**Odell Thompson**

**1/28/24**

**Nova Southeastern University**

**Capstone**

**Phishing and Machine Learning**

**Overview**

This capstone project embarks on a critical mission to harness the power of machine learning in the fight against one of the most pervasive cybersecurity threats: phishing emails. The project is rooted in the goal of developing an advanced machine learning model capable of accurately identifying and classifying phishing attempts among email communications. Phishing, a deceptive practice that tricks email recipients into divulging sensitive information, poses a significant risk to personal and organizational security. By automating the detection of such emails, the project aims not only to enhance cybersecurity defenses but also to foster a safer online environment for individuals and organizations alike. I’ve selected the Naive Bayes algorithm for its proven efficiency in text classification and its suitability for identifying phishing attempts. Moving forward, the project is geared towards refining the model with a keen eye on testing and evaluation, aiming to present a functioning prototype by March 15th and to reach a 90% completion milestone by April 1st. The project's progress will be continuously evaluated through scheduled consultations with my advisor, ensuring that the early success in dataset acquisition translates into significant contributions to cybersecurity efforts, particularly in phishing detection.

**Project Introduction and Objectives**

**Objective 1: Comprehensive Dataset Acquisition and Preparation**

The foundation of any successful machine learning project lies in the quality and volume of its data. Recognizing this, a primary objective for this capstone project was to secure a large and diverse dataset of email communications, containing both phishing and legitimate (ham) emails. With over 10,000 examples collected, the dataset's richness allows for the nuanced training of the model, enabling it to learn and identify the subtle markers that distinguish phishing attempts from legitimate messages. Preparing this dataset involves meticulous cleaning and formatting to ensure that the data fed into the model is consistent, relevant, and reflective of real-world email communications.

**Objective 2: Selection and Implementation of an Appropriate Machine Learning Algorithm**

The project is centered around the implementation of the Naive Bayes algorithm, chosen for its effectiveness in text classification tasks and its suitability for probabilistic analysis in distinguishing between phishing and legitimate emails. This objective involves not just the technical application of the algorithm but also a deep understanding of its workings, ensuring that the model can accurately learn from the dataset and improve its detection capabilities over time.

**Objective 3: Development of a Robust Machine Learning Model**

Building upon the selected algorithm and prepared dataset, the project seeks to develop a machine learning model that stands out for its accuracy, efficiency, and adaptability. This model aims to serve as a reliable tool in the cybersecurity arsenal, capable of quickly and accurately identifying phishing emails. To achieve this, the project will involve iterative training, testing, and refinement of the model, leveraging both the vast dataset and the Naive Bayes algorithm's strengths to optimize performance.

**Objective 4: Real-world Application and Impact**

Beyond the technical achievements of dataset preparation, algorithm selection, and model development, the project is driven by a broader objective to make a tangible impact in the realm of cybersecurity. The ultimate goal is to create a model that can be integrated into existing email systems, providing an additional layer of security that helps users navigate their inboxes safely. By reducing the risk of phishing attacks, the project contributes to the ongoing efforts to secure digital communications, protect sensitive information, and prevent the financial and reputational damage associated with these cyber threats.

The goal of this project is to develop a machine learning model capable of accurately identifying phishing emails. By leveraging a dataset containing thousands of email examples, the model aims to learn and distinguish between legitimate (ham) and malicious (spam) emails, thus enhancing email security and user protection against phishing attacks.

**Current Level of Completion**

**Objective 1: Comprehensive Dataset Acquisition and Preparation**

* **Achieved**: A dataset containing over 10,000 emails has been successfully acquired. This extensive collection of data surpasses initial expectations and provides a solid foundation for training the machine learning model.
* **In Progress**: The dataset is currently undergoing meticulous cleaning and formatting to ensure data consistency and relevance. While significant strides have been made in preparing the data for model training, this process is ongoing and critical for the success of subsequent objectives.

Given the completion of dataset acquisition and the substantial progress in data preparation, this objective can be considered approximately **80%** complete.

**Objective 2: Selection and Implementation of an Appropriate Machine Learning Algorithm**

* **Achieved**: The Naive Bayes algorithm has been selected based on its suitability for email classification tasks and its efficacy in handling probabilistic data. This decision was informed by a thorough evaluation of various machine learning algorithms, highlighting a deep understanding of the project's technical requirements.
* **In Progress**: The implementation phase, involving the integration of the Naive Bayes algorithm within the project's framework, has begun. Initial conceptualization and planning have laid the groundwork for model development.

The selection of the algorithm and the initial steps toward its implementation suggest this objective is around **70%** complete.

**Objective 3: Development of a Robust Machine Learning Model**

* **In Progress**: With the dataset preparation underway and the machine learning algorithm selected, the development of the model has commenced. This phase includes setting up the initial code structure, integrating the algorithm with the dataset, and beginning the iterative process of training, testing, and refining the model.
* **Pending**: Comprehensive model evaluation and refinement based on performance metrics are critical next steps. These steps are essential for optimizing the model's accuracy and efficiency before its final application and deployment.

Considering the early stages of model development, this objective is approximately **50%** complete.

**Objective 4: Real-world Application and Impact**

* **Pending**: The application of the developed model in real-world settings and its integration into existing email systems remain future goals. The project's success in achieving a tangible impact on cybersecurity through phishing email detection will depend on the completion of the model development and its subsequent evaluation and refinement.

As the real-world application and impact of the project hinge on the completion of prior objectives, this area is yet to be initiated, making it **0%** complete at this stage.

**Milestones and Deliverables**

- Developed a preliminary understanding of how to structure the machine learning model, including the workflow from data preprocessing to model training and evaluation.

Evaluated several machine learning algorithms and selected Naive Bayes due to its suitability for text classification tasks and its efficient handling of probabilistic data, making it ideal for email filtering. Identified and obtained a comprehensive dataset containing over 10,000 emails, including both phishing and legitimate messages, setting a strong foundation for model training.

**Current Accomplishments**

Successfully identified and obtained a dataset with over 10,000 phishing and legitimate email examples, providing a solid foundation for training the machine learning model.

After evaluating various machine learning algorithms, selected Naive Bayes for its effectiveness in classification tasks and its suitability for handling the probabilistic nature of email filtering.

**Scheduled completions**

* Complete the preprocessing of the dataset, including cleaning and formatting the data to ensure it is correctly parsed and ready for model training.
* Implement the Naive Bayes algorithm to develop the machine learning model, focusing on accurately classifying emails as phishing or legitimate.

**Missed Targets**

* Initially, the project anticipated challenges in acquiring a sufficiently large and diverse dataset of phishing emails within the planned timeline. The concern was that the scarcity of phishing-specific datasets might delay the model's development phase.
* Surpassed Expectations: Contrary to these concerns, the project successfully secured a dataset containing over 20,000 phishing emails ahead of schedule. This accomplishment not only exceeded initial expectations but also provided an early boost to the project's timeline.

**Strategies for Addressing Missed Targets**

**Leveraging the Early Start**: The early success in dataset acquisition presents an unexpected opportunity to devote additional time to subsequent phases of the project, particularly data preprocessing and model refinement. By reallocating time saved from the dataset acquisition phase, the project can focus on enhancing the quality of data cleaning, feature extraction, and model training processes.

**Enhanced Focus on Data Quality and Model Complexity**: With the primary concern of dataset acquisition resolved sooner than anticipated, efforts can be redirected towards ensuring the high quality of the data through thorough cleaning and preprocessing. Furthermore, the project can explore more complex model architectures or feature engineering techniques that could potentially improve the model's accuracy and robustness against phishing threats.

**Strategic Use of Additional Resources**: The project will utilize the additional time and resources to conduct a deeper exploration of the datasets provided by Kaggle and Zendo. This involves not just the integration of these datasets into the training process but also leveraging the community and tools available on these platforms for insights into best practices in phishing email detection.

**Issues and Changes**

**Open Issues**

* Need to clean and preprocess the data in the CSV files, as the current format is not fully compatible with the desired analysis and model training requirements.
* Must refine the code structure to effectively sort emails and prepare the dataset for model training, ensuring the machine learning model can learn from the data efficiently.

**Open Issues and Their Solutions**

* **Automated Cleaning Scripts**: Develop Python scripts that automatically clean the data by removing duplicates, handling missing values, and correcting formatting errors. Libraries such as pandas will be instrumental in these tasks.
* **Text Normalization**: Implement text preprocessing techniques, including tokenization, stemming, and lemmatization, to normalize the email texts. This process simplifies the dataset, making it more uniform and easier for the model to process.

**Open Change Request**

The project's trajectory experienced a pivotal adjustment in its approach to dataset composition and sourcing strategy. Initially, the emphasis was on acquiring a broad spectrum of spam emails to train the machine learning model. This strategy was rooted in the assumption that a wider variety of spam examples would enrich the model's learning process and enhance its detection capabilities. However, this approach encountered a significant challenge: the generic spam datasets available lacked the specificity and depth of phishing email examples necessary for the model to effectively learn and distinguish the nuanced characteristics of phishing attempts.

**Strategic Shift in Dataset Focus:**

* **From Generic Spam to Phishing Emails**: Recognizing the critical need for specificity in training data to effectively combat phishing threats, the project shifted its focus solely towards phishing emails. This change was motivated by the understanding that the success of the machine learning model hinged on its ability to accurately identify phishing-specific cues and patterns, which were underrepresented in general spam datasets.

**Change in File Types and Data Sources:**

* **Adapting to Varied File Types**: Initially, the project was prepared to handle datasets primarily in CSV format, typical of many spam email records. The shift towards phishing-specific datasets introduced a need to adapt to different file types, including Excel files, which were more prevalent among phishing email datasets. This necessitated a flexible approach to data parsing and preparation, ensuring the model could be trained on data regardless of its initial format.
* **Utilization of New Data Sources**: To accommodate this shift, the project tapped into specialized resources:
* **Kaggle**: Leveraged for its versatile infrastructure and community-driven datasets, Kaggle provided an invaluable platform for accessing, sharing, and discussing datasets relevant to phishing email detection.
* **Zendo**: A pivotal resource in this strategic shift, Zendo offered access to a curated collection of phishing email records. The website's focus on anonymized data collection aligned perfectly with the project's revised objectives, providing a rich source of specific and actionable phishing email examples.

**Next Phase Schedule**

**Deepening Python Proficiency**: Beginning immediately, focus on enhancing Python coding skills and understanding of machine learning libraries. This is crucial for implementing the model and preprocessing the data effectively.

**Library Exploration and Integration**: Plan to research and integrate essential Python libraries, such as pandas for data manipulation, scikit-learn for implementing the Naive Bayes algorithm, and Count Vectorizer for text vectorization, into the project by March 10th. This step is vital for developing the prototype and preparing for the advisory meeting.

**Final Model Refinement and Evaluation**: Following the prototype presentation and advisor feedback, dedicate the remainder of March to refining the model. This includes adjusting the model based on feedback, improving accuracy, and preparing for a comprehensive evaluation.

**Summary**

The project has made significant progress, with 60% completion and key milestones in dataset acquisition and algorithm selection achieved ahead of schedule. The immediate next steps involve addressing open issues related to data cleaning and preparation, which are crucial for the successful development and training of the Naive Bayes model. By overcoming these challenges, the project will move closer to its goal of creating an effective tool for phishing email detection, contributing to enhanced cybersecurity measures.

***References***

Phishing Email Curated Datasets. (n.d.). Zenodo.org. <https://doi.org/10.5281/zenodo.8339691>

MachineLearning-SVM-LinearClassifier. (n.d.). Kaggle.com. Retrieved February 15, 2024, from <https://www.kaggle.com/code/ujanpradhan/machinelearning-svm-linearclassifier>

*Email Spam Classification*. (n.d.). Kaggle.com. Retrieved February 18, 2024, from https://www.kaggle.com/code/animeshshukla44/email-spam-classification