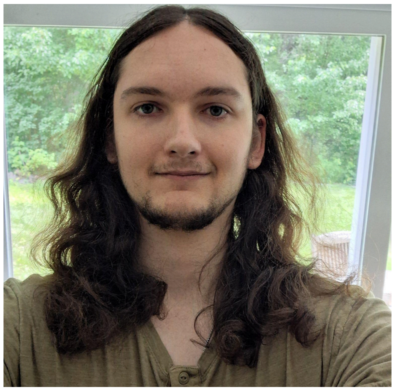
***Please note the following:***

- ***I do my initial writing in a separate document, and then copy-paste into here***
- ~~***Current citation format is MLA: I am aware that it needs to be APA. My current plan with that is to copy the citations into LaTeX and have it auto-generate the APA style for me, then copy-paste those back to here. That will be done once the writing is done.***~~
- ~~***I have chosen two datasets: one email based, one URL based. I will be double checking the requirements for and actually writing part d later today.***~~
  - ~~***I think the plan, as of now, is the following:***~~
    - ~~***Split data into train/test/validate***~~
    - ~~***DistilBERT to extract meaningful features from the data (not TF-IDF since DistilBERT can extract context and meaning and TF-IDF can't. If we have time, it could be interesting to compare them. But the focus should be on DistilBERT off the bat (I think))***~~
    - ~~***Random Forest, XGBoost, and MLP are passed the output from DistilBERT***~~

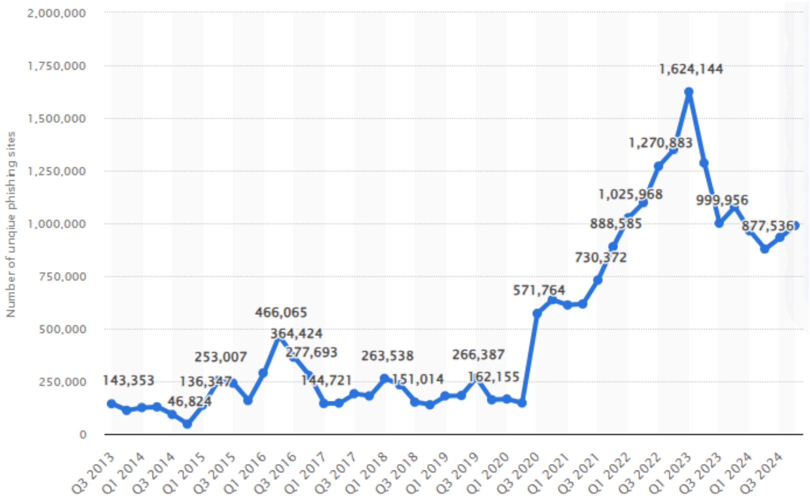
Rodrigo Arguello, Maurisa Dacosta, Nathan Poch
Professor Ben Ngan
CS 534: Introduction to Artificial Intelligence
06/11/2025

Project Proposal: Artificial Intelligence and Machine Learning Solutions to Phishing Detection

| Rodrigo Arguello | Maurisa Dacosta | Nathan Poch |
|---|---|---|
| | |  |
| Masters in Computer Science | Masters in Computer Science | Masters in Cybersecurity |

Phishing, in cybersecurity, refers to an attempt by hackers to obtain a user's private information, such as passwords, credit cards, or bank account data (Burita et. al., 2021). Typically, an attacker will use email or sms to trick a user into clicking on harmful links or visiting harmful sites, and entering their credentials (GeeksforGeeks, 2024). Types of phishing include but are not limited to General Phishing, where an email is spoofed and sent to a large number of people, Spear Phishing, which is targeted at specific individuals working specific jobs at specific organisations (depending on the intended end goal), and Whaling, which is Spear Phishing targeting exclusively CEO's and other top executives (Rashid & Leyden, 2025). Phishing can target corporations or individuals, with consequences ranging from identity theft to data exfiltration, and often come at a financial cost to the victim

(Admin, 2024). Email-based attacks lead to 91% of system breaches (Alkhalil et. al., 2021). Stolen credentials are the most common cause of data breaches, which cost victim organizations an average of over 4 million dollars (Imber, 2025). In 2015, a Whaling attack on FACC cost the aerospace company 47 million dollars (Imber, 2025). With an estimated 3.4 billion spam emails sent every day, and Google alone blocking approximately 100 million phishing emails daily, phishing is the most common form of cyber crime (Imber, 2025). With nearly one million unique phishing sites detected in the fourth quarter of 2024 alone, phishing remains a prevalent and dangerous form of social engineering (Statista, 2025).



The number of detected phishing sites across the world in each quarter from the 3rd quarter of 2013 to the 4th quarter of 2024 (Statista, 2025). The latest peak ended a sharp two year climb in Q1 2023, with 1.6 million unique phishing assaults. The 4th quarter of 2024 saw about five eighths as many, with only 0.989 million unique phishing sites (Statista, 2025). These numbers represent the number of unique base URLs of phishing sites (Statista, 2025).

The three SOTA methods that were chosen as the most effective to address our problem are: XGBoost, Random Forest, and MLP (Multilayer Perceptron). Despite similarities between XGBoost and Random Forest, we justify the exploration of both by their score similarity in previous studies (Singh & Maurya, 2024): we aim to replicate that result, and compare it to the MLP models. XGBoost (Extreme Gradient Boosting) is an optimized implementation of the gradient boosting algorithm, known for its high performance and efficiency (Shahrivari, Darabi, & Izadi, 2020). It builds an ensemble of weak learners, typically decision trees, and incrementally improves them to form a strong predictive model. One of the key advantages of XGBoost is its scalability; it can handle large-scale datasets with billions of examples, making it suitable for both distributed systems and memory-constrained environments (Shahrivari, Darabi, & Izadi, 2020). Its support for parallel and distributed computing significantly accelerates the training process, enabling faster experimentation and model refinement These reasons—along with its ability to utilize faster training and a regularization parameter that effectively reduces variance, helping to avoid overfitting—make it well-suited for phishing detection(Gu & Xu, 2022). Its use of a learning rate improves its generalization ability, which is beneficial when dealing with noisy or imbalanced datasets common in phishing scenarios, and its flexibility allows for more precise performance optimization.

Random forest is a supervised machine learning method that has been widely utilized in classification and regression problems(Dinesh et al., 2023). It combines multiple decision trees to improve prediction accuracy and reduce overfitting by allowing the trees to correct each other's errors. It also handles large datasets well and estimates generalization error effectively. Random forest will be effective for phishing detection because it is highly accurate, robust against noise and outliers, and performs well with imperfect data which is common in phishing scenarios. Its ability to reduce variance through ensemble averaging, as explained by the Central Limit Theorem, and its implicit feature selection make it both efficient and reliable for identifying complex patterns in phishing emails (Shahrivari, Darabi, & Izadi, 2020). The Multilayer Perceptron (MLP) model is a widely used light weight feedforward neural network in machine learning, serving purposes such as classification, non-linear pattern recognition and regression (Asani et al., 2024). MLPs are capable of extracting features from input data, identifying anomalies, and creating non-linear decision boundaries, making them especially suited for recognizing sophisticated phishing strategies that might evade simpler models (Reddy, 2023). Their adaptability allows them to continually learn from updated datasets, improving accuracy over time and reducing false positives, which contributes to a more reliable and responsive phishing detection system. The implementation of MLP supports real-time analysis which is crucial in phishing detection. With the proper parameter tuning and training the MLP model can be very successful.

To evaluate the effectiveness of these SOTA approaches, we will use the Phishing Email Dataset (Rokibulroni, 2025) (containing emails), with the Phishing Detection Dataset (Tamal, 2023) (containing URLs) as a backup. The dataset will be split into training, testing, and validation sets, and then fed to DistilBERT, a more lightweight version of the self-attention-based BERT designed to be used in real-time scenarios, for feature extraction (Uddin et. al., 2024). DistilBERT was chosen over BERT due to it being 40% smaller and 60% faster while still retaining 97% of the language understanding capabilities (Sanh et. al., 2019). We will perform sample-by-sample mean pooling on DistilBERT's output to preserve computational efficiency and balanced representation (Xing et. al., 2024). The resulting data will then be fed to the Random Forest, XGBoost, and MLP algorithms, which act as downstream classifiers. As in Owen & White's work from May of 2024, we will use Accuracy, False Positive Rate, and False Negative Rate, each defined below (Owen & White, 2024):

Key: $TP = True\ Positive,\ TN = True\ Negative,\ FP = False\ Positive,\ FN = False\ Negative$

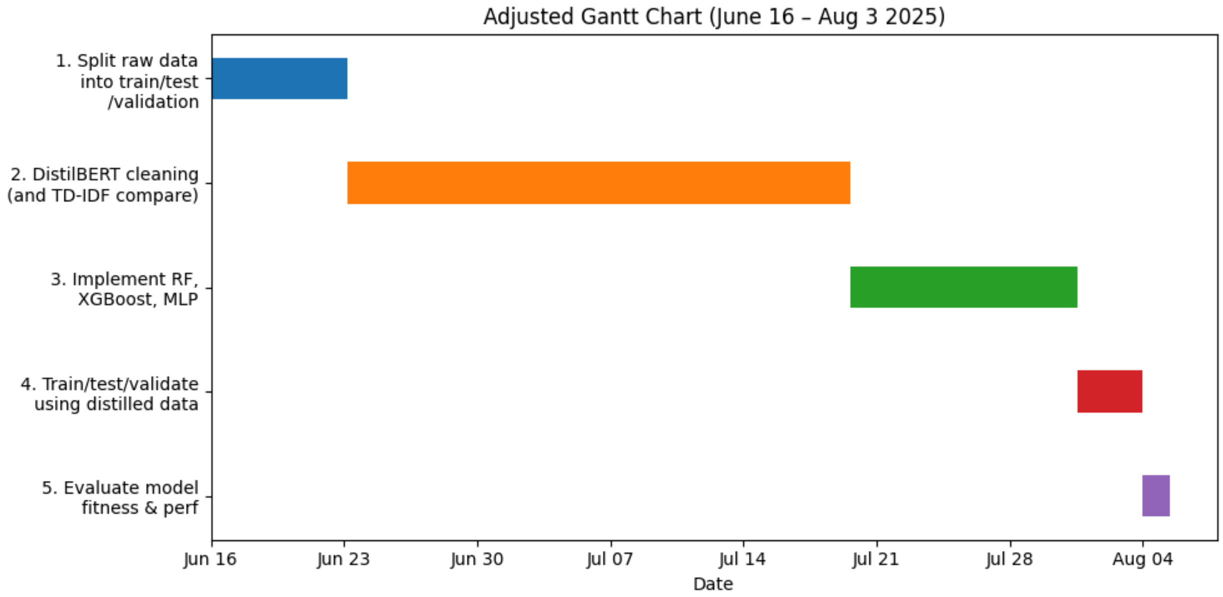
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN + TP} = 1 - Recall$$

Owen & White also evaluate the model's Processing Latency, which is the time (measured in milliseconds) that it takes for the model to analyze incoming data and produce a classification result. We will simulate this by comparing the total testing time for each model on only the testing data.

| Name: | Task Contributions |
|---|---|
| Rodrigo Arguello | Contribute to project ideas, relevant SOTAs, Gantt Chart, and Project Coordinator. |
| Maurisa Dacosta | - Project ideas<br>- Wrote SOTA descriptions and model justifications<br>- Designed and developed the PowerPoint presentation<br>- Translated technical research into visual and written content for presentation |
| Nathan Poch | - Project ideas<br>- Acquisition of research papers<br>- Motivational background<br>- Evaluation methods/metrics |



## Citations

Admin. (2024, April 23). *Phishing Attacks and Its Consequences - NTT Security - EN*. NTT Security - EN. https://se.security.ntt/en/understanding-the-consequences-of-a-phishing-attack/

GeeksforGeeks. (2024, June 13). *What Is Phishing?* GeeksforGeeks. https://www.geeksforgeeks.org/what-is-phishing/

Imber, D. (2025, June 3). The Latest Phishing Statistics (Updated June 2025) | AAG IT Support. *AAG IT Services*. https://aag-it.com/the-latest-phishing-statistics/

Rashid, F., & Leyden, J. (2025, April 24). *9 Types of Phishing Attacks and How to Identify Them*. CSO Online. https://www.csoonline.com/article/563353/8-types-of-phishing-attacks-and-how-to-identify-them.html

Rokibulroni. (n.d.). *GitHub - Rokibulroni/Phishing-Email-Dataset: A Comprehensive Dataset of Phishing and Legitimate Emails Curated for Cybersecurity Research and Applications. This Dataset Is Designed to Help Researchers, Data Scientists, and Cybersecurity Professionals Develop, Train, and Evaluate Models for Phishing Detection, Email Filtering, and Threat Analysis.* GitHub. https://github.com/rokibulroni/Phishing-Email-Dataset/tree/main

Statista. (2025, April 23). *Number of Global Phishing Attacks Q3 2013- Q4 2024*. https://www.statista.com/statistics/266155/number-of-phishing-attacks-worldwide/

Tamal, M. (2023). Phishing Detection Dataset. *Mendeley Data*. https://doi.org/10.17632/6tm2d6sz7p.1

Amaz Uddin, M., & Sarker, I. H. (2024). An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach. arXiv e-prints, arXiv-2402.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Singh, P., & Maurya, H. ML TECHNIQUES FOR PHISHING DETECTION.

Asani, E. O., Babalola, O. S., Akinola, A. E., Barnabas, A. S., Adams, O. V., & Odumesi, J. O. (2024, April). A server-side phishing detection api using long-short term memory (lstm) and multi-layer perceptron (mlp). In 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG) (pp. 1-8). IEEE.

REDDY, K. T. Enhancing Phishing Detection Through Multilayer Perceptron in Cybersecurity.

Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. arXiv preprint arXiv:2009.11116.

Dinesh, P. M., Mukesh, M., Navaneethan, B., Sabeenian, R. S., Paramasivam, M. E., & Manjunathan, A. (2023). Identification of phishing attacks using machine learning algorithm. In E3S Web of Conferences (Vol. 399, p. 04010). EDP Sciences.

Gu, J., & Xu, H. (2022, January). An ensemble method for phishing websites detection based on xgboost. In 2022 14th international conference on computer research and development (ICCRD) (pp. 214-219). IEEE.

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.

Burita, L., Matoulek, P., Halouzka, K., & Kozak, P. (2021). Analysis of phishing emails. AIMS electronics and electrical engineering, 5(1), 93-116.

Xing, J., Luo, D., Xue, C., & Xing, R. (2024). Comparative analysis of pooling mechanisms in llms: A sentiment analysis perspective. arXiv preprint arXiv:2411.14654.