National College of Ireland

Configuration Manual

MSc Research Project
MSc in Cybersecurity

Bejoy Asha Shajilal
Student ID: 22227067

School of Computing
National College of Ireland

Supervisor: Niall Heffernan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Bejoy Asha Shajilal		
Student ID:	22227067		
Programme:	MSc. Cybersecurity	**Year:**	2023-2024
Module:	MSc. Research Project		
Lecturer:	Niall Heffernen		
Submission Due Date:	16-09-2024		
Project Title:	Configuration Manual		
Word Count:	755 **Page Count:** 5		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Bejoy Asha Shajilal

Date: 16/09/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Bejoy Asha Shajilal
Student ID: 22227067

1 Section 1

This configuration manual provides step-by-step instructions to reproduce the phishing detection system described in the study. It covers the necessary software installations, dataset preparation, and code execution. By following this guide, you will be able to replicate the results and further experiment with the models and methods used in this project.

2 Section 2

To reproduce this project, you will need the following system specifications:

- Operating System: Windows, macOS, or Linux
- Processor: Intel i5 or equivalent
- Memory: 16 GB RAM (minimum)
- Storage: At least 20 GB of free disk space
- GPU: NVIDIA GPU with CUDA support (optional, but recommended for faster training).

3 Section 3

Ensure that the following software and libraries are installed:

Python: Version 3.8 or higher
IDE: Jupyter Notebook or any Python-compatible IDE (e.g., PyCharm, VS Code)
Package Manager: pip or conda (for managing Python packages)

Install the required Python libraries by running the following commands:

>> pip install numpy pandas scikit-learn transformers torch matplotlib seaborn jupyter

Key Libraries:

- numpy: For numerical computations
- pandas: For data manipulation and analysis
- scikit-learn: For machine learning algorithms and model evaluation
- transformers: For BERT model and tokenization
- torch: For working with PyTorch models (required by transformers)
- matplotlib and seaborn: For data visualization

4 Datasets

The project uses two datasets:

1. **Enron Corpus**: Used for legitimate emails.
2. **Figshare Curated Nigerian Dataset**: Used for phishing emails.

Downloading the Datasets:

- Enron Corpus: Available from Kaggle or the official Enron repository.(*Enron Email Dataset*, no date)
- Figshare Nigerian Dataset: Available from Figshare. ('Curated Dataset - Phishing Email', 2023)

Data Preparation:

1. Download the datasets and extract them to a directory of your choice.
2. Ensure that the datasets are structured as follows:
    - data/
        - enron/
            - legitimate_emails/
        - figshare/
            - phishing_emails/
3. Convert any non-UTF-8 encoded files to UTF-8 to avoid encoding issues
4. Run add_urls.py to extract urls from Nigerian_5.csv dataset
5. Run 'phishing_emails_all_features.py' to extract features from the phishing dataset generated in the previous set.
6. Run 'clean.py' on 'emails.csv' (enron dataset) and then 'enron_all_features.py' to extract features from enron dataset.
7. Combine required number of emails from both datasets generated in the previous steps to curate the final dataset.

5 Running the code

Open the Jupyter Notebook:

1. Launch Jupyter Notebook in the project directory:
2. Open the notebook file (final_analysis.ipynb) provided with this project.

Data Preprocessing:

1. The first step in the notebook involves loading and preprocessing the data. Ensure the file paths in the notebook match the location of your datasets.
2. Run the preprocessing cells to clean the data, generate BERT embeddings, and extract content-based features.

Model Training:

1. Proceed to the model training section. Here, you will train individual models (KNN, Decision Tree, Random Forest, SVM) using the preprocessed data.
2. Use the provided Grid Search and cross-validation scripts to optimize the models.

Ensemble Learning:

After training the individual models, run the ensemble learning cells to create and evaluate the Stacking and Soft Voting models.

Evaluation and Visualization:

1. Finally, execute the cells that generate evaluation metrics and visualizations such as confusion matrices, precision-recall curves, and AUC plots.
2. Review the outputs to compare the performance of the models.

References

‘Curated Dataset - Phishing Email’ (2023). figshare. Available at: https://doi.org/10.6084/m9.figshare.24899952.v2.

Enron Email Dataset (no date). Available at: https://www.cs.cmu.edu/~./enron/ (Accessed: 12 August 2024).