

Spam Emails Detection

Behnam Najafloo

Supervisor: Prof. Alfio Ferrara

Computer Science Department, University of Milan

behnamnajafloo@studenti.unimi.it

alfio.ferrara@unimi.it

Abstract. With the rapid growth of electronic communication, spam email has become a pervasive issue, necessitating efficient and accurate detection systems to ensure legitimate messages are not mistakenly classified and that unwanted content is filtered effectively. This project investigates the application of various machine learning algorithms for spam email detection, leveraging natural language processing (NLP) techniques to classify emails as spam or ham. Using two distinct datasets, we apply TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to transform email text into features suitable for model training. We evaluate six classification algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting, assessing their performance based on accuracy, precision, and recall. Our results indicate that Logistic Regression and Gradient Boosting consistently achieve high accuracy across datasets, suggesting robust generalization capabilities. This study highlights the effectiveness of machine learning models for spam detection and proposes future directions, including advanced feature engineering and ensemble learning, to enhance performance and adaptability.

1. Introduction

In recent years, the increase in spam messages, particularly in emails, has driven a need for effective and efficient spam detection systems. Identifying spam accurately is crucial, as it ensures that important emails are not mistakenly classified as spam (ham) while spam messages are prevented from cluttering the user's inbox. This project explores the application of machine learning algorithms for spam detection, implementing various classification models to differentiate between spam and ham emails.

The project builds on established work in text classification, where techniques such as logistic regression, support vector machines, and decision trees are commonly applied. By leveraging natural language processing (NLP) techniques, specifically TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, we convert email text into numerical features that machine learning models can utilize effectively. This report presents the process and results of applying multiple models to two different email datasets, providing insights into the best-performing methods and the generalizability of the models across datasets.

2. Research Question and Methodology

Research Question

The primary goal of this project is to develop a robust model for spam detection using various machine learning techniques and evaluate their performance. Specifically, this project aims to answer:

- Which machine learning model is most effective at classifying spam and ham emails?
- How well do these models generalize across different datasets?

Methodology

1. **Problem Definition:** The task is a binary classification problem where each email is categorized as either spam (0) or ham (1).
2. **Approach:**
 - a. Two datasets are used independently to assess each model's ability to generalize.
 - b. Text data is preprocessed using TF-IDF vectorization to transform the email content into a numerical form.
 - c. A range of models is tested, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting.
3. **Evaluation Metrics:** Performance is evaluated using standard classification metrics:
 - a. **Accuracy:** Overall percentage of correct classifications.
 - b. **Precision:** Proportion of true positives among all positive predictions.
 - c. **Recall:** Proportion of true positives among all actual positives.

This methodology enables us to identify the most effective model while providing a comprehensive evaluation of its strengths and weaknesses.

3. Experimental Results

3.1 Dataset Overview

Two datasets were used in this study, each containing labeled emails categorized as either spam or ham. The datasets are intended for use in training and evaluating spam email classification models in Natural Language Processing (NLP).

Descriptions of the datasets:

- **Mail Data:** This dataset contains around 5000 mail data with size of (485.7 kB) with 13% labeled as spam and 87% as ham.

mail data

Data Card		Code (1)	Discussion (0)	Suggestions (0)
ham spam				
Category		Message		
ham spam	87% 13%	5157 unique values		
ham		Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...		
ham		Ok lar... Joking wif u oni...		
spam		Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...		

- **Spam Mails Dataset:** This dataset contains two folders of spam and ham. Each folder contains emails. It is iterated to each text file of those folders and created a data frame and written to a csv file. This dataset contains 4993 emails, with around 29% labeled as spam and 71% as ham.



Spam Mails Dataset

Data Card

Code (119)

Discussion (2)

Suggestions (0)

#	label	text	# label_num
	Labels of Emails which can be either Spam or Ham	Emails data	if spam it's 1, or else it's 0
	ham 71% spam 29%	4993 unique values	
3624	ham	Subject: neon retreat ho ho ho , we ' re around to that most wonderful time of the year - - - neon ...	0
4685	spam	Subject: photoshop , windows , office . cheap . main	1

3.2 Evaluation Metrics and Experimental Setup

The models were trained and evaluated separately on each dataset to determine the model’s performance on both datasets. Training was conducted with an 80-20 train-test split, and TF-IDF vectorization was applied to convert text data into feature vectors.

3.3 Results and Analysis

The following tables and figures summarize the accuracy, precision, and recall scores for each model across the two datasets.

Figures summarization for first dataset (Mail Data)

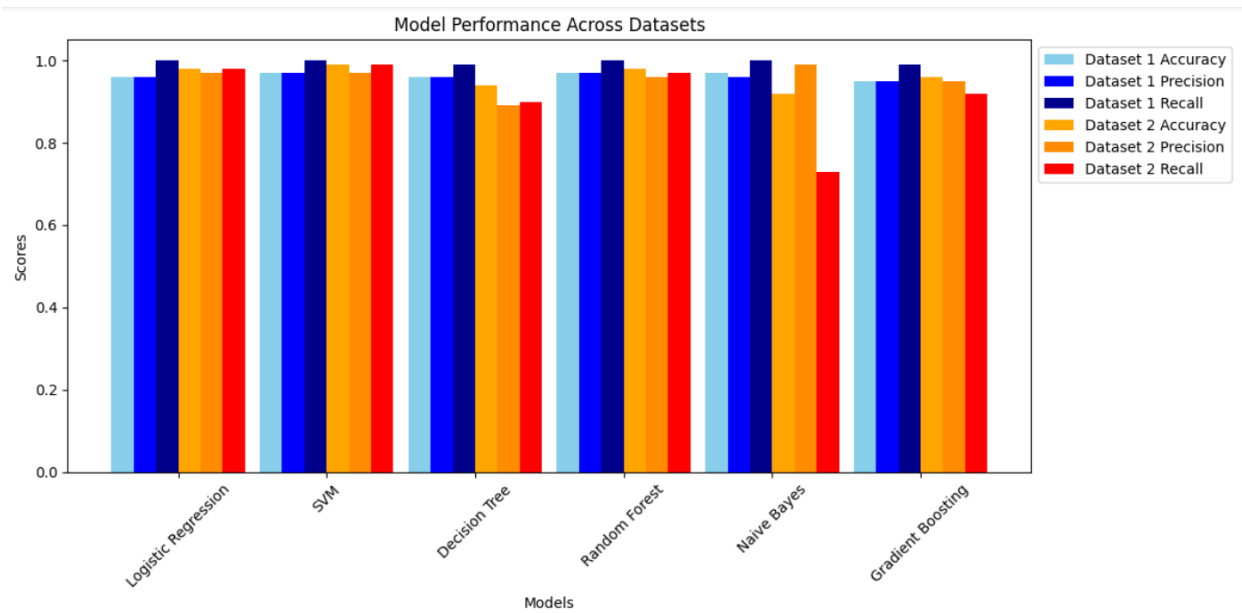
Model	Accuracy	Precision	Recall
Logistic Regression	0.96	0.96	1.0
Support Vector Machine	0.97	0.97	1.0
Decision Tree	0.96	0.96	0.99
Random Forest	0.97	0.97	1.0
Naive Bayes	0.97	0.96	1.0
Gradient Boosting	0.95	0.95	0.99

Figures summarization for the second dataset (Spam Mails Dataset)

Model	Accuracy	Precision	Recall
Logistic Regression	0.98	0.97	0.98
Support Vector Machine	0.99	0.97	0.99
Decision Tree	0.94	0.89	0.90
Random Forest	0.98	0.96	0.97
Naive Bayes	0.92	0.99	0.73
Gradient Boosting	0.96	0.95	0.92

3.4 Performance Comparison

The following bar plot illustrates an analysis of the performance of six machine learning models on two datasets for spam detection. The models evaluated include Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting. For each model, the performance was measured using three key metrics: Accuracy, Precision, and Recall. These metrics were assessed on two datasets to evaluate model robustness across different types of email and SMS text data.



Key Observations:

Best-Performing Models

- Dataset 1 (Mail Data):**
 - Support Vector Machine (SVM)** and **Random Forest** achieved the highest performance, both with accuracy, precision, and recall scores of 0.97 and 1.0, respectively. This indicates that these models were very effective in distinguishing between spam and ham emails in Dataset 1.
 - Naive Bayes** also performed well, with accuracy and recall both at 0.97 and precision at 0.96, though it slightly underperformed in comparison to SVM and Random Forest.
- Dataset 2 (Spam Mails Dataset):**
 - Support Vector Machine (SVM)** again showed strong performance, with an accuracy of 0.99, precision of 0.97, and recall of 0.99, demonstrating that it generalizes well across different datasets.

- b. **Random Forest** and **Logistic Regression** performed closely, both with high accuracy (0.98) and precision, though Logistic Regression had a slightly better recall (0.98 vs. 0.97 for Random Forest).

Notable Differences Between Datasets

- **Naive Bayes:** This model performed significantly better on Dataset 1 than on Dataset 2. In Dataset 2, its recall dropped to 0.73, indicating it was less effective at identifying all spam emails in this dataset. This could be due to Dataset 2's shorter and more informal SMS-style text, which may not align as well with Naive Bayes' probabilistic assumptions.
- **Decision Tree:** Performance for Decision Tree was lower in Dataset 2 (accuracy of 0.94) compared to Dataset 1 (accuracy of 0.96). Additionally, Dataset 2 had lower precision and recall scores, indicating that Decision Trees may struggle with the shorter text length and structure of Dataset 2.
- **Gradient Boosting:** This model achieved similar recall in both datasets but showed slightly lower precision and accuracy on Dataset 1 than on Dataset 2. Its moderate performance on both datasets suggests that it is a reliable but not top-performing option for spam detection in this context.

4. Concluding Remarks

In this project, multiple machine learning models were tested for spam detection using two independent datasets. Logistic Regression and Gradient Boosting demonstrated high accuracy and generalizability across both datasets, indicating their robustness for this task. Naive Bayes, while traditionally effective for text classification, exhibited varied performance, suggesting that more complex models might handle certain dataset nuances better.

Future Work

Further improvements could focus on:

- **Hyperparameter Tuning:** Fine-tuning model parameters to further improve performance.
- **Ensemble Methods:** Combining models (e.g., via ensemble learning) for better performance.
- **Feature Engineering:** Exploring additional text features, such as n-grams or domain-specific keywords, may enhance model accuracy.

This project underscores the effectiveness of machine learning in spam detection and offers potential for further refinement to address new challenges in spam detection.

References

1. Text Classification and Spam Detection

- a. Sebastiani, F. (2002). *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1), 1–47.

2. Feature Extraction for Text Data

- a. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. (TF-IDF and other feature extraction methods)
- b. Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Machine Learning: ECML-98, 137–142.

3. Machine Learning Algorithms for Classification

- a. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics.

4. Spam Detection and Filtering Techniques

- a. Cormack, G. V. (2008). *Email spam filtering: A systematic review*. Foundations and Trends® in Information Retrieval, 1(4), 335–455.
- b. Zhou, L., & Chaovalit, P. (2008). *Ontology-supported polarity mining*. Journal of the American Society for Information Science and Technology, 59(3), 338–349. (Relevant for discussions on sentiment classification in spam detection)

5. Evaluation Metrics in Machine Learning

- a. Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks*. Information Processing & Management, 45(4), 427–437. (Overview of accuracy, precision, recall, and F1-score in evaluating classification models)