# Email Spam Detection

## Comprehensive Project Documentation

Generated: 2025-06-28 23:30

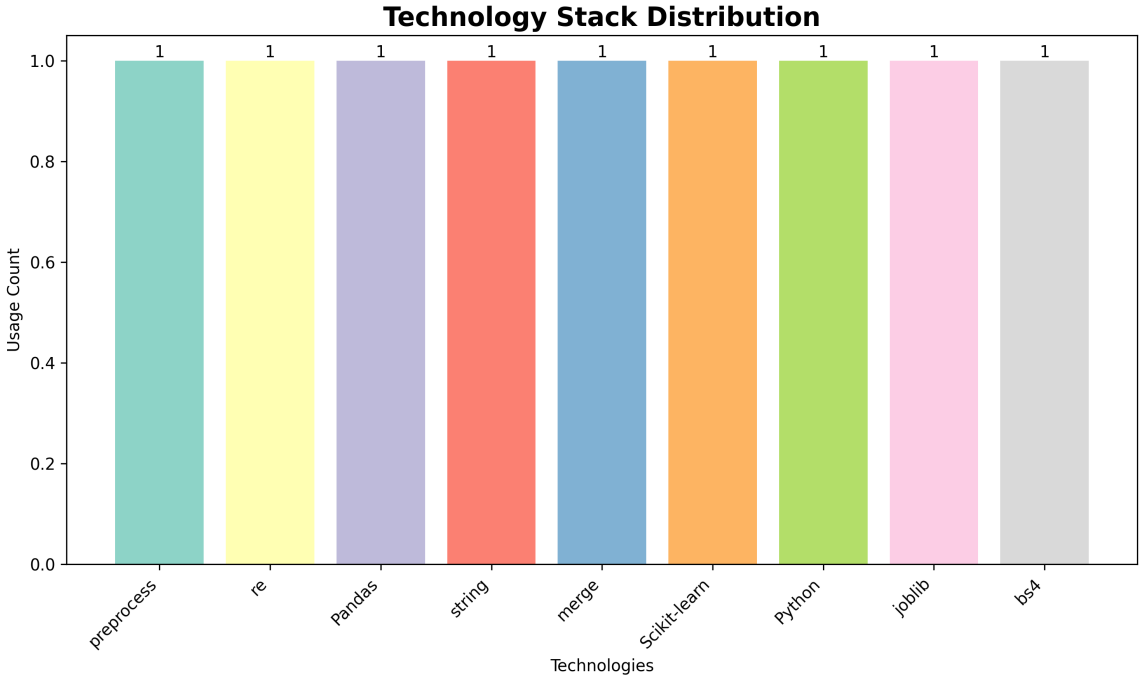
# Table of Contents

# 1. Project Overview

{'merge.py': "This Python script uses the pandas library to merge four CSV files ('combined_data.csv', 'emails.csv', 'spam_assassin.csv', 'spam.csv') containing text data and spam/ham labels. It cleans and renames columns to create a single unified dataframe named `mergedDf` with 'text' and 'target' columns, then prints the column names and dataset size before returning the merged dataframe.\n", 'predict_custom.py': 'This Python script uses a pre-trained model (`spam_model.pkl`) and vectorizer (`tfidf_vectorizer.pkl`) to classify user-input email text as spam or not spam. It preprocesses the input (removes HTML, lowers case, cleans text), vectorizes it, makes a prediction, and prints the result.\n', 'preprocess.py': "This Python file preprocesses text data for machine learning. It uses BeautifulSoup to remove HTML tags, regular expressions to remove URLs, email addresses, numbers, and punctuation, and then applies TF-IDF vectorization to convert the cleaned text into numerical features suitable for a model. The `preprocess` function encapsulates these steps, returning the feature matrix (X), target variable (df['target']), and the trained TF-IDF vectorizer.\n", 'train.py': "This Python script preprocesses text data, trains a logistic regression model to classify it (likely spam detection), evaluates the model's performance using accuracy and a classification report, and then saves both the trained model and the text vectorizer using joblib for later use.\n"}

| Metric | Value |
|---|---|
| Total Files | 7 |
| Total Lines of Code | 2313 |
| Total Size | 100.06 KB |
| Technologies Used | 9 |

## 2. Technology Stack Analysis




| Technology | Category | Usage |
|---|---|---|
| Pandas | Data Science | Active |
| Python | Backend | Active |
| Scikit-learn | Data Science | Active |
| bs4 | Other | Active |
| joblib | Other | Active |
| merge | Other | Active |
| preprocess | Other | Active |
| re | Other | Active |
| string | Other | Active |

# 3. Architecture Overview

**Project Architecture Overview**

## 4. Detailed File Analysis

### *File: Email Spam Detection\merge.py*

| Property | Value |
|---|---|
| Type | Python |
| Lines of Code | 20 |
| Size | 837 bytes |
| Functions | 1 |
| Classes | 0 |

### *File: Email Spam Detection\predict_custom.py*

| Property | Value |
|---|---|
| Type | Python |
| Lines of Code | 16 |
| Size | 520 bytes |
| Functions | 0 |
| Classes | 0 |

### *File: Email Spam Detection\preprocess.py*

| Property | Value |
|---|---|
| Type | Python |
| Lines of Code | 33 |
| Size | 965 bytes |
| Functions | 4 |
| Classes | 0 |

### *File: Email Spam Detection\README.md*

| Property | Value |
|---|---|
| Type | Markdown |
| Lines of Code | 186 |
| Size | 3622 bytes |
| Functions | 0 |
| Classes | 0 |

### *File: Email Spam Detection\spam_model.pkl*

| Property | Value |
|---|---|

| | |
|---|---|
| Type | Unknown |
| Lines of Code | 510 |
| Size | 13843 bytes |
| Functions | 0 |
| Classes | 0 |

### *File: Email Spam Detection\tfidf_vectorizer.pkl*

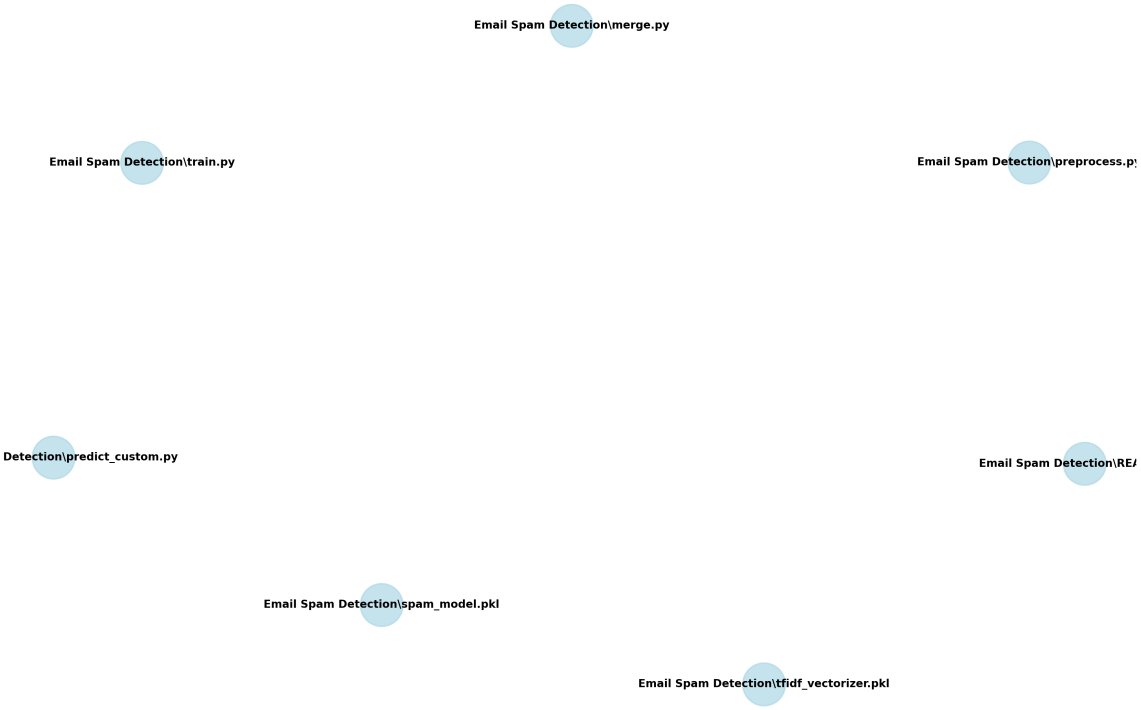
| Property | Value |
|---|---|
| Type | Unknown |
| Lines of Code | 1522 |
| Size | 82012 bytes |
| Functions | 0 |
| Classes | 0 |

### *File: Email Spam Detection\train.py*

| Property | Value |
|---|---|
| Type | Python |
| Lines of Code | 26 |
| Size | 667 bytes |
| Functions | 0 |
| Classes | 0 |

# 5. Dependency Analysis

## File Dependency Graph



# 6. Code Statistics