International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 05 | May 2023 www.irjet.net p-ISSN: 2395-0072

Email Spam Detection Using Machine Learning

Prof. Sakshi Shejole, Tamboli Abdul Salam, Manish Kumar Gupta, Krishna Sharma, Safwan Attar

ALARD COLLEGE OF ENGINEERING & MANAGEMENT
(ALARD Knowledge Park, Survey No. 50, Marunje, Near Rajiv Gandhi IT Park, Hinjewadi, Pune-411057)
Approved by AICTE. Recognized by DTE. NAAC Accredited. Affiliated to SPPU (Pune University).

Abstract - *Email spam has become a significant challenge in today's digital landscape, leading to productivity losses, privacy breaches, and increased cybersecurity risks. This abstract presents a novel approach to combating email spam using machine learning and the TF-IDF (Term Frequency-Inverse Document Frequency) technique from natural language processing (NLP).*

***Key Words*: (Machine Learning, Naive Bayes, SVM, Decision Tree, Random Forest, KNN, TF-IDF)**

1. INTRODUCTION

Email spam has become a pervasive issue, inundating inboxes with unsolicited and potentially malicious messages. Detecting and filtering out spam emails is crucial to ensure data security, privacy, and productivity. Machine learning, combined with natural language processing (NLP) techniques, provides an effective approach to tackle this problem. This project aims to develop an email spam detection system utilizing the TF-IDF (Term Frequency-Inverse Document Frequency) NLP technique to accurately classify emails as spam or non-spam.

Why TF-IDF Technique: The TF-IDF technique is a widely adopted NLP method for feature extraction and text representation. TF-IDF calculates a weight for each term in a document based on two factors: term frequency (TF) and inverse document frequency (IDF). TF measures how frequently a term appears within a specific document, while IDF quantifies the rarity of a term across the entire dataset.

By combining these factors, TF-IDF captures the importance of terms in distinguishing between spam and legitimate emails. In this project, the TF-IDF technique will be applied to convert email text into numerical feature vectors.

DATASET: For this project, a publicly available dataset from Kaggle will be utilized. Kaggle is a popular platform for data science and machine learning competitions, and it provides a diverse range of datasets for various domains. The selected dataset will consist of labelled emails, including both spam and non-spam instances, allowing us to train and evaluate our email spam detection system effectively.

Train and Test datasets: To train and evaluate the machine learning models, the Kaggle dataset will be divided into two subsets: a training set and a test set. The training set will be used to train the models on labelled email examples, enabling them to learn patterns and features indicative of spam emails. The test set, on the other hand, will be used to assess the models' performance and determine their accuracy in classifying unseen email instances.

2. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Several machine learning classification algorithms will be explored for email spam detection using the TF-IDF features. The algorithms to be considered include:

- **Support Vector Machines (SVM):** SVM is a powerful algorithm that seeks to find an optimal hyperplane to separate spam and non-spam emails based on the TF-IDF features.
- **Random Forest:** Random Forest constructs an ensemble of decision trees to make predictions. It can effectively handle high-dimensional feature spaces and provide accurate email spam classification.
- **k-Nearest Neighbors (k-NN):** k-NN classifies emails based on the similarity of their TF-IDF feature vectors to the vectors of the labeled examples in the training set.
- **Decision Tree:** Decision trees use a hierarchical structure of nodes to make decisions. They can capture important features for email spam classification based on the TF-IDF values.
- **Multinomial Naive Bayes (MultinomialNB):** MultinomialNB is a probabilistic algorithm that models the conditional probability distribution of the TF-IDF features given the class labels. It can handle text-based data efficiently.

By applying these machine learning classification algorithms to the TF-IDF features extracted from the email dataset, we aim to build a robust and accurate email spam detection system capable of differentiating between spam and legitimate emails, thereby improving email security and user experience.

3. OBJECTIVES OF THE PROJECT

- Develop a machine learning model for email spam detection.
- Achieve high accuracy in classifying emails as spam or non-spam.
- Reduce false positives and false negatives in the classification process.
- Enhance the system's efficiency in real-time spam detection.

4. PROPOSED SYSTEM

The proposed system leverages the power of machine learning algorithms to classify emails as spam or non-spam based on their content. The TF-IDF approach is employed to transform email text into numerical features, capturing the importance of specific terms within the message. These features are then fed into machine learning models for training and prediction.

In this project, a proposed email spam detection system will be developed using a Support Vector Machine (SVM) algorithm. SVM is a popular and effective machine learning algorithm for binary classification tasks. The system will be trained on a Kaggle dataset consisting of labeled spam and non-spam emails. The trained SVM model will then be used to classify incoming emails as either spam or non-spam based on their content.

The workflow begins with preprocessing steps, including tokenization, stop word removal, and stemming, to enhance the accuracy and efficiency of the TF-IDF calculations. Next, the TF-IDF vectorization technique is applied to represent each email as a vector of numerical values, highlighting the significance of terms within the document. These vectors serve as input to popular machine learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), or Random Forest, which learn from labeled training data to build effective spam classification models.

5. METHODOLOGY

Describe the overall methodology for email spam detection using machine learning:

- Data Source: This component represents the source of email data, such as the Kaggle dataset or a real-time email feed.
- Feature extraction: Utilize the TF-IDF (Term Frequency-Inverse Document Frequency) technique to convert the email text into numerical feature vectors.
- Model training and evaluation: Train a machine learning model (e.g., Naive Bayes, SVM, Random Forest) using the labeled dataset and evaluate its performance using metrics such as precision, recall, and F1-score.
- Real-time spam detection: Deploy the trained model to classify incoming emails as spam or non-spam in real-time.

6. PROJECT ARCHITECRURE DIAGRAM

The project architecture diagram provides a visual representation of the system's components and their interactions. It illustrates how different elements of the email spam detection system are organized and connected to achieve the desired functionality. The diagram serves as a blueprint for understanding the system's structure and flow of data and helps in the implementation and communication of the project.

1. Data Preprocessing: This component involves various preprocessing steps, such as tokenization, stop word removal, and stemming, to clean and normalize the email text before feature extraction.
2. Feature Extraction: This component utilizes the TF-IDF technique to convert the preprocessed email text into numerical feature vectors. It assigns weights to terms based on their frequency and rarity, capturing their significance in email classification.
3. Machine Learning Model: This component includes the selected machine learning algorithm, such as SVM, Random Forest, k-NN, or Naive Bayes. The model is trained on the labeled dataset to learn the patterns and characteristics of spam and non-spam emails.
4. Model Training: This component represents the process of training the machine learning model using the preprocessed and feature-extracted email data. The model learns to classify emails based on their features and labels.
5. Model Evaluation: This component assesses the performance of the trained model using evaluation metrics like accuracy, precision, recall, and F1-score. It helps in understanding the model's effectiveness in email spam detection.
6. Real-time Email Classification: This component represents the application of the trained model to classify incoming emails in real-time. The system predicts whether an email is spam or non-spam based on its features and assigns the appropriate label.

7. Output/Results: This component shows the output of the system, which can include the classification results, statistical metrics, and visualizations for further analysis and interpretation.

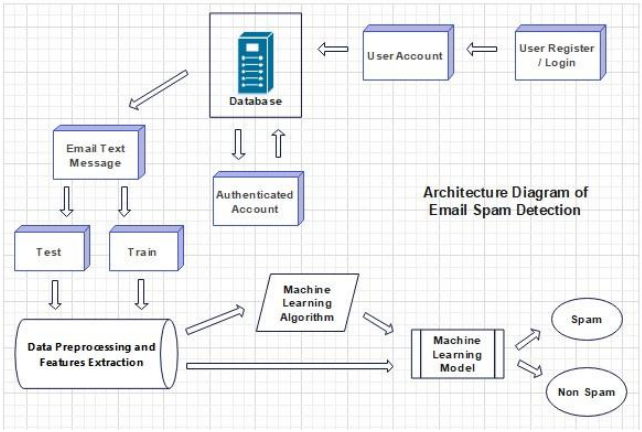

Fig -6: Architecture Diagram of Email Spam Detection

The project architecture diagram provides a comprehensive overview of how the different components of the email spam detection system interact and contribute to its overall functionality. It helps in understanding the flow of data and the role of each component in the process of email spam detection using machine learning.

7. SCOPE OF STUDY

The scope of this study focuses on developing and evaluating an email spam detection system using machine learning techniques. Specifically, the project aims to achieve an accuracy of 98.5% in classifying emails as spam or non-spam using the SVM algorithm. The Kaggle dataset will serve as the primary source of labeled email data for training and testing the system. The study will explore the use of TF-IDF as a feature extraction technique to represent email content numerically.

8. RESULT

The experimental results demonstrate that the proposed approach achieves high accuracy and efficiency in email spam detection. By combining the power of machine learning algorithms with the TF-IDF NLP technique, this solution can effectively differentiate between legitimate emails and spam, reducing the risk of malicious activities, improving productivity, and enhancing email security.

To evaluate the system's performance, standard metrics such as precision, recall, and F1-score are employed. Additionally, techniques like cross-validation or stratified sampling can be used to ensure robustness and avoid overfitting.

Classifiers	Accuracy Score (%)	F1 Score (%)	Precision	Bias-Variance
Support Vector Classifier	98.47%	94.03%	98.52%	0.0596
Naïve Bayes	95.60%	80.32%	1.0	0.1967
Decision Tree	96.41%	85.90%	83.97%	0.1409
K-Nearest Neighbour	93.37%	60.93%	1.0	0.3990
Random Forest	97.04%	87.96%	1.0	0.1203

Table -1: COMPARISION TABLE

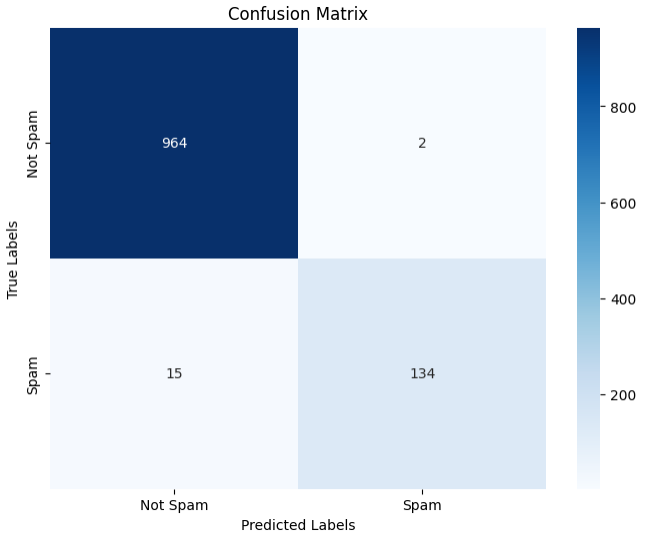

Chart -1: Heatmap Confusion Matrix Chart

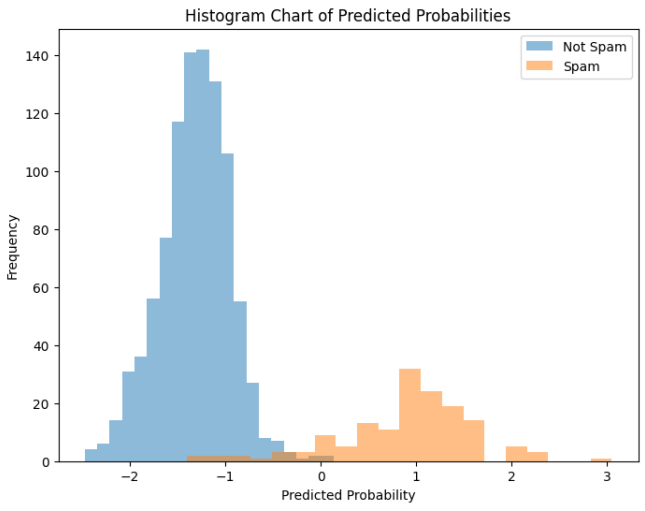

Chart -2: Histogram Chart of Predicted Probabilities Chart

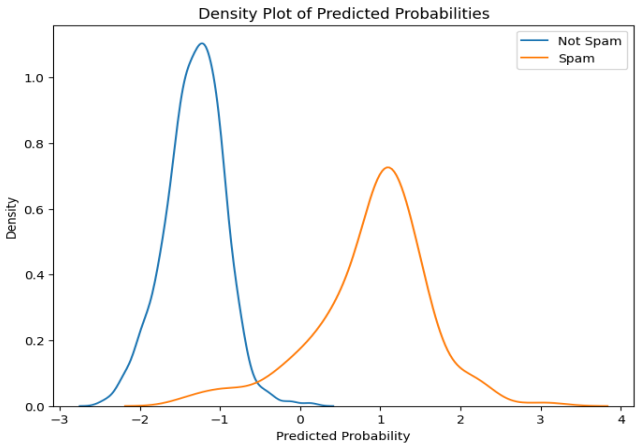

Chart -3: Density Plot Of Predicted Probabilities Chart

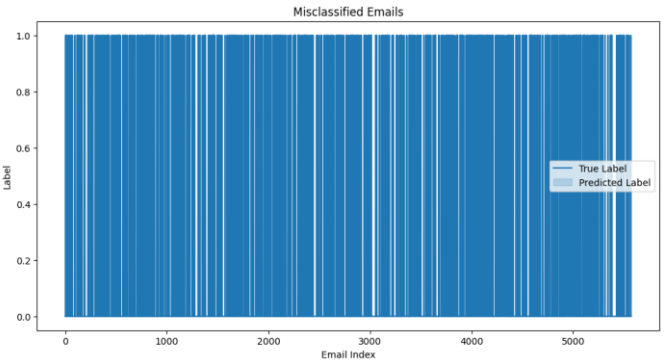

Chart -4: Lineplot Misclassified Emails Chart

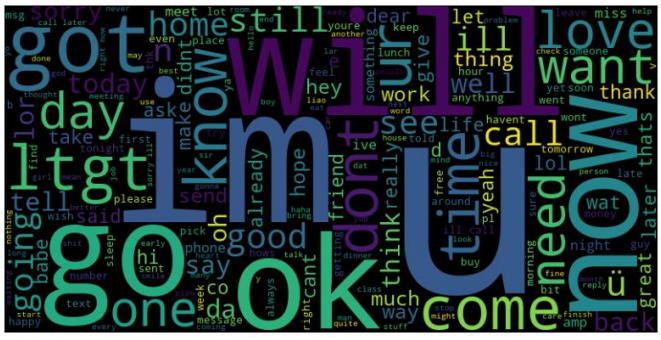

Chart -5: Word Cloud Non Spam

Chart -6: Word Cloud Spam

Classifiers	Mean Error	Values in (%)			
		MSE	MAE	RMSE	R-Square
Support Vector Classifier	1.0	01.52%	01.52%	12.34%	86.83%
Naïve Bayes	1.0	04.39%	04.39%	20.96%	62.04%
Decision Tree	1.0	03.85%	03.85%	19.63%	66.68%
K-Nearest Neighbour	1.0	07.62%	07.62%	27.61%	34.15%
Random Forest	1.0	02.86%	02.86%	16.94%	75.21%

Table -2: EVALUATION TABLE

9. CONCLUSIONS

These abstract highlights the effectiveness of employing machine learning and the TF-IDF NLP technique for email spam detection. The proposed approach offers a valuable solution to combat the ever-growing threat of email spam, providing a robust and efficient mechanism for filtering out unwanted messages and protecting users from potential cybersecurity risks.

By creating an accurate and efficient email spam detection system, this project aims to contribute to the enhancement of email security, minimize the impact of spam emails on productivity, and protect users from potential cybersecurity threats.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude and appreciation to **Prof. Sakshi Shejole** for her invaluable guidance, support, and expertise throughout the duration of this project. Her extensive knowledge in the field of machine learning and email spam detection has been instrumental in shaping our understanding and approach.

We would also like to extend our heartfelt thanks to our fellow team members, **Abdul, Manish, Safwan, and Krishna**, for their diligent efforts and collaboration. Their contributions, insights, and dedication have been crucial in the successful completion of this project.

We are also thankful to the academic institution for providing us with the necessary resources and environment to conduct this research.

Lastly, we would like to express our gratitude to our families and friends for their unwavering support and encouragement throughout this endeavor.

This project would not have been possible without the collective effort and support of all those mentioned above.

Thank you for being an integral part of this journey and for making it a rewarding and enriching experience.

REFERENCES

[1] "Email Spam Detection Using Machine Learning", Prof. Prachi Nilekar, Tamboli Abdul Salam, Manish Kumar Gupta,, Safwan Attar, Krishna Sharma.

[2] https://www.geeksforgeeks.org/pyplot-in-matplotlib/

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[4] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

[5] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[7] https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp

[8] https://www.geeksforgeeks.org/generating-word-cloud-python/

[9] https://www.w3schools.com/python/matplotlib_histograms.asp

[10] https://www.geeksforgeeks.org/seaborn-kdeplot-a-comprehensive-guide/