UPES

UNIVERSITY OF TOMORROW

Report

Project Topic:
Email spam classifier

Submitted by

Name	SAP ID	Batch
Advaitesha Gupta	500095942	B2(H)
Akshiti Agarwal	500093633	B1(H)
Riddhi Jain	500096443	B2(H)
Srishti Bhatnagar	500093666	B1(H)

Contents

1. Abstract

This project tackles the challenge of email spam filtering by implementing a logistic regression classifier. Unwanted spam emails can significantly hinder workflow and introduce security vulnerabilities. Our approach leverages logistic regression, a machine learning algorithm adept at binary classification tasks like spam detection. Emails will undergo pre-processing using Natural Language Processing techniques for feature extraction. This may involve tokenization, stop-word removal, and the creation of features that capture the essence of spam emails. The trained logistic regression model will then analyze incoming emails and predict the probability of them being spam. The project will evaluate the model's effectiveness using metrics like accuracy, precision, and recall. This project contributes to a more streamlined email experience by filtering out unwanted spam messages.

2. Introduction

The ever-increasing influx of spam emails into our inboxes poses a significant challenge. These unsolicited messages not only disrupt productivity and waste valuable time but can also harbour phishing attempts or malware, jeopardizing our security. This project aims to combat this issue by developing a robust spam classifier utilizing the Random Forest algorithm, a machine learning technique well-suited for classification tasks.

Random Forest offers a powerful approach to categorize emails as either spam or legitimate. Our strategy involves analysing email content to identify key features that differentiate spam emails from legitimate ones. These features might encompass urgency indicators, unusual sender addresses, specific keywords or phrases commonly associated with spam, or even stylistic elements. By meticulously extracting these features through Natural Language Processing techniques like tokenization and stop-word removal, we can create a comprehensive profile of a typical spam email.

The heart of the project lies in the Random Forest model. This model will be trained on a large dataset of labelled emails, allowing it to learn the intricate relationships between the extracted features and the corresponding spam/not-spam classifications. Once trained, the model will be able to analyse incoming emails and predict the probability of them being spam.

This project will evaluate the effectiveness of the Random Forest model using industry-standard metrics like accuracy, precision, and recall. Ultimately, we aim to develop a reliable spam classifier that significantly reduces the volume of spam emails reaching our inboxes, thereby enhancing productivity, minimizing security risks, and fostering a more streamlined email experience.

3. Motivation

- **Combatting Productivity Drain:** Spam emails are a major time waster, forcing us to constantly delete unwanted messages. A logistic regression classifier can significantly reduce this burden, boosting productivity.
- **Enhanced Security:** Spam often acts as a gateway for phishing scams and malware. By filtering them out, logistic regression safeguards users from online threats, protecting sensitive information.
- **Improved User Experience:** Imagine an inbox free from the clutter of spam. A logistic regression classifier fosters a more streamlined email experience, allowing for quicker access to important messages and eliminating frustration.
- **Random Forest** is a powerful machine learning algorithm that is well-suited for spam classification due to its:
- **Efficiency**: Random Forest is a fast and efficient algorithm that can handle large datasets with ease, making it ideal for real-time email filtering.
- **Interpretability**: Random Forest provides insights into the features most indicative of spam, helping to understand spam tactics.
- **Accuracy**: Random Forest can achieve high accuracy in spam filtering, minimizing false positives and negatives.
- **Robustness**: Random Forest is a robust algorithm that can handle noisy data and outliers, making it a reliable choice for spam classification.

By training a Random Forest model on a large dataset of labeled emails, it can learn the intricate relationships between the extracted features and the corresponding spam/not-spam classifications. Once trained, the model can analyze incoming emails and predict the probability of them being spam, providing a reliable, user-centric spam classification system that empowers users and simplifies email management.

4. Objective

- **Leverage Random Forest for Spam Classification:** Utilize Random Forest for Spam Classification: This project is designed to harness the capabilities of Random Forest, a robust machine learning algorithm well-suited for binary classification tasks. Random Forest offers efficiency, interpretability, and high accuracy, making it an excellent choice for spam classification.

- **Multi-faceted Approach to Spam Filtering:** This project goes beyond a simple script that filters out spam based on a predefined list of keywords. It incorporates Natural Language Processing (NLP) techniques for feature extraction.
    - Feature Extraction with NLP
    - Model Training and Spam Prediction

- **Rigorous Evaluation for Real-World Performance:** The project success hinges not only on the model's ability to classify spam emails but also on its effectiveness in a real-world setting. To achieve this, the project will employ industry-standard metrics like accuracy, precision, and recall to evaluate the model's performance.

5. Literature Review

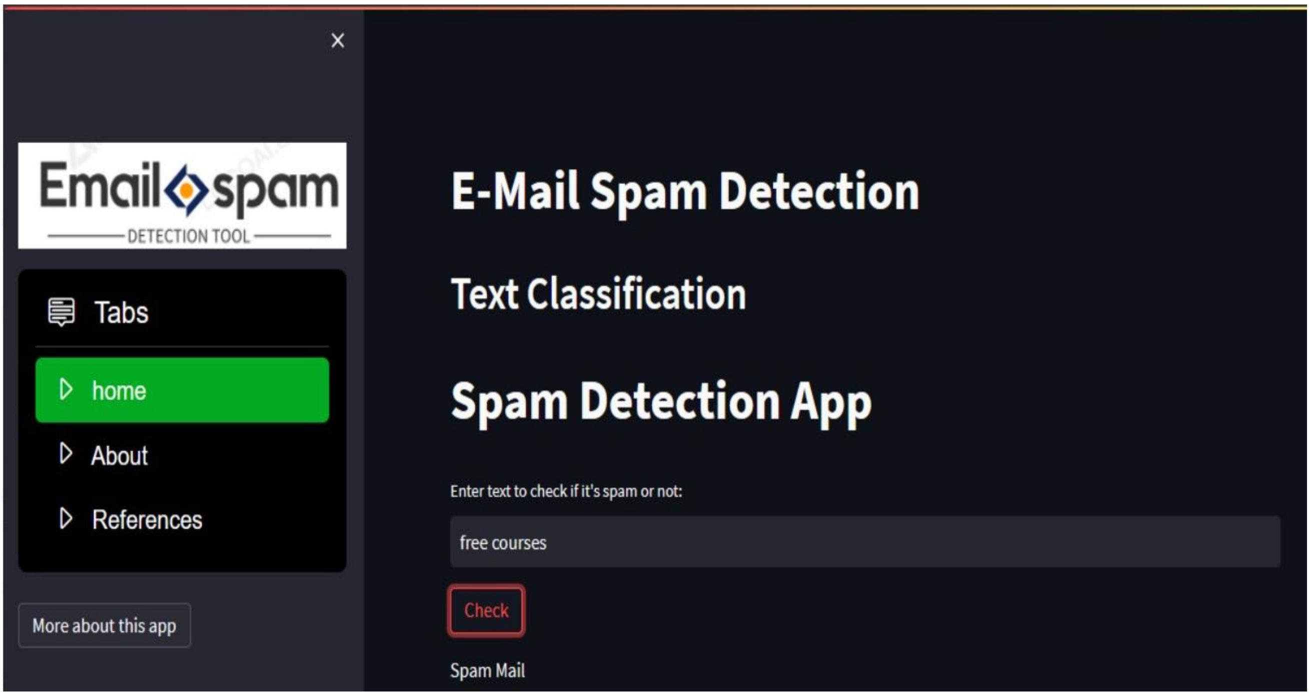
5.1. [1] This paper compares multiple classifiers including KNN, Decision Tree, Naïve Bayes, Random Forest, SVM, LDA.
The Random Forest classification algorithm applied on relevant features produced more than 99% accuracy in spam detection. This classifier is also tested with test dataset which gives accurate results than other classifiers for this spam dataset.

5.2. [2] The comparison has been done among different machine learning classifiers (such as Bayesian, Naïve Bayes, SVM, decision tree, Bayesian with Adaboost, Naïve Bayes with Adaboost). The concerned classifiers are tested and evaluated on metric (such as F-measure (accuracy), False Positive Rate, and training time). It has been found that SVM is the best classifier to be used. It has the high accuracy and the low false positive rate.

5.3. [3] This paper evaluates different machine learning classifiers like Naive Bayes, SVM, KNN, Bagging and Boosting (Adaboost), and Ensemble Classifiers. Different accuracy measures like Accuracy Score, F measure, Recall, Precision, Support and ROC are used. The preliminary result shows that Ensemble Classifier with a voting mechanism is the best to be used. It gives the minimum false positive rate and high accuracy.

5.4. [4] This paper presents a combining classifiers approach where the main objective is to combine individual decisions of the good classifiers for utmost classification outcome in spam classification domain. In this context, three different classifiers have been selected i.e. "Boosted Bayesian", "Boosted Naïve Bayes and Support Vector Machine (SVM). Results show the best results of novel combining classifier approach in compression with individual classifiers compared in terms of good performance accuracy and low false positives.

6. Results



E-Mail Spam Detection

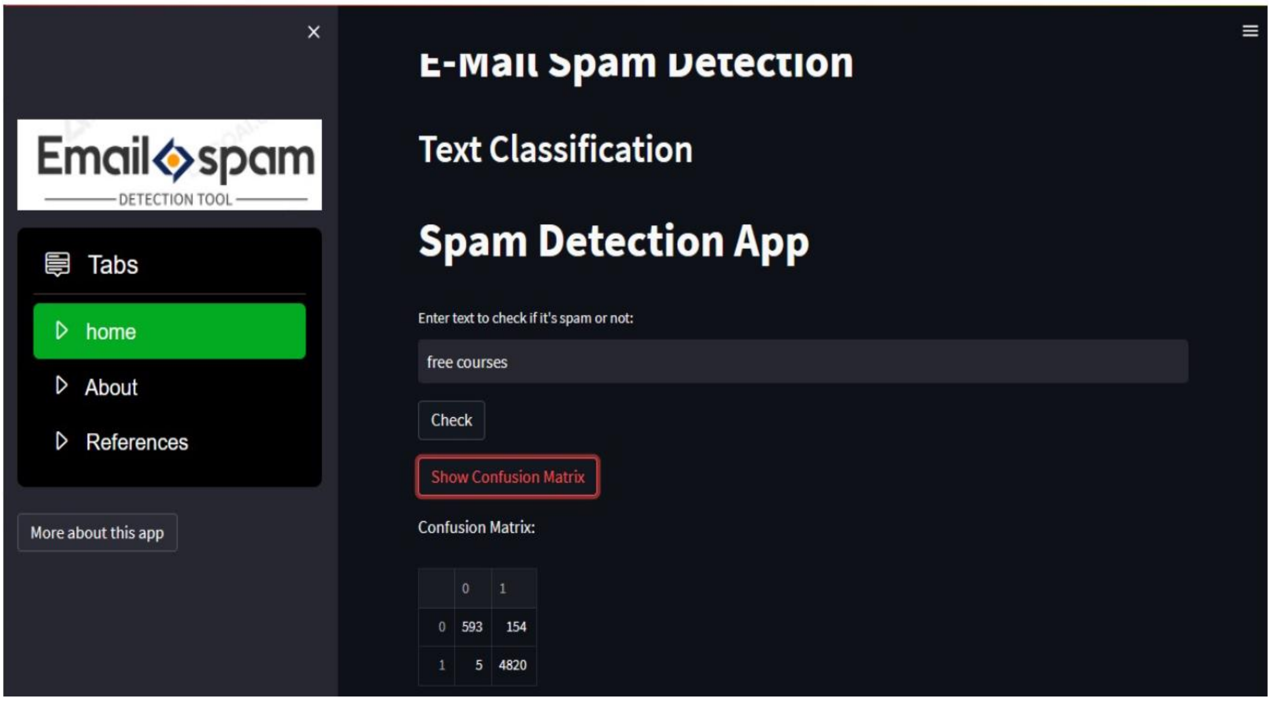
Text Classification

Spam Detection App

Enter text to check if it's spam or not:

free courses

Spam Mail



E-Mail Spam Detection

Text Classification

Spam Detection App

Enter text to check if it's spam or not:

free courses

Confusion Matrix:

	0	1
0	593	154
1	5	4820

7. SWOT Analysis

Strengths:

- **Efficiency:** Logistic regression is known for its fast training and implementation, making it suitable for real-time spam filtering.
- **Interpretability:** The model provides insights into the features most indicative of spam, aiding in understanding spam tactics and adapting to new ones.
- **Accuracy:** Logistic regression can achieve high accuracy in spam filtering, minimizing misclassifications.
- **Feature Extraction with NLP:** Utilizing NLP techniques allows for a more nuanced understanding of email content, capturing diverse spam indicators beyond just keywords.

Weaknesses:

- **Data Dependence:** The model's performance relies heavily on the quality and size of the training data. Insufficient or imbalanced data (too much spam or legitimate emails) can lead to poor performance.
- **Feature Engineering:** Selecting the most relevant features for spam classification requires expertise and can impact the model's effectiveness.
- **Potential for Bias:** Biases present in the training data can be reflected in the model's predictions. Careful data selection and monitoring are crucial.
- **Limited to Binary Classification:** Logistic regression classifies emails as spam or not-spam. It cannot categorize spam into sub-types (e.g., phishing, marketing).

Opportunities:

- **Integration with Email Clients:** The classifier can be integrated with email clients to provide real-time spam filtering directly within the inbox.
- **Adaptability:** The model can be continuously improved by incorporating new features, training on evolving spam data, and staying updated with the latest spam tactics.
- **Customization:** The system can be customized for individual users by allowing them to define specific filters based on their preferences.
- **Cloud Deployment:** The classifier can be deployed in the cloud for scalability and accessibility from any device.

Threats:

- **Evolving Spam Techniques:** Spammers continuously develop new tactics to bypass filters. The model needs to be adaptable to stay effective.
- **Computational Resources:** Training and deploying the model might require significant computational resources, especially for large datasets.
- **Privacy Concerns:** Data privacy considerations are crucial when handling email content. User trust and transparency are essential.
- **Alternative Spam Filtering Techniques:** Other spam filtering techniques, like content filtering or heuristic rules, might pose competition.

8. References

[1] Kumar, R. K., Poonkuzhali, G., & Sudhakar, P. (2012, March). Comparative study on email spam classifier using data mining techniques. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, pp. 14-16). Newswood Limited, Hong Kong.

[2] Trivedi, S. K. (2016, September). A study of machine learning classifiers for spam detection. In *2016 4th international symposium on computational and business intelligence (ISCBI)* (pp. 176-180). IEEE.

[3] Suryawanshi, S., Goswami, A., & Patil, P. (2019, December). Email spam detection: an empirical comparative study of different ml and ensemble classifiers. In *2019 IEEE 9th International Conference on Advanced Computing (IACC)* (pp. 69-74). IEEE.

[4] Trivedi, S. K., & Dey, S. (2016, March). A combining classifiers approach for detecting email spams. In *2016 30th international conference on advanced information networking and applications workshops (WAINA)* (pp. 355-360). IEEE.