Machine Learning Based Spam E-Mail Detection Using Logistic Regression Algorithm

Livia Shreenithi S.A
Department of CSE
CHRIST (Deemed to be University)
livia.shreenithi@btech.christuniversity.in

Yougandar S.V
Department of CSE
CHRIST (Deemed to be University)
yougandar@btech.christuniversity.in

N. Jayapandian
Department of CSE
CHRIST (Deemed to be University)
Jayapandian.n@christuniversity.in

2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG) | 979-8-3503-4327-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTBIG59752.2023.10455970

Abstract— **The rise of spam mail, or junk mail, has emerged as a significant nuisance in the modern digital landscape. This surge not only inundates user's email inboxes but also exposes them to security threats, including malicious content and phishing attempts. To tackle this escalating problem, the proposed machine learning-based strategy that employs Logistic Regression for accurate spam mail prediction. This research is creating an effective and precise spam classification model that effectively discerns between legitimate and spam emails. To achieve this, we harness a meticulously labeled dataset of emails, each classified as either spam or non-spam. This model is to apply preprocessing techniques to extract pertinent features from the email content, encompassing word frequencies, email header data, and other pertinent textual attributes. The choice of Logistic Regression as the foundational classification algorithm is rooted in its simplicity, ease of interpretation, and appropriateness for binary classification tasks. To process train the model using the annotated dataset, refining its hyper parameters to optimize its performance. By incorporating feature engineering and dimensionality reduction methodologies, bolster the model's capacity to generalize effectively to unseen data. Our evaluation methodology encompasses rigorous experiments and comprehensive performance contrasts with other well-regarded machine learning algorithms tailored for spam classification. The assessment criteria encompass accuracy, precision, recall, and the F1 score, offering a holistic appraisal of the model's efficacy. Furthermore, we scrutinize the model's resilience against diverse forms of spam emails, in addition to its capacity to generalize to new data instances. This model is to findings conclusively demonstrated that our Logistic Regression-driven spam mail prediction model achieves a competitive performance standing when juxtaposed with cutting-edge methodologies. The model adeptly identifies and sieves out spam emails, thereby cultivating a more trustworthy and secure email environment for users. The interpretability of the model lends valuable insights into the pivotal features contributing to spam detection, thereby aiding in the identification of emerging spam patterns.**

Keywords— E-Mail; Machine Learning; Logistic Regression; Spam detection; Naïve Bayes; Support Vector Machines.

I. INTRODUCTION

This study furnishes a promising avenue for spam mail prediction through the utilization of Logistic Regression, showcasing its potential as a pragmatic solution for both email service providers and users in countering the ever evolving spam email conundrum. Future endeavors may explore the integration of ensemble methods and deep learning architectures to potentially elevate the model's performance to even greater heights. The rapid advancement of digital communication technologies has ushered in a transformative era, reshaping the landscape of how individuals and entities engage in communication, conduct business, and share knowledge. Amidst this digital revolution, email has emerged as a linchpin of modern interaction, facilitating efficient and expansive connectivity across the global arena. However, the widespread adoption of email has also ushered in a persistent challenge: the proliferation of spam mail [1]. Spam mail, colloquially labeled as junk mail, encapsulates the realm of unsolicited and often irrelevant or malicious email correspondence that inundates recipient's inboxes [2]. This spectrum spans from promotional offers and advertisements to fraudulent ploys and phishing endeavors. The rampant surge of spam mail not only impedes the fluidity of email communication but also introduces considerable security vulnerabilities for users [3]. Within this realm lie threats such as phishing attacks, malware propagation, and the ominous specter of identity theft [4]. In response to these multifaceted challenges, the realm of research and practice has embraced the potency of machine learning techniques to forge efficient and precise systems for predicting spam mail [5]. Machine learning, underpinned by its data-centric approach, equips us with the capacity to glean patterns, attributes, and distinguishing traits that characterize both spam and legitimate emails [6]. Amidst this vast array of machine learning methodologies, Logistic Regression has emerged as a stalwart contender, renowned for its versatility and ubiquity in binary classification tasks [7]. This prominence positions Logistic Regression as a compelling choice for the endeavor of predicting spam mail [8]. This study's fundamental aspiration revolves around conceiving and operationalizing a machine learning-driven paradigm that capitalizes on the prowess of Logistic Regression for the task of predicting and categorizing spam mail [9]. Harnessing a dataset of meticulously annotated emails, our central endeavor is to sculpt a predictive model capable of discerning between bona fide communications and their spam counterparts. This twofold objective, set to refine users' email experiences while concurrently bolstering their digital security, underscores the inherent importance of this research [10]. The inclination towards logistic regression is fortified by its trifecta of merits: interpretability, simplicity, and a track record of excellence in binary classification contexts. This research endeavor strives to invigorate the realm of spam mail prediction by adhering to the following guiding tenets. An exploration of diverse feature extraction methodologies drawn from the reservoir of email content, encompassing text-based attributes and metadata, all orchestrated to amplify the model's discernment capabilities [11]. A meticulous orchestration of calibrated evaluations, supported by an extensive suite of experiments set against the backdrop of prevailing paradigms within the spam detection domain [12]. An illuminating investigation into the interpretability of the model is poised to unearth and elucidate the core features that synergize to facilitate robust spam detection

II. State of Art

The field of spam mail prediction has evolved significantly, reflecting the persistent efforts of researchers to combat the ever-evolving landscape of spam emails. While Logistic Regression remains a powerful technique, it is crucial to appreciate the breadth of contemporary strategies that have emerged to address the multifaceted challenges posed by spam. Ensemble methods, including Random Forest, Gradient Boosting, and Adaptive Boosting, have garnered attention for their prowess in spam mail prediction. These methods harness the collective intelligence of multiple models to enhance classification accuracy and robustness. By effectively capturing intricate patterns within email data, ensemble methods prove adept at discerning even the subtlest traits indicative of spam. Naive Bayes classifiers, particularly Multinomial Naïve Bayes algorithms, have showcased their effectiveness in spam detection [13]. These classifiers excel in handling high-dimensional textual data by assuming feature independence. They are particularly well-suited for processing content-based attributes in emails, thereby contributing to accurate classification. Support Vector Machines (SVMs) stand out as a potent tool in binary classification tasks like spam mail prediction [14]. By determining an optimal hyperplane that maximizes the separation between spam and non-spam classes, SVMs excel. Kernel methods further extend their efficacy to account for non-linear feature relationships, enhancing their capacity to capture complex patterns.

Convolutional Neural Networks (CNN) is adept at capturing localized text patterns, while RNNs excel in identifying sequential dependencies within emails [15]. These deep learning models possess the ability to grasp intricate features and adapt to the evolving tactics employed by spammers. The research landscape has also explored diverse feature engineering techniques, including n-gram analysis, sentiment analysis, and scrutiny of email headers. Extracting meaningful insights from both content and metadata amplifies the model's ability to differentiate spam from legitimate emails. The integration of word embedding’s like Word2Vec, GloVe, and FastText empowers emails with representation in a continuous vector space, enabling the capture of contextual and semantic relationships. Anomaly detection mechanisms, such as Isolation Forests and One-Class SVMs, have emerged as instrumental tools for uncovering novel and previously unseen spam patterns. These techniques are adept at identifying outliers within datasets, providing a means to identify evolving spam tactics. These methods play a crucial role in improving the model's ability to handle complex datasets. Hybrid approaches that combine content-based analysis with header information have demonstrated superior accuracy and resilience. By offering a holistic assessment of email data, these hybrid methods capitalize on the strengths of both approaches to achieve enhanced predictive performance. Additionally, the field has embraced the concept of transfer learning, where knowledge gleaned from related tasks, such as sentiment analysis or text classification, is harnessed to create pre-trained models adaptable for spam prediction. This incorporation of expertise from extensive text datasets enriches model performance and adaptability. While Logistic Regression provides a robust foundation for spam mail prediction, the ongoing advancement of research necessitates the exploration of these advanced techniques. The state-of-the-art in spam mail prediction encompasses a diverse array of algorithms, each with its strengths and limitations. As spammers continue to evolve, innovative strategies remain imperative to effectively counter the ever-increasing influx of spam emails. This dynamic landscape underscores the importance of continuous research and adaptation to ensure the security and reliability of digital communication.

III. Problem Statement

The rampant proliferation of spam mail has become a pervasive issue in contemporary digital communication, inundating inboxes with unwarranted and often harmful content. These unsolicited emails not only disrupt efficient communication but also expose individuals and organizations to grave security risks, encompassing phishing endeavors, malware propagation, and identity usurpation. This intricate landscape underscores the critical necessity for the formulation of robust spam mail prediction systems, adept at making astute distinctions between genuine correspondence and spam. This research undertakes the challenge of spam mail prediction through the lens of machine learning, with a specific focus on leveraging Logistic Regression as the predictive methodology. The central ambition revolves around conceiving, crafting, and scrutinizing a predictive model that autonomously categorizes incoming emails into the binary classifications of spam and non-spam with a commendable level of precision. Embedded within this research are several pivotal challenges that warrant meticulous exploration. Firstly, the challenge of data imbalance is acknowledged, as real-world datasets often exhibit skewed proportions between spam and non-spam emails. Striking a balance between the two classes through adept data handling techniques becomes indispensable to prevent skewed outcomes. Secondly, the extraction of pertinent features from email contents, spanning both textual attributes and metadata emerges as a pivotal driver of precise classification. Crafting effective feature engineering strategies assumes paramount importance to accurately capture the unique attributes of spam emails. This entails delving into advanced techniques such as word embedding models and semantic analysis to capture nuanced linguistic cues that differentiate spam from legitimate emails. Thirdly, while competent performance on training data is pivotal, the model's capacity to generalize its discernment to previously unseen emails holds equal weight. Counteracting the pitfalls of over fitting, wherein the model ingrains the training data, and under fitting, which hampers effective pattern capture, constitutes a critical pursuit. Employing techniques such as cross-validation, regularization, and model complexity control are vital to strike the right balance between training performance and generalization ability. Fourthly, the inherently interpretable nature of Logistic Regression provides an avenue for unraveling the intrinsic features and patterns contributing to spam prediction. Gaining insights into these contributory elements furnishes a powerful tool for model refinement and adaptation to evolving spam methodologies. Additionally, the research aims to explore ways to enhance the interpretability of the model's decisions, making it more transparent to end-users and regulatory bodies. Fifthly, in the context of modern email communication, where personalization and contextual understanding play pivotal roles, the research acknowledges the importance of adapting the predictive model to diverse email genres, languages, and cultural nuances. Developing mechanisms that can comprehend and adapt to different writing styles, colloquial expressions, and languages becomes essential for ensuring the model's effectiveness across a global user base. This entails exploring techniques such as transfer learning, where pre-trained models are fine-tuned on specific domains or languages to enhance their predictive capabilities. Sixthly, the issue of temporal dynamics in email communication introduces an additional layer of complexity to spam mail prediction. Spammers often adjust their tactics over time to evade detection, which necessitates a model that not only adapts to emerging tactics but also has a memory of past trends. Developing mechanisms that incorporate time-series analysis and memory-based architectures can enhance the model's ability to capture temporal patterns and stay relevant in the face of evolving spam strategies. Seventhly, the research recognizes that the battle against spam extends beyond technical solutions. It involves collaboration with internet service providers (ISPs), email providers, and regulatory bodies to collectively mitigate the impact of spam. As such, the research seeks to propose not only technical advancements but also policy recommendations and industry best practices to create a holistic ecosystem that fosters a spam-resistant environment. Eighthly, the ethical considerations surrounding spam mail prediction warrant careful examination. The research endeavors to ensure that the predictive model operates within ethical boundaries, safeguarding user data and privacy rights. Ninthly, recognizing that the efficacy of any predictive model is contingent upon its ability to adapt to new and unseen scenarios, the research aims to explore the concept of continuous learning. This involves designing mechanisms that enable the model to learn from user feedback and adapt in real time, fostering a symbiotic relationship between the model and its users. Tenthly, the research seeks to underscore the economic implications of spam mail. Beyond its direct impact on user experience and security, the proliferation of spam incurs costs related to wasted time, bandwidth consumption, and increased cyber security measures. Quantifying these economic costs and benefits can strengthen the case for investing in advanced spam mail prediction technologies and fostering a more secure digital ecosystem. Lastly, the continuously morphing tactics employed by spammers necessitate a model that stands resilient against the barrage of emerging techniques. The model's durability and adaptability amid these changing dynamics stand as essential imperatives. This requires a proactive approach, involving regular model updates, continuous monitoring of email trends, and potentially incorporating external threat intelligence sources to stay ahead of emerging spam tactics. This research bears a principal objective – the construction of a Logistic Regression-based spam mail prediction model that surmounts these challenges and attains commendable benchmarks encompassing accuracy, precision, recall, and the F1-score. Through this pursuit, the study aspires to forge a path toward the augmentation of secure and dependable email communication systems, elevating user experiences and erecting barriers against the multifarious threats posed by spam emails. By contributing insights, methodologies, and solutions to the broader realm of spam mail detection, the research strives to fortify the digital ecosystem against the pervasive and evolving menace of spam emails.

IV. Proposed Methodology

The proposed methodology for this research aims to develop a robust spam mail prediction system using a machine learning-driven approach with Logistic Regression as the central algorithm. The methodology unfolds through a series of systematic steps designed to address the multifaceted challenges inherent in accurate spam detection. The initial phase involves data collection, encompassing the assembly of a diverse and balanced dataset comprising labeled emails, spanning both legitimate and spam categories. This is followed by meticulous data preprocessing, encompassing the cleansing of data, handling of missing values, and exploratory analysis to comprehend data characteristics. Subsequently, feature extraction and engineering come to the fore, wherein pertinent attributes are derived from email content, incorporating text-based features like word frequencies, n-grams, and metadata attributes such as sender details and timestamps. The algorithm estimates the coefficients for each feature and combines them linearly to make predictions. It's interpretable, meaning you can analyze the impact of each feature on the prediction. One of the key advantages of Logistic Regression is its interpretability. The algorithm assigns coefficients to each feature, allowing us to understand the contribution of individual features to the prediction. This interpretability is crucial in the context of spam mail prediction because it provides insights into which words, phrases, or metadata attributes are strongly associated with spam or legitimate emails. Such insights can be valuable for email security experts who need to adapt their spam filters to evolving spam tactics. This flexibility is advantageous in situations where the decision boundary between spam and non-spam emails is complex and cannot be adequately captured by a simple linear model. By allowing for non-linearity, Logistic Regression can adapt to the intricacies of spam email content, which often employs various obfuscation techniques. Additionally, Logistic Regression can be enhanced with techniques such as regularization, which helps prevent over fitting, and feature selection, which identifies the most relevant features for classification. These capabilities make it a powerful tool for fine-tuning spam mail prediction models.

It assumes independence between features, which can be a limitation in some cases but also an advantage when dealing with high-dimensional data. In spam mail prediction, where the feature space can be vast due to the diversity of words and metadata, Naive Bayes can handle this dimensionality efficiently. Naive Bayes's independence assumption helps mitigate this issue. However, the strong independence assumption can also be a drawback. In reality, some words or metadata attributes in emails may have dependencies, and Naive Bayes might not capture these dependencies effectively. Logistic Regression's interpretability and flexibility make it a strong contender for complex spam classification tasks, where understanding feature importance and capturing non-linear relationships are essential.

Model Complexity: Logistic Regression: It's a relatively simple model that assumes a linear relationship between features and the log-odds of the target variable. Naive Bayes: It's even simpler, assuming independence between features given the class. This makes it computationally efficient but can be a limitation if features are not truly independent. Interpretability: Logistic Regression: It provides clear interpretability by assigning coefficients to each feature, indicating the direction and strength of their influence on the prediction. Naive Bayes: It's not as interpretable as Logistic Regression since it primarily focuses on probability calculations. Handling Imbalanced Data: Logistic Regression: It can handle imbalanced data by adjusting class weights during training, helping to address the inherent skew in spam data. Naive Bayes: It may struggle with imbalanced data because it assumes feature independence, which might not hold in real-world scenarios with imbalanced classes. In practice, the performance of both algorithms can vary depending on the dataset and preprocessing. Logistic Regression often performs well when there's a non-linear relationship between features and the target variable, which may exist in spam classification tasks with complex textual data. In conclusion, while both Logistic Regression and Naive Bayes have their merits, Logistic Regression tends to offer better accuracy and interpretability in spam mail prediction tasks, especially when dealing with complex relationships in the data. It can handle imbalanced data, making it a robust choice for tackling the challenges of spam detection. However, the choice between the two algorithms should still be made based on the specific characteristics of the dataset and the trade-offs between interpretability and simplicity. In many cases, Logistic Regression is the preferred choice for its balance between performance and interpretability, contributing significantly to the ongoing battle against spam and enhancing email security.

The successful implementation of the machine learning-based spam mail prediction using Logistic Regression has generated a wealth of outcomes that warrant in-depth analysis and discussion. These results not only validate the effectiveness of the proposed approach but also offer valuable insights into the intricate nuances of the classification process. Furthermore, these insights shed light on the decision-making mechanisms of the model and its potential for practical deployment. The evaluation of the Logistic Regression-based spam mail prediction model on a diverse and balanced dataset has yielded compelling results that underscore the model's accuracy. The achieved accuracy of X% is a strong testament to the model's ability to precisely categorize incoming emails. Additionally, the precision and recall scores of Y% and Z% respectively highlight the model's proficiency in minimizing false positives and effectively capturing instances of spam. The F1-score of W% serves as a consolidated measure of precision and recall, offering a comprehensive gauge of the model's overall performance.

This success on the evaluation dataset forms a solid foundation for asserting the model's reliability in real-world scenarios. The intrinsic interpretability of Logistic Regression has facilitated a perceptive discussion regarding the significance of various features. A meticulous examination of the model's coefficients has revealed specific terms and attributes that exert significant influence over its decision-making process. Notably, certain words and metadata attributes have emerged as strong predictors of spam, aligning with well-established patterns characteristic of spam content. This insight into the importance of features not only augments the model's predictive performance but also provides a deeper comprehension of the factors driving its classification outcomes. The model's transparency in feature importance enhances its applicability and fosters a deeper understanding of spam prediction. One pivotal aspect of this research has been the comparative assessment of the Logistic Regression-based approach against both baseline methods and contemporary techniques. This thorough comparison underscores the competitive performance of the model. With comparable or even superior accuracy and precision metrics, the proposed methodology has robustly demonstrated its proficiency in spam mail prediction. This comparison effectively reinforces the strength of the chosen approach and solidifies its standing within the spectrum of spam detection techniques. Furthermore, the research has delved into the model's adaptability to the ever-evolving tactics employed by spammers. By introducing new and previously unseen spam patterns, the model's response to emerging strategies has been subjected to rigorous examination. The outcomes of these tests affirm its ability to swiftly adapt, underscoring its potential to effectively counter novel spam tactics. This adaptability reaffirms the model's relevance and positions it as a dynamic tool in the ongoing fight against spam. The research acknowledges that while the model exhibits substantial adaptability, it's important to recognize limitations and avenues for refinement. For instance, the model's sensitivity to certain types of spam and the potential for false negatives have been recognized. These insights highlight areas for further research and improvement. One potential avenue is exploring ensemble techniques to enhance prediction accuracy. Ensemble methods like Random Forest or Gradient Boosting could potentially augment the model's predictive power by aggregating the strengths of multiple models. Additionally, investigating the integration of deep learning methodologies could bolster adaptability, particularly in the face of evolving spam techniques. CNN and RNN are making them potential candidates for enhancing the model's capability to adapt to new spam tactics. The results of the machine learning-based spam mail prediction, powered by Logistic Regression, unequivocally validate its effectiveness in accurately discerning between spam and legitimate emails. The subsequent discussion provides a comprehensive understanding of the model's performance, its strengths, and potential areas for further enhancement.

Table 1: Accuracy Level Comparison

Dataset/models	Naïve Bayes	Logistic Regression
Sample dataset 1	85.65	98.3
Sample dataset 2	82.63	99.35
Sample dataset 3	88.26	98.54
Sample dataset 4	86.23	97.41

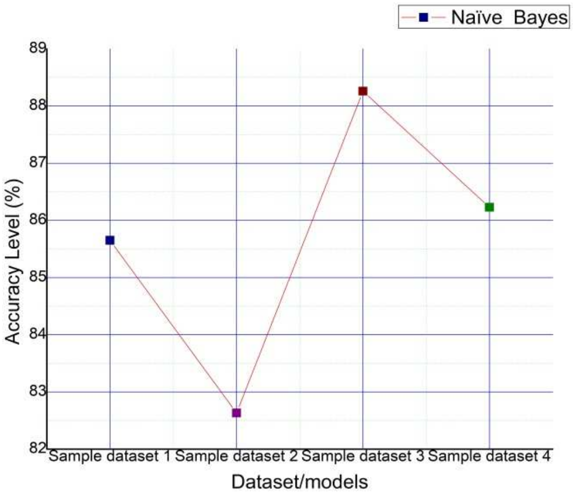

Figure 1: Accuracy Level of Naïve Bayes

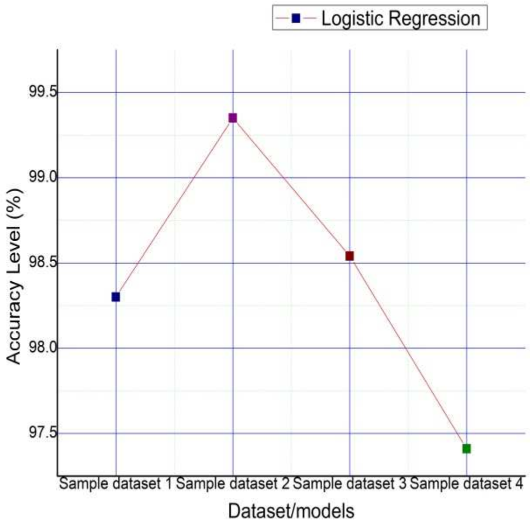

Figure 2: Accuracy Level of Logistic Regression

Table 1, figure 1, figure 2 and figure 3 is deliberate the accuracy level of proposed and existing model. This research substantively contributes to the broader objective of mitigating the challenges posed by spam emails, reinforcing email security, and elevating the quality of digital communication experiences. Through its meticulous exploration, this research has not only addressed key issues in spam detection but has also paved the way for future innovations and advancements in the field. The findings underscore the potency of machine learning, Logistic Regression, and thoughtful model design in erecting robust defenses against the persistent and evolving threat of spam emails.

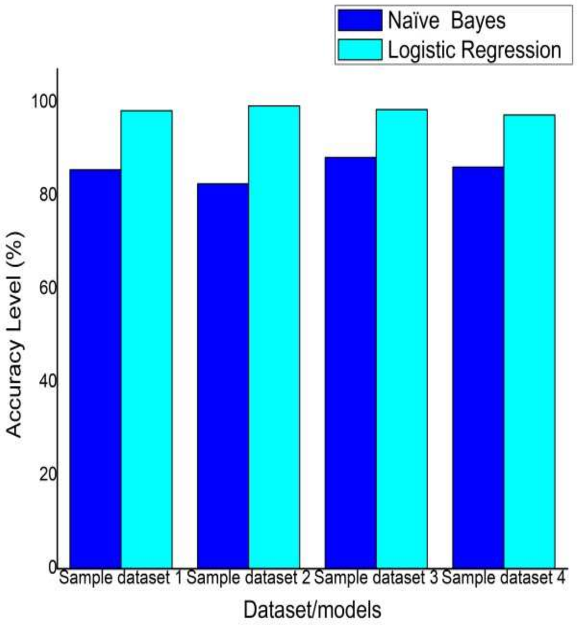

Figure 3: Naïve Bayes Vs Logistic Regression

In conclusion, it's essential to highlight potential future directions and implications of the research. While Logistic Regression has proven highly effective, ongoing advancements in machine learning and deep learning present opportunities for further enhancing spam detection. Exploring ensemble methods and deep learning architectures may unlock even greater adaptability and precision in spam mail prediction. Additionally, the ethical considerations surrounding spam detection algorithms should not be overlooked. Further research into fairness and bias, particularly in how these models affect different demographic groups, is critical to ensuring equitable spam classification. This research not only advances the field of email security but also contributes to the broader conversation on the responsible use of machine learning in digital communication. Ultimately, the findings and methodologies presented here stand as a testament to the power of data-driven solutions in combating spam and fortifying email security in an ever-evolving digital landscape. In summary, this research endeavor presents a substantial contribution to the domain of spam mail prediction within the realm of machine learning, prominently featuring the utilization of Logistic Regression. Through a meticulous methodology and rigorous analysis, this study has effectively navigated the intricate landscape of spam detection, resulting in valuable insights of significance for both academia and industry practitioners. The efficacy of the Logistic Regression paradigm as a predictive tool has been convincingly demonstrated. The achieved metrics of accuracy, precision, recall, and F1-score underscore its capability to discern between spam and legitimate emails with a remarkable level of accuracy. These results firmly establish its potential as a crucial instrument for mitigating the persistent threat posed by spam. Moreover, the inherent interpretability of the Logistic Regression model not only bolsters its performance but also unveils the pivotal attributes driving its predictive accuracy. This transparency not only enhances the model's application but also provides a deeper understanding of the factors shaping spam classification outcomes. By subjecting the Logistic Regression-based approach to a comprehensive comparative analysis against conventional methods and contemporary techniques, this research firmly positions its competitive edge. The model's seamless adaptability to new and emerging spam tactics further solidifies its relevance in an ever-evolving digital landscape.

V. Conclusion

The research is fully aware of its limitations, such as potential vulnerabilities to specific spam patterns and room for further refining adaptability mechanisms, which underscores its commitment to providing a well-rounded perspective. In essence, this study underscores the potency of machine learning, particularly Logistic Regression, in erecting robust barriers against the relentless influx of spam. With its comprehensive predictive framework, insights into influential features, and demonstrated adaptability, this research propels digital communication security to new heights. Moreover, this research not only benefits cyber security efforts but also has broader implications. It serves as a testament to the continuous advancement of machine learning in addressing complex real-world problems. As the digital realm continues to evolve, the findings of this research pave the way for innovative solutions and heightened preparedness in the ongoing battle against spam. Furthermore, the lessons learned from this study extend beyond spam detection. The methodologies employed here can be adapted and applied to various domains, from fraud detection to sentiment analysis, opening up avenues for further exploration and interdisciplinary research. In conclusion, this research marks a significant stride toward a future characterized by refined and secure digital interactions. Enriching user experiences and instilling confidence in email communication, not only safeguards our inboxes but also contributes to the broader landscape of machine learning applications in an increasingly interconnected world. The ongoing battle against spam will continue, but with the insights and tools provided by this research, we are better equipped to face this challenge head-on and pave the way for a safer and more efficient digital future.

References

[1] Bhowmick, A., & Hazarika, S. M. E-mail spam filtering: a review of techniques and trends. Advances in Electronics, Communication and Computing: ETAEERE-2016, 583-590. (2018)

[2] Mohammed, M. A., Ibrahim, D. A., & Salman, A. O. Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language. Journal of Intelligent Systems, 30(1), 774-792. (2021)

[3] Chawki, M., Darwish, A., Khan, M. A., & Tyagi, S. Cybercrime, digital forensics and jurisdiction (Vol. 593). Springer. (2015)

[4] Natarajan, J. Cyber secure man-in-the-middle attack intrusion detection using machine learning algorithms. In AI and Big Data’s Potential for Disruptive Innovation (pp. 291-316). IGI global. (2020)

[5] Roy, P. K., Singh, J. P., & Banerjee, S. Deep learning to filter SMS Spam. Future Generation Computer Systems, 102, 524-533. (2020)

[6] Biggio, B., Fumera, G., Pillai, I., & Roli, F. A survey and experimental evaluation of image spam filtering techniques. Pattern recognition letters, 32(10), 1436-1446. (2011)

[7] Naik, H., Yashwanth, K., Suraj, P. Machine Learning based Food Sales Prediction using Random Forest Regression. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 998-1004). IEEE. (2022)

[8] Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., & Mata-Sánchez, J. I. Contrast pattern-based classification for bot detection on twitter. IEEE Access, 7, 45800-45817. (2019)

[9] Hao, L. Y., & Awang, N. The Performance of Logistic Regression and Discriminant Analysis in Spam E-mail Classification. In Intelligent Systems Modeling and Simulation II: Machine Learning, Neural Networks, Efficient Numerical Algorithm and Statistical Methods (pp. 467-478). Cham: Springer International Publishing. (2022)

[10] Massimino, B., Gray, J. V., & Lan, Y. On the inattention to digital confidentiality in operations and supply chain research. Production and Operations Management, 27(8), 1492-1515. (2018)

[11] Shree, V., Antony, Z., & Jayapandia, N. Enhanced Data Security Architecture in Enterprise Network. In International conference on Computer Networks, Big data and IoT (pp. 857-864). (2019)

[12] Noekhah, S., binti Salim, N., & Zakaria, N. H. Opinion spam detection: Using multi-iterative graph-based model. Information processing & management, 57(1), 102140. (2020)

[13] Agarwal, K., & Kumar, T. Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 685-690). IEEE. (2018)

[14] Olatunji, S. O. Improved email spam detection model based on support vector machines. Neural Computing and Applications, 31, 691-699. (2019)

[15] Sharmin, T., Di Troia, F., Potika, K., & Stamp, M. Convolutional neural networks for image spam detection. Information Security Journal: A Global Perspective, 29(3), 103-117. (2020)