Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

SPAM EMAIL DETECTION SYSTEM

Ashutosh Dash, 4th Year, Department of CSE, Gandhi Institute for Technology, BPUT, India
ashutosh2021@gift.edu.in
Ashish Kumar Sahoo, 4th Year, Department of CSE, Gandhi Institute for Technology, BPUT, India
asahoo2021@gift.edu.in
Er. Jagannath Ray, Assistant Professor, Department of CSE, Gandhi Institute for Technology, BPUT, India

Abstarct-
Email communication is integral to modern life, yet the proliferation of unsolicited spam emails presents significant challenges to security, productivity, and privacy. Traditional spam filters often fail against evolving spam tactics, necessitating intelligent solutions. This paper details the development of a robust spam email detection system using machine learning. The system employs a stacking classifier model, integrating Naive Bayes, Support Vector Machine (SVM), and Random Forest algorithms as base learners, with Logistic Regression as a meta-learner to enhance classification accuracy. Email text is preprocessed and converted into numerical features using TF-IDF vectorization. A user-friendly interface developed with Streamlit allows for real-time email classification. This work highlights the practical application of ensemble machine learning techniques in cybersecurity for effective text classification.

Keywords:
Spam Detection, Machine Learning, Stacking Classifier, Naive Bayes, SVM, Random Forest, TF-IDF, Natural Language Processing.

1. INTRODUCTION

Electronic mail is a cornerstone of personal and professional communication. However, its widespread use has made it a primary channel for spam, which includes phishing attempts, malware distribution, and fraudulent schemes. Effective spam detection is crucial for safeguarding email security, maintaining user productivity, and optimizing network resources. This project focuses on developing a machine learning model to accurately classify emails as spam or legitimate (ham). The motivation is to enhance email security by mitigating risks from malicious content, improve user productivity by automating spam management, and protect users from deception and fraud.

2. LITERATURE REVIEW

Spam detection has evolved from early keyword matching and rule-based systems, which were easily bypassed, to more sophisticated machine learning approaches.

- **Naive Bayes:** An early and popular probabilistic classifier for text, effective and computationally efficient. It's used as a base estimator in this project.
- **Support Vector Machines (SVM):** Powerful for high-dimensional data, finding an optimal hyperplane to separate classes. SVM (SVC) is another base estimator here.
- **Random Forests:** An ensemble method using multiple decision trees to reduce overfitting and improve generalization. It's also a base estimator in our model.
- **Logistic Regression:** A statistical model for binary classification, often used as a meta-estimator in stacking ensembles, as in this project.

- **Ensemble Learning (Stacking):** Combines predictions from multiple base learners using a meta-learner to achieve better performance than any single model. This project utilizes a Stacking Classifier.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** A technique to evaluate word relevance in documents, used here for feature extraction from email text.

3. SYSTEM DESIGN

The system architecture comprises distinct layers:

- **Data Layer:** Utilizes a primary dataset (dataset.csv) of labeled spam/ham emails. It also allows user contributions to this dataset and uses a MySQL database (spam_detection) for storing user credentials and prediction history.
- **Machine Learning Pipeline:** Involves data ingestion, preprocessing (label encoding, text cleaning), TF-IDF feature extraction, training a Stacking Classifier (SVC, RandomForestClassifier, MultinomialNB as base; LogisticRegression as meta-model), model evaluation, and persistence using joblib.
- **Application Layer:** A Streamlit-based web UI for user authentication, real-time prediction, data contribution, and viewing prediction history.

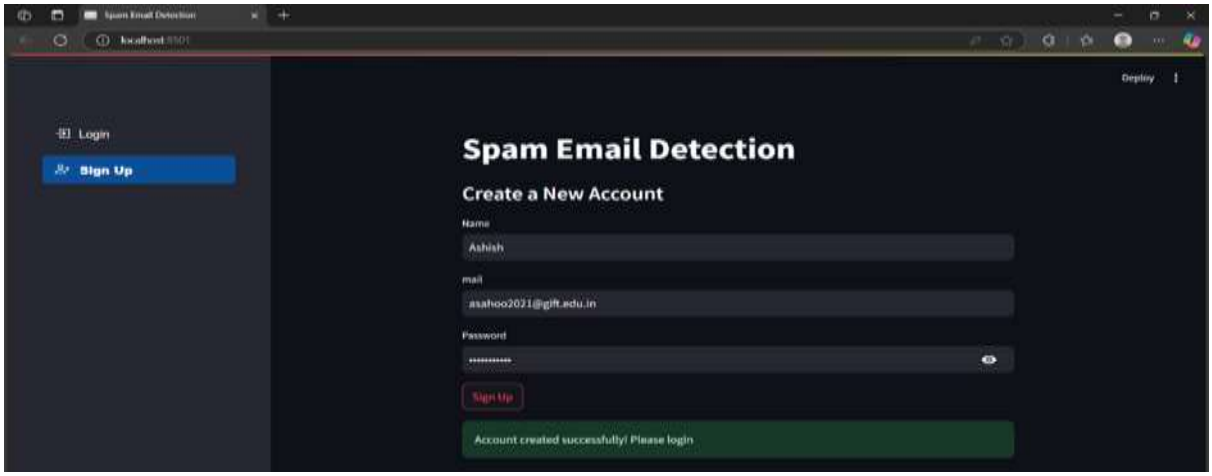
The dataset (dataset.csv) contains 5563 email messages with 'Category' (ham/spam) and 'Message' (email text). There's a class imbalance with approximately 86.5% ham and 13.5% spam messages. Data preprocessing includes loading data, handling missing values (though specifics beyond pd.notnull(df) aren't detailed, it is an initial step), label encoding the 'Category' column, and text preparation (lowercasing, removing punctuation/stopwords are standard before TF-IDF). The data is split into 80% training and 20% testing sets. TF-IDF vectorization converts email text into numerical feature vectors.

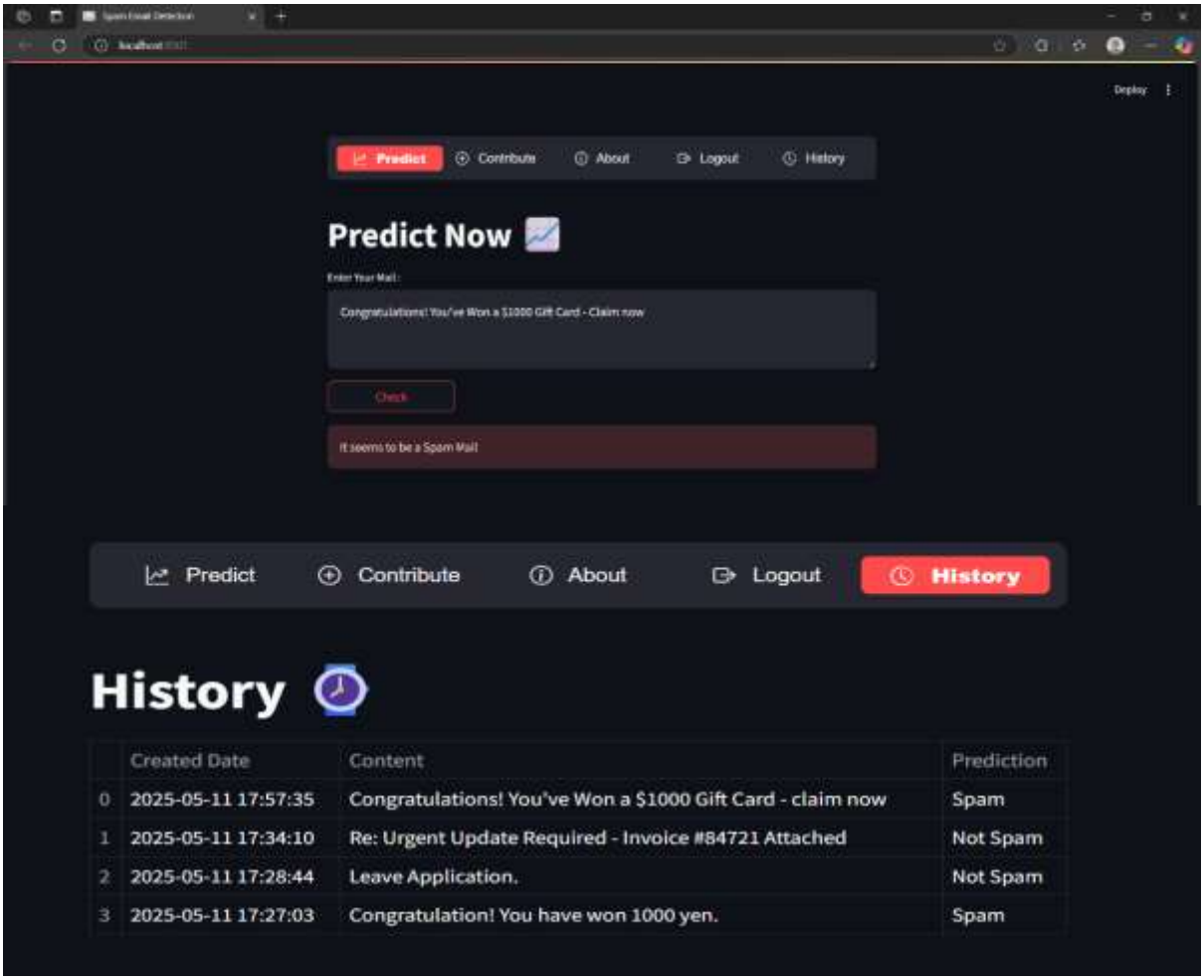
4. IMPLEMENTATION

The implementation involves:

- **Data Acquisition and Preparation:** Loading and parsing dataset.csv, cleaning data, and transforming 'Category' labels using Label Encoding.
- **Feature Engineering:** Applying TF-IDF to convert email text into numerical features.
- **Model Development:** Constructing a Stacking Classifier with SVC, RandomForestClassifier, and MultinomialNB as base learners, and LogisticRegression as the meta-learner. The model is trained on 80% of the data.
- **Model Persistence:** Saving the trained model and TF-IDF vectorizer using joblib.
- **Web Application:** Developing a Streamlit application for user registration/login, real-time classification, new sample contribution, and prediction history viewing.
- **Database Integration:** Using MySQL (spam_detection) to store user credentials and prediction history.

Python is the core language, with libraries like NumPy, Pandas for data handling; Scikit-learn for machine learning; Joblib for model persistence; Streamlit for the web app; and MySQL Connector for database interaction.





5. RESULTS

The Stacking Classifier model achieved an accuracy of 98.8% on the test dataset (20% of over 5000 rows). This high accuracy indicates the model's robust performance in distinguishing spam from ham emails. The success is attributed to the ensemble approach, which combines the diverse strengths of SVC, Random Forest, and Naive Bayes, with Logistic Regression effectively weighing their outputs. While accuracy is high, other metrics like precision, recall, and F1-score are also important, especially with class imbalance, though specific values for these were not detailed in the immediate summary. The system provides real-time prediction via a user-friendly interface, allows dataset augmentation through user contributions, and maintains user-specific prediction history.

6. CONCLUSION

This project successfully developed an effective machine learning-based spam email detection system using a Stacking Classifier, achieving 98.8% accuracy. The system integrates data preprocessing, TF-IDF feature extraction, ensemble modeling, and a user-friendly Streamlit application with database support for user management and history logging. It provides a practical solution to enhance email security and productivity.

Future work could include advanced feature engineering (n-grams, metadata), comprehensive model tuning, addressing class imbalance more explicitly (e.g., SMOTE), exploring alternative models like Gradient Boosting or Deep Learning, implementing online learning for continuous adaptation, and enhancing application security (password hashing) and scalability.

ACKNOWLEDGEMENT

We are grateful to Er. Jagannath Ray for guidance and support throughout this project. We also thank Dr. Sujit Kumar Panda, H.O.D, Department of Computer Science and Engineering, for their support.

REFERENCES

- **https://scikit-learn.org/**
- **https://pandas.pydata.org/docs/**
- **https://numpy.org/doc/**
- **https://docs.streamlit.io/**
- **https://joblib.readthedocs.io/**
- **https://www.deeplearningbook.org/**
- **https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-008.pdf**