International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Spam E-mail Detection Using Machine Learning Techniques

Karan Katariya[1], Aniket Desai[2], Mohmedvaaris Ghaswala[3], Monali Parikh[4]

[123]*Department of Information Technology, Institute of Information Technology, Krishna School Of Emerging Technology & Applied Research, KPGU University , Varnama, Vadodara, Gujarat, India*

[4] *Assistant Professor, Department of Information Technology and Engineering, Krishna School Of Emerging Technology & Applied Research, KPGU University. Varnama, Vadodara, Gujarat, India*

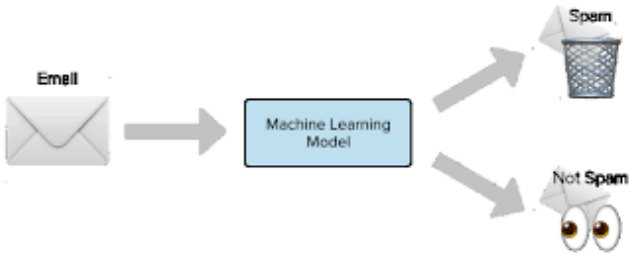
Abstract -In recent years, email spams have emerged as a significant concern, coinciding with the rapid expansion of Internet usage. Certain individuals exploit this medium for unlawful activities, including phishing and fraudulent schemes. These malicious actors often distribute harmful links via spam emails, potentially compromising computer systems and infiltrating personal data. The implementation of email spam detection mechanisms is crucial for preventing unwanted messages from cluttering users' inboxes, thereby enhancing their overall experience. This project aims to identify spam emails using machine-learning techniques. Machine learning, which is a subset of Artificial Intelligence, enables systems to learn and improve based on experience without explicit programming. This study focuses on the logistic regression algorithm, which is a probabilistic classifier that predicts outcomes based on the likelihood of an object. This method was chosen for email spam detection because of its superior precision and accuracy.

1.INTRODUCTION

Electronic mail remains a crucial and widely adopted tool for both personal and business communications. Nevertheless, this vital medium continues to be susceptible to spam — unwanted and potentially dangerous messages that seek to deceive recipients into revealing confidential information, buying undesired goods, or accessing harmful websites. As the quantity and complexity of spam emails increase, the demand for precise, effective, and automated spam identification systems has become critical.

Spam detection typically involves classifying emails as either spam or legitimate based on their content and associated metadata. Conventional rule-based filtering methods are becoming less effective due to the constantly evolving nature of spam tactics. Machine learning techniques, especially supervised learning algorithms, have emerged as powerful tools for identifying spam by recognizing patterns in labeled datasets.

This study utilizes logistic regression, a commonly employed binary classification algorithm, to tackle the issue of spam detection. Logistic regression is a statistical approach that estimates the likelihood of a binary outcome using one or more predictor variables. In the realm of spam detection, logistic regression offers simplicity, interpretability, and effectiveness, making it an excellent choice for differentiating between spam and legitimate emails.



The objective of this research is to develop a logistic regression-based model for spam detection, evaluate its accuracy, and assess its performance on real-world email datasets. By investigating the predictive capabilities of logistic regression and comparing it to other algorithms, we aim to contribute to the advancement of efficient, dependable email filtering mechanisms that can improve security and user experience in digital communication..

2.Aim and Objectives of the Study

The primary goal of this project is to improve the accuracy and efficiency of detecting and managing spam emails. To achieve this, the system employs two distinct filtering models designed to identify and classify spam emails more effectively.

The first approach uses an "Opinion Rank" mechanism to evaluate a sender's credibility based on their email address. This evaluation is informed by a combination of high and inverse page rank measures, with the Opinion Rank algorithm calculating an average score to establish an overall credibility ranking. Following this, the system applies Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique that categorizes email content based on specific topics. This process aids in more effectively filtering spam emails, helping to reduce the volume of unwanted messages.

In summary, this project focuses on enhancing the efficiency of spam email identification and organization. By integrating Opinion Rank and LDA, the system aims to improve its ability to accurately detect and manage spam, resulting in a more streamlined and effective email communication experience.

3. LITRETURE REVIEW

Email: Email, also known as electronic mail, is a widely used method of sending electronic messages across computer networks. It is essential for both personal and professional communication, with accessibility provided by any internet connection. Emails are commonly used in business, formal, and personal contexts.

Categories of Spam:

- **Health**: Proliferation of bogus drug advertisements.
- **Products Promotion**: Fake promotions for counterfeit goods like clothing, watches, etc.
- **Adult Content**: Spam related to pornography and prostitution.
- **Marketing and Accounts**: Excessive emails related to loans, tax strategies, and financial offers.
- **Fraud**: Emails aimed at scamming individuals to access personal wealth.

Spam: Spam refers to unsolicited bulk emails that can overwhelm inboxes and reduce system performance. These emails often carry irrelevant content, and in some cases, can be harmful, containing malware or phishing attempts.

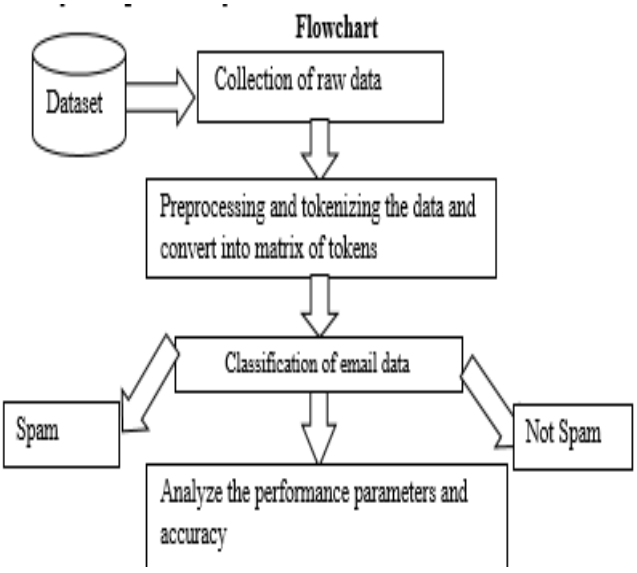
Spam Detection: Spam detection involves using spam filters to identify and block unwanted emails, including virus-infected ones. Various methods exist for detecting spam, such as blacklists, whitelists, and advanced machine learning techniques like Naive Bayes, Support Vector Machines (SVM), and neural networks.

Machine Learning Algorithms: Machine learning algorithms are a subset of artificial intelligence (AI) that utilize statistical models to make predictions. These algorithms analyze input data and generate the most accurate output, making them ideal for spam detection tasks by learning from labeled datasets of spam and non-spam emails.

Logistic Regression: Logistic regression is a widely used classification algorithm based on Bayes' theorem. It is particularly effective for solving binary classification problems, such as spam detection. The algorithm estimates the probability of an email belonging to a particular class (spam or ham) and classifies it accordingly. Despite being a simple linear model, logistic regression can be effective for spam detection when combined with proper feature engineering and preprocessing.

4 METHODOLOGY

In this research, we focus on applying Logistic Regression for email spam detection, leveraging machine learning techniques to classify emails as either "spam" or "ham" (non-spam).



The dataset used in this study, contains 5,572 email entries labelled as either **spam** or **ham** (non-spam), providing a balanced foundation for training and evaluating a machine learning model for email spam detection. The data is essential for supervised learning, where the model can learn to distinguish between

legitimate and unwanted emails based on labeled examples. Each entry includes two main attributes:

- **Category**: A label identifying whether an email is **spam** or **ham**.
- **Message**: The actual text content of each email, serving as the primary source for feature extraction and analysis.

Data Preprocessing

The dataset is pre-processed to ensure that the text data is clean, standardized, and ready for analysis:

- **Loading and Initial Cleaning**: The data is loaded from mail_data.csv, and any duplicate or irrelevant information is removed. The text content is cleaned by converting it to lowercase and removing any special characters or punctuation.
- **Tokenization and Stop Word Removal**: Each message is tokenized, breaking the text into individual words, and common stop words (e.g., "the," "and") are removed to focus on meaningful words.
- **Stemming/Lemmatization**: Words are reduced to their root forms, standardizing variations like "running" and "run" into a single form.
- **Vectorization**: To convert text into a numerical format, we use techniques like **TF-IDF** (Term Frequency-Inverse Document Frequency), which weighs each word based on its frequency and relevance, or **Bag of Words (BoW)**, which represents the text as word occurrence matrices.

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

2. Feature Engineering

The model is designed to learn from various features derived from the email content:

- **Word Frequency Analysis**: Certain words common in spam emails (e.g., "free," "win," "prize") are given more significance.
- **Message Length**: The length of each email message, as spam messages often differ in length compared to ham messages.
- **Presence of Links and Special Keywords**: Spam emails are more likely to contain URLs, which can be a strong feature for spam identification.

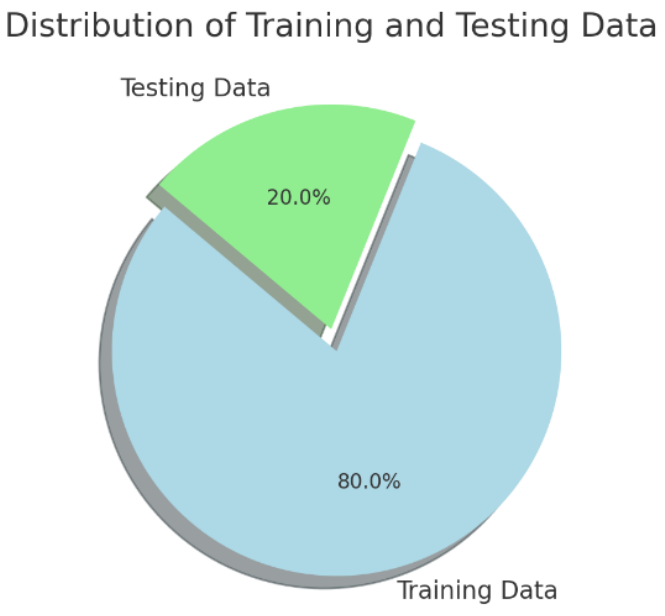
```
print(X_train)
```

```
3075                  Don know. I did't msg him recently.
1787    Do you know why god created gap between your f...
1614                          Thnx dude. u guys out 2nite?
4304                                        Yup i'm free...
3266    44 7732584351, Do you want a New Nokia 3510i c...
                              ...
789     5 Free Top Polyphonic Tones call 087018728737,...
968     What do u want when i come back?.a beautiful n...
1667    Guess who spent all last night phasing in and ...
3321    Eh sorry leh... I din c ur msg. Not sad alread...
1688    Free Top ringtone -sub to weekly ringtone-get ...
Name: Message, Length: 4457, dtype: object
```

3. Model Training and Testing

A **Logistic Regression** model is applied to classify each email as spam or ham. This model uses a linear combination of the input features, which are passed through a sigmoid function to output a probability score for each category:

- **Training and Validation Split**: The dataset is split into training and validation sets, allowing for an unbiased evaluation of model performance.
- **Parameter Optimization**: Hyperparameters like regularization strength are tuned to optimize the model's generalizability.



4. Model Evaluation

The model's performance is evaluated using key metrics, including:

Performance Metrics: The model exhibited exceptional performance, with a 96.77% accuracy rate on training data and 96.68% on test data. These high percentages indicate the model's strong ability to correctly categorize emails as either spam or legitimate (ham).

Precision and Recall Analysis: Precision evaluates the model's ability to correctly identify spam among all emails flagged as such, while recall assesses its capacity to detect all actual spam messages. Strong scores in both metrics would demonstrate the model's effectiveness in minimizing false positives and capturing the majority of spam, thereby reducing the risk of overlooking potential threats.

5. CONCLUSIONS

In contemporary society, email has become the predominant form of communication, facilitating the transmission of messages globally through internet connectivity. More than 270 billion emails are sent and received daily, with 57% classified as spam. Spam emails, also referred to as "non-self," are unsolicited commercial or malicious communications that compromise personal information, such as bank account details, financial data, or other sensitive information that may harm individuals, businesses, or organizations. In addition to advertisements, these emails may contain links to phishing websites or malware that are designed to extract personal data. Spam presents a significant issue, not only as an annoyance to end-users, but also as a financial liability and security threat. The spam detection mechanism in this project was designed to identify emails containing specific information. The use of reputable and verified domain names can aid in the identification of fraudulent emails. Classification of spam emails is crucial for categorizing messages and determining their status. Logistic regression, with its low false-positive spam detection rates that are generally acceptable to users, serves as a baseline technique for regulating spams according to individual user email requirements. The parameters of the Logistic Regression approach were further optimized, which improved the accuracy of the entire classification process. The Logistic regression can improve the accuracy of spam detection.

REFERENCES

1. Elchouemi, P. W. C. Prasad, A. Alsadoon, and M. K. Chae. Gain and the graph mining method are used in spam filtering and email categorization (sfecm). The 7th IEEE Annual Computing and Communication Workshop and Conference will be held in 2017.

2. "Logistic Regression for Machine Learning," by Jason Brownlee April 1, 2016, The Machine Learning Mastery. Logistic regression using machine learning is available at https://machinelearningmastery.com.

3. ianying Zhou, Wee-Yung Chin, Rodrigo Roman, and Javier Lopez, (2007) "An Effective MultiLayered Defense Framework against Spam", Information Security Technical Report 01/2007.

4. For email spam classification in a distributed context, K.R. Dhanaraj, V. Palaniswami, Firefly, and Bayes classifier, Aust.J. BasicAppl 5. A review of machine learning techniques to spam filtering by Guzella, T. S., and Caminhas, W. M. Appl. Expert Syst.