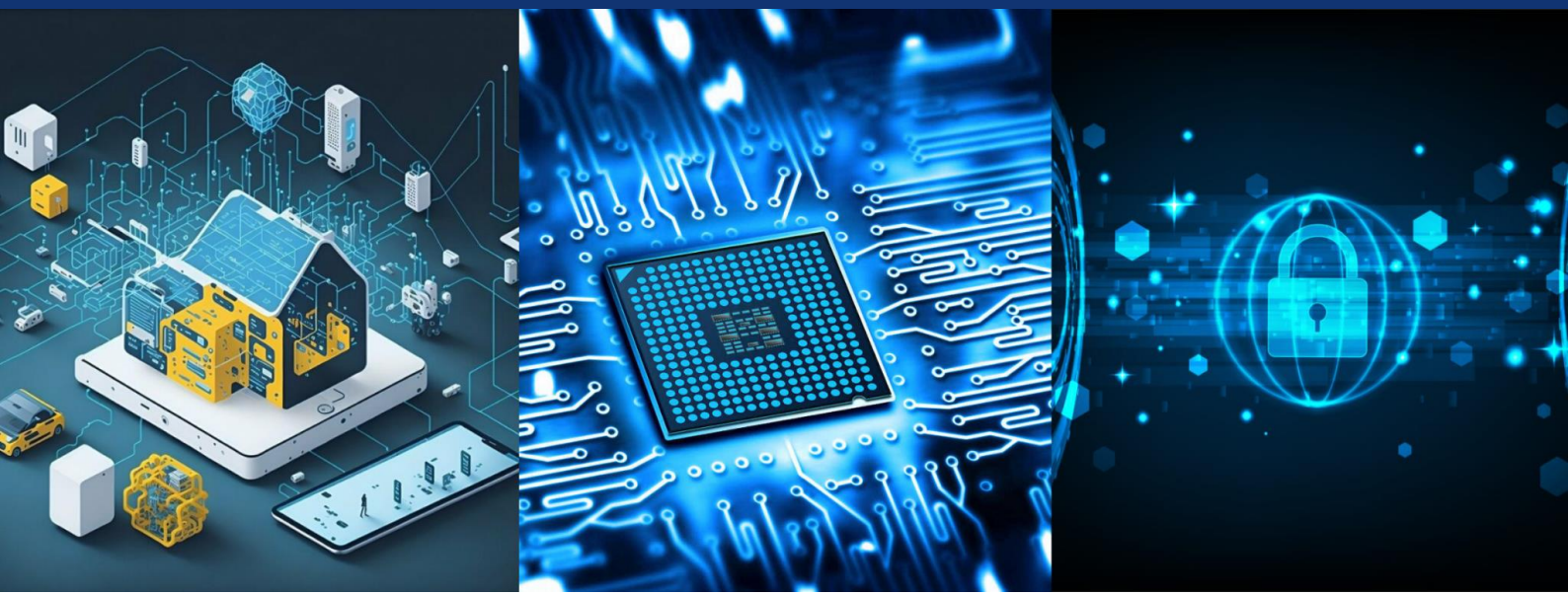
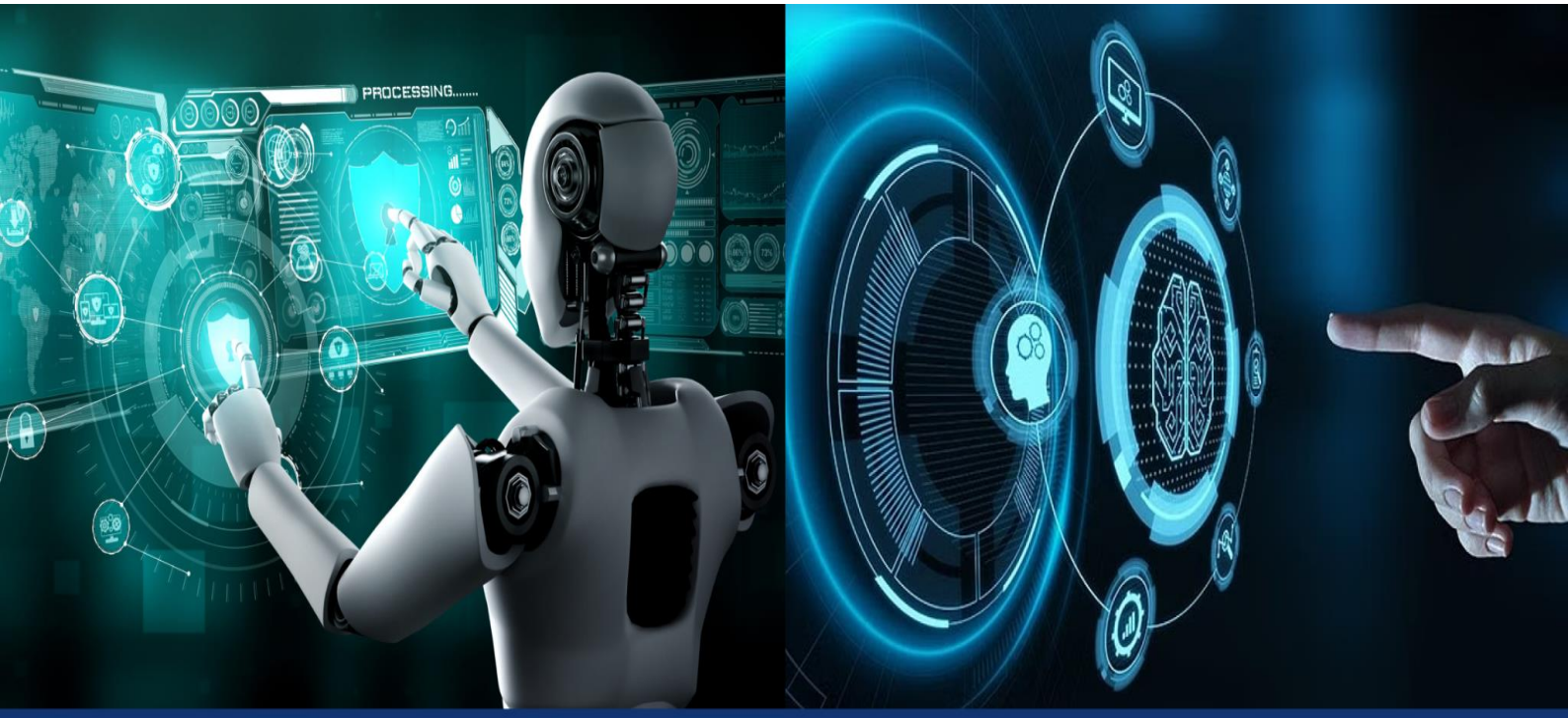
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771 **Volume 13, Issue 3, March 2025**

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Public and Private Email Services with Spam Spoiler Detection System

Sathiyapriya, Kaviya S, Jananisri K, Selciya M

Assistant Professor, Department of Cyber Security, Muthayammal Engineering College, Rasipuram, India

Student, Department of Cyber Security, Muthayammal Engineering College, Rasipuram, India

ABSTRACT: In recent years, cyber security incidents have occurred frequently. In most of these incidents, attackers have used different types of spam email as a knock-on to successfully invade government systems, well-known companies, and websites of politicians and social organizations in many countries. The detection of spam mail from big email data has been paid public attention. However, the camouflage technology of spam mail is becoming more and more complex, and the existing detection methods are unable to confront the increasingly complex deception methods and the growing number of emails. In this project, we proposed to design a novel efficient approach named Spam Spoiler for big e-mail data classification into four different classes: Normal, Fraudulent, Harassment, and Suspicious E-mails by using LSTM-based GRU.

KEYWORDS: Email Services, Spam Detection, Spoiler Detection Systems, Naive Bayes Classifier, Support Vector Machines (SVM), spam and non- spam, Deep Learning (Neural Networks), Natural Language Processing (NLP), Long Short- Term Memory (LSTM), Neural Network (RNN), Email services (whether public or private) and the specific mechanisms used to detect and prevent spam and spoilers.

I. INTRODUCTION

Dispatch stands for Electronic Correspondence. It's a system to shoot dispatches from one computer to another computer through the Internet. It's substantially used in business, education, specialized communication, and document relations. Globally, It allows communicating with people all over the world without bothering them. In 1971, a test dispatch was transferred to Ray Tomlinson to himself containing the textbook. E-mail dispatches are conveyed through dispatch waiters; it uses multiple protocols within the TCP/IP suite. Long Short-Term Memory (LSTM) is a specialized type of Recurrent Neural Network (RNN) designed to overcome the limitations of traditional RNNs when dealing with sequential data The main challenge that LSTM addresses is the vanishing gradient problem, a common issue faced by traditional RNNs when learning long-range dependencies in sequences. In standard RNNs, information from previous time steps is propagated through the network to influence predictions for future time steps. However, this propagation weakens over time, making it difficult for the model to retain important information from earlier steps. This is particularly problematic for tasks where long-term memory is required, such as language modeling, time-series forecasting, and many other applications involving sequences of data.

LSTM networks are designed to capture long- term dependencies by using a more advanced structure. They include memory cells that can store information over time and special gating mechanisms to regulate the flow of information into and out of these memory cells. These gates enable the LSTM to decide what to "remember" and "forget" from one time step to the next. which makes them particularly powerful for tasks such as natural language processing (NLP), speech recognition, and time- series forecasting.

Public and private email services are designed to provide users with a secure platform to send, receive, and manage their emails. Public email services, such as Gmail, Yahoo, and Outlook, are free-to-use and widely accessible, offering a range of features for general users. To ensure a better and safer email experience, these email services integrate sophisticated spam detection systems and, in some cases, advanced technologies to handle detection.

1.1 SCOPE OF THIS PROJECT

The scope of this project is to develop and Develop a system capable of detecting unsolicited emails, promotional

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

messages, phishing attempts, and other forms of spam. Implement a mechanism for detecting spoilers, which may include accidental or intentional disclosure of sensitive or time-sensitive information (e.g., sports results, movie endings). Ensure compatibility with both public and private email systems, allowing seamless deployment in various environments. Allow users to define what constitutes spam and spoilers through customizable settings. The project utilizes several Design algorithms that minimize the likelihood of falsely flagging legitimate emails as spam or spoilers. Ensure the system can process high volumes of emails in real-time or near real-time without significant delays. Work with stakeholders to create clear definitions and boundaries for spam and spoilers.

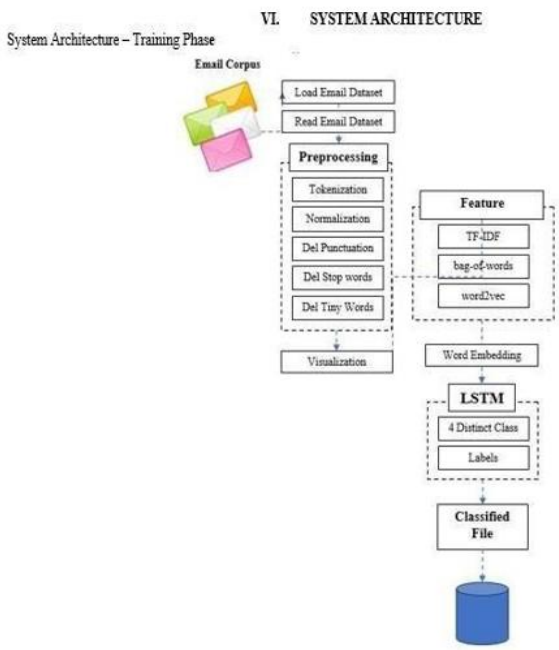
The scope also involves evaluating the Design an API or module that integrates seamlessly with existing public email services (e.g., Gmail API, Outlook API) and custom/private email servers .Use machine learning models, keyword-based filtering, and heuristics (e.g., Bayesian classifiers, blacklists). Develop NLP-based algorithms for identifying spoilers, such as context-sensitive keyword analysis (e.g., movie titles, sports events). Ensure proper storage of user preferences, email content history, and spam/spoiler categories. Implement a filter that flags emails based on content, sender reputation, and context. Continuously update spam patterns using feedback loops.

II. PROPOSED WORK

The proposed system for fraud mail identification aims to leverage machine learning algorithms to automatically detect and block fraudulent emails. The system uses Natural Language Processing (NLP) techniques to analyze the content of incoming emails, identifying patterns indicative of phishing attempts, spam, or other fraudulent behavior.

2.1 System Architecture & Flowchart

The system follows a Designing a architecture for spam spoiler detection in both public and private email services requires careful consideration of various components, security, scalability, and integration points.



III. DATA PREPROCESSING TECHNIQUES

Data preprocessing is crucial to improve the accuracy of detection models. Below are key data preprocessing techniques:

- Removing Duplicates: Eliminate duplicate emails to prevent biased training.
- Handling Missing Values: Replace missing subject lines or body text with placeholders.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Removing Special Characters & Punctuation: Normalize text by removing symbols like #, @, and * that do not contribute to spam detection.
- Email Header Analysis: Extract relevant header features such as sender, subject, and IP addresses.
- Stopword Removal: Remove common words like "the," "is," "and," to reduce noise.
- Handling HTML & Special Formatting: Strip unnecessary HTML tags and metadata.

Additionally, data augmentation techniques, such as oversampling and undersampling, are employed to address class imbalance, ensuring that the model is trained with an adequate representation of both the majority and minority classes. This step improves the robustness of the model and helps prevent overfitting, ensuring better generalization to unseen data. By combining these preprocessing strategies, the model's performance is optimized, leading to more accurate and reliable predictions in disease classification tasks.

3.1 Machine Learning

Machine learning is a branch of artificial intelligence that enables systems to learn from data and improve their performance over time without being explicitly programmed. By utilizing algorithms and statistical models, machine learning allows computers to identify patterns and make decisions or predictions

3.1.1 Support Vector Machine (SVM)

- Purpose: SVM is a supervised learning algorithm that classifies emails as spam or non-spam (ham) based on features extracted from the email content.
- Hyperparameters:
  - Keyword Frequency – Common spam words (e.g., "free," "win," "click here").
  - N-grams & Word Embeddings – Identifies patterns in text.
  - Header Analysis – Examines sender information and metadata.
  - Embedded Links & Attachments – Detects suspicious URLs or file types.
  - TF-IDF Scores – Measures the importance of words in an email compared to a dataset.

3.1.2. Recurrent Neural Networks (RNN)

- Purpose: Recurrent Neural Networks (RNN) are powerful for spam and spoiler detection.
- Step 1: Data Preprocessing:
- Tokenization: Convert email text into word or character sequences.
- Word Embedding (e.g., Word2Vec, GloVe): Represent words in a dense vector space.
- Stopword Removal & Lemmatization**:** Clean unnecessary words.
- Feature Engineering**:** Extract sender metadata, subject line, email structure, etc.

Step 2: Training RNN on Labeled Emails:

- Dataset: Spam/non-spam or spoiler/non- spoiler emails.
- Model Architecture: Simple RNN, LSTM (Long Short-Term Memory), or GRU (Gated Recurrent Unit) for better performance.

Step 3: Classification & Deployment:

- RNN models classify incoming emails in real-time.
- Integrated with cloud-based filtering systems.
- Continuous learning via user feedback (mark as spam).

3.1.3 LSTM (Long-Short Term Memory)

- Purpose: Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that overcomes the limitations of traditional RNNs by handling long-term dependencies effectively.
- LSTM prevents the vanishing gradient problem, allowing it to capture important context in long sequences, such as email messages**.**
- It is widely used in text classification tasks**,** including spam detection and spoiler filtering in emails.
- Memory Cells & Gates: LSTM has a forget gate, input gate, and output gate that help it selectively remember or discard information.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.2 Tools and Libraries

3.2.1 Python

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990. Like Perl, Python source codeis also available under the GNU General Public License (GPL). This tutorial gives enough understanding on Python programming language.

3.2.2 Tensor Flow

Tensor Flow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in ML, and gives developers the ability to easily build and deploy ML-powered applications. Tensor Flow provides a collection of workflows with intuitive, high-level APIs for both beginnersand experts to create machine learning models in numerous languages.

Developers have the option to deploy models on a number of platforms such as on servers, in the cloud devices, in browsers, and on many other JavaScript platforms. This enables developers to go from model building and training to deployment much more easily. Spam emails often contain misleading content or spoilers, which can be harmful or annoying to recipients. Detecting such messages requires advanced machine learning techniques.

3.2.3 Pandas

Pandas is a fast, powerful, flexible and easy to use open source data language.. It is used for tasks such as data cleaning, data exploration, and feature extraction, making it easier

to handle the input data for the disease prediction system. pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.

3.2.4 NumPy

NumPy is a fundamental library for numerical computing in Python, offering support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions. It is used to handle numerical data and perform mathematical operations on the dataset, including matrix operations required for training machine learning models. NumPy provides efficient storage and manipulation of large data arrays and supports comprehensive mathematical and statistical operations.

3.2.5 Matplotlib

Matplotlib and Seaborn are visualization libraries that provide functions to create a wide range of static, animated, and interactive plots. These libraries are used to visualize the data, model performance, and evaluation metrics, making it easier to interpret and present the results of the disease prediction system. Matplotlib allows the creation of various types of charts like line charts, bar charts, and histograms, while Seaborn offers higher-level functions for statistical visualizations and data exploration.

3.2.6 Keras

Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation. Allows the same code to run on CPU or on GPU, seamlessly prototype deep learning.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. PROGRAM

```
▼<configuration name="main" type="PythonConfigurationType" factoryName="Python" temporary="true">
    <module name="Spam Email Detection"/>
    <option name="INTERPRETER_OPTIONS" value=""/>
    <option name="PARENT_ENVS" value="true"/>
  ▼<envs>
      <env name="PYTHONUNBUFFERED" value="1"/>
    </envs>
    <option name="SDK_HOME" value=""/>
    <option name="WORKING_DIRECTORY" value="$PROJECT_DIR$"/>
    <option name="IS_MODULE_SDK" value="true"/>
    <option name="ADD_CONTENT_ROOTS" value="true"/>
    <option name="ADD_SOURCE_ROOTS" value="true"/>
    <option name="SCRIPT_NAME" value="$PROJECT_DIR$/main.py"/>
    <option name="PARAMETERS" value=""/>
    <option name="SHOW_COMMAND_LINE" value="false"/>
    <option name="EMULATE_TERMINAL" value="false"/>
    <option name="MODULE_MODE" value="false"/>
    <option name="REDIRECT_INPUT" value="false"/>
    <option name="INPUT_FILE" value=""/>
    <method v="2"/>
  </configuration>
▼<component name="FileTemplateManagerImpl">
  ▼<option name="RECENT_TEMPLATES">
    ▼<list>
       <option value="HTML File"/>
       <option value="Python Script"/>
     </list>
   </option>
  </component>
  <component name="ProjectId" id="2Zc1jruhQXdJY7bIVVMkLJu0RQ9"/>
▼<component name="ProjectViewState">
    <option name="hideEmptyMiddlePackages" value="true"/>
    <option name="showExcludedFiles" value="true"/>
    <option name="showLibraryContents" value="true"/>
  </component>
▼<component name="PropertiesComponent">
    <property name="DefaultHtmlFileTemplate" value="HTML File"/>
    <property name="RunOnceActivity.ShowReadmeOnStart" value="true"/>
    <property name="last_opened_file_path" value="$PROJECT_DIR$/../lstm"/>
    <property name="settings.editor.selected.configurable" value="com.jetbrains.python.configuration.PyActiveSdkModuleConfigurable"/>
  </component>
▼<component name="RecentsManager">
  ▼<key name="CopyFile.RECENT_KEYS">
      <recent name="F:\PYTHON PROJECT\Spam Email Detection"/>
      <recent name="F:\PYTHON PROJECT\Spam Email Detection\templates"/>
    </key>
  </component>
▼<component name="RunManager" selected="Python.dataapp">
  ▼<configuration name="dataapp" type="PythonConfigurationType" factoryName="Python" temporary="true">
      <module name="Spam Email Detection"/>
      <option name="INTERPRETER_OPTIONS" value=""/>
      <option name="PARENT_ENVS" value="true"/>
    ▼<envs>
        <env name="PYTHONUNBUFFERED" value="1"/>
      </envs>
```

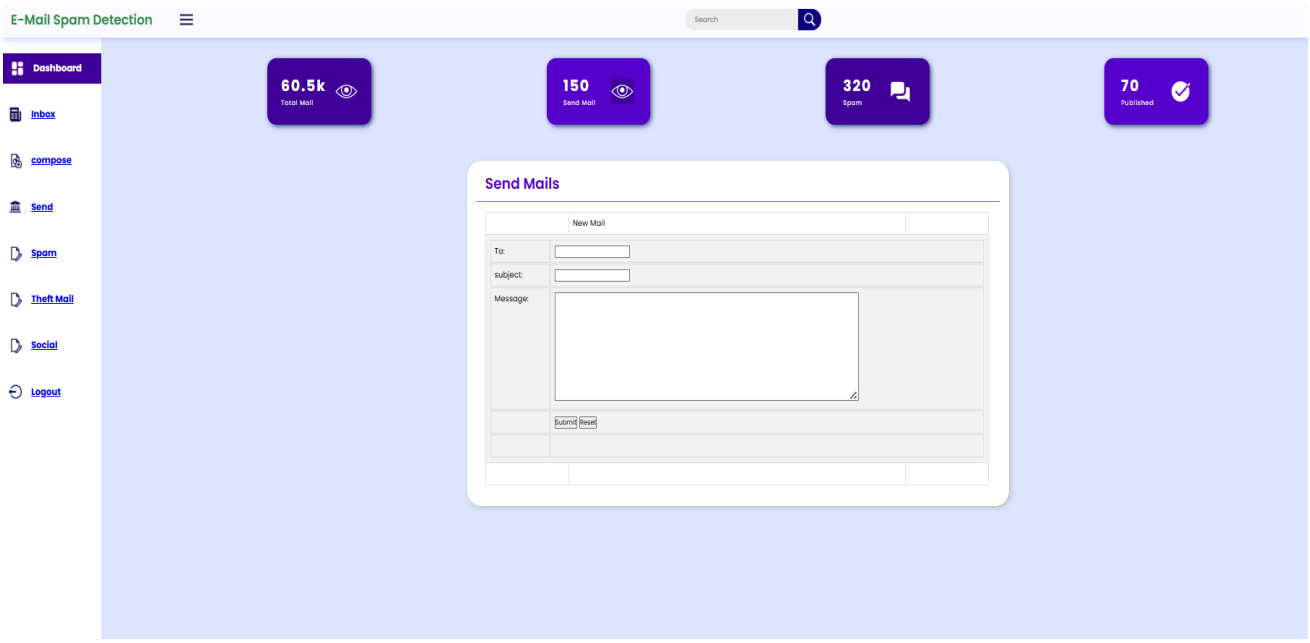
© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

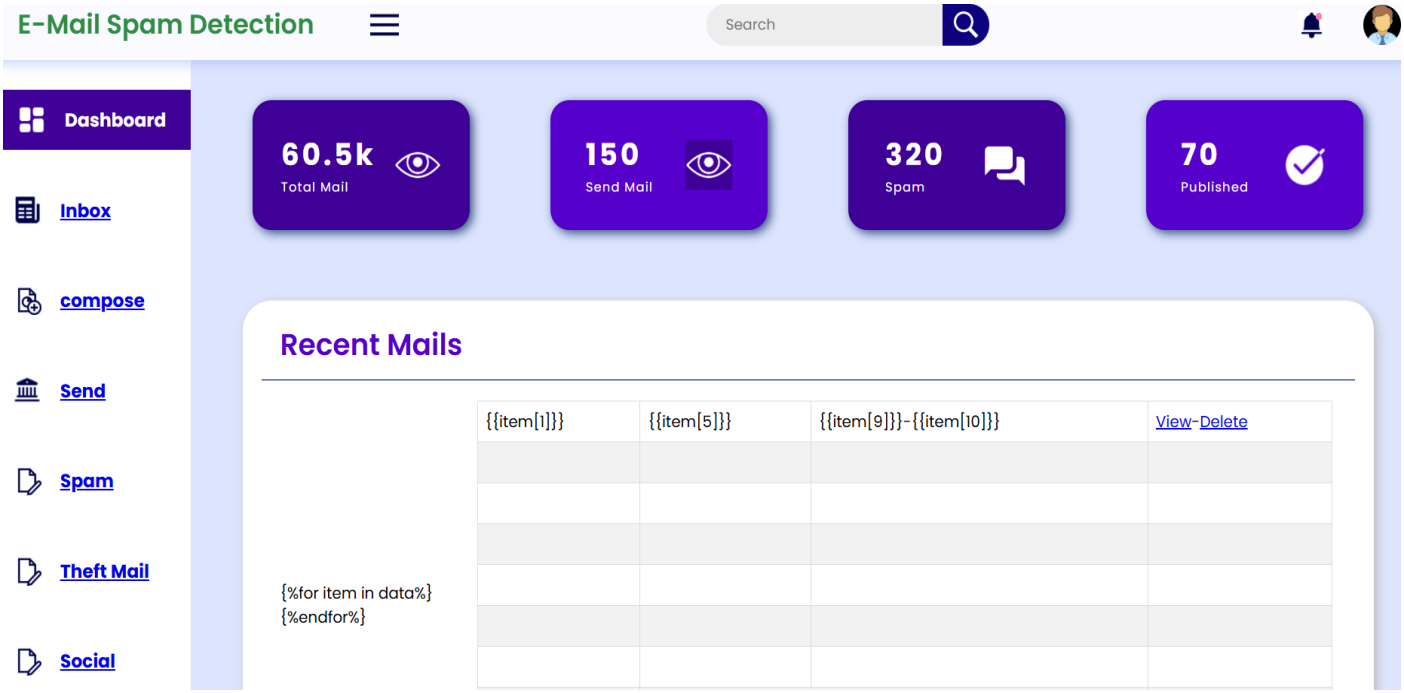
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. OUTPUT





© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303098

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. CONCLUSION

In conclusion, both public and private email Services with spam spoiler detection system offer Distinct advantages and drawbacks.

Future work includes the incorporation of real-time monitoring data and refining these algorithms for Prioritize security, encryption, and data privacy, they offer more control over spam and spoiler detection but may require higher cost, tehnical expertise.

REFERENCES

1. W.-F. Chen, M. Hagen, B. Stein, and M. Potthast, ‘‘A user study on snippet generation: Text reuse vs. paraphrases,’’ in Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jun. 2018, pp. 1033–1036, doi: 10.1145/3209978.3210149.
2. D. Maxwell, L. Azzopardi, and Y. Moshfeghi, ‘‘A study of snippet length and informativeness,’’ in Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Aug. 2017, pp. 135– 144, doi: 10.1145/3077136.3080824.
3. J. Sachse, ‘‘The influence of snippet length on user behavior in mobile Web search,’’ Aslib
4. J. Inf. Manage., vol. 71, no. 3, pp. 325–343, May 2019, doi: 10.1108/AJIM-07-2018-0182.
5. J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon, ‘‘What snippet size is needed in mobile Web search?’’ in Proc. Conf. Hum. Inf. Interact. Retr. (CHIIR), Mar. 2017, pp. 97– 106, doi: 10. 1145/3020165.3020173.[7]T.
6. Berners-Lee, J. Hendler, and O. Lassila, D. L. McGuinness and F. Van Harmelen,‘‘Owl Web ontology language. Carroll and G. Klyne. (Feb. 2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C, W3C Recommendation. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf- concepts20040210/
7. M. Schmachtenberg, C. Bizer, and H. Paulheim, ‘‘Adoption of the linked data best practices in different topical domains,’’ in Proc. Int. Semantic Web Conf. Cham, Switzerland: Springer, 2014, pp. 245–260.
8. J. Ker, L. Wang, J. Rao, and T. Lim, ‘‘Deep learning applications in medical image analysis,’’ IEEE Access, vol. 6, pp. 9375–9389, 2018.
9. K. Doi, ‘‘Computer-aided diagnosis in medical imaging: Historical review, current status and future potential,’’ Computerized Med. Imag. Graph., vol. 31, nos. 4–5, pp. 198–211, Jun. 2007.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details