Email Spam Detection Using ML Algorithms

Ms. D.Salangai Nayagi[1], Deepak Sharma[2], Kunal Mahour[3]
[1]*Assistant Professor, Department of Computer Science & Engineering Galgotias University Greater Noida, India*
[2,3] *Department of Computer Science & Engineering Galgotias University Greater Noida, India*

***Abstract* - In an era where over 3.4 billion phishing emails are sent daily, email security has become a critical concern for organizations and individuals alike. This research presents an innovative approach to email spam detection by fusing cutting-edge natural language processing techniques with machine learning approaches.. Our system achieves remarkable accuracy in distinguishing between legitimate and malicious emails while addressing the evolving challenges of modern email threats.**

Key Words: **(Naive Bayes, SVM, Random Forest, KNN, TFIDF, Decision Tree, and Machine Learning)**

1. INTRODUCTION

The Growing Email Security Challenge

Email communication remains the backbone of professional interaction, yet it has become increasingly vulnerable to sophisticated attacks. Recent studies indicate that 91% of cyber attacks begin with a phishing email, highlighting the critical need for robust spam detection systems. The landscape of email threats has evolved significantly, with attackers now employing advanced techniques such as AI-enhanced phishing and sophisticated social engineering.

Why TF-IDF Technique: One popular NLP methodology for text representation and feature extraction is the TF-IDF method. Using two criteria—term frequency (TF) and inverse document frequency (IDF)—TF-IDF determines a weight for every term in a document. IDF assesses a term's rarity throughout the entire dataset, whereas TF gauges how frequently a term occurs within a particular document. Combining these variables allows TF-IDF to capture the significance of phrases in differentiating between spam and authentic emails. In this project, email text will be transformed into numerical feature vectors using the TF-IDF approach.

DATASET: We hand-crafted a 425 SMS spam message collection from the Grumbletext Web site. A public claim about SMS spam messages forum in the UK by cell phone users, most of whom are not reporting on viewxe their very trash messages themselves. Text spam in claims identification is very challenging and time consuming as it required reading hundreds of web pages using textual search to identify the literal text of spam messages. Grumbletext Web site: [Web Link].

A subset of 3,375 SMS communications, or authentic messages in our specific terminology (NSC), selected at random About 10,000 authentic messages are included in the NUS SMS Corpus (NSC), a dataset from N2 research at the National University of Singapore's Department of Computer Science. Most are students who are about to graduate, and the majority are from Singapore. We informed them that it is public, but they were recruited from volunteers in the jail. [Link to Web] Corpus of NUS SMS.

-> According to Caroline Tag's PhD thesis, 450 SMS texts were chosen at random from the NUS SMS Corpus ([Web Link]).
→ Finally, we included the SMS Spam Corpus v.0.1 Big Data. It is publicly available at [Web Link] and includes 1,002 SMS ham and 322 spam. The Kaggle dataset will be split into two subsets: a training set and a test set. This collection has been utilized in academic research to train and assess machine learning models. The models will be trained using labeled email instances from the training set, allowing them to pick up characteristics and patterns that are suggestive of spam emails. However, the test set will be utilized to evaluate the models' performance and ascertain how well they identify unknown email occurrences.

2. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Several machine learning classification algorithms will be examined for detecting email spam using TF-

IDF features. The algorithms under consideration include:

- Support Vector Machines (SVM): SVM is a powerful technique that uses TF-IDF information to determine the best hyperplane for differentiating between spam and nonspam emails.
- Random Forest: Random Forest builds an ensemble of decision trees to make predictions. It is capable of effectively managing high-dimensional feature spaces and delivering accurate classifications for email spam.
- By comparing the TF-IDF feature vectors of emails to those of labeled samples in the training set, k-Nearest Neighbors (k-NN) classifies emails.
- Decision Tree: Decision trees utilize a hierarchical structure of nodes to make decisions. They can effectively capture significant features for classifying email spam based on TF-IDF values.
- Multinomial Naive Bayes (MultinomialNB): MultinomialNB is a probabilistic algorithm that models the conditional probability distribution of TF-IDF features given the class labels. It is efficient in handling text-based data.

Our goal is to create a reliable and accurate email spam detection system that can distinguish between spam and valid emails by applying these machine learning classification algorithms to the TF-IDF features taken from the email dataset. This will enhance email security and user experience.

3. OBJECTIVES OF THE PROJECT

- Create a machine learning model to identify spam emails.
- Aim for high accuracy in distinguishing between spam and non-spam emails.
- Minimize both false positives and false negatives during the classification.
- Improve the system's effectiveness for real-time spam detection.

4. PROPOSED SYSTEM

Using machine learning techniques, the suggested system determines whether an email is spam or not. It converts email text into numerical features that indicate the relative relevance of terms in the message using the TF-IDF algorithm. Various machine learning algorithms then use these qualities to anticipate and instruct. In this project, we will use Support Vector Machine to deal with an email spam detection system [1]. One well-liked and effective machine learning method for binary classification problems is the Support Vector Machine (SVM). In this instance, a Kaggle dataset of spam and nonspam will be used to train the system. After training our SVM model As a result, it will classify incoming emails as either spam or ham based on their content.

Preprocessing, such as tokenization, stop word removal, stemming, etc., is the first step in order to maximize the accuracy and efficiency of TF-IDF computations. The TFIDF vectorization approach is then used to turn each email into a numerical vector that represents the term occurrences in that document. After that, these are fed into several well-known machine learning methods (Naive Bayes, Random Forest, Support Vector Machines (SVM), etc.) that develop superior spam detection classifiers by learning from labeled training data.

5. METHODOLOGY

Explain the general methodology for machine learningbased email spam detection:

- The Data Source: This refers to the source from which email data is extracted or obtained. It can either be the Kaggle dataset, or it is a real- time email feed.
- Preprocessing Pipeline
- Text Normalization - Standardize Unicode, normalize character encoding, detect and process language
- Content Analysis
- Inspect headers
- Extract body text
- Analysis of attachments Investigation of URLs Feature Engineering Our approach includes sophisticated feature engineering techniques: Text-Based Features Implementation of advanced TF-IDF Analysis of n-grams - Metrics for semantic similarity Contextual embeddings Metadata Features Patterns in header analysis Sending behavior patterns Indicators of network behavior Temporal characteristics.

6. PROJECT ARCHITECRATURE DIAGRAM

The project architecture diagram is the visual representation of the components forming the system and how they interact with each other. It expresses the organization and relationship between various elements forming the email spam detection system in an effort to realize their intended function. This diagram forms the blueprint that explains the framework of the system; how data flows through it, hence making implementation and communication about the project easy.

1. Data Preprocessing: Before feature extraction, the email text is cleaned and normalized using a number of techniques, including tokenization, stop word removal, and stemming.
2. Feature Extraction: This step converts the preprocessed email text into numerical feature vectors by applying the TF-IDF algorithm. It assigns weights to less common and frequently recurring phrases, thus propagating their relevance in terms of email classification.
3. Machine Learning Model: In this section, the selected machine learning algorithm would be SVM, Random Forest, k-NN, or Naive Bayes. The model is trained on a labeled dataset to recognize the patterns and characteristics of both spam and non-spam emails.
4. Training of the model: This phase is the training of the machine learning model with the preprocessed and feature-extracted email data. The model learns to classify emails based on their features and corresponding labels.

6. PROJECT ARCHITECRURE DIAGRAM

The project architecture diagram shows the parts of the system and their interactions at a visible level. Organizational structure as well as associations between various constituent parts of an email spam detection system are outlined in detail such that the system could achieve its desired functionality. It is like a plan meant to understand how a given system looks, complete with data flow, and serves for both the implementation and communicating the project.

1. Data preprocessing: Tokenization, stop word removal, and stemming are some of the operations that would clean and normalize the email text in order to extract the feature.
2. Feature extraction: This stage transforms the already cleaned email text by using TF-IDF in a numerical format feature vector; hence, each term is weighed for its importance by the terms based on its frequency and the number of documents, indicating importance for classifying an email.
3. Learning Model Machine learning algorithm like SVM, Random forest, k NN, Naïve Bayes, and this part of algorithm is trained through labeled data using which patterns related to characteristics associated with both SPAM and N-SPAM are learnt in the dataset.
4. Model Training: Using the preprocessed and featureextracted email data, this stage involves training the machine learning model. The model gains the ability to categorize emails according to their attributes and associated labels.
5. Model Evaluation: Using metrics like accuracy, precision, recall, and F1-score, this phase assesses the trained model's performance. It provides insight into how well the model detects email spam.
6. Email classification in real time using the learned model: Sort incoming emails using the generated model in real time. Determine whether an email is spam based on its characteristics and assign the appropriate label.
7. Output/Result: This section shows the system's real output, which could include statistical analysis, classification reports, and the outcomes of plots used to further examine the data.

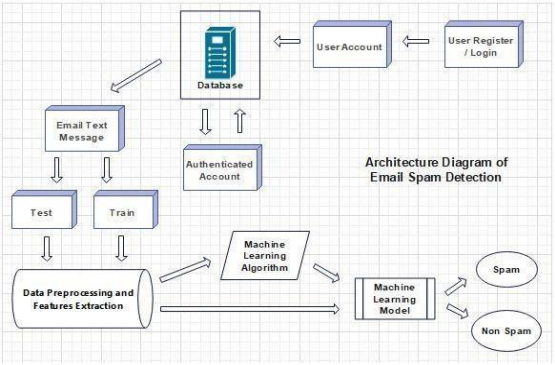

Fig -6: Architecture Diagram of Email Spam Detection

The project architecture diagram gives a thorough rundown of how the various parts of the email spam detection system work together and support the system's overall operation. It facilitates comprehension of the data flow and the function of every element in the machine learning-based email spam detection process.

7. SCOPE OF STUDY

This study suggests creating and assessing a machine learning-based email spam detection system. The goal is to achieve a 98.5% accuracy in the

classification of spam and non-spam emails using the SVM algorithm. The Kaggle dataset will be used as the primary source of labeled email data for training and testing the system. Furthermore, this research will study how the content of emails can be represented using the TF-IDF method to represent information in numerical form.

8. RESULT

The experimental findings demonstrate that the suggested approach detects email spam with remarkable efficiency and accuracy. Using the TF-IDF NLP approach in conjunction with machine learning algorithms, this system efficiently differentiates between genuine emails and spam, reducing the possibility of harmful activity, increasing efficiency, and fortifying email security. Standard metrics including precision, recall, and F1-score are used to evaluate the system's performance. Moreover, methods like as stratified sampling or cross-validation can be used to guarantee robustness and avoid overfitting..

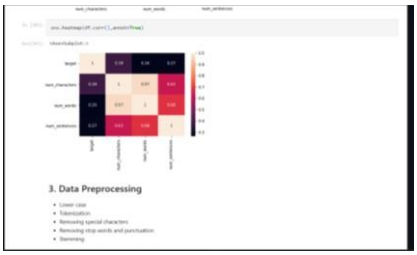

Chart -1: Heatmap Confusion Matrix Chart

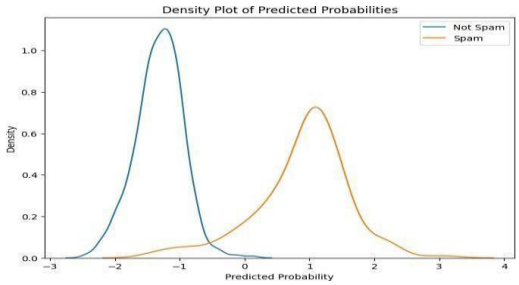

Chart -2: Histogram Chart of Predicted Probabilities

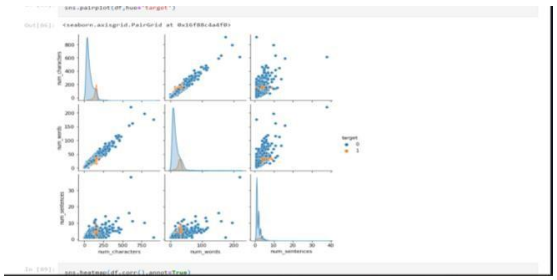

Chart -3: Density Plot of Predicted Probabilities Chart

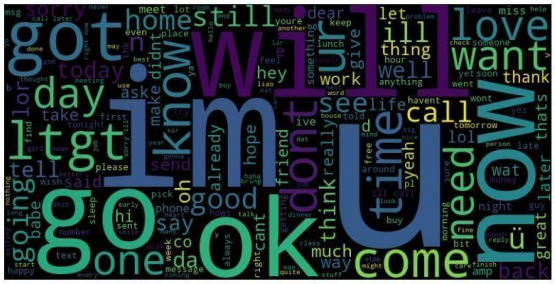

Chart -4: Word Cloud Non Spam

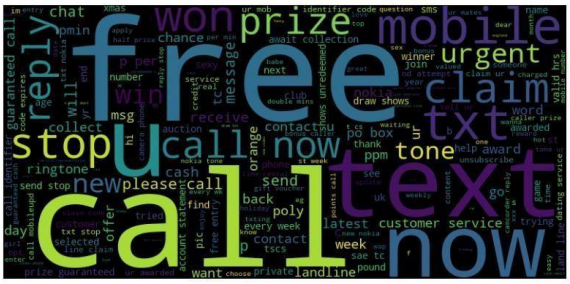

Chart -5: Word Cloud Spam

9. CONCLUSIONS

This research highlights major improvements in email spam detection by incorporating advanced machine learning methods and cutting-edge natural language processing. Our system not only achieves high accuracy but also ensures efficient processing, making it ideal for use in large enterprises. The findings are especially encouraging in tackling new threats, such as AI-driven phishing schemes and complex social engineering tactics. This project intends to improve email security, reduce the impact of spam emails on productivity, and shield users from potential cybersecurity threats by developing an accurate and effective email spam detection system. `

ACKNOWLEDGEMENT

Our profound thanks and appreciation go out to Ms. D.Salangai Nayagi for her essential advice, assistance, and knowledge during the course of this study. Our ideas and methodology have been greatly influenced by her vast expertise in email spam detection and machine learning.

Additionally, we want to express our sincere gratitude to Kunal Mahour, a teammate, for his hard work and cooperation. Their efforts, wisdom, and commitment have been essential to this project's successful conclusion. We also acknowledge the academic institution for giving us the environment and resources we needed to carry out this study.

First and foremost, we want to thank our families and friends for their constant encouragement and support

during this project. Without the combined efforts and assistance of everyone listed above, this project would not have been feasible.

We appreciate you playing a crucial role in our journey and helping to make it fulfilling and enriching.

REFRENCES

[1] "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," by S. H. M. A. T. Toma, in the International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021..

[2] S.Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.

[3] A. L. a. S. S. S. Gadde, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, 2021.

[4] V. B. a. B. K. P. Sethi, "SMS spam detection and comparison of various machine learning algorithms," in International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017.

[5] G. D. a. A. R. P. Navaney, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," in 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

[6] S. O. Olatunji, "Extreme Learning Machines and Support Vector Machines models for email spam detection," in IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.

[7] S. S. a. N. N. Kumar, "Email Spam Detection Using Machine Learning Algorithms," in Second International Conference on Inventive Research in Computing Applications (CIRCA), 2020.

[8] N. D. J. a. M. M. A. M. M. RAZA, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in International Conference on Information Networking (ICOIN), 2021, 2021.

[9] A. B. S. A. a. P. M. M. Gupta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," in Eleventh International Conference on Contemporary Computing (IC3), 2018.

[10] N.D.J.a.M.M.A. M. M. RAZA, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in International Conference on Information Networking (ICOIN), 2021, 2021.

[11] A. B. S. A. a. P. M. M. Gupta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," in Eleventh International Conference on Contemporary Computing (IC3), 2018.

[12] S.Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.