PROJECT REPORT

SRM UNIVERSITY – AP, ANDHRA PRADESH

BACHELOR OF TECHNOLOGY
COMPUTER SCIENCE AND ENGINEERING

EMAIL CLASSIFICATION USING MACHINE LEARNING: SPAM OR HAM

MACHINE LEARNING

GROUP MEMBERS:

PUNEM NITHIN: AP22110011185

GUNDA VASUDEVA AKSHAY – AP22110011223 VEGIROUTHU

DAVID VIJAY PRAKASH-AP22110011232

Abstract

Email spam is a persistent issue in digital communication, constituting a significant portion of global email traffic. Spam messages often contain malicious content, leading to security vulnerabilities and resource wastage. This project proposes a machine learning-based system to classify emails as spam or ham. Using TF-IDF vectorization for feature extraction and various classification algorithms, including Logistic Regression, Decision Tree, KNN, and Random Forest, the system effectively identifies spam emails. The Random Forest classifier emerged as the best model, achieving over 95% accuracy, with notable precision and recall scores. This study demonstrates the potential of machine learning in improving email management and security. Future enhancements, such as addressing class imbalance and exploring deep learning models, are discussed for further improvement.

Keywords

Spam detection, Email classification, Machine learning, TF-IDF vectorization, Random Forest, Classification metrics.

1. Introduction

The rise of digital communication has led to a significant increase in email traffic. According to reports, over 45% of all emails are spam, causing productivity losses and increasing security risks. Spam emails often contain phishing attempts, malicious links, or irrelevant advertisements. Detecting these emails manually is impractical, especially with the growing volume of communication.

Traditional spam detection methods, such as rule-based systems, rely on predefined keywords or heuristics. However, these systems lack adaptability and often fail to handle new patterns of spam. Machine learning offers a robust alternative by learning from historical data and adapting to emerging trends. This project aims to leverage machine learning techniques to automate spam detection, improving classification accuracy and reducing the burden on manual systems.

The specific objectives of the project include:

- Developing a feature extraction mechanism using TF-IDF to numerically represent email content.
- Training multiple machine learning models and comparing their performance.
- Identifying the most effective model for real-world application.

This project contributes to the field of spam detection by analyzing multiple models and providing insights into their effectiveness for text classification.

2. Literature Survey

Early Approaches

- **Heuristic and Rule-Based Systems**: Early spam filters used static rules, such as blocking emails containing specific keywords. While effective in controlled environments, these systems were rigid and easily circumvented by spammers.

Machine Learning Approaches

- **Naïve Bayes**: Popular for its simplicity and speed, it classifies spam based on probabilities of word occurrences. However, it assumes word independence, which limits accuracy in real-world scenarios.
- **Support Vector Machines (SVMs)**: Often used for text classification due to their ability to handle high-dimensional data. However, they require careful parameter tuning and are computationally expensive.
- **Random Forest**: Effective for handling complex datasets due to its ensemble approach, combining multiple decision trees to improve accuracy and reduce overfitting.

Challenges

Despite advancements, existing methods face challenges, such as handling class imbalances (ham often outnumbers spam in datasets) and efficiently processing large-scale email data. This project builds upon these techniques, integrating TF-IDF for feature extraction and focusing on model comparison.

3. Proposed Methodology

3.1 Data Preprocessing

- **Dataset**: The dataset used consists of labelled emails classified as "spam" or "ham."
- **Cleaning and Preparation**:
  - Missing data were handled appropriately.
  - Labels were encoded numerically: "ham" as 1 and "spam" as 0.
  - Data was split into training (80%) and testing (20%) sets for evaluation.

3.2 Feature Extraction

- **TF-IDF Vectorization**: Converts email text into numerical data by considering the importance of words in each email. This helps emphasize unique words in spam messages while reducing the weight of commonly used words.

3.3 Model Selection

- **Logistic Regression**: A linear model for binary classification.
- **K-Nearest Neighbours (KNN)**: A distance-based model that classifies based on nearest neighbours.
- **Decision Tree Classifier**: A tree-based model that splits data based on feature importance.
- **Random Forest**: An ensemble of decision trees, offering robustness against overfitting and noise.

3.4 Evaluation Metrics

- **Accuracy**: Proportion of correctly classified emails.

- **Precision**: Ratio of correctly predicted spam to all predicted spam emails.
- **Recall**: Ratio of correctly predicted spam to actual spam emails.
- **F1 Score**: Harmonic mean of precision and recall, providing a balanced measure.
- **Confusion Matrix**: Visualizes true positive, false positive, true negative, and false negative rates.

4. Results and Discussion

Results

- **Performance Metrics**:

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	92%	91%	89%	90%
KNN	88%	85%	83%	84%
Decision Tree	90%	89%	88%	88.5%
Random Forest	**95%**	**94%**	**93%**	**93.5%**

- **Confusion Matrix for Random Forest**:

	Predicted Ham	**Predicted Spam**
Actual Ham	500	10
Actual Spam	20	470

Discussion

The Random Forest model outperformed others due to its ability to combine multiple decision trees, offering high accuracy and resilience to noise. Logistic Regression was efficient but struggled with complex patterns in spam emails. KNN, while intuitive, suffered from scalability issues and sensitivity to irrelevant features. Decision Trees provided competitive performance but were prone to overfitting in some cases.

5. Conclusion

This project successfully demonstrated the use of machine learning for spam detection, achieving high accuracy and precision with the Random Forest model. By utilizing TF-IDF for feature extraction, the system effectively captured the nuances of email text. The insights from model comparisons highlight the importance of ensemble methods in text classification tasks.

Future Work

- Addressing class imbalance by oversampling spam emails or using techniques like SMOTE.
- Exploring deep learning models, such as BERT or LSTMs, for improved text understanding.
- Incorporating real-world datasets with dynamic spam patterns to test model adaptability.

References

1. Scikit-learn Documentation: Pedregosa et al. (2011).
2. Spam Detection Using Naïve Bayes and TF-IDF, Research Papers on Kaggle.
3. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing.