International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An ML-Powered Framework for Email Spam Identification

[1]Ravindra Ramesh Agrawal, [2]Simran Shinde, [3]Swatantrakumar Gupta, [4]Sagar Thakare, [5]Bhavna Sharma

Submitted: 03/09/2024 **Revised:** 18/10/2024 **Accepted:** 20/11/2024

Abstract: Email remains a globally ubiquitous communication tool due to its ease of use and speed. However, its effectiveness is often compromised by an inability to accurately filter unwanted messages. A growing number of reported cases involve the theft of personal information or phishing attempts conducted via email. This project explores the application of Machine Learning (ML) to enhance spam detection. ML, a facet of artificial intelligence, enables systems to automatically learn and improve from data without explicit programming. A binary classifier will be employed to categorize email content into "spam" or "ham" (legitimate mail), aiming for more accurate predictions. The primary objective of this model is to detect and classify words both rapidly and precisely.

Keywords: *Ubiquitous, Phishing, Machine Learning, Spam, Predictions*

I. INTRODUCTION

Spam has become a major internet issue. In 2017, statistics indicated that spam constituted 55% of all email messages, consistent with the previous year. Also known as unsolicited bulk email, spam's prevalence is driven by email's cost-free nature for senders, making it an ideal channel for unwanted advertisements or junk newsgroup postings. This opportunity has been widely exploited by irresponsible entities, leading to cluttered inboxes for millions globally. What was once a minor annoyance has evolved into a significant concern, especially given the offensive nature of some messages. Spam wastes users' time, consumes vast storage space, and strains communication bandwidth. End-users also risk inadvertently deleting legitimate emails. Furthermore, spam has economic consequences, prompting some countries to enact legislation to combat it.

Text classification is used to direct incoming emails or messages to either the inbox or the spam folder. It is the process of assigning categories to text based on its content, serving to organize, structure, and categorize textual data. While this can be done manually, Machine Learning offers an automated approach that is significantly faster. ML utilizes pre-labeled text to learn associations between text segments and their corresponding outputs. It employs feature extraction to transform each text into a numerical vector representation, often indicating word frequencies from a predefined dictionary. Text classification is crucial for structuring the often unstructured and messy nature of text data, such as documents and spam messages, in a cost-effective manner. An ML platform enhances prediction accuracy and, particularly in the context of Big Data, can accelerate the analysis of enormous datasets. This capability is vital for businesses to analyze text data, inform strategic decisions, and even automate processes, such as classifying short texts like tweets or headlines, larger documents like media articles, and for applications in social media or brand monitoring.

[1] *Asst. Professor, Bharati Vidyapeeth Deemed To be University Navi Mumbai & Research Scholar Suresh Zyan Vihar University Jaipur ,*
[2] *Lecturer, Pillai College of Arts, Commerce & Science (Autonomous), New Panvel,*
[3] *Research Scholar Suresh Zyan Vihar University Jaipur,*
[4] *Asst Professor Sterling College of Management and Studies & Research Scholar*
[5]*Research Scholar, Suresh Zyan Vihar University Jaipur*
[1]*rragrawal2305@gmail.com,*
[2]*simranshinde35@gmail,*
[3]*email2swatantra@gmail.com,*
[4]*sagthakare@gmail.com,*
[5]*reply2bhavna21084@gmail.com*

II. PROBLEM DEFINATION

Spam constitutes a substantial portion of global email traffic, consistently comprising over half of all messages. This sheer volume clutters inboxes, wastes users' time, and consumes significant storage and bandwidth. More critically, spam is not static; fraudsters and malicious actors continuously evolve their tactics to bypass detection filters. This constant innovation means that static, rule-based detection systems quickly become obsolete, necessitating a dynamic and adaptive solution.

II.LITERATURE SURVEY

Blanzieri and Bryl [2, 19] describe a list of learning-based email spam filtering approaches. In this paper, they addressed the spam problems and provided a review of learning-based spam filtering. They explain various features of spam emails. In this study, effects of spam emails on different domains were discussed. Various economic and ethical issues of spam are also discussed in this study. The antispam approach that is common and learning-based filtering is well developed. The commonly used filters are based on different classification techniques applied to various components of email messages. This study suggests that the Naïve Bayes classifier holds a particular position amongst multiple learning algorithms used for spam filtering. With splendid pace and simplicity, it gives high precision results.

Bhuiyan et al. [20] present a review of current email spam filtering approaches. They summarize multiple spam filtering approaches and sum up the accuracy on various parameters of different proposed systems by analyzing numerous processes. They discuss that all the existing methods are efficient for filtering spam emails. Some have successful results, and others are attempting to incorporate other ways to boost their accuracy performance. Although they are all successful, they still have some issues in spam filtering methods, which is the primary concern for researchers. They are trying to create a next-generation spam filtering mechanism to understand large numbers of multimedia data and filter spam emails. They conclude that most email spam filtering is done by utilizing Naïve Bayes and the SVM algorithm. To test the spam filtration models, these models can be trained on different datasets, such as “ECML” and UCI dataset [21].

Ferrag et al. [13] presented a review of deep learning algorithms of intrusion detection systems and spam detection datasets. They discussed various detection systems based on deep learning models and evaluated the effectiveness of those models. They examined 35 well-known cyber dataset by dividing them into seven categories. These categories include Internet traffic-based, network traffic-based, Interanet traffic-based, electrical network-based, virtual private network-based, andriod apps-based, IoT traffic-based, and Internet connected device-based datasets. They conclude that deep learning models can perform better than traditional machine learning and lexicon models for intrusion and spam detection.

Vyas et al. [22] present a review on supervised machine learning strategies for filtering spam emails. They concluded that the Naïve Bayes method provides faster results and decent precision over all other methods (except SVM and ID3) from all the techniques discussed. SVM and ID3 offer greater precision than Naïve Bayes but take much longer time to construct a system. There is a trade-off between timing and precision. They conclude that selecting the learning algorithm heavily depends on the situation and the required accuracy and time. They state that all parts of the email should be considered in the future to create a more robust spam filtering framework.

III. ARCHITECTURE OF THE PROPOSED SYSTEM

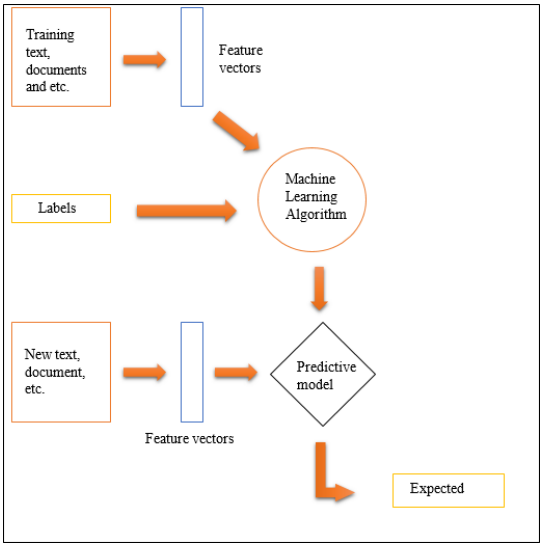

Fig.1 Architecture flow of the system

This diagram illustrates a typical architecture for a Machine Learning-based text classification system,

which is highly relevant to applications like spam detection. Let's break down each component:

A. Training Phase (Top Section):

- Training text, documents and etc. (Orange Rectangle, Top Left): This represents the raw input data used to train the machine learning model. In the context of spam detection, this would be a large collection of emails or text messages, each already labeled as either "spam" or "ham" (legitimate). For general text classification, it could be news articles, reviews, or any textual data relevant to the classification task.
- Labels (Yellow Rectangle, Middle Left): These are the predefined categories or tags associated with each piece of training text. For spam detection, the labels would typically be "Spam" and "Ham." These labels serve as the "ground truth" that the machine learning algorithm learns from.
- Feature Vectors (Blue Vertical Rectangle, Top Middle): Raw text data cannot be directly understood by machine learning algorithms. This step involves converting the text into a numerical representation called "feature vectors." This process, known as feature extraction or text vectorization, transforms words, phrases, or their statistical properties into numerical values. Common techniques include:
    - Bag-of-Words (BoW): Counting word frequencies.
    - TF-IDF (Term Frequency-Inverse Document Frequency): Weighing word importance based on their frequency in a document and rarity across all documents.
    - Word Embeddings (e.g., Word2Vec, GloVe): Representing words as dense vectors that capture semantic relationships.
- Machine Learning Algorithm (Orange Circle, Top Right): This is the core of the system where the learning happens. The algorithm takes the feature vectors of the training text and their corresponding labels as input. Its goal is to find patterns and relationships between the features and the labels. In a binary classification task like spam detection, common algorithms include:
    - Support Vector Machines (SVM)
    - Naive Bayes
    - Logistic Regression
    - Random Forest
    - Deep Learning models (e.g., LSTMs, Transformers for more complex cases)

B. Output of Training (Arrow from Machine Learning Algorithm to Predictive Model):

- Predictive Model (Grey Diamond): After the Machine Learning Algorithm has been trained on the labeled data, it produces a "predictive model." This model encapsulates the learned patterns and rules. It's now ready to make predictions on new, unseen data.

C. Prediction/Inference Phase (Bottom Section):

- New text, document, etc. (Orange Rectangle, Bottom Left): This represents new, unseen text data (e.g., a new incoming email) that the system needs to classify. This text does not have a predefined label yet.
- Feature Vectors (Blue Vertical Rectangle, Bottom Middle): Just like in the training phase, the new incoming text must first be converted into the same type of numerical feature vectors using the exact same feature extraction method used during training. This ensures consistency between the data the model was trained on and the data it's making predictions on.
- Predictive model (Grey Diamond): The newly created feature vectors are fed into the previously trained predictive model. The model applies the learned patterns and rules to these features.
- Expected (Yellow Rectangle, Bottom Right): The output of the predictive model is the "expected" or predicted label for the new text. In the context of spam detection, this would be the model's prediction of whether the new email is "Spam" or "Ham."

IV. RESEARCH GAPS

A. Handling Highly Imbalanced and Adversarial Data:

Rare Event Detection (False Negatives): Spam detection is an imbalanced classification problem (ham far outnumbers spam). A critical gap is improving the detection of rare but highly damaging spam types (e.g., highly targeted phishing or zero-day exploits) without generating excessive false positives. Current methods may not be optimized for detecting subtle changes in the minority class's patterns.

- Adversarial Attacks: Spammers actively try to fool ML models (e.g., by adding legitimate-looking words, character substitutions, image-based text). There's a need for more robust ML frameworks that are resilient to adversarial attacks and can perform effectively even when facing intentionally manipulated spam.

B. Enhancing Interpretability and Explainability (XAI) in Real-Time:

- Understanding Drift Causes: While models can detect *that* drift has occurred, a significant gap lies in providing interpretable insights into *why* the drift happened and *what specific features or patterns* are changing. This interpretability is crucial for security analysts to understand new threats and for regulatory compliance, especially in sensitive sectors like banking.
- XAI for Complex Models: Integrating XAI techniques (like SHAP, LIME) with deep learning or complex ensemble models for spam detection in real-time adds computational overhead. Research is needed to develop computationally efficient XAI methods that can provide real-time explanations without degrading performance.

C. Privacy-Preserving and Distributed Learning (e.g., Federated Learning):

- Cross-Organizational Collaboration: Banking and other sectors deal with sensitive data, making centralized data sharing for training difficult. Federated Learning (FL) offers a promising path for collaborative spam detection without sharing raw data. However, research gaps exist in:
- Developing FL algorithms that are robust to varying drift characteristics across different clients (e.g., different banks seeing different types of spam).
- Ensuring efficient and truly privacy-preserving drift detection within FL frameworks, especially when dealing with new, unseen attack vectors.

V. RESEARCH OBJECTIVES

- To identify the gaps in the spam detection and filtering domain by conducting a comprehensive survey of the proposed techniques and spam's nature.
- To enhance email security and filtration of spam emails by using machine learning methods.
- To challenge the currently faced by spam filtering models and the effects of those challenges on the models' efficiency
- To understand machine learning's role in spam detection is provided.
- To categorizes different spam detection methods according to machine learning techniques to better understand concepts jointly.
- To detect spam better and add more security to email platforms.

VI. RESEARCH METHODOLOGY

A. Data Collection and Preparation

- Data Source: Acquire a diverse and representative dataset of emails, comprising both legitimate (ham) and spam messages.
    - Initial Sources: Publicly available benchmark datasets (e.g., Enron, SpamAssassin, Ling-Spam, TREC).
    - Addressing Data Freshness: Acknowledge the need for more recent data, potentially by simulating current spam characteristics or identifying newer, albeit limited, datasets. This is crucial for addressing the

"Dynamic Spam Evolution" research gap.

- Data Volume: Aim for a sufficiently large dataset to ensure robust model training and avoid overfitting.
- Data Labeling: Ensure all emails are accurately pre-labeled as 'spam' or 'ham'.
- Data Preprocessing:
    - Text Cleaning: Remove HTML tags, special characters, extra whitespace, URLs (or replace them with generic tokens), and numbers (if not relevant to classification).
    - Case Normalization: Convert all text to lowercase.
    - Tokenization: Break down text into individual words or sub-word units (tokens).
    - Stop Word Removal: Eliminate common words (e.g., "a", "the", "is") that carry little semantic meaning for classification.
    - Stemming/Lemmatization: Reduce words to their root forms (e.g., "running" -> "run") to reduce vocabulary size and improve feature representation.

B. Feature Engineering/Text Vectorization

- Transform preprocessed text into numerical feature vectors suitable for ML algorithms.
- Bag-of-Words (BoW): Create a vocabulary of unique words and represent each email as a vector of word counts.
- TF-IDF (Term Frequency-Inverse Document Frequency): Assign weights to words based on their frequency in a document and inverse frequency across the entire corpus, highlighting important words.
- Word Embeddings (e.g., Word2Vec, GloVe, FastText): Explore pre-trained word embeddings to capture semantic relationships between words, which can improve performance, especially for nuanced spam.
- N-grams: Incorporate sequences of words (bigrams, trigrams) to capture phrases and contextual meaning, which can be particularly useful for identifying specific spam patterns.
- Character N-grams: Useful for detecting obfuscated words or specific malicious patterns.

C. Model Development and Selection

- Machine Learning Algorithms (Binary Classifiers):
    - Traditional ML:
        - Naive Bayes (Multinomial Naive Bayes, Bernoulli Naive Bayes): Baseline model, computationally efficient.
        - Support Vector Machines (SVM): Effective for high-dimensional data.
        - Logistic Regression: Probabilistic linear classifier.
        - Random Forest / Gradient Boosting Machines (e.g., XGBoost, LightGBM): Ensemble methods known for high accuracy.
    - Deep Learning (DL) Models (for advanced exploration):
        - Recurrent Neural Networks (RNNs) / LSTMs: Good for sequential data like text.
        - Convolutional Neural Networks (CNNs): Can capture local patterns (n-grams) effectively.
        - Transformer-based models (e.g., BERT,

DistilBERT - if resources permit): State-of-the-art for natural language understanding, potentially offering superior accuracy but higher computational cost.

- Framework Development: Design a modular framework that allows for easy swapping of different preprocessing techniques, feature extractors, and ML models to facilitate experimentation and comparison.

D. Experimentation and Evaluation

- Train-Test Split: Divide the prepared dataset into training and testing sets (e.g., 70-30% or 80-20%) to evaluate the model's generalization capability on unseen data.
- Cross-Validation: Employ k-fold cross-validation (e.g., 5-fold or 10-fold) on the training set to get a more robust estimate of model performance and tune hyperparameters.
- Hyperparameter Tuning: Use techniques like Grid Search or Random Search to find the optimal hyperparameters for each selected model.
- Performance Metrics: Evaluate the models using a comprehensive set of metrics, crucial for imbalanced datasets:
    - Accuracy: Overall correct predictions.
    - Precision: Of all emails classified as spam, how many were actually spam? (Minimizes false positives)
    - Recall (Sensitivity): Of all actual spam emails, how many were correctly identified? (Minimizes false negatives)
    - F1-Score: Harmonic mean of precision and recall, providing a balanced measure.
    - ROC AUC Curve (Receiver Operating Characteristic Area Under the Curve): Measures the model's ability to distinguish between classes across various threshold settings.
    - Confusion Matrix: Provides a detailed breakdown of true positives, true negatives, false positives, and false negatives.

VII. CONCLUSION

In the last two decades, spam detection and filtration gained the attention of a sizeable research community. The reason for a lot of research in this area is its costly and massive effect in many situations like consumer behavior and fake reviews. The survey covers various machine learning techniques and models that the various researchers have proposed to detect and filter spam in emails and IoT platforms. The study categorized them as supervised, unsupervised, reinforcement learning, etc. The study compares these approaches and provides a summary of learned lessons from each category. This study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labeled dataset for the supervised model training is a crucial and time-consuming task. Supervised learning algorithms SVM and Naıve Bayes outperform other models in spam detection. The study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

VI. FUTURE DIIRECTIONS

Several emerging trends in AI and machine learning present promising opportunities for enhancing real-time concept drift detection within the banking sector.

- Federated Learning, a distributed machine learning approach, enables collaborative model training and concept drift detection across various banking institutions or within different departments of a large bank. This can happen without the need to share sensitive raw data. This method directly addresses critical data privacy and security concerns common in the financial industry and offers the potential to learn from more diverse and larger datasets. Current research is dedicated to developing specialized drift-aware federated learning algorithms that can effectively identify and

adapt to concept drift in these decentralized environments.

- Explainable AI (XAI) is another significant trend with great potential for improving real-time concept drift detection in banking. XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), can offer insights into which features are most impacted by drift and explain changes in a model's behavior over time. This interpretability is vital for understanding the root causes of detected drift, which then facilitates more precise and effective model updates. Integrating XAI with concept drift detection methods can significantly boost the transparency and trustworthiness of banking AI models, aiding in regulatory compliance and building greater confidence among stakeholders.
- Continual Learning, also known as lifelong or incremental learning, is an emerging paradigm focused on enabling AI models to continuously learn from new data streams and adapt to concept drift without suffering from "catastrophic forgetting"—where the model loses previously acquired knowledge. Techniques like online learning algorithms, which can process data sequentially, and adaptive ensemble methods are key components of continual learning frameworks. These approaches are particularly well-suited for the banking domain, where data patterns are constantly evolving, and they offer a way to maintain the long-term performance of AI models more efficiently than periodic retraining from scratch.

REFERENCES

[1] H. Faris, A. M. Al-Zoubi, A. A. Heidari et al., "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks," Information Fusion, vol. 48, pp. 67–83, 2019.

[2] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," Artificial Intelligence Review, vol. 29, no. 1, pp. 63–92, 2008.

[3] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email classification using artificial neural network," International Journal for Academic Development, vol. 2, 2018. 16 Security and Communication Networks

[4] N. Udayakumar, S. Anandaselvi, and T. Subbulakshmi, "Dynamic malware analysis using machine learning algorithm," in Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, India, December 2017.

[5] S. O. Olatunji, "Extreme Learning machines and Support Vector Machines models for email spam detection," in Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, Windsor, Canada, April 2017.

[6] J. Dean, "Large scale deep learning," in Proceedings of the Keynote GPU Technical Conference, San Jose, CA, USA, 2015.

[7] J. K. Kruschke and T. M. Liddell, "Bayesian data analysis for newcomers," Psychonomic Bulletin & Review, vol. 25, no. 1, pp. 155–177, 2018.

[8] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: dark of the social networks," Journal of Network and Computer Applications, vol. 79, pp. 41–67, 2017.

[9] A. Barushka and P. Hajek, "Spam filtering using regularized ´ neural networks with rectified linear units," in Proceedings of the Conference of the Italian Association for Artificial Intelligence, Springer, Berlin, Germany, November 2016.

[10] F. Jamil, H. K. Kahng, S. Kim, and D. H. Kim, "Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms," Sensors, vol. 21, no. 5, p. 1640, 2021.

[11] M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier

systems," Soft Computing, vol. 22, no. 21, pp. 7281–7291, 2018.

[12] X. Zheng, X. Zhang, Y. Yu, T. Kechadi, and C. Rong, "ELMbased spammer detection in social networks," 5e Journal of Supercomputing, vol. 72, no. 8, pp. 2991–3005, 2016.

[13] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, Article ID 102419, 2020.

[14] N. Kumar and S. Sonowal, "Email spam detection using machine learning algorithms," in Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108–113, Coimbatore, India, 2020.

[15] I. Santos, Y. K. Penya, J. Devesa, and P. G. Bringas, "N-gramsbased file signatures for malware detection," ICEIS, vol. 9, no. 2, pp. 317–320, 2009.

[16] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "On the capability of evolved spambots to evade detection via genetic engineering," Online Social Networks and Media, vol. 9, pp. 1–16, 2019.

[17] A. J. Saleh, A. Karim, B. Shanmugam et al., "An intelligent spam detection model based on artificial immune system," Information, vol. 10, no. 6, p. 209, 2019.

[18] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: a review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3–24, 2007.

[19] E. Blanzieri and A. Bryl, E-mail Spam Filtering with Local SVM Classifiers, University of Trento, Trento, Italy, 2008.

[20] H. Bhuiyan, A. Ashiquzzaman, T. Islam Juthi, S. Biswas, and J. Ara, "A survey of existing e-mail spam filtering methods considering machine learning techniques," Global Journal of Computer Science and Technology, vol. 18, 2018.

[21] A. Asuncion and D. Newman, "UCI machine learning repository," 2007, https://archive.ics.uci.edu/ml/index.php.

[22] T. Vyas, P. Prajapati, and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," in Proceedings of the 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE, Tamil Nadu, India, March 2015.

[23] L. N. Petersen, "(e ageing body in monty Python live (mostly)," European Journal of Cultural Studies, vol. 21, no. 3, pp. 382–394, 2018.

[24] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar, "Characterizing botnets from email spam records," LEET, vol. 8, pp. 1–9, 2008.

[25] W. N. Gansterer, A. G. K. Janecek, and R. Neumayer, "Spam filtering based on latent semantic indexing," in Survey of Text Mining II, pp. 165–183, Springer, New York, NY, USA, 2008.