International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Smart Spam Detection: An AI-Based Machine Learning

Janakiraman S[1], Prithika S[2]

[1]Assistant Professor, Department of Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur
[2]II MCA, Department of Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur

Abstract
Email has become an essential communication tool, seamlessly facilitating information exchange across personal and professional spheres. While its convenience and global accessibility are unparalleled, email systems have increasingly become targets for cybercriminals exploiting sophisticated spam tactics to breach government networks, corporate systems, and individual accounts. These threats are characterized by their complexity and scale, outpacing traditional detection mechanisms and emphasizing the need for innovative and adaptive solutions to combat emerging cyber risks effectively. This project proposes an advanced system for classifying large-scale email datasets into four distinct categories: Normal, Fraudulent, Harassment, and Suspicious. The approach integrates Natural Language Processing (NLP) with Bidirectional Long Short-Term Memory (BiLSTM) networks to capture nuanced patterns and semantic meanings within email content. The methodology includes a sample expansion phase to enhance training data diversity and a robust testing stage to ensure high accuracy under varied conditions. This innovative system enables effective forensic analysis by extracting and analysing meaningful information from email communications. Through extensive experimentation, the proposed system demonstrates a significant improvement over existing machine learning techniques, achieving a remarkable classification accuracy of 99.1%. The use of BiLSTM with recurrent gradient units ensures reliable performance across diverse email topics and complex scenarios. By offering a highly accurate and robust solution, this project contributes to advancing email security and strengthens the defence mechanisms against evolving cyber threats in today's interconnected digital environment.

Keywords: Natural Language Processing(NLP), Supervised Learning, Spam Classification, Feature Extraction, Text Preprocessing Machine Learning Algorithms

INTRODUCTION

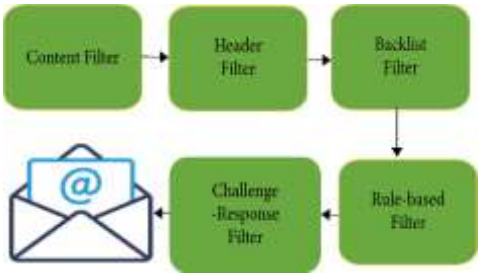
In the digital age, spam messages—ranging from unsolicited advertisements to malicious phishing attempts—pose significant challenges to communication platforms and email services. To address this, we present an AI-based machine learning framework for efficient spam detection. Leveraging supervised learning algorithms such as Naïve Bayes, Support Vector Machines, and deep learning models, the system can analyze textual patterns and classify messages as spam or legitimate. The framework preprocesses data using techniques like tokenization and TF-IDF vectorization to enhance accuracy. Designed for

scalability and real-time performance, this model aims to reduce false positives and improve user safety through intelligent, automated detection.

EXISTING ALGORITHM

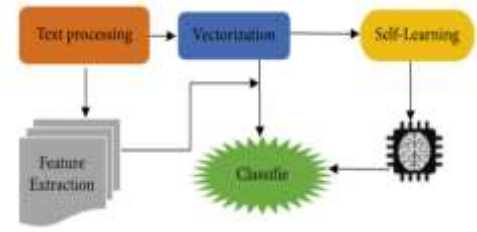
1. **Content-Based Filtering Technique**

Algorithms analyse words, the occurrence of words, and the distribution of words and phrases inside the content of e-mails and segregate them into spam non-spam categories.



2. Case Base Spam Filtering Method

Algorithms trained on well-annotated spam/non-spam marked emails try to classify the incoming mails into two categories.



3. Heuristic or Rule-Based Spam Filtering Technique

Algorithms use pre-defined rules in the form of a regular expression to give a score to the messages present in the e-mails. Based on the scores generated, they segregate emails into spam non-spam categories.

4. The Previous Likeness Based Spam Filtering Technique

Algorithms extract the incoming mails' features and create a multi-dimensional space vector and draw points for every new instance. Based on the KNN algorithm, these new points get assigned to the closest class of spam and non-spam.

5. Adaptive Spam Filtering Technique

Algorithms classify the incoming mails in various groups and, based on the comparison scores of every group with the defined set of groups, spam and non-spam emails got segregated.

6. Machine learning classifiers

The machine learning models are selected based on their group, diversity and acceptance in the machine learning community. SVM, Naive Bayes (NB) and DT are from three different groups of classifiers.

Support vector machine

SVM are based on the assumption that the input data can be linearly separable in a geometric space. This is often not the case when working with real word data. To solve this problem SVM map the input to a

high dimension feature space, i.e hyperplane, where a linear decision boundary is constructed in such a manner that the boundary maximises the margin between two classes. SVM is introduced as a binary classifier intended to separate two classes when obtaining the optimal hyperplane and decision boundary.

Decision tree

A DT classifier is modelled as a tree where rules are learned from the data in a if-else form. Each rule is a node in the tree and each leaf is a class that will be assigned to the instance that fulfil all the above nodes conditions. For each leaf a decision chain can be created that often is easy to interpret. The interpretability is one of the strengths of the DT since it increases the understanding of why the classifier made a decision, which can be difficult to achieve with other classifiers. The telecommunication company is today using a manually created decision tree, in which the rules are based on different combinations of words.

Naive bayes

The NB classifier is considered to perform optimal when the features are independent of each other, and close to optimal when the features are slightly dependant. Real world data does often not meet this criterion but researchers have shown that NB perform better or similar to C4.5, a decision tree algorithm in some settings. The researchers argue that NB performs well even when there is a clear dependency between the features, making it applicable in a wide range of tasks.

Disadvantages

- Predefined rule-based filtering lacks adaptability to evolving spam tactics, diminishing effectiveness over time.
- Content-based filtering suffers from limited feature extraction, missing subtle spam indicators.
- Rule-based filtering often results in high false positive rates, flagging legitimate emails as spam.
- Machine learning classifiers offer improved performance but introduce complexity and reduced interpretability.
- Classifier performance is heavily dependent on the quality and representativeness of training data.
- Machine learning classifiers are vulnerable to adversarial attacks, compromising system security.

PROPOSED SYSTEM

The proposed approach comprises data collection, pre-processing, feature extraction, parameter tuning, and classification using the NLP and BiLSTM model. In this project, E-mail datasets are divided into normal, harassing, suspicious, and fraudulent classes. The E-mail is divided into word levels of the E-mail body, and the embedding layer is applied to train and obtain the sequence of vectors.

Real-Time Spam Detection: The system continuously monitors incoming emails and applies machine learning algorithms to detect spam emails in real-time.

Multi-Class Classification: Emails are classified into multiple categories including Normal, Fraudulent, Harassment, and Suspicious, allowing for more nuanced detection and handling of spam.

NLP-Based Analysis: Natural Language Processing techniques are employed to analyze email content, extract meaningful features, and identify patterns indicative of spam or fraudulent activity.

BiLSTM Model: The system utilizes Bidirectional Long Short-Term Memory (BiLSTM) networks, a type of recurrent neural network, to capture contextual information and improve the accuracy of email classification.

User-Friendly Interface: The system features an intuitive user interface for both administrators and end-users, providing easy access to spam detection results, notifications, and configuration settings.

Automated Deletion and Alerting: Detected spam emails are automatically deleted from the user's inbox, reducing the risk of user exposure to malicious content. Users also receive notifications/alerts via email about the detected spam mail and deletion actions.

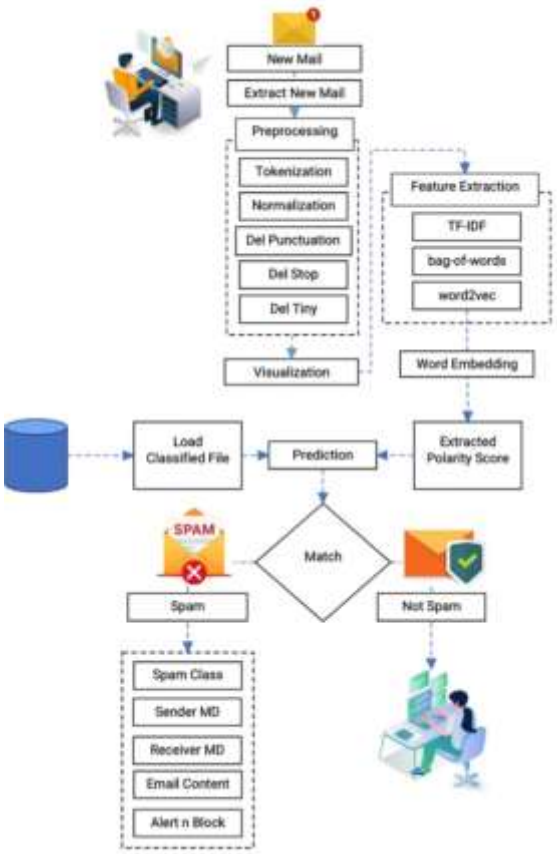
Customizable Settings: Administrators have the ability to customize system settings, including model parameters, data sources, and notification preferences, to suit the specific requirements of their organization.

Scalable Architecture: The system is designed with scalability in mind, allowing for seamless integration with existing email servers and infrastructure, and supporting high volumes of email traffic.

Advantages

- Enhanced email security through real-time spam detection.
- Reduction in the risk of users falling victim to phishing scams or fraudulent activities.
- Time and resource savings with automated spam detection and deletion.
- Improved user experience with hassle-free email management.
- Customizable settings and adaptable architecture to meet diverse organizational needs.
- Seamless integration with existing email servers and infrastructure.
- Higher accuracy in identifying and filtering out spam emails using advanced technologies.
- Increased productivity by freeing users from manual handling of spam emails.

SYSTEM ARCHITECTURE



MODULES DESCRIPTION

1.E-Mail Screener Web App

The E-Mail Screener Web App module serves as a vital interface, providing users with a centralized platform to interact with the sophisticated spam mail detection system driven by Natural Language Processing (NLP) and Bidirectional Long Short-Term Memory (BiLSTM) technologies. Through this intuitive interface, users gain access to a suite of powerful features designed to enhance email security and streamline their inbox management process. At the core of the module is its user-friendly design, which prioritizes simplicity and ease of use. Users can effortlessly navigate through the app's various sections, accessing key functionalities with just a few clicks. Whether it's reviewing spam detection results, adjusting filtering settings, or seeking assistance, users can accomplish tasks efficiently within the app. One of the module's key functionalities is its ability to provide users with real-time insights into spam detection outcomes. By displaying classification results directly within the app, users gain immediate visibility into which emails have been identified as spam and which ones are legitimate. This transparency enables users to take swift action, ensuring that spam emails are promptly addressed and mitigated. Furthermore, the module offers robust customization options, allowing users to tailor spam filtering settings to suit their specific preferences and needs. From setting confidence thresholds to defining rules for handling detected spam emails, users have full control over how the system processes incoming messages. This level of customization empowers users to fine-tune the spam detection process, optimizing it for their unique requirements. In addition to its core functionalities, the module also provides comprehensive support resources to assist users as they navigate the platform. Whether it's accessing FAQs, knowledge base articles, or reaching out to customer support, users can easily find the assistance they need to resolve any issues or inquiries they may encounter. Thus, the E-Mail Screener Web App module represents a powerful tool for users to enhance email security and streamline their inbox management process. With its intuitive design, real-time insights, customization options, and robust support resources, the module empowers users to take control of their email security with confidence and ease.

2. End User

Admin Module

Login: This module allows administrators to securely log in to the E-Mail Screener system using their credentials. Authentication ensures that only authorized personnel can access administrative functionalities.

Upload E-Mail Dataset: Administrators can upload email datasets for training and fine-tuning the spam detection model. The module supports various data formats and provides validation mechanisms to ensure data integrity.

Build and Train the Model: In this module, administrators can initiate the process of building and training the spam detection model using the uploaded email dataset. The system employs advanced machine learning algorithms, including NLP and BiLSTM, to train the model effectively.

User Management: Administrators have the authority to manage user accounts within the system. This includes creating new user accounts, updating user information, resetting passwords, and deactivating or deleting user accounts as needed.

User Module

Register: New users can register for an account by providing necessary information such as username, email address, and password. The registration process includes validation steps to ensure the integrity of

user data.

Login: Registered users can securely log in to their accounts using their credentials. Authentication mechanisms protect user accounts from unauthorized access.

Configure Personal Email: This module allows users to configure their personal email accounts with the E-Mail Screener system. Users provide their original email usernames and passwords, enabling the system to access and monitor their email activity for spam detection purposes.

Receive Spam Mail and Deletion Alert: Users receive notifications/alerts via their personal email accounts whenever spam mail is detected. Additionally, users are notified when detected spam emails are automatically deleted from their inbox. These alerts help users stay informed about the status of their email security and take necessary actions if required.

These modules collectively facilitate seamless interaction between administrators and end users, enabling efficient management of email datasets, model training, user accounts, and spam detection functionalities within the E-Mail Screener system.

3. EmailNet Model: Build and Train

3.1. Dataset Annotation

In this project, E-mails are divided into normal, harassing, suspicious, and fraudulent classes. The E-mail is divided into word levels of the E-mail body, and the embedding layer is applied to train and obtain the sequence of vectors.

3.2. E-mail Data Set Preparation and Exploration

Import Dataset

In this module the admin uploads an email dataset (CSV) file. This will be used to train your email forensics analysis model.

Read Dataset

The EmailSinkAI reads email dataset to output the purpose or objective of the project.

Explore Dataset – EDA

Data visualization tool that brings the entirety of data together into a striking and easy-to-follow view.

3.3. Data Pre-processing

The data pre-processing phase consists of natural language-based steps that standardize the text and prepare it for analysis.

Tokenization

Breaking up the original text into component pieces is the tokenization step in natural language processing. There are predefined rules for tokenization of the documents into words. The tokenization step is performed in Python by using the SpaCy library.

Normalization

These are the steps needed for translating text from human language (English) to machine-readable format for further processing. The process starts with:

- changing all alphabets to lower or upper case
- expanding abbreviations
- excluding numbers or changing those to words
- removing white spaces
- removing punctuation, inflection marks, and other circumflexes
- removing stop words, sparse terms, and particular words

Stop Words Removal

Words like ``a'' and ``the'' that appear so frequently are not relevant to the context of the E-mail and create noise in the text data. These words are called stop words, and they can be filtered from the text to be processed. We utilized the``NLTK'' Python library to remove stop words from the text.

Punctuation Removal

Punctuation includes (e.g., full stop (.), comma (,), brackets) to separate sentences and clarify meaning. For punctuation removal, we utilize the ``NLTK'' library.

3.4. Feature Extraction

After eliminating irrelevant information, the elaborated list of words is converted into numbers. The TF-IDF method is applied to accomplish this task. Term Frequency is several occurrences of a word in a document, and IDF is the ratio of a total number of documents and the number of documents containing the term. A popular and straightforward method of feature extraction with text data is called the bag-of-words model of text. A bag-of-words model, or BoW for short, is a way of extracting features from the text for use in modelling, such as machine learning algorithms. A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things (1) A vocabulary of known words, (2) A measure of the presence of known words. We extract features on the basis of Equations Here tf represents term frequency and df represents document frequency.

4. E-Mail Prediction

This module of the E-Mail Screener system is responsible for predicting whether incoming emails are spam or legitimate, notifying users of detected spam, and automatically deleting spam emails from their inbox. The process involves real-time monitoring of users' personal email accounts, extraction of new emails, classification of emails as spam or normal, and sending notifications to users regarding detected spam. Below is a detailed description of the components and functionalities of the E-Mail Prediction module:

4.1. Email Monitoring and Extraction

The E-Mail Prediction module continuously monitors users' personal email accounts for new incoming emails. Upon detection of new emails, the module extracts relevant information such as sender, subject, body, and attachments for further processing.

4.2. Spam Classification

Extracted email data is processed through the spam detection model trained using NLP and BiLSTM algorithms. The model predicts whether each email is spam or normal based on learned patterns and features extracted from the email content. Classification labels (e.g., "spam" or "normal") are assigned to each email based on the model's predictions.

4.3. Notification of Detected Spam

If an email is classified as spam, the module generates a notification to alert the user about the detected spam. Notifications may include details such as the sender, subject, and classification label of the detected spam email.

4.4. Automatic Deletion of Spam Emails

Upon classification of an email as spam, the module automatically deletes the spam email from the user's inbox to prevent it from reaching the user. Deleted spam emails are permanently removed from the user's email account to ensure that they do not pose any further risk.

4.5. Notification to User

Following the deletion of a spam email, the module sends a notification to the user's email address to inform them about the detected spam and the action taken. Notification provide users with transparency regarding the spam detection process and help them stay informed about their email security.

5. Notification

The Notification Module enables users to stay informed about their email security status and system updates through timely alerts. Users can configure their notification preferences, receiving notifications via email or other preferred channels. When a spam email is detected, users receive alerts detailing the detected spam and any actions taken, such as automatic deletion. Additionally, notifications about system updates and maintenance activities ensure users are aware of changes to the system's functionalities. With personalized settings and real-time delivery, the Notification Module enhances user engagement and awareness within the E-Mail Screener system.

RESULTS

CONCLUSION

In conclusion, the project marks a significant advancement in email security technology. Leveraging natural language processing (NLP) methods and bidirectional Long Short-Term Memory (BiLSTM) networks, the system effectively identifies and filters out spam messages in real-time. The project has yielded a robust email screening system capable of accurately categorizing emails into various classes, including normal, fraudulent, harassment, and suspicious. NLP and BiLSTM technologies empower the system to discern intricate patterns within email content, enhancing its spam detection capabilities. The

project's success is demonstrated through its strong performance metrics, including accuracy, precision, recall, and F1-score, as well as the informative insights provided by the confusion matrix. These results validate the effectiveness of the approach in combating email spam and ensuring the security of email communication channels. Looking forward, there are opportunities for further refinement and improvement of the E-Mail Screener system. Future iterations may focus on optimizing model parameters, exploring additional classification features, and implementing real-time monitoring capabilities to adapt to evolving spamming techniques. Thus, the E-Mail Screener project represents a significant contribution to email security technology, offering a valuable tool for organizations and individuals to protect their inboxes from malicious spam and fraudulent activities. With continued development and innovation, the system holds great potential to become an essential asset in the ongoing battle against email-based threats.

REFERENCES

1. W. Park, N.M.F. Qureshi and D.R. Shin, "Pseudo nlp joint spam classification technique for big data cluster", Computers Materials & Continua, vol. 71, no. 1, pp. 517-535, 2022.
2. S. Sinha, I. Ghosh, and S. C. Satapathy, ``A study for ANN model for spam classification," in Intelligent Data Engineering and Analytics. Singapore: Springer, 2021, pp. 331-343.
3. Q. Li, M. Cheng, J. Wang, and B. Sun, ``LSTM based phishing detection for big email data," IEEE Trans. Big Data, early access, Mar. 12, 2020, doi: v10.1109/TBDATA.2020.2978915.
4. T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, ``Applicability of machine learning in spam and phishing email filtering: Review and approaches," Artif. Intell. Rev., vol. 53, no. 7, pp. 5019-5081, Oct. 2020, doi: 10.1007/s10462-020-09814-9.
5. E. Bauer. 15 Outrageous Email Spam Statistics That Still Ring True in 2018, RSS. Accessed: Oct. 10, 2020. [Online]. Available: https://www.propellercrm.com/blog/email-spam-statistics.
6. A.Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, ``A comprehensive survey for intelligent spam email detection," IEEE Access, vol. 7, pp. 168261-168295, 2019.
7. K. Singh, S. Bhushan, and S. Vij, ``Filtering spam messages and mails using fuzzy C means algorithm," in Proc. 4th Int. Conf. Internet Things, Smart Innov. Usages (IoT-SIU), Apr. 2019, pp. 1-5.
8. R. S. H. Ali and N. E. Gayar, ``Sentiment analysis using unlabeled email data," in Proc. Int. Conf. Comput. Intell. Knowl. Economy (ICCIKE), Dec. 2019, pp. 328-333.
9. K. Agarwal and T. Kumar, ``Email spam detection using integrated approach of naïve Bayes and particle swarm optimization," in Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), Jun. 2018, pp. 685-690.
10. M. Shuaib, O. Osho, I. Ismaila, and J. K. Alhassan, ``Comparative analysis of classification algorithms for email spam detection," Int. J. Comput. Netw. Inf. Secur., vol. 10, no. 1, pp. 60-67, Aug. 2018.
11. G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, ``Email classification research trends: Review and open issues," IEEE Access, vol. 5, pp. 9044-9064, 2017.
12. Z. Chen, Y. Yang, L. Chen, L.Wen, J.Wang, G. Yang, and M. Guo, ``Email visualization correlation analysis forensics research," in Proc. IEEE 4th Int. Conf. Cyber Secur. Cloud Comput. (CSCloud), Jun. 2017, pp. 339-343.
13. N. Moradpoor, B. Clavie, and B. Buchanan, ``Employing machine learning techniques for detection and classification of phishing emails," in Proc. Comput. Conf., Jul. 2017, pp. 149-156.
14. A.S. Aski and N. K. Sourati, ``Proposed efficient algorithm to filter spam using machine learning techniques," Pacific Sci. Rev. A, Natural Sci. Eng., vol. 18, no. 2, pp. 145-149, Jul. 2016.

15. Y. Kaya and Ö. F. Ertu§rul, ``A novel approach for spam email detection based on shifted binary patterns," Secur. Commun. Netw., vol. 9, no. 10, pp. 1216-1225, Jul. 2016.
16. I. Idris, A. Selamat, N. T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, ``A combined negative selection algorithm-particle swarm optimization for an email spam detection system," Eng. Appl. Artif. Intell., vol. 39, pp. 33-44, Mar. 2015.