JINKA UNIVERSITY

COLLEGE OF NATURAL COMPUTATIONAL SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

COURSE TITLE:Introduction to Artificial Intelligence

COURSE CODE:CoSc3072

Project Title:Email spam Detection in Machine Learning

Group -3

GROUP MEMBERS	Id_number
1,Elias Tadele	13776
2,Kasim Logita	14158
3,Yadeta Egazu	15480
4,Abeje Ashenafi	13145
5,Rediat Demile	14744

Summited Date: 16 /10/2016E.C

Summited to Inst:Lencho Desalegn
Jinka JKU Ethiopia

Email Spam Detection Using Machine Learning Algorithms

Abstract

Nowadays, a big part of people rely on available email or messages sent by the stranger.
The possibility that anybody can leave an email or a message provides a golden opportunity for spammers to write spam message about our different interests .Spam fills inbox with number of ridiculous emails . Degrades our internet speed to a great extent .Steals useful information like our details on our contact list.Identifying these spammers and also the spam content can be a hot topic of research and laborious tasks. Email spam is an operation to send messages in bulk by mail .Since the expense of the spam is borne mostly by the recipient ,it is effectively postagedue advertising. Spam email is a kind of commercial advertising which is economically viable because email could be a very cost effective medium for sender .With this proposed model the specified message can be stated as spam or not using Bayes' theorem and Naive Bayes' Classifier and Also IP addresses ofthe sender are often detected.

Table of Contents

1.Introduction

1.1Background of Information

Email or electronic mail spam refers to the "using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. "The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet.

Fig.2 Email Spam

Emails are important because they creates a fast, reliable form of communication that is free and easily accessible. They allow people to foster long-lasting, long-distance communication . Spam e-mails can be not only annoying but also dangerous to customers.Spam e-mails can be defined as

- Anonymity
- Mass-Mailing
- Unsolicited

Spam e-mail messages are randomly sent to multiple addresses by all sort of groups , but mainly by lazy advertisers and criminals who wish to lead

you to phishing sites.

Spam and Ham: According to Wikipedia "the use of electronic messaging systems to send unsolicited bulk messages, especially mass advertis ement, malicious links etc."are called as spam. "Unsolicited means that those things which you didn't asked for messages from the sources. So, if you do not know about the sender the mail can be spam. People generally don't realize they just signed in for those mailers when they download any free services, software or while updating the software. "Ham" this term was given by Spam Bayes around 2001 and it is defined as "Emails that are not generally desired and is not considered spam"

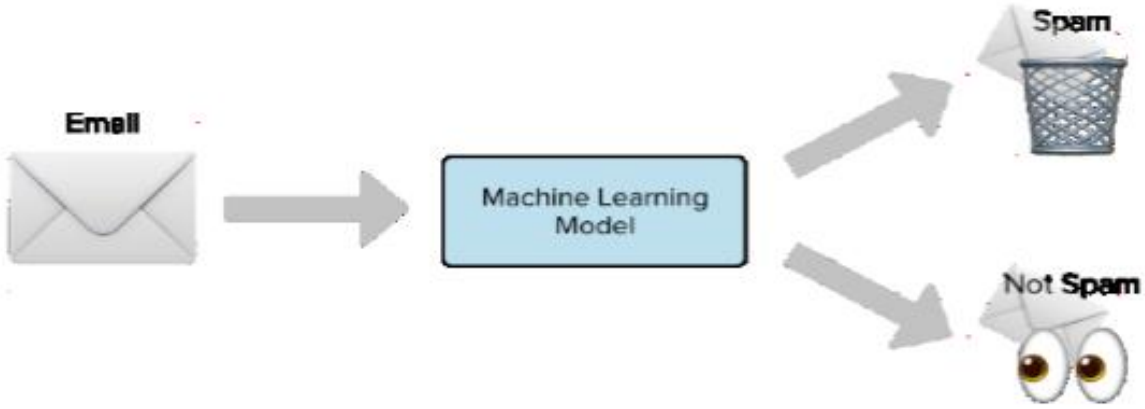

Fig.1. Classification into Spam and non-spam

1.2 Objective

The objective of identification of spam e-mails are :

- To give knowledge to the user about the fake e-mails and relevant e-mails.
- To classify that mail is spam or not.

1.3 Statement of Problem:

Nowadays there are lots of people trying to fool you just by sending you fake e-mails like you have won a 1000 dollars, this much amount is deposited in your account once you open this link and then they will try to hack your information,

- Unwanted emails irritating internet consumers.
- Critical e-mail messages are missed and/or delayed.
- Identity Theft.
- Spam can crash mail servers and fill up hard drive.
- Billions of dollars lost worldwide.

1.4 Scope

Detect spam and ham mail which are received within inbox and filter it.

1.5 Research Methodology

There are a number of techniques available for detecting and limiting spam emails. Our primary motivation for doing so is to investigate existing spam text detection and categorization methods. Here, we'll discuss the survey methodology we used to gather data for our in-depth analysis of spam filtersIn this section, we provide an overview of the methodology used in our email spam detection project. The process involves several key steps data collection and preprocessing ,feature extraction ,data splitting ,model evaluation and deployemnt.

1.6 Architecture

The architecture provides a high-level overview of the email spam detection system, illustrating the flow of data and processes from data collection to model evaluation and spam detection.

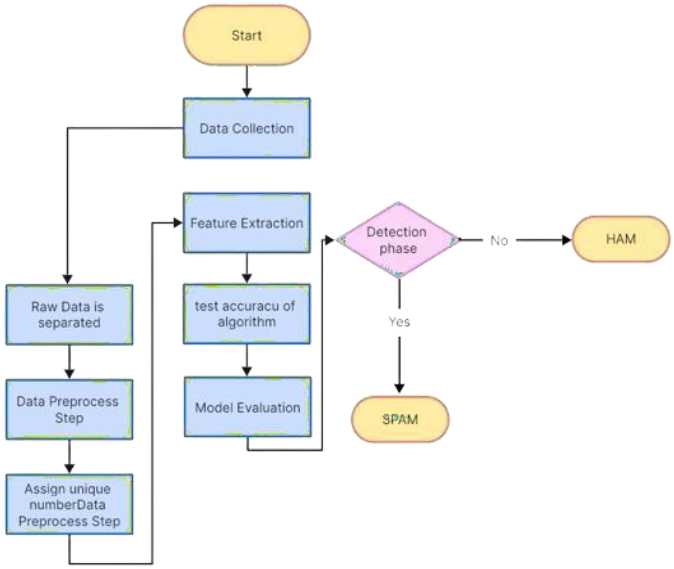

Fig.3 Architectural Design for Email Spam Detection

2.Literature Review

There is some related work that apply machine learningmethods in email spam detection, A. Karim, S. Azam, B.Shanmugam, K. Kannoorpatti and M. Alazab.[ii] They describe a focused literature survey of Artificial Intelligence Revised (AI) and Machine learning methods for email spam detection. K. Agarwal [3] and T. Kumar. Harisinghaney et al. (2014) [4]and Mohamad & Selamat (2015) [v] have used the
"image and textual dataset for the e-mail spam detection with the use of various methods. Harisinghaney et al. (2014) [iv] have used methods of KNN algorithm, Naïve Bayes, and Reverse DBSCAN algorithm with experimentation on dataset.
For the text recognition, OCR library" [iii] is employed but this OCR doesn't perform well. Mohamad & Selamat (2015) [v] uses the feature selection hybrid approach of TF-IDF (Term Frequency Inverse Document Frequency) and Rough
pure mathematics.

2.1 Related work

In In this section, we review several notable works in this area, highlighting their methodologies, findings, and limitations.

1. Efficient Spam Email Classification using Machine Learning Algorithms (Pallavi N & Jayarekha, 2023): Pallavi N and Jayarekha presents a study on email spam detection utilizing various machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM) and Decision Trees. Overall, the paper offers a comprehensive analysis of ML-based email spam classification. Challenges may arise from concentrating on a particular subset of machine learning algorithms, which could restrict its relevance to broader contexts.

2. Spam Detection System Using Supervised ML(Abhila & Delphin, 2021): Abhila and Delphin present a proposed spam detection system which utilizes the Naive Bayes method, a mining approach, for identifying spam and ham messages in an inbox. The system employs a series of steps, including data collection, pre-processing, feature extraction, training, and testing, to classify messages accurately. Overall, while Naïve Bayes classifiers offer simplicity and ease of implementation, they may exhibit limitations in terms of model complexity, feature handling, adaptability to evolving spamming techniques.

3. A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms (Mansoor & Muhana, 2021): Mansoor and Muhana summarize and emphasize the importance of machine learning algorithms in enhancing email spam

3. Methodology

In this section, we provide an overview of the methodology used in our email spam detection project.
The process involves several key steps:

3.1 Data Collection and Preprocessing Module:

Responsible for gathering labeled email datasets containing both spam and non-spam examples, the system proceeds to clean the data by removing duplicates, irrelevant information, and formatting inconsistencies. Following this preprocessing step, the dataset is divided into training and testing sets, facilitating model training and evaluation seamlessly.

1. Feature Extraction and Engineering Module:

The system extracts relevant features from the email data, encompassing text-based, structural, and metadata features. These extracted features undergo further refinement through feature engineering techniques, aimed at enhancing their discriminative power and improving the overall effectiveness of the spam detection model.

Model Training and Hybridization Module:

The system trains Random Forest and Gradient Boosting classifiers separately on the training dataset. Subsequently, it employs ensemble methods or feature integration techniques to combine predictions from both classifiers. Additionally, the system fine- tunes hyperparameters and explores various feature combinations to optimize the hybrid model's performance, ensuring robust and effective email spam detection.

2. Evaluation and Validation Module:

The system evaluates the performance of the hybrid model on the testing dataset using appropriate evaluation metrics. It further compares the performance of the hybrid model with individual Random Forest and Gradient Boosting classifiers. To ensure the model's robustness, the system conducts cross-validation and sensitivity analysis, validating its effectiveness and reliability in email spam detection.

3. Deployment and Monitoring Module:

After training, the hybrid model is deployed for real-world email spam detection. Continuous monitoring ensures its effectiveness in detecting evolving spam tactics or data changes. A user-friendly interface allows easy interaction and spam detection setting management, ensuring ongoing optimization and adaptability for effective spam email combat.

The proposed system aims to provide an effective and adaptable solution for email spam detection, leveraging the complementary strengths of Random Forest and Gradient Boosting algorithms to improve detection accuracy and robustness.

3.2 Model Selection and Training

Model selection plays a pivotal role in the success of an email spam detection system.

We opted for the following choices:

- **Logistic Regression**: We selected Logistic Regression as the primary classification algorithm. It is known for its simplicity, efficiency, and interpretability, making it a suitable choice for binary classification tasks.

Training the Model:

The selected Logistic Regression model was trained on the TF-IDF transformed training data.

This process involved learning the model's parameters and decision boundary based on the provided email features and labels

Performance Evaluation

To assess the effectiveness of our email spam detection system, we performed the following steps:

- **Dataset Split**: We divided the dataset into training and testing sets using the train_test_split function from scikit-learn. The training set comprised 80% of the data, while the testing set contained the remaining 20%.

- Model Evaluation: We evaluated the model's performance using standard metrics, including accuracy, precision, recall, and F1-score, on both the training and testing data.

- **Real-Time Prediction**: We demonstrated the practical application of our trained model by making predictions on a sample email.

Through this methodology, we aimed to develop an effective and adaptable email spam detection system capable of distinguishing between spam and ham emails, ultimately enhancing email security and user experience.

1. **Data Collection:**

- Responsible for gathering labeled email datasets containing both spam and non-spam examples.

2. **Raw Data Separation:**

- Separates raw email data into spam and non-spam categories.

3. **Preprocessing:**

- Cleans the data by removing duplicates, irrelevant information, and formatting inconsistencies.

4. **Tokenization:**

- Splits the cleaned text into tokens (words or phrases) for further processing.

5. **Feature Extraction:**

- Extracts relevant features from the email data, such as word frequency, structural features, and metadata features.

6. **Data Splitting:**

- Divides the dataset into training and testing sets to prepare for model training and evaluation.

Model Evaluation:

- Evaluates the performance of the trained model on the testing dataset using appropriate evaluation metrics.

7. **Accuracy of Algorithm:**

- Measures the accuracy of the algorithm in correctly classifying emails as spam or non-spam.

8. **Spam Mail Found or Not:**

- Determines whether a given email is classified as spam or non-spam based on the output of the model

3.3 workflow

The outlined workflow offers a comprehensive and systematic approach for developing a hybrid email spam detection system using Random Forest and Gradient Boosting algorithms. Beginning with data collection and preprocessing, labelled email datasets are gathered, cleaned, and divided into training and testing sets. Feature extraction follows, where relevant features are extracted from the email data and engineered to enhance discriminative power. Subsequently, Random Forest and Gradient Boosting classifiers are trained on the training dataset, with hyperparameters tuned for optimal performance. The hybridization step combines predictions from both classifiers, either through ensemble methods or feature integration. Model evaluation assesses the hybrid model's performance using appropriate metrics, comparing it with individual classifiers. Finally, deployment of the hybrid model for real-world spam detection applications ensures ongoing monitoring and updates to adapt to evolving spamming tactics and data distributions. Overall, this structured workflow provides a robust foundation for the development of an effective and adaptable email spam detection system.

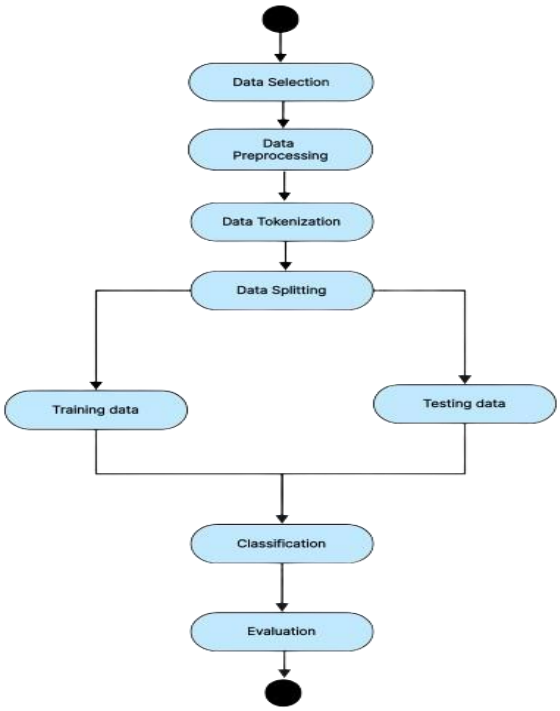

Fig.2 Workflow of Proposed syst

4.Results And Discussion

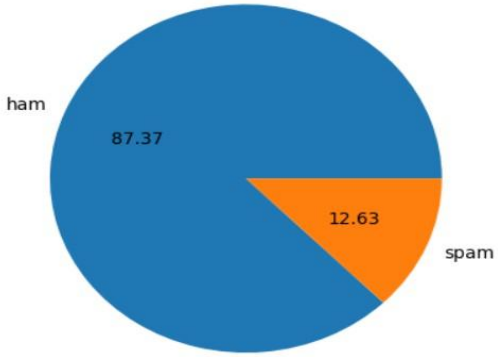

Fig.4 Total count of spam and ham mails

The pie chart depicting 87.37% ham and 12.63% spam emails in mail spam detection using a hybrid of Random Forest and Gradient Boosting provides a concise visual summary of the email classification results. The chart effectively communicates the dominance of ham emails over spam in the dataset, reflecting the success of the hybrid model in accurately identifying non-spam messages. This visual representation aids in quickly understanding the distribution of email types and underscores the effectiveness of the spam detection system**.**

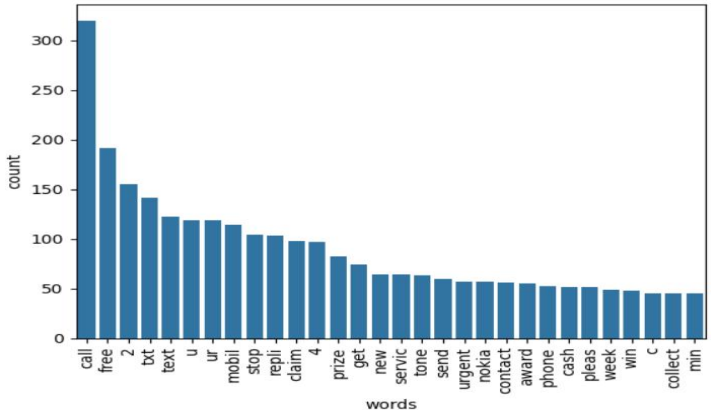

Fig.5 Spam word count

The graph depicting the spam word count in mail spam detection using a hybrid of Random Forest and Gradient Boosting algorithms showcases the frequency of spam words identified by the model. Each bar on the graph represents a spam word, while the height of the bar indicates the word's count. This visualization offers insights into the most prevalent spam words detected by the hybrid model, providing valuable information for refining and enhancing email spam detection algorithms.

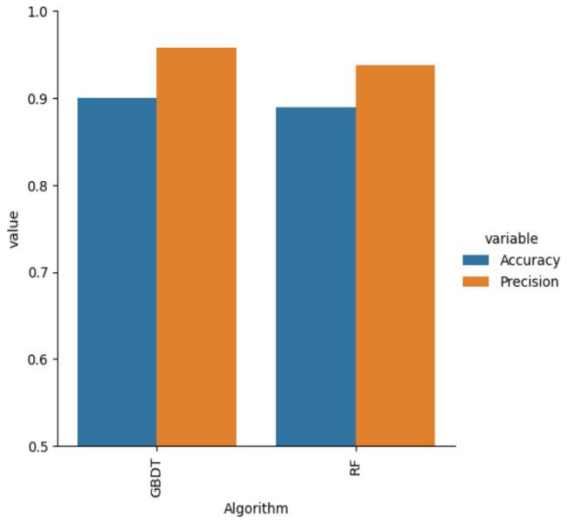

Fig.6 Measuring Performance of the ML Algorithms

When evaluating machine learning (ML) algorithms for email spam detection, accuracy and precision are two crucial performance metrics.

9. **Accuracy**: This metric measures the overall correctness of the model's predictions. It calculates the ratio of correctly predicted emails (both spam and non-spam) to the total number of emails in the dataset. High accuracy indicates that the model correctly identifies most emails, regardless of their classification.

10. **Precision**: Precision focuses on the accuracy of positive predictions, i.e., how many of the emails predicted as spam are actually spam. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives. Precision helps understand the model's ability to avoid misclassifying non-spam emails as spam.

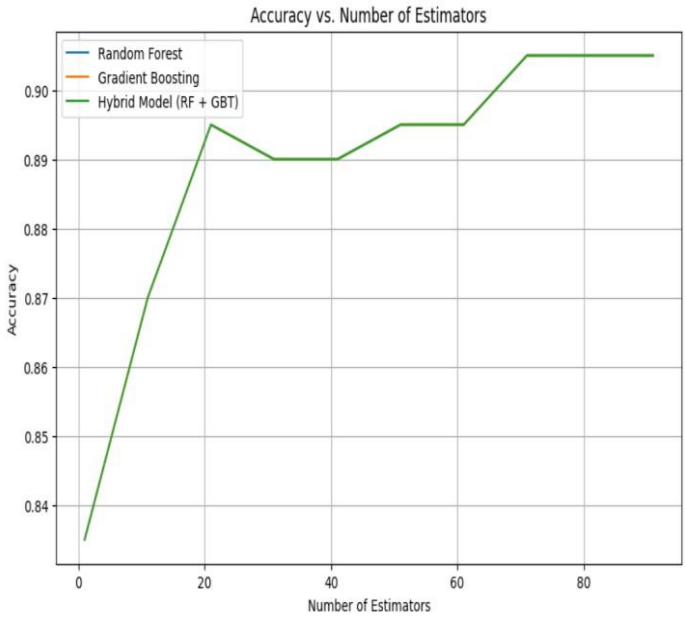

Fig.7. Comparison of Hybrid accuracy with the total no. of estimators

After the comparison, it is found that Hybrid model has obtained the highest accuracy as compared to Random Forest and Gradient Boosting Algorithms. The accuracy of hybrid model came out to be 91.00%. This model combines multiple machine learning techniques or algorithms to leverage the strengths of each.

4.2 Limitations

The limitation of this email spam detection is only language dependet that means only work for english we must develop for other language .

5.Conclusion

The implementation of a hybrid approach combining Random Forest and Gradient Boosting algorithms for mail spam detection demonstrates promising results. The hybrid model leverages the strengths of both algorithms, resulting in enhanced accuracy and precision in distinguishing between spam and non-spam emails. By effectively combining the predictive power of Random Forest's ensemble learning and Gradient Boosting's sequential learning, the hybrid model achieves robust performance in detecting spam emails while minimizing false positives. The accuracy of this hybrid approach turned out to be 91.00% overall. This approach offers

a versatile solution that adapts well to varying types of spam and evolving spamming tactics. Additionally, the hybrid model's ability to efficiently process and analyse email data makes it suitable for real-world applications where timely and accurate spam detection is crucial. Future research could focus on further optimizing the hybrid model, exploring additional ensemble techniques, and integrating advanced features to improve detection performance and adaptability to emerging spam threats. Overall, the mail spam detection system using a hybrid of Random Forest and Gradient Boosting demonstrates effectiveness and potential for addressing the persistent challenge of email spam in today's digital landscape.

5.1 FUTURE SCOPE

The mail spam detection system using a hybrid of Random Forest and Gradient Boosting shows considerable promise, and there are several avenues for future research and development:

- **Enhanced Feature Engineering:** Utilize advanced techniques like semantic analysis and sentiment analysis to extract more meaningful features from email data, enhancing the model's ability to differentiate between spam and legitimate emails.
- **Ensemble Methods:** Explore additional ensemble methods like stacking or bagging to further enhance the performance and resilience of the spam detection model by combining the strengths of different algorithms.
- **Deep Learning Integration:** Incorporate neural networks and other deep learning techniques to handle complex data patterns and improve detection accuracy, particularly for sophisticated spamming tactics.
- **Real-time Detection:** Develop mechanisms for real-time spam detection to swiftly identify and filter out spam emails as they are received, improving user experience and email security.

REFERENCES

[1] P. N and P. Jayarekha, "Efficient Spam Email Classification Using Machine Learning Algorithms," 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023.

[2] T. Toma, S. Hassan and M. Arifuzzaman, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, 2021.

[3] A. B, D. P. M, K. M, M. N. Joseph and D. R, "Spam Detection System Using Supervised ML," 2021 International Conferenceon System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021.

[4] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019.

[5] W. Hijawi, H. Faris, J. Alqatawna, A. M. Al-Zoubi and I. Aljarah, "Improving email spam detection using content based feature engineering approach," 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Aqaba, Jordan, 2017.

Appendix

IP :imternet protocol

ML:machine Learning