**Open Program Proposal**

**Spam Email Classifier**

**Kaloyan Rakov**

****

****

**Table of Contents:**

1. **Introduction**
   1. **What?**
   2. **Why?**
   3. **Who?**
   4. **When?**
   5. **How?**
2. **Analytic Approach**
   1. **Defining the target variable(s)**
3. **Domain Understanding**
4. **Data Requirements**
   1. **Data Sources**
   2. **Data Legality and Ethics**
5. **Introduction**
   1. **What?**

This project involves building a Spam Email Classifier — a machine learning model that can automatically detect whether an email is spam or not. The system will analyze the content of emails and classify them as either spam or actual mail, using classification algorithms.

* 1. **Why?**

Spam emails are a daily annoyance for most users and can also pose serious security risks, such as attacks, scams, or malware. While email providers offer basic spam filters, they are not always perfect. This project explores how machine learning can be used to improve spam detection by learning from real email data. It is also a practical and well-known classification task, which makes it ideal for learning and demonstrating skills in text processing, data cleaning, and supervised learning.

* 1. **Who?**

This tool would be useful by:

* Everyday email users who want fewer spam emails reaching their inbox.
* Organizations looking to protect employees from attacks.
* Email service providers (like Gmail or Outlook) that want to improve their spam filtering.
  1. **When?**

The project will be developed over the course of weeks 17 and 18.

* 1. **How?**

**Data Collection:** Use a publicly available dataset like the Enron Spam dataset or the SMS Spam Collection.

**Preprocessing:** Clean the text data by removing punctuation, converting to lowercase, removing useless information, etc.

**Feature Extraction:** Convert text into numerical values.

**Model Training**: Train a classifier (e.g., Random Forest) on the labeled dataset.

**Model Boosting**: Using tools like AdaBoost to ensure better quality model.

**Evaluation**: Use metrics like accuracy and precision to assess performance.

1. **Analytic Approach**
   1. **Defining the target variable(s)**

The target variable in this classification problem is the email label, which indicates whether a given email is spam or not. This variable can be represented as binary as:

* 1 = Spam
* 0 = Not Spam

This target label is assigned to each email in the dataset and is used during training to teach the model to distinguish between spam and non-spam messages. The classifier will learn to recognize patterns, keywords, and characteristics common to spam emails, and use this knowledge to predict the correct label for new, unseen emails.

1. **Domain Understanding**

To ensure a broad understanding of the domain, comprehensive research will be conducted on the nature of spam emails, how they are detected, and the impact they have on users and organizations. This research will be centered around expert insight and guided by a number of relevant research questions, such as:

* What are the most common characteristics of spam emails?
* How do spammers/ attackers attempt to bypass spam filters?
* What are the legal and security implications of spam?
* How do email service providers classify and filter emails?
* What are the current limitations of spam detection systems?

**Research Methods:**

* Literature review of existing spam detection models and techniques.
* Analysis of public email datasets.
* Review of cybersecurity guidelines and spam filtering systems used by major email providers.

While historical email data will form the foundation of the technical model, this research ensures that the classifier is developed with an understanding of real-world email systems, spam tactics, and user impact.

1. **Data Requirements**

To train and evaluate the spam classifier, the following data attributes are required:

* **Email Text** (body content): to detect common spam phrases or patterns
* **Subject Line**: often contains promotional or misleading content in spam
* **Sender Address/Domain**: some addresses are associated with spam networks
* **Presence of URLs or Attachments**: often found in phishing or advertising spam
* **Email Metadata** (e.g., time sent, email length, formatting)
* **Spam Label**: a binary classification label indicating spam (1) or not spam (0)
  1. **Data Sources**

Several public datasets are available for building and testing spam classifiers, including:

* SpamAssassin Public Corpus – A widely used benchmark dataset containing both spam and legitimate emails
* Enron Email Dataset – A large set of real business emails (requires labeling for spam detection)
* Kaggle Datasets – Various user-contributed spam datasets available for machine learning tasks
* Apache SpamAssassin Project – Often used for filtering rule validation
* Data Scraping – We can get the data ourselves.
  1. **Data Legality and Ethics**

Working with email data involves handling sensitive and potentially personal information. Ethical and legal considerations include:

* **Anonymization**: Removing personal identifiers such as names, addresses, or private correspondence
* **Compliance privacy regulations**: Ensuring data usage aligns with legal frameworks
* **Use of Public Datasets**: Only datasets with public, non-restricted licenses will be used
* **No Personal Email Use**: Under no circumstances will personal or private email data be collected or processed