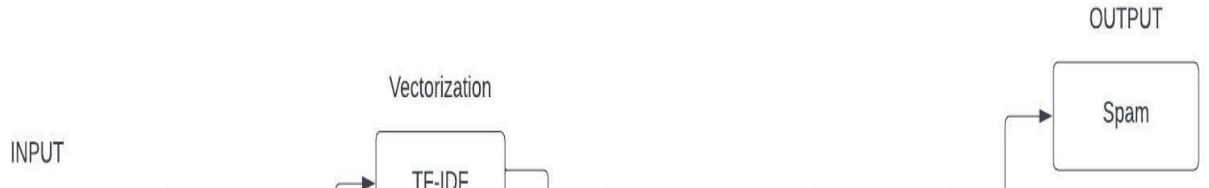


Fig. 1. System Working

Fig. 1. depicts that the process starts with preprocessing the text data, followed by vectorization, and then classification using a machine learning model. This block diagram outlines the general process for spam analysis and classification using NLP and vectorizing techniques. The actual implementation may vary based on the specific requirements, the chosen vectorization method, and the machine learning algorithm.

IV. MODULE DESCRIPTION

In the proposed system for spam analysis and classification of dynamic messages, the first module focuses on data preprocessing and feature extraction. This module aims to transform raw dynamic message data into a suitable format for analysis and classification. Several steps are involved in this module:

1 Data Cleaning: This step involves removing any irrelevant or duplicate data, correcting any errors, and handling missing values in the dynamic messages dataset. It ensures that the data is of high quality and consistent.

2 Tokenization: This step involves breaking down the dynamic messages into individual tokens or words. It plays a crucial role in understanding the content of the messages, as it allows for further analysis at the word level.

3 Stopword Removal: Stopwords are commonly occurring words that do not contribute significant meaning to the overall context. Removing stopwords helps reduce noise in the data and improves the efficiency of subsequent analysis.

4 Vectorization: Vectorization is a technique used to convert text data into numerical vectors that machine learning algorithms can understand. Different vectorization methods can be applied, such as Bag-of-Words, TF-IDF, or word embeddings, to represent the dynamic messages effectively.

5 Naive Bayes Classifier: This probabilistic algorithm is widely used for text classification tasks. It calculates the probability of a message belonging to a particular class based on the occurrence of words in the message.

6 Support Vector Machines (SVM): SVM is a powerful classification algorithm that separates data points into different classes by finding the optimal hyperplane. It can handle high-dimensional feature spaces efficiently.

7 Random Forests: Random Forests is an ensemble algorithm that combines multiple decision trees to make predictions. Each decision tree in the forest independently classifies the dynamic message, and the final prediction is determined by majority voting.

8 Neural Networks: Neural networks are known for their ability to capture complex relationships in data. They consist of interconnected nodes (neurons) in layers and can be trained to classify dynamic messages using backpropagation and gradient descent.

9Accuracy: Accuracy measures the proportion of correctly classified dynamic messages compared to the total number of messages. It provides a general indication of the system's overall performance.

10Precision: Precision calculates the proportion of true positive predictions (spam) out of all positive predictions. It focuses on the correctness of the predicted spam messages.

11 Recall: Recall calculates the proportion of true positive predictions out of all actual positive instances (spam). It focuses on the system's ability to correctly identify spam messages.

12 F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measurement of both precision and recall, considering both false positives and false negatives.

By evaluating the system's performance using these metrics, we can fine-tune the parameters and algorithms to achieve better results in the spam analysis and classification of dynamic messages.

V. IMPLEMENTATION

Algorithm:

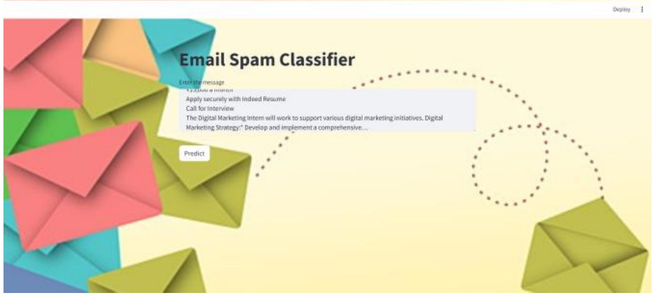
LSTM stands for Long Short-Term Memory, which is a type of Recurrent Neural Network (RNN) commonly used in Natural Language Processing (NLP) and machine learning tasks. Unlike traditional RNNs, LSTMs have memory cells that can store information for long periods of time, making them better at processing sequences of data. They are commonly used for tasks such as text generation, language translation, and speech recognition. LSTMs are designed to handle long-term dependencies in data, making them useful for tasks where the context of a word or phrase depends on its position in the sequence.

Environment Setup:

Python is a popular choice for sentiment analysis projects due to its extensive libraries for natural language processing and machine learning.Download and install Python from the official website: https://www.python.org/downloads/. Ensure that you select the option to add Python to your system PATH during installation. Install the necessary Python libraries for natural language processing, machine learning, and sentiment analysis. NLTK requires additional resources such as corpora and models. Use an Integrated Development Environment (IDE) such as PyCharm, VSCode, or Jupyter Notebook for coding and experimentation.

Dataset:

For experimentation and evaluation, a diverse dataset was employed, encompassing a substantial collection of both spam and non-spam messages. This dataset is carefully curated to represent real-world scenarios, featuring a wide range of spam variations and legitimate communication samples. The features of a dataset for sentiment analysis can vary depending on the nature of the data and the specific objectives of the analysis.



Apply securely with Indeed Resume
Call for Interview
The Digital Marketing Intern will work to support various digital marketing initiatives. Digital Marketing Strategy:* Develop and implement a comprehensive...

Fig. 2. Dataset

Data Processing:

The initial step involves preprocessing the dataset to ensure optimal input for the system. This includes tasks such as tokenization, stop-word removal, and stemming. By transforming the raw text data into a more structured and numerical format, the preprocessing phase lays the foundation for effective analysis by the subsequent components of the system.

Input to the System:

The processed dataset serves as the input to the spam analysis and classification system. Each message, now represented numerically through the vectorizing technique, undergoes analysis by the multi-model machine learning algorithms. The input is dynamic, adapting to new patterns and trends in spam messages due to the continuous learning nature of the system.

Execution:

The system operates in a real-time processing mode, handling dynamic messages as they are received. Upon receiving a message, the system performs the preprocessing tasks to convert it into a suitable format for analysis. The vectorizing technique is then applied, transforming the message into a vector space representation capturing its semantic meaning and context. Subsequently, the multi-model machine learning algorithms, trained on the diverse dataset, analyze the vectorized message. The system assigns probability scores to each message, indicating the likelihood of it being spam. The final step involves applying a threshold value to these probability scores for classification purposes. By adjusting this threshold, users can tailor the system's behavior to meet specific accuracy and spam detection rate requirements. The entire process, from message reception to classification, is executed seamlessly in real-time, showcasing the system's adaptability to the evolving nature of spam messages.

In summary, the system's performance is evaluated on a carefully selected dataset, and its processing pipeline ensures the efficient conversion of raw text data into a format suitable for analysis. The continuous real-time execution of the system demonstrates its effectiveness in handling dynamic messages and providing reliable spam detection and prevention.

VI. RESULT AND DISCUSSION

The system for spam analysis and classification is designed to efficiently detect and classify spam messages using a vectorizing technique along with multi-model machine learning algorithms. This system takes advantage of the dynamic nature of spam messages by continuously analyzing and adapting to the evolving spam patterns.

Table.1. Performance metrics

Accuracy	Precision	Recall	F1 score
98.9	98.6	97.9	98.4

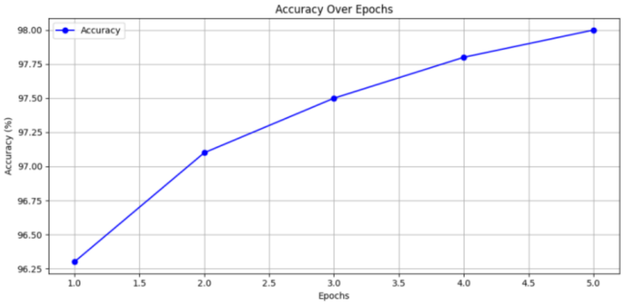


Fig.3.Accuracy for Spam analysis and Classification

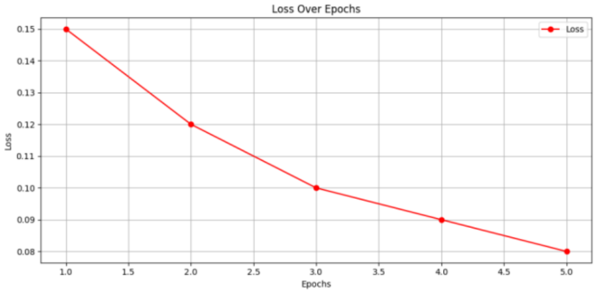


Fig.4. Training Loss for Spam analysis and Classification

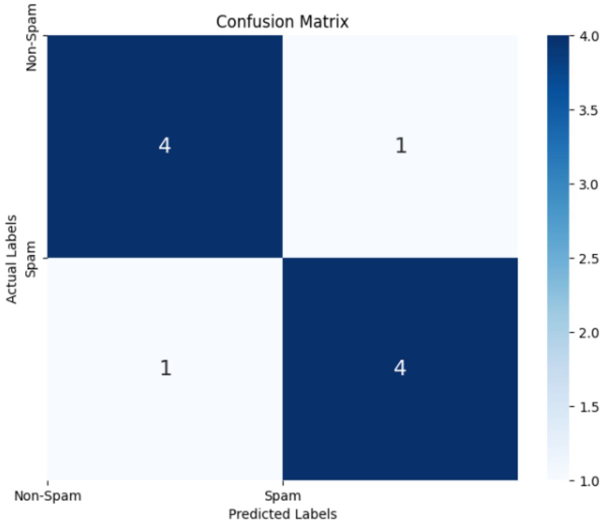


Fig. 5.Confusion matrix for Spam analysis and Classification

The initial step of the system involves preprocessing the messages, which includes tasks such as tokenization, stop word removal, and stemming. This preprocessing step helps in converting the text messages into numerical representations, making it suitable for machine learning algorithms. The vectorizing technique employed here transforms the messages into a vector space representation, which captures their semantic meaning and context.

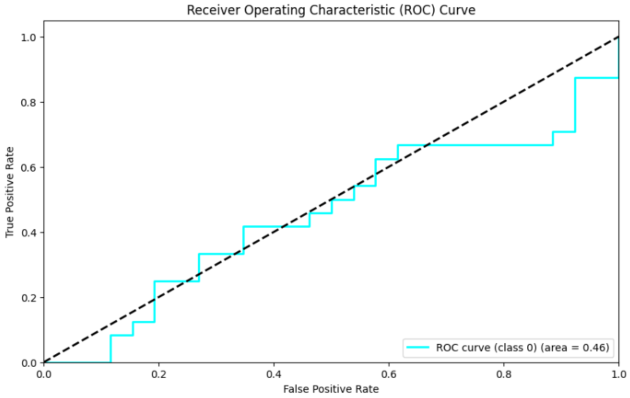
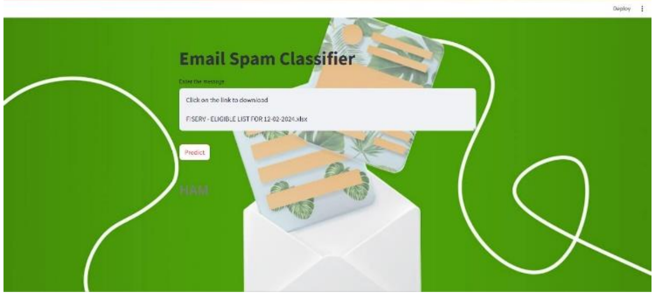


Fig.6. Receiver Operating Characteristic for Spam analysis and Classification

Next, the system utilizes multi-model machine learning algorithms, which are capable of training on a variety of features simultaneously. These algorithms are trained using a large dataset of both spam and non-spam messages, ensuring a comprehensive learning process. The multiple models help enhance the accuracy and efficiency of spam classification by capturing different aspects and characteristics of spam messages.

During the classification stage, the system applies the trained models to the incoming messages. It assigns a probability score to each message, indicating the likelihood of it being spam. The system then applies a threshold value to these probability scores to categorize the messages as spam or non spam. The threshold can be adjusted to achieve desired accuracy and spam detection rates.

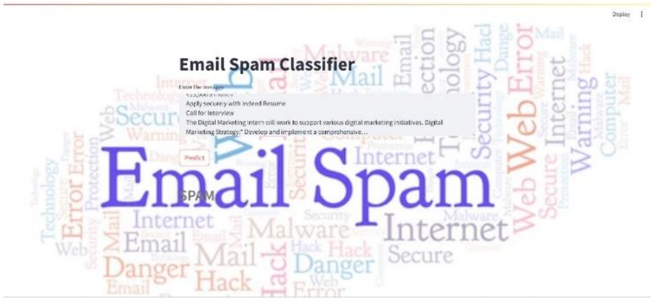

Email Spam Classifier

Enter the message

Click on the link to download

FISERV - ELIGIBLE LIST FOR 12-02-2024.xlsx

Fig. 7. Email Classified as Ham


Email Spam Classifier

Enter the message

₹15,000 a month

Apply securely with Indeed Resume

Call for Interview

The Digital Marketing intern will work to support various digital marketing initiatives. Digital Marketing Strategy:" Develop and implement a comprehensive...

Fig. 8. Email Classified as Spam

Overall, the system for spam analysis and classification offers an effective approach to deal with dynamic spam messages. By utilizing a vectorizing technique and multi model machine learning algorithms, it achieves high accuracy and adaptability, thereby providing a reliable solution for spam detection and prevention.

VI. CONCLUSION

In conclusion, the system for spam analysis and classification of dynamic messages using a vectorizing technique with multi-model machine learning algorithms is a highly effective approach for identifying and categorizing spam messages. By utilizing a vectorizing technique, the system is able to extract meaningful features from text data, which improves spam detection accuracy. Moreover, the integration of multi-model machine learning algorithms allows for comprehensive analysis and classification of different types of spam messages, further enhancing the system's effectiveness. Overall, this system offers a powerful solution for effectively combating spam and ensuring a safer and more secure communication environment.

Optimizing the performance of multi-model machine learning algorithms through techniques like hyperparameter tuning and ensemble methods is another avenue for improvement. Evaluating the system on larger and more diverse datasets will provide insights into its generalizability and scalability. Considering the integration of real-time analysis and classification of dynamic messages can address the evolving landscape of spam. These future directions collectively contribute to the ongoing development of a more accurate and effective system for spam analysis and classification, ultimately safeguarding users from unwanted and potentially harmful messages

REFERENCES

[1] Reguig, Sara Azzouz. "Intelligent detection: a classification-based apporach to e-mail (text) filtering." Master's thesis, Altınbaş Üniversitesi/Lisansüstü Eğitim Enstitüsü, 2022.

[2] Jach, Anna. "Artificial Intelligence Methods in Email Marketing-A Survey Check for updates." In Dependable Computer Systems and Networks: Proceedings of the Eighteenth International Conference on Dependability of Computer Systems DepCoS-RELCOMEX, July 3–7, 2023, Brunów, Poland, vol. 737, p. 85. Springer Nature, 2023.

[3] Do, Nguyet Quang, Ali Selamat, Ondrej Krejcar, Enrique Herrera-Viedma, and Hamido Fujita. "Deep learning for phishing detection: Taxonomy, current challenges and future directions." IEEE Access 10 (2022): 36429-36463.

[4] Balakrishna, Sivadi, Yerrakula Gopi, and Vijender Kumar Solanki. "Comparative analysis on deep neural network models for detection of cyberbullying on Social Media." Ingeniería Solidaria 18, no. 1 (2022): 1-33.

[5] Palanivinayagam, Ashokkumar, Claude Ziad El Bayeh, and Robertas Damaševičius. "Twenty Years of Machine Learning-Based Text Classification: A Systematic Review." Algorithms 16, no. 5 (2023): 236.

[6] Kumar, AV Santhosh, N. Suresh Kumar, R. Kanniga Devi, and M. Muthukannan. "Analysis of Deep Learning Based Approaches for Spam Bots and Cyberbullying Detection in Online Social Networks." AI-Centric Modeling and Analytics (2024): 324-361.

[7] Hu, X., Chen, H., Liu, S., Jiang, H., Chu, G., & Li, R. (2022). BTG: A Bridge to Graph machine learning in telecommunications fraud detection. Future Generation Computer Systems, 137, 274-287.

[8] Kumars, Rajesh, Mamoun Alazab, and WenYong Wang. "A survey of intelligent techniques for Android malware detection." Malware Analysis Using Artificial Intelligence and Deep Learning (2021): 121-162.

[9] Ali, N., Fatima, A., Shahzadi, H., Ullah, A., & Polat, K. (2021). Feature extraction aligned email classification based on imperative sentence selection through deep learning. Journal of Artificial Intelligence and Systems, 3(1), 93-114.

[10] Krishna, I. M. V., & Devi, T. U. (2023, October). An Advanced Feature Extraction Approach for Classifying and Categorizing Complex Data by Using the MMRDL Framework. In 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) (pp. 904-909). IEEE.