School of Computing

UNIVERSITY OF LEEDS

COMP5200M Scoping and Planning Document

Student Name: Long Yue
Programme of Study: MSc Advanced Computer Science (Data Analytics)
Provisional Title of Project: Development and Evaluation of Machine Learning-Based Spam Filtering Models: An Improvement Study of Algorithm Performance
Name of External Company (if any)**:**
Supervisor Name: Sebastian Ordyniak
Type of Project: Exploratory Software
NOTE to student*: ensure you have discussed the content with the supervisor before submitting this document to Minerva. Submit an* ***electronic version*** *of this report in pdf via the appropriate link in Minerva; with filename of the format <surname><year>-SP (e.g. SMITH15-SP.pdf).*

Signature of Student: Long Yue	**Date:** 2023/03/23

Contents

1. Background Research for the project

1.1 Context

Spam emails have been a pervasive problem since the inception of email communication. Despite various measures taken to combat spam, such as blacklisting and content filtering, the sheer volume and complexity of spam messages continue to be a challenge. Machine learning offers a promising solution for spam filtering, as it can learn to distinguish between spam and legitimate emails based on patterns and characteristics in the data.

Python is a popular programming language for implementing machine learning algorithms, thanks to its ease of use and the availability of various third-party libraries. Some of the commonly used libraries for machine learning in Python include scikit-learn, TensorFlow, and Keras. These libraries provide a wide range of algorithms and tools for tasks such as data preprocessing, feature selection, and model training and evaluation.

The motivation for this project is to develop and evaluate a machine learning-based spam filtering model using Python and relevant libraries. The goal is to compare the performance of different machine learning algorithms and identify the most effective approach for spam filtering. The project aims to contribute to the existing research on spam filtering and provide a practical solution for individuals and organizations that face spam-related issues.

1.2 Problem statement

The problem addressed in this project is the high volume and complexity of spam emails, which can negatively impact individuals and organizations in various ways, such as reduced productivity, compromised security, and increased risk of phishing attacks. Traditional methods of spam filtering, such as content-based filtering and blacklisting, have limitations in terms of accuracy and effectiveness, as spammers constantly evolve their tactics to evade detection for example: third-party ghost writer advertisement, this kind of service is not academic or legal, it has become popular in recent years and cannot be filtered by traditional email filtering models.

The aim of this project is to develop and evaluate a machine learning-based spam filtering model using Python and relevant libraries. The main research question is: How effective are different machine learning algorithms in spam filtering, and which approach provides the best performance?

To address this question, the following hypotheses are proposed:

H1: Machine learning-based spam filtering models can achieve higher accuracy than traditional spam filtering methods. H2: The choice of machine learning algorithm significantly impacts the performance of the spam filtering model. H3: Feature selection and data

preprocessing techniques can improve the performance of machine learning-based spam filtering models.

By refining the problem and formulating these hypotheses, this project aims to contribute to the existing research on spam filtering and provide practical solutions for individuals and organizations that face spam-related issues. The outcomes of this project can help improve the accuracy and efficiency of spam filtering and enhance the overall security and productivity of email communication specially to filter the new spam like ghost writer advertisement.

1.3 Possible solution

The solution proposed for this project is to develop and evaluate a machine learning-based spam filtering model using Python and relevant libraries. The development of the solution will involve the following steps:
Data collection and preprocessing: The first step will be to collect a representative dataset of spam and legitimate emails. The data will be preprocessed to extract relevant features and remove noise and irrelevant information. The preprocessing techniques may include tokenization, stemming, and stop word removal.

Feature selection and engineering: The next step will be to select and engineer relevant features that can distinguish between spam and legitimate emails. The feature selection techniques may include chi-squared test, mutual information, and correlation-based feature selection.

Model selection and training: Several machine learning algorithms will be evaluated for their performance in spam filtering, including Naive Bayes, Decision Trees, Random Forest, Support Vector Machines (SVM), and Deep Learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The model training will involve splitting the dataset into training and validation sets, and evaluating the performance of each algorithm using metrics such as accuracy, precision, recall, and F1-score.

Model evaluation and tuning: The best performing model(s) will be further evaluated and tuned to improve their performance. This may involve hyperparameter tuning, ensemble learning, and cross-validation techniques.

The development of the solution will involve the use of various modules and computing topics, including data manipulation and visualization libraries such as pandas and matplotlib, machine learning libraries such as scikit-learn, TensorFlow, and Keras, and deep learning frameworks such as PyTorch and TensorFlow. The flow path of the solution will involve a pipeline approach, where data will be loaded, preprocessed, and transformed before being used for model training and evaluation.

1.4 How to demonstrate the quality of the solution

The success of the proposed solution will be judged based on its ability to accurately classify spam and legitimate emails. To demonstrate the quality of the solution, the following metrics will be used:

Accuracy: The overall accuracy of the model in classifying spam and legitimate emails. This metric will be calculated as the ratio of correctly classified emails to the total number of emails in the test set.

Precision: The precision of the model in classifying spam emails. This metric will be calculated as the ratio of correctly classified spam emails to the total number of emails classified as spam.

Recall: The recall of the model in classifying spam emails. This metric will be calculated as the ratio of correctly classified spam emails to the total number of spam emails in the test set.

F1-score: The harmonic means of precision and recall. This metric provides a balanced measure of the model's performance in classifying spam and legitimate emails.

The proposed solution will be considered successful if it achieves high accuracy, precision, recall, and F1-score values on a held-out test set. Additionally, the performance of the proposed solution will be compared with state-of-the-art spam filtering models reported in the literature.

2. Scope for this project

2.1 Aim

Collecting and pre-processing the dataset: A dataset of spam and legitimate emails will be collected and preprocessed for training and testing the proposed model. The dataset will be cleaned, formatted, and feature-engineered to extract relevant information.

Exploratory data analysis: An exploratory data analysis will be conducted to gain insights into the dataset's characteristics and identify patterns and correlations.

Model development and evaluation: Several machine learning algorithms will be explored and evaluated for their performance in classifying spam and legitimate emails. The proposed solution will be iteratively developed and evaluated using techniques such as hyperparameter tuning, feature engineering, and model selection.

Model comparison: The proposed solution's performance will be compared with state-of-the-art spam filtering models reported in the literature.

Deployment: The final model will be deployed and tested on a held-out test set. The model's performance will be evaluated based on various metrics such as accuracy, precision, recall, and F1-score.

This task model judges whether the email is normal (ham) or spam email (spam) according to the text content contained in the email, to realize automatic spam filtering especially the newest spam will be filtered.

2.2 Objectives

The main objective of this project is to develop an effective spam filtering model using machine learning algorithms. The model should be able to accurately classify spam and legitimate emails with high precision, recall, and F1-score values. Additionally, the project aims to explore and evaluate different machine learning algorithms and techniques for feature engineering and model selection, and Improve algorithm accuracy, and train to adapt to the latest spam.

After the new model training is completed, it will be compared with the filtering models used in the past, and the model will be further optimized through comparison and evaluation, and the principles of machine learning used in this project will be compared and discussed.

2.3 Deliverables

The following deliverables will be produced as part of this project:
A cleaned and formatted dataset of spam and legitimate emails for training and testing the proposed model.

A report on the exploratory data analysis conducted on the dataset.

A spam filtering model developed using machine learning algorithms.

A report on the performance evaluation of the proposed model.

A comparison of the proposed solution's performance with state-of-the-art spam filtering models reported in the literature.

A final report on the project, summarizing the objectives, scope, methodology, results, and conclusions.

An accurate, reproducible and reuse filtering model. And machine learning filtering models are studied and discussed through evaluation and comparative studies. The model through data processing and massive training will have reliable mail filtering accuracy, and it is

expected that in the future, the interface and application port can be further designed and imported into the mail application to realize the mail filtering function for users. And a report discussing in-depth model design, training, comparison and implementation process

3. Project schedule

3.1 Methodology

The chosen approach for this project is the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining). This methodology is a widely-used and recognized approach to data mining projects and provides a structured framework for the entire data mining process. The CRISP-DM methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In this project, we will adapt the CRISP-DM methodology as follows:

Business Understanding: Define the business problem and objectives, as well as identify the stakeholders and requirements.

Data Understanding: Collect and analyze the dataset to gain insights into its characteristics and identify patterns and correlations.

Data Preparation: Clean, format, and feature-engineer the dataset to extract relevant information and prepare it for modeling.

Modeling: Develop and evaluate several machine learning algorithms for classifying spam and legitimate emails.

Evaluation: Evaluate the performance of the proposed solution using various metrics such as accuracy, precision, recall, and F1-score.

Deployment: Deploy the final model and test it on a held-out test set to ensure its effectiveness.

3.2 Tasks, milestones and timeline

The project schedule will be presented using a Gantt chart, which outlines the tasks, milestones, and timeline for the project. The Gantt chart for this project is presented in Figure 1.

Figure 1: Gantt chart for the project schedule

The Gantt chart outlines the major tasks and milestones for the project, along with their start and end dates. The project will start with data collection and preprocessing, followed by exploratory data analysis, model development, performance evaluation, and deployment.

In addition, a Google questionnaire will be developed to ask people to evaluate the existing mail filtering model, and the feedback of receipt users. And learn to specify the corresponding model optimization direction, guide this project aim, and further collect some spam columns in the latest era background from users, such as ghost-writing advertisements, etc.

3.3 Risk assessment (if appropriate)

The following risks have been identified for this project:

Data availability: The availability and quality of the dataset may affect the performance of the proposed solution. Mitigating strategy: Plan for data collection and preprocessing early in the project timeline, and consider alternative datasets if necessary.

Time constraints: The project should be completed within the allocated time frame, which may limit the number of algorithms and techniques explored and evaluated. Mitigating strategy: Prioritize the most promising algorithms and techniques based on initial exploratory data analysis.

Performance requirements: The proposed solution should achieve high performance in classifying spam and legitimate emails while minimizing false positives and false negatives. Mitigating strategy: Conduct thorough performance evaluation and iteration of the solution to optimize performance.

In conclusion, this chapter has outlined the methodology, tasks, milestones, and timeline for the project, as well as identified potential risks and their mitigating strategies. The methodology will guide the order of the activities/tasks, and a Gantt chart has been presented to provide a visual representation of the project schedule. The risk assessment has identified potential risks and their mitigating strategies to ensure the project's success.

References

Appendix A. How ethical issues are addressed

The project will begin with training using existing open-source datasets on the Internet, and this part will have no ethical implications. In addition, I will assign Google questionnaire to conduct user project research and new spam collection. This part involves relevant user ethics privacy and security issues. I will sign the corresponding ethics setting list and negotiate with my tutor