MCA Final Year Project (Review I)

E-Mail Spam Detection Using Machine Learning

Submitted to the Presidency University, Bengaluru in partial fulfillment for the award of the degree of Master of Computer Applications(MCA)

Project Number : 215

Name	Roll Number
Ashwini Hosamani	20232MCA0263

Under the supervision of

Mr. Sakthi S

Asst. Prof, Department of SCSE

School of Computer Science and Engineering

PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

Content

- Problem Statement
- Literature Survey
- Tools and Technologies to be used
- GitHub
- Timeline of the Project
- References

Problem Statement

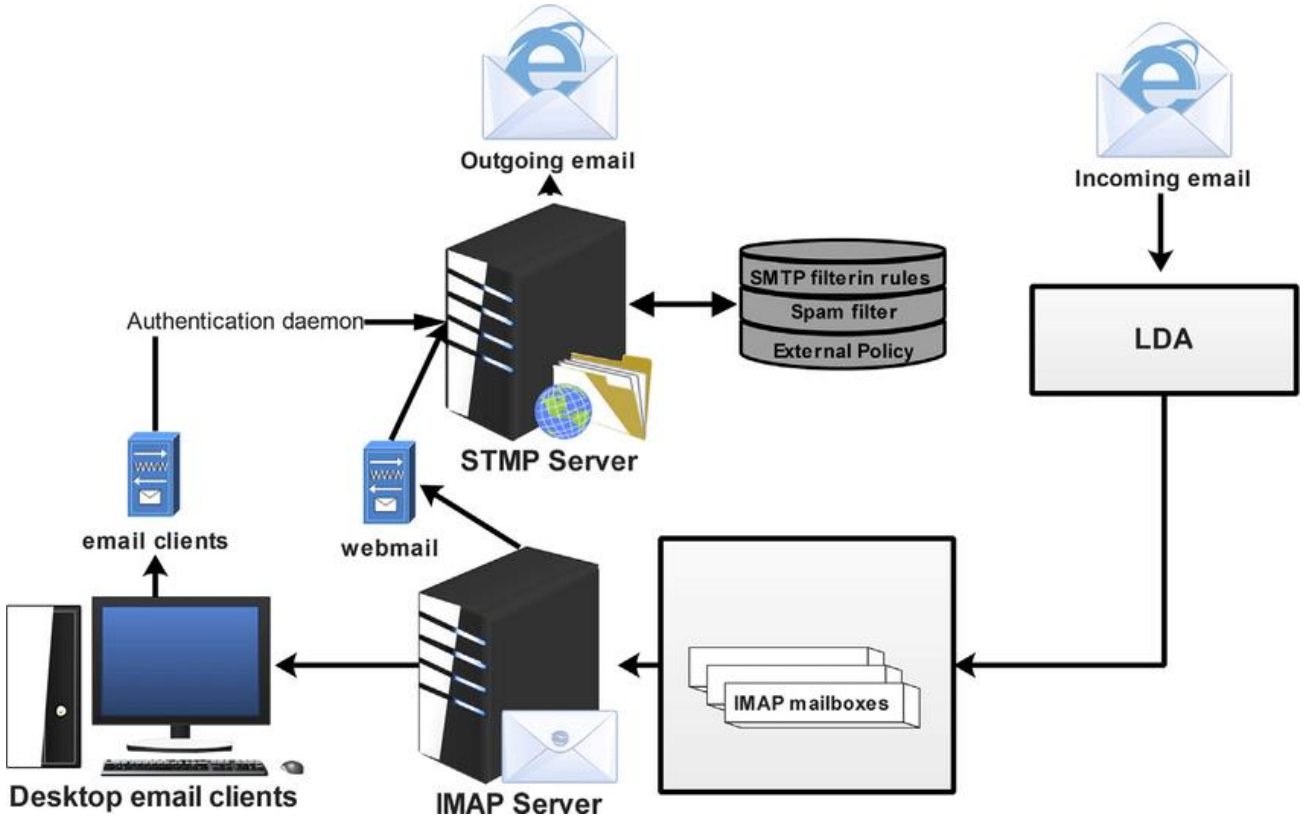
- Spam detection helps filter out unwanted emails, but built-in filters are not always reliable, sometimes misclassifying important messages.
- Spammers continuously evolve their techniques, making it challenging to maintain effective filtering systems.
- Spam emails waste storage, consume time, cause financial losses for businesses, and pose cybersecurity threats like malware and phishing.
- This project evaluates multiple AI models on the same dataset, comparing their performance based on accuracy and efficiency.

Literature Review

SL No	Year	Authors	Title	Methodology	Advantage	Disadvantage
1	2023	Maxime Labonne, Sean Moran	Spam-T5: Benchmarking LLMs for Email Spam Detection	Compares BERT-like, Sentence Transformers, and Seq2Seq models for spam detection. Introduces Spam-T5.	High accuracy, effective in few-shot scenarios.	High computational cost.
2	2023	Suhaima Jamal, Hayden Wimmer	Improved Transformer-Based Spam Detection	Fine-tunes BERT models for spam and phishing detection.	Performs well on unbalanced datasets.	Limited adaptability to evolving spam tactics.
3	2022	Sultan Zavrak, Seyhmus Yilmaz	Hybrid Deep Learning for Email Spam Detection	Uses CNNs, GRUs, and attention mechanisms for classification.	Outperforms traditional models.	High computational requirements.
4	2022	Vijay Srinivas Tida, Sonya Hsu	Universal Spam Detection with Transfer Learning	Fine-tunes BERT on multiple datasets for improved spam classification.	High accuracy (97%), robust model.	May struggle with domain-specific nuances.

5	2023	P. Charanarur, H. Jain, G.S. Rao, et al.	ML-Based Spam Mail Detector	Applies various ML models for email classification.	Enhances security, improves accuracy.	Computational resource requirements.
6	2021	M. Al-Sarem, M. Al-Hadhrami, A. Alshomrani, et al.	Deep Learning for Spam Detection	Fine-tunes BERT, compares with BiLSTM, k-NN, and Naive Bayes.	High accuracy (98.67%), strong spam detection.	Dependence on pre-trained models.

Module Design



Modular Breakdown

Module 1: Email Data Preprocessing

Functionality:

- Extracts email content (subject, body, metadata) for analysis.
- Removes unnecessary elements like HTML tags, stopwords, and special characters.

Importance:

- Ensures high-quality input data for accurate spam classification.
- Reduces noise and enhances meaningful feature extraction.

Module 2: Spam Classification using Machine Learning

Functionality:

- Applies Supervised Learning Models (e.g., SVM, Random Forest) for spam detection.
- Labels emails as spam or legitimate based on trained models.

Importance:

- Forms the core of the system by identifying and filtering spam emails.
- Improves detection accuracy, reducing false positives and false negatives.

Module 3: Feature Extraction & Selection

Functionality:

- Extracts key features from emails such as **word frequency, presence of URLs, metadata analysis**.
- Uses N-grams, TF-IDF, and Word Embeddings for text representation.

Importance:

- Enhances model efficiency by reducing irrelevant data.
- Improves spam classification accuracy by focusing on key indicators.

Module 4: Model Deployment & System Integration

Functionality:

- Integrates with **email clients (e.g., Gmail API, Outlook API)** for seamless filtering.
- Continuously updates the model using feedback from user-labeled emails

Importance:

- Provides an adaptive system that improves over time based on new spam patterns.

Tools And Technologies To Be Used

1. Development Tools:

- **Google Colab / Jupyter Notebook** – For coding and testing the machine learning model.

2. Programming Language:

- **Python** – For implementing facial recognition and the music recommender system.

3. Frameworks & Libraries:

- **Scikit-learn** – For machine learning algorithms and feature selection.
- **NLTK / SpaCy** – For text preprocessing and natural language processing tasks.
- **TensorFlow / Keras** – For building and training deep learning models.

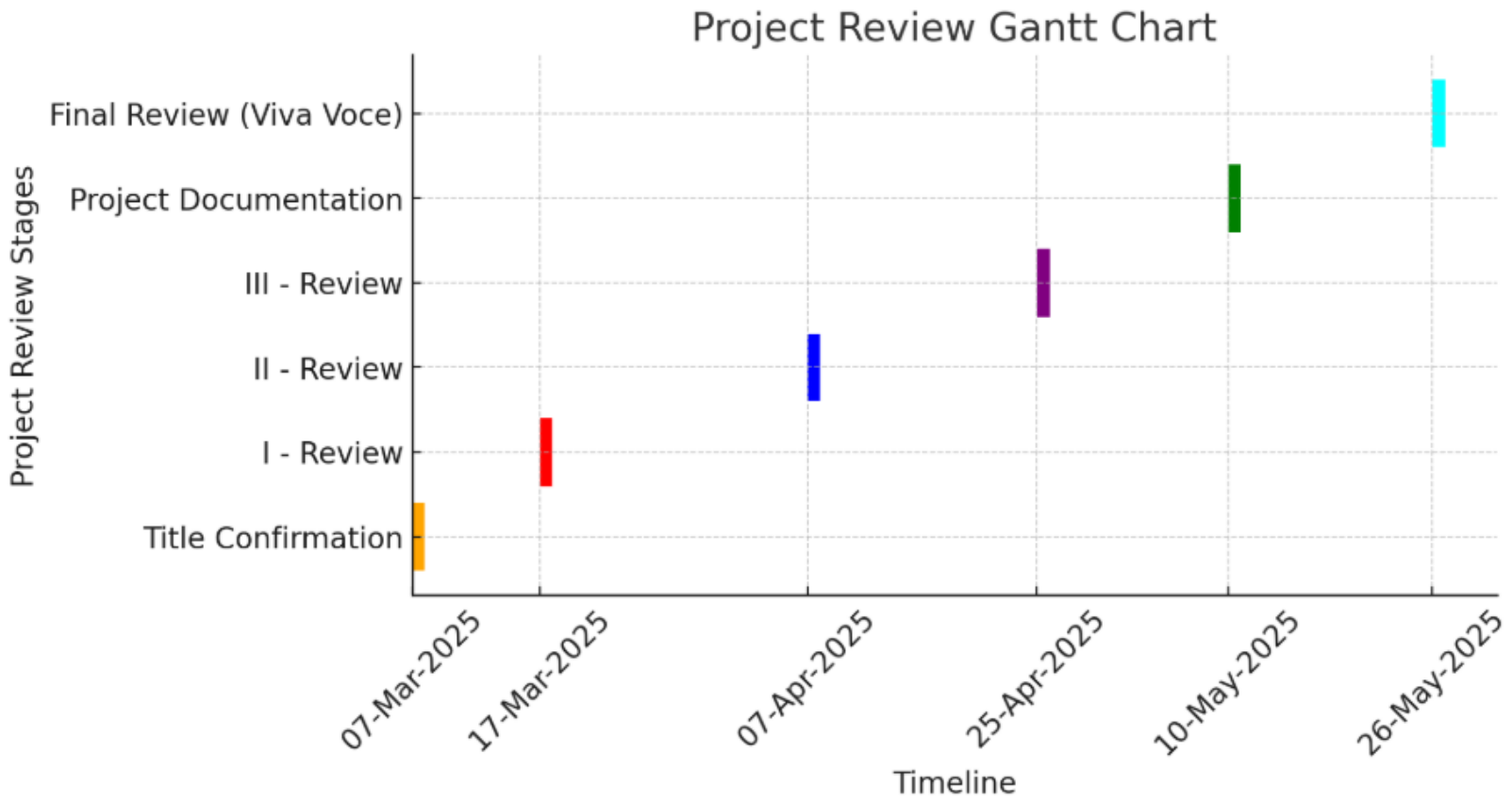
4. Additional Tools:

- **Gmail API / Outlook API** – For integrating the spam detection system with email clients.
- **Flask / FastAPI** – For deploying the trained spam detection model as an API.

GitHub Link

- https://github.com/ashwinihosamani/Email-Spam-Detection-Using-Machine-Learning

Timeline of the Project



References

1. M. Labonne and S. Moran, "Spam-T5: Benchmarking LLMs for Email Spam Detection," in Proceedings of the International Conference on Computational Linguistics (COLING), 2023.

2. S. Jamal and H. Wimmer, "Improved Transformer-Based Spam Detection," Journal of Artificial Intelligence Research (JAIR), vol. **35**, pp. **120-135**, 2023.

3. S. Zavrak and S. Yilmaz, "Hybrid Deep Learning for Email Spam Detection," IEEE Transactions on Neural Networks and Learning Systems, vol. **34**, no. **6**, pp. **987-999**, 2022.

4. V. S. Tida and S. Hsu, "Universal Spam Detection with Transfer Learning," in Proceedings of the ACM Conference on Machine Learning (ACM-ML), pp. **230-242**, 2022.

5. narur, H. Jain, G. S. Rao, et al., "ML-Based Spam Mail Detector," Springer Journal of Machine Learning and Applications, vol. **27**, pp. **89-104**, 2023.

6. *M. Al-Sarem, M. Al-Hadhrami, A. Alshomrani, et al., "Deep Learning for Spam Detection," Expert Systems with Applications, Elsevier, vol. **167**, pp. **113872**, 2021.*

7. *M. A. Shafi, H. Hamid, E. G. Chiroma, J. S. Dada, and B. Abubakar, "Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems," in Proceedings of the International Conference on Artificial Intelligence and Machine Learning (AIML), pp. **45-56**, 2018.*

8. *M. Almeida, T. A. Almeida, and A. Silva, "Spam Email Detection Using Deep Learning Techniques," in Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. **92-105**, 2021..*


Thank
You!