***Phishing-Email Detection with Supervised Machine Learning***

**PROJECT OVERVIEW**

Summary:

Email phishing is a major cyber threat that targets both individuals and organizations. These attacks trick people into giving away personal information or downloading malware by pretending to be from trustworthy sources. To combat this, machine learning can help spot phishing emails by recognizing patterns in their content (Abbasi et al., 2021; Alam et al., 2020; Boustead and Kugler, 2023; Butt et al., 2023)This project will build a supervised machine learning model to classify emails based on their content as either phishing or legitimate. We’ll start with a logistic regression classifier using TF-IDF for email vectorization. The project will also examine how n-gram character features and different sampling methods affect performance metrics like recall and F1-score. The aim is to create a reliable model that works well, even with tricky or imbalanced cases.

Research Questions:

1. Can we achieve 95 percent accuracy or more with a TF-IDF + Logistic Regression setup in distinguishing phishing from safe emails?
2. Does adding n-gram character features help improve recall for obfuscated phishing emails compared to just using word-level features?
3. How much do class-weighting or random under-sampling improve the F1-score for the less common phishing category?

Project Objectives:

1. Train a TF-IDF + Logistic Regression model and check how accurate it is at classifying emails.
2. Compare the effectiveness of word-level versus character-level n-gram features in recalling phishing emails.
3. Evaluate how different sampling techniques affect class imbalance using metrics like precision, recall, and F1-score.

References

Abbasi, M.A., Khan, A.Q., Mustafa, G., Abid, M., Khan, A.S., Ullah, N., 2021. Data-Driven Fault Diagnostics for Industrial Processes: An Application to Penicillin Fermentation Process. IEEE Access 9, 65977–65987. <https://doi.org/10.1109/ACCESS.2021.3076783>

Alam, M.N., Sarma, D., Lima, F.F., Saha, I., Ulfath, R.E., Hossain, S., 2020. Phishing attacks detection using machine learning approach. Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020 1173–1179. <https://doi.org/10.1109/ICSSIT48917.2020.9214225>

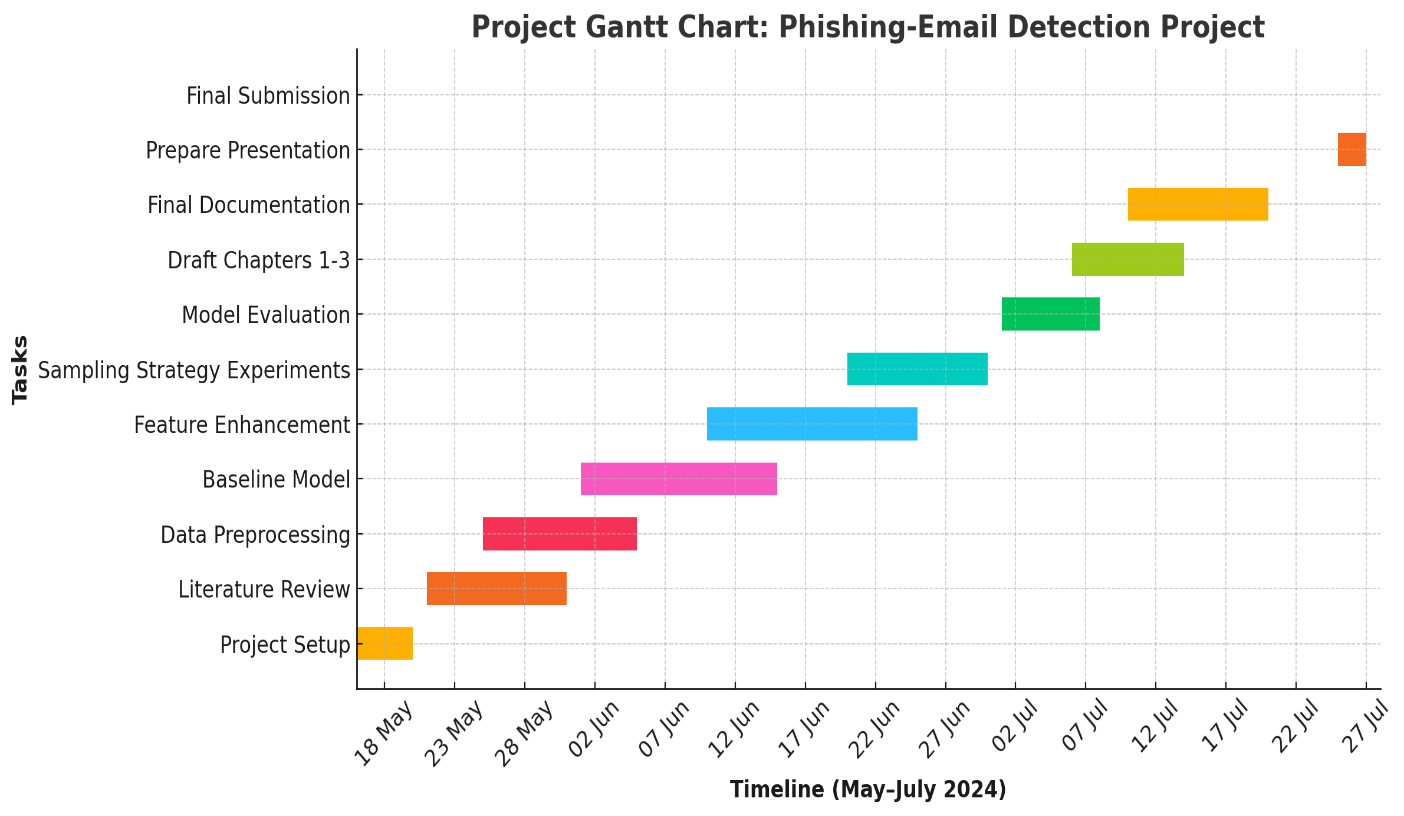
Boustead, A.E., Kugler, M.B., 2023. Juror interpretations of metadata and content information: implications for the going dark debate. J Cybersecur 9, 1–13. <https://doi.org/10.1093/CYBSEC/TYAD002>

Butt, U.A., Amin, R., Aldabbas, H., Mohan, S., Alouffi, B., Ahmadian, A., 2023. Cloud-based email phishing attack using machine and deep learning algorithm. Complex and Intelligent Systems 9, 3043–3070. <https://doi.org/10.1007/S40747-022-00760-3/FIGURES/20>

PROJECT PLAN: TASK LIST AND TIMELINE

Task List

|  |  |  |
| --- | --- | --- |
| **Date Range** | **Task Name** | **Description** |
| 16 May – 20 May | Project Setup | Create a GitHub repo, download the dataset, and review the project scope |
| 21 May – 31 May | Literature Review | Review related work, summarize relevant research papers |
| 25 May – 05 June | Data Preprocessing | Clean and structure data, handle missing values |
| 01 June – 15 June | Baseline Model | Train TF-IDF + Logistic Regression and evaluate |
| 10 June – 25 June | Feature Enhancement | Implement n-gram character and word-level comparisons |
| 20 June – 30 June | Sampling Strategy | Apply under-sampling and class-weighting techniques |
| 01 July – 08 July | Model Evaluation | Compare metrics across all variants |
| 06 July – 14 July | Draft Chapter 1–3 | Write project introduction, methodology, and literature review |
| 10 July – 20 July | Final Documentation | Write results, conclusion, and references |
| 25 July – 27 July | Prepare Presentation | Prepare final slides and practice delivery |
| 28-Jul | Final Submission | Submit the PDM Plan and presentation materials |



***DATA MANAGEMENT PLAN***

Dataset Overview

We’re working with the “Phishing Email Detection” dataset from Kaggle. This collection includes labeled emails that are either phishing attempts or legitimate emails. Thanks to Subhajournal, it was put together to help with NLP studies in cybersecurity and offers real examples of phishing.

Data Collection

We’ll download the dataset from Kaggle (<https://www.kaggle.com/datasets/subhajournal/phishingemails>) using their API and organize it in a data/raw/folder. Preprocessing scripts will prepare it by turning the emails into tokenized formats that can be used to train our model.

Metadata

The dataset comes in CSV format and has around 5,000 emails. Each email contains its subject line, body, and label. Once processed, the TF-IDF vectors and related logs will be about 500 MB. If we save different model checkpoints, those files might go up to about 1 GB.

Document Control

All our code will be saved in a GitHub repo: https://github.com/username/phishing-email-detection. We’ll push updates weekly using semantic versioning like v1.0, v1.1, etc. Protected branches will require reviews before any changes are merged.

ReadMe File

The README will cover everything: an overview of the project, details about the dataset, a description of the model, instructions for running the code, what you need to get started, and how to cite the work. This will help future users or contributors understand the project.

Security and Storage

We'll back up the data weekly on both GitHub and Google Drive. All data will be anonymized. We'll share access to the GitHub repo through a link. No sensitive or identifiable information will be included or stored.

Ethical Requirements

GDPR Compliance:

The dataset is fully anonymized and contains no personal information, so it’s not subject to GDPR.

UH Ethical Policies:

This research follows the University of Hertfordshire’s ethical guidelines as it uses anonymized, open-access data strictly for academic and non-commercial use.

Permission to Use:

Per Kaggle’s terms, the dataset can be freely used for educational and research purposes, so we have complete permission to use it for our project.

Ethical Collection:

The data was ethically collected from publicly available emails, and the original contributor to community machine learning research labeled it.