



Sri Lanka Institute of Information Technology

# Email Spam Classification Using Machine Learning

## The patent

### **Individual Assignment**

Secure Software Systems - IE3042

IT20618872- Thisitha K.L.D

Date of submission  
24-05-2023

## Title - Email Spam Classification Using Machine Learning

It should Before proceed with the patent application, conduct a thorough search to ensure that your Email Spam Classification system is unique and not already patented by someone else.

I include a new data set and it is updated application.

1. Document the Invention:
2. Detailed documentation of Email Spam Classification system-

Scope: The Email Spam Classification System is designed to classify incoming emails as spam or non-spam (ham) using machine learning techniques. This documentation covers the system's architecture, components, data collection, preprocessing, feature extraction, machine learning models, model training, evaluation, deployment, system performance, monitoring, maintenance, and updates.

Audience: This documentation is intended for developers, data scientists, and technical stakeholders involved in the design, development, deployment, and maintenance of the Email Spam Classification System.

- System Overview

1.1 Architecture: The Email Spam Classification System follows a modular architecture comprising several components, including data collection, preprocessing, feature extraction, machine learning models, and deployment.

1.2 Components:

- Data Collection: Collects email data for training and testing purposes.
- Data Preprocessing: Cleans and prepares the collected email data for further processing.
- Feature Extraction: Extracts relevant features from the preprocessed email data.
- Machine Learning Models: Trains and utilizes machine learning models to classify emails as spam or ham.
- Deployment: Integrates the system with email systems for real-time email classification.

1.3 Workflow: The system workflow involves the following steps:

1. Data collection
2. Data preprocessing
3. Feature extraction
4. Model training
5. Model evaluation
6. Deployment
7. System performance monitoring
8. Maintenance and updates
9. Data Collection and Preprocessing

Data Collection: The system collects email data from various sources, including public datasets or user-provided labeled email data.

- Machine Learning Models

1.1 Naive Bayes: Naive Bayes is a probabilistic classification algorithm that calculates the probability of an email being spam or ham based on the occurrence of words in the email.

1.2 Support Vector Machines (SVM): SVM is a supervised machine learning algorithm that constructs hyperplanes to separate spam and ham emails in a high-dimensional feature space.

1.3 Random Forest: Random Forest is an ensemble learning method that combines multiple decision trees to classify emails as spam or ham based on various features.

## 2. Model Training

2.1 Data Split: The collected email data is divided into training and testing sets to evaluate the model's performance.

2.2 Model Selection: Different machine learning models are evaluated using suitable evaluation metrics, and the best-performing model is selected for further training and deployment.

2.3 Model Training Process: The selected model is trained using the training dataset and the chosen feature extraction technique. Model parameters are optimized using techniques such as grid search or cross-validation.

## 3. Model Evaluation

3.1, and receiver operating characteristic (ROC) curve.

3.2 Cross-Validation: Cross-validation is performed to validate the model's performance by dividing the training dataset into multiple subsets and iteratively training and evaluating the model on different subsets.

#### 4. Maintenance and Updates


4.1 Retraining: The system should be periodically retrained with new labeled data to improve its classification accuracy and adapt to changing spam patterns.

4.2 Model Updating: The trained model can be updated by incorporating new features, exploring different algorithms, or retraining on a larger and more diverse dataset.

features and advantages-

1. **Real-Time Classification:** The patented system offers real-time email classification, ensuring incoming emails are classified as spam or non-spam immediately upon arrival. This provides users with instant protection against spam messages.
  2. **High Accuracy:** The patented system achieves high accuracy in spam classification, significantly reducing false positives (legitimate emails classified as spam) and false negatives (spam emails classified as legitimate). This enhances user experience by minimizing the risk of important emails being missed or legitimate emails being marked as spam.
  3. **Low Resource Consumption:** The patented system is designed to optimize resource utilization, requiring minimal computational power and memory resources. This allows for efficient deployment on various hardware platforms and reduces operational costs.
  4. **Scalability:** The patented system is scalable, capable of handling large volumes of emails without compromising performance. It can easily accommodate growing email traffic, making it suitable for enterprise-level applications.
- ❖ **Consult with a Patent Attorney:** It is highly recommended to seek guidance from a qualified patent attorney or patent agent experienced in software patents. They can help you navigate the patent application process and ensure that your application adheres to the necessary legal requirements.

This is the example when pass the patent the company issued license



datamagic2020/Fake\_News\_Detection is licensed under the  
**MIT License**

A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

**Permissions**

- ✓ Commercial use
- ✓ Modification
- ✓ Distribution
- ✓ Private use

**Limitations**

- ✗ Liability
- ✗ Warranty

**Conditions**

- ① License and copyright notice

This is not legal advice. [Learn more about repository licenses.](#)

```
1 MIT License
2
3 Copyright (c) 2021 Data Magic
4
5 Permission is hereby granted, free of charge, to any person obtaining a copy
6 of this software and associated documentation files (the "Software"), to deal
7 in the Software without restriction, including without limitation the rights
8 to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
9 copies of the Software, and to permit persons to whom the Software is
10 furnished to do so, subject to the following conditions:
11
12 The above copyright notice and this permission notice shall be included in all
13 copies or substantial portions of the Software.
14
15 THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
16 IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
17 FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
18 AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
19 LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
20 OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
21 SOFTWARE.
```