**Odell Thompson**

**3/22/24**

**Nova Southeastern University**

**Capstone**

**Phishing and Machine Learning**

**Overview**

**Project Introduction and Objectives**

This capstone project is at the forefront of leveraging machine learning to combat phishing, one of the most persistent cybersecurity threats. With a focus on developing a machine learning model adept at accurately identifying phishing emails, the project aims to significantly bolster email security and create a safer digital space. Utilizing the Naive Bayes algorithm for its proven text classification prowess, we are now in the process of refining the model, ensuring rigorous testing and evaluation. Thanks to Professor Mary Hayward's recommendation, valuable resources from LinkedIn have been instrumental in overcoming coding challenges and enhancing our model's development. With the ambitious goals of showcasing a functional prototype by March 15th and reaching 90% completion by April 1st, the project is under continuous review to ensure alignment with the overarching objective of improving phishing detection methods. In the latest phase of my project, I've made significant progress in using machine learning to address the challenge of phishing emails. With a carefully selected dataset and advanced algorithmic techniques, I've been refining a model that aims to set new standards in phishing detection. My journey into enhancing email security led me to dive into Python's scikit-learn library, utilizing its extensive features to develop, test, and evaluate my model. The choice of specific tools like train\_test\_split, CountVectorizer, MultinomialNB, and Pipeline was motivated by my need for precision, efficiency, and scalability in distinguishing phishing attempts from legitimate communications.

**Utilization of Python's Scikit-learn Library**

I turned to scikit-learn, a foundational library in Python's data science toolkit, to build the core of my machine learning model. This decision was based on several crucial considerations:

* **Data Preparation and Evaluation**: Using the **train\_test\_split** function allowed me to divide my dataset into training and testing segments effectively, ensuring that my model learns from a wide range of examples and is evaluated fairly.
* **Feature Extraction**: **CountVectorizer** has been invaluable in converting the textual content of emails into numerical data that my model can understand, a critical step for analyzing the diverse linguistic cues in phishing emails.
* **Algorithm Implementation**: I chose **MultinomialNB** for its excellence in text classification challenges. Implementing this version of the Naive Bayes algorithm through scikit-learn offered a straightforward and potent approach to applying probabilistic analysis for email classification.
* **Workflow Optimization**: The **Pipeline** functionality has significantly streamlined my development process, integrating preprocessing steps with model training. This not only boosts efficiency but also ensures consistency across my work.

**Current Level of Completion**

As of today, I estimate that the project is approximately 85% complete. This assessment reflects the significant progress made in several key areas:

**Dataset Acquisition and Preparation (95% Complete):** The dataset, comprising over 10,000 emails, has been fully acquired and is nearing the end of the cleaning and formatting process. The meticulous preparation to ensure data quality and relevance has been more time-consuming than initially expected, but it is now almost finished. The enhanced data analysis techniques I've implemented have provided deeper insights into the dataset, allowing me to prioritize the most impactful features for model training.

**Machine Learning Algorithm Implementation (80% Complete**): The selection of the Naive Bayes algorithm and its initial implementation within the project's framework marked a significant milestone. Following this, the exploration of alternative machine learning models, including deep learning approaches, has been initiated. While this exploratory phase has extended the timeline slightly, it is crucial for ensuring the model's robustness and adaptability to various phishing email tactics.

**Model Development and Optimization (75% Complete):** With the dataset nearly ready and the foundational algorithm in place, the focus has shifted to refining and optimizing the machine learning model. This phase involves integrating the preprocessing tools, fine-tuning the model's parameters, and beginning the iterative process of training and evaluation. Despite the challenges in enhancing the model's accuracy to meet the project's high standards, substantial progress has been made, with promising initial results.

**Preparation for Real-world Application (50% Complete):** While the development of the model has been my primary focus, considerations for its integration into existing email systems and its real-world applicability have begun. This includes preliminary planning for pilot testing and discussions on how best to implement the model to provide users with effective protection against phishing attacks.

**Milestones and Deliverables**

* **Current Accomplishments**: The transformation of data within Excel sheets for model compatibility and the utilization of LinkedIn resources have marked significant milestones in overcoming coding challenges. I've successfully acquired a dataset with over 10,000 emails, surpassing my initial expectations.
* I've delved deep into Python's scikit-learn library, mastering tools essential for my project's success.
* **Scheduled Completions**: I plan to finish preprocessing the dataset soon, ensuring it's perfectly set up for model training.
* Following that, my focus will shift to fine-tuning the Naive Bayes model to achieve optimal accuracy.

**Missed Targets**

Initially, I faced concerns over acquiring a sufficiently large and diverse dataset for my project. To my surprise, I managed to secure an extensive dataset containing over 20,000 phishing emails well ahead of schedule. This early success allowed me to shift my focus to other critical aspects of the project sooner than anticipated. However, I encountered unexpected challenges in parsing and cleaning the data to ensure its accuracy and compatibility with my model's requirements. This tedious process of data preparation proved more time-consuming than I had estimated, leading to a missed target: I had hoped to have the training phase of my project underway by now. The complexity of ensuring the dataset's integrity, while maintaining its relevance for accurate model training, has been a significant hurdle. It has required a careful balance between thoroughness in data preparation and the timeline of my project's development phases.

**Strategies for Addressing Missed Targets**

Implement Automated Data Cleaning Tools: To accelerate the data preparation process, I'm adopting more sophisticated automated tools for cleaning and parsing the data. These tools can identify and correct inconsistencies more efficiently than manual methods, helping to speed up the preparation phase without sacrificing accuracy.

Prioritize Critical Data Features: I'm refining my approach to feature extraction, prioritizing the most impactful data features that contribute to model accuracy. By focusing on these key features, I can streamline the preprocessing phase, ensuring that the model training is not only timely but also based on the most predictive elements.

Parallel Processing of Data Tasks: To overcome the delay, I'm employing parallel processing techniques for data cleaning and preparation tasks. This approach allows multiple data processing tasks to run simultaneously, significantly reducing the time required to get the dataset ready for training.

Engage in Incremental Training: Instead of waiting for the entire dataset to be perfectly cleaned and prepared, I'm initiating incremental training sessions. This method allows the model to start learning from the data that's already processed, adjusting and improving as more data becomes available. It's a flexible approach that keeps the project moving forward while still ensuring comprehensive learning from the dataset.

Leverage External Expertise: Acknowledging the complexity of the task at hand, I'm seeking guidance and support from external experts in data science and machine learning. This includes engaging with my network on professional platforms and utilizing resources recommended by Professor Mary Hayward. External expertise can provide valuable insights into efficient data handling and model training strategies.

**Issues and Changes**

**Open Issues**

One of the open issues I'm currently facing is the challenge of enhancing the model's accuracy. Despite the progress in developing and training the model, achieving the high standards of accuracy I've set for detecting phishing emails remains a hurdle. This issue is primarily due to the complexities involved in accurately parsing and preprocessing the dataset, which is crucial for the model to learn effectively.

Another significant issue is the time required for data preparation. The extensive process of cleaning, parsing, and fixing the dataset for accuracy has proven to be more time-consuming than initially anticipated. This has impacted my project timeline, particularly delaying the start of the model training phase.

**Open Issues and Their Solutions**

**Advanced Feature Engineering**: I plan to dive deeper into feature engineering, exploring more sophisticated methods to extract and select the most predictive features from the dataset. This might involve utilizing natural language processing (NLP) techniques to better understand the textual content of emails and identify subtle cues indicative of phishing attempts.

**Model Optimization and Hyperparameter Tuning**: I'll be experimenting with different model configurations and tuning the hyperparameters of the Naive Bayes and Support Vector Machine algorithms. The aim is to find the optimal settings that improve the model's ability to generalize from the training data and perform accurately on unseen data.

**Cross-validation Techniques**: To ensure that the model's performance is robust and reliable, I'll implement cross-validation techniques. This will help in evaluating the model's accuracy across different subsets of the dataset, providing a more comprehensive understanding of its predictive capabilities.

**Open Change Request**

In response to the challenges encountered during the project, I've initiated several change requests aimed at refining our methodology and ensuring the project's success:

**Adopting Advanced Data Analysis Techniques**: Recognizing the need for a deeper analysis of the phishing email dataset, I've requested the integration of advanced data analysis techniques. This includes leveraging machine learning-powered text analytics to uncover patterns and features not immediately apparent through conventional analysis.

**Exploring Alternative Machine Learning Models**: While the Naive Bayes and Support Vector Machine algorithms have been central to our approach, I'm now exploring the feasibility of incorporating additional machine learning models. Specifically, I'm interested in evaluating the potential of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for their ability to process sequential data and capture the context within email communications more effectively.

**Enhancing Data Preprocessing Tools**: Given the intricacies involved in preparing the dataset, I've proposed the adoption of more sophisticated data preprocessing tools. These tools are designed to automate and streamline the cleaning and formatting processes, potentially reducing the time required to prepare the data for training while improving its overall quality.

**Next Phase Schedule**

With the project's culmination in sight and the deadline on April 26th rapidly approaching, the following condensed schedule outlines the focused efforts required to refine and finalize the machine learning model for phishing detection:

**Week 1: Advanced Data Analysis and Model Evaluation:**

In the first week, I'll swiftly implement advanced data analysis techniques to deepen my understanding of the dataset. Concurrently, I will begin evaluating alternative machine learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for their potential to enhance the model's accuracy in detecting phishing emails.

**Week 2: Integration of Enhanced Data Preprocessing Tools:**

The second week will be dedicated to integrating enhanced data preprocessing tools that have been identified as crucial for automating and improving the efficiency of the data cleaning and formatting processes. This step is critical for ensuring the data is optimally prepared for the final round of model training.

**Week 3: Model Optimization:**

During the third week, my focus will shift to the optimization of the machine learning model. This includes fine-tuning the hyperparameters of our Naive Bayes and SVM models based on the insights gained from the advanced data analysis and the initial evaluations of alternative models. I will also refine the feature selection to ensure the model is trained on the most predictive attributes.

**Week 4: Final Testing and Refinement:**

The final week leading up to the submission deadline will be reserved for comprehensive testing and any necessary refinements. This period will allow for the assessment of the model's performance, ensuring it meets the high accuracy standards set for the project. I'll also prepare the final report and presentation materials, summarizing the methodologies, findings, and the model's implications for cybersecurity efforts against phishing.

**Summary**

As I approach the final stages of my capstone project, I'm driven by the significant progress we've made and the strategic plan in place for these last crucial weeks. With the deadline of April 26th in sight, my aim is not just to meet this date but to complete the project a week or two early, allowing for additional refinement and testing. The advanced tools and methodologies at my disposal are key to achieving this goal, especially as I fine-tune the machine learning model for optimal phishing detection. This effort is not just about fulfilling academic requirements; it's about contributing a meaningful solution to the pressing issue of cybersecurity, marking my entry into the field with a project that has real-world implications. I look forward to the final evaluation and the opportunity to share how this project can enhance digital safety.

***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>

Scikit-learn developers. (n.d.). Scikit-learn: Machine learning in Python. Retrieved March 18, 2023, from <https://scikit-learn.org/stable/>

Akanbi, O., & Amiri, I. S. (n.d.). A machine learning approach to phishing detection and defense. University of Colorado Colorado Springs. Retrieved March 21, 2023, from https://www.researchgate.net/publication/267156776\_A\_Machine\_Learning\_Approach\_to\_Phishing\_Detection\_And\_Defense