CSC311 Project Report

UofT Scam-Busters - Email Fraud Detection

Helena Glowacki, Lucia Kim, Alessia Ruberto

August 9, 2024

Access the full dataset from our ***GitHub Repository***

Introduction

Phishing continues to be the most common form of cyber crime, with 3.4 billion phishing emails sent daily, as of June 2024 (Griffiths, 2024). Cyber criminals often use deceptive techniques to trick individuals into disclosing sensitive information or downloading malware by clicking on email attachments.

According to a study conducted by the National Cyber Security Alliance, Millennials and Gen-Z internet users are more susceptible to phishing attacks compared to their Gen-X peers. (National Cybersecurity Alliance, 2023). This puts college and university students at particular risk. Sure enough, in recent years, students at the University of Toronto have been receiving an increasing number of fraudulent emails concerning the security of their personal information. In 2021, 40,000 University of Toronto student email accounts were targeted by phishing attempts (Burns, 2021). Students are vulnerable as it is common for cyber criminals to pose as reputable sources, such as another student or the university department offering job opportunities.

Our project aims to investigate the pressing issue of phishing emails targeting University of Toronto student email accounts. The findings from our survey with 36 participants indicate that 52.8% of respondents receive phishing emails from UofT several times a month, and 8.3% have fallen victim to these scams. Interestingly, 80.6% of these students believe they are more than likely to be able to differentiate between a UofT and non-UofT phishing email, indicating a noticeable difference in these attacks. Understanding the unique aspects of UofT-specific phishing emails can be crucial for developing more robust and tailored detection systems. Our goal is to highlight the importance of context-specific analysis in cybersecurity and give broader insight into phishing detection.

Literary Review

Article 1: *Phishing Websites Detection using Machine Learning (Kulkarni & Brown, 2019)*

This article focuses on phishing detection for website URLs with domain name features while our research concentrates on phishing detection within email content. Similar machine learning algorithms were utilized in both studies, including decision trees, Naïve Bayes, and Neural Networks.

A notable difference is the performance of decision trees, which were most accurate in Kulkarni and Brown's study, but performed the worst among all our algorithms in our research. This discrepancy may be attributed to the use of pruned decision trees in their study, compared to unpruned decision trees in ours.

While both studies address phishing detection, they differ in focus: Kulkarni and Brown's study emphasizes the domain-specific features for detecting phishing websites, whereas our study emphasizes the content of emails, specifically in university settings. As algorithms' effectiveness may vary depending on the data context, it is essential to tailor approaches to phishing detection to address the unique characteristics of different attack vectors.

Article 2: ***A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms (Raza & Muslam, 2021)***

This article discusses different methodologies for classifying spam emails using machine learning algorithms. The overarching problem domain is similar to ours; we are both interested in utilizing machine learning to classify spam emails and performing an analysis on which methods perform best. Additionally, we both use ensembling and some of the same machine learning algorithms. However, this article places an emphasis on determining whether unsupervised, supervised, or semi-supervised learning algorithms perform best in general. In contrast, we only focus on supervised learning algorithms, and instead emphasize applying the problem domain to the specific “brand” of spam emails that target UofT students.

Article 3: ***Classifying Phishing Email Using Machine Learning and Deep Learning (Bagui et al., 2019)***

This article examines how machine learning and deep learning techniques can analyze email text to determine if an email is phishing, with a focus on classifiers such as Naive Bayes, Decision Trees, Support Vector Machines, and Convolutional Neural Networks.

Similar to our study, Sikha Bagui, Nandi, Subhash Bagui, and White focused exclusively on the text of emails. However, their dataset included samples from a variety of industries, such as insurance, law, medical, hotel, school, and banking. In contrast, our study aims to first evaluate these models on a general dataset and then assess their accuracy on a more specific dataset, like emails from the University of Toronto. Additionally, their research investigated the impact of word phrasing on model performance, finding that incorporating word phrasing significantly enhanced accuracy.

In their findings, without word phrasing, the Convolutional Neural Network performed best with a score of 95.97%, followed by the Decision Tree at 94.26%, while Naive Bayes lagged behind at 75.72%. The results with word phrasing were scored higher, with 97.20%, 97.50%, and 93.27% respectively. This suggests potential alignment between our research and theirs, particularly in model selection, and a valuable direction for future investigation of text preprocessing techniques.

Problem Formulation

Our goal is to investigate the effectiveness of different machine learning models—such as Neural Networks, Naive Bayes, Decision Trees, and Logistic Regression—in detecting phishing emails. At the core, our problem is an otherwise classic Machine Learning problem: determining whether the content of an email is phishing or safe. However, a critical aspect of our investigation is researching whether the traditional machine learning approaches using a non-University of Toronto centric dataset can correctly identify University of Toronto phishing emails.

Data Processing

The phishing email detection model was trained using a dataset from *Kaggle* that contains emails labeled as either "Phishing" or "Safe." Although phishing detection models can consider various features—such as the sender's address, embedded URLs, or images—we decided to focus specifically on the content of the emails. This decision was made because, at the University of Toronto, many emails are sent from student email addresses, making the content of the email the primary factor in identifying phishing attempts.

To prepare the email content for analysis, we performed several text preprocessing steps:

1. **Removing URLs:** Any URLs present in the email content were removed to prevent their influence on the model.
2. **Removing Special Characters:** All special characters were stripped from the text, leaving only letters and numbers.
3. **Normalizing Whitespace:** Multiple, trailing, and leading spaces were removed with a single space to clean up the text and make it consistent.
4. **Converting to Lowercase**: Finally, all the text was converted to lowercase, to ensure that the model treats words like "Email" and "email" as the same word.

After preprocessing, the cleaned email content was vectorized, converting the words into numerical features that the models could use to identify patterns. To prevent overfitting and mitigate the curse of dimensionality, we also performed dimensionality reduction, limiting the features to a maximum of 10,000. Finally, we split the data into 70% for training, 15% for testing, and 15% for validation.

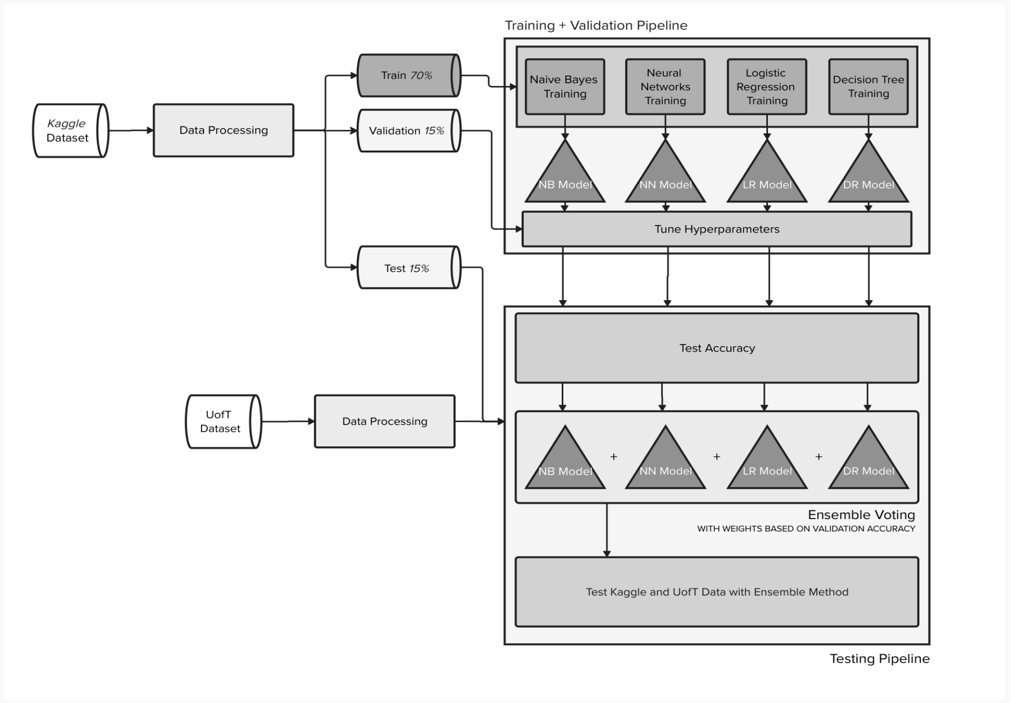
To test the model specifically with emails related to the University of Toronto, we created a custom dataset. This dataset was built by collecting emails from our own inboxes, as well as phishing emails shared on Discord servers. In total, we gathered 60 emails: 30 phishing emails and 30 safe emails.

Architecture

After preparing the data, we trained Naive Bayes, SGD Neural Network, Logistic Regression, and Decision Tree models on the training set from the Kaggle dataset. To build the models, we used PyTorch for the Neural Network, and scikit-learn for Naive Bayes, Logistic Regression, and Decision Trees. We then fine-tuned the hyperparameters of each model using the validation set to optimize their performance.

Next, we tested the accuracy of these models and compared them with an ensemble model that uses a voting algorithm, with weights assigned based on the validation set accuracy. Finally, we evaluated the performance of both the individual models and the ensemble model on the UofT dataset to see if their scores remained consistent as well as to explore any discrepancies.

Figure: Training Pipeline



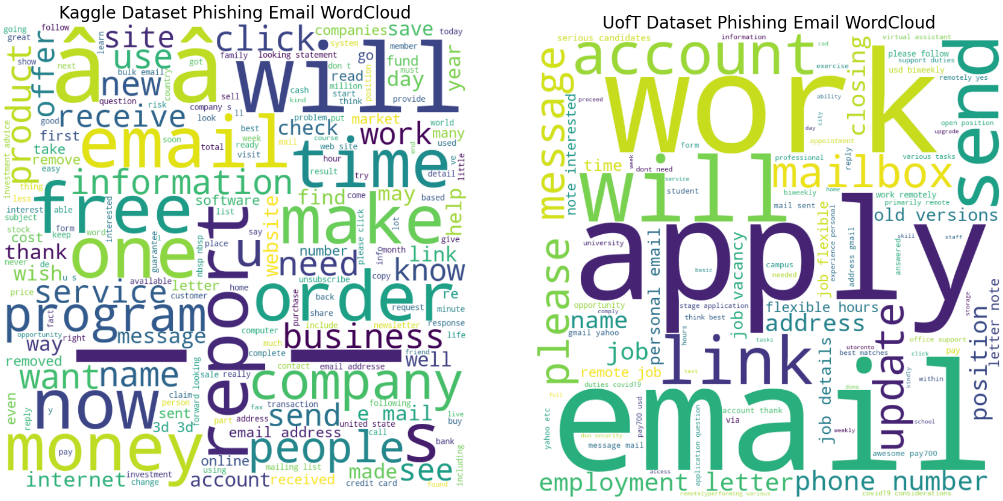
Results

Our models were trained on the phishing email dataset found on *Kaggle*. We were interested in seeing how well the models would generalize when tested on UofT-specific emails without being trained on them. Below are the results for the accuracy of each model under each set of inputs:

ML Algorithm	Training Accuracy	Validation Accuracy	Test Accuracy (Non-UofT Data)	Test Accuracy (UofT Data)
Decision Tree	0.9282	0.8981	0.8848	0.4500
Neural Network (SGD)	0.9549	0.9475	0.9498	0.5833
Naive Bayes	0.9793	0.9733	0.9711	0.6667
Logistic Regression	0.9866	0.9730	0.9738	0.7167
Ensemble	0.9838	0.9734	0.9692	0.6667

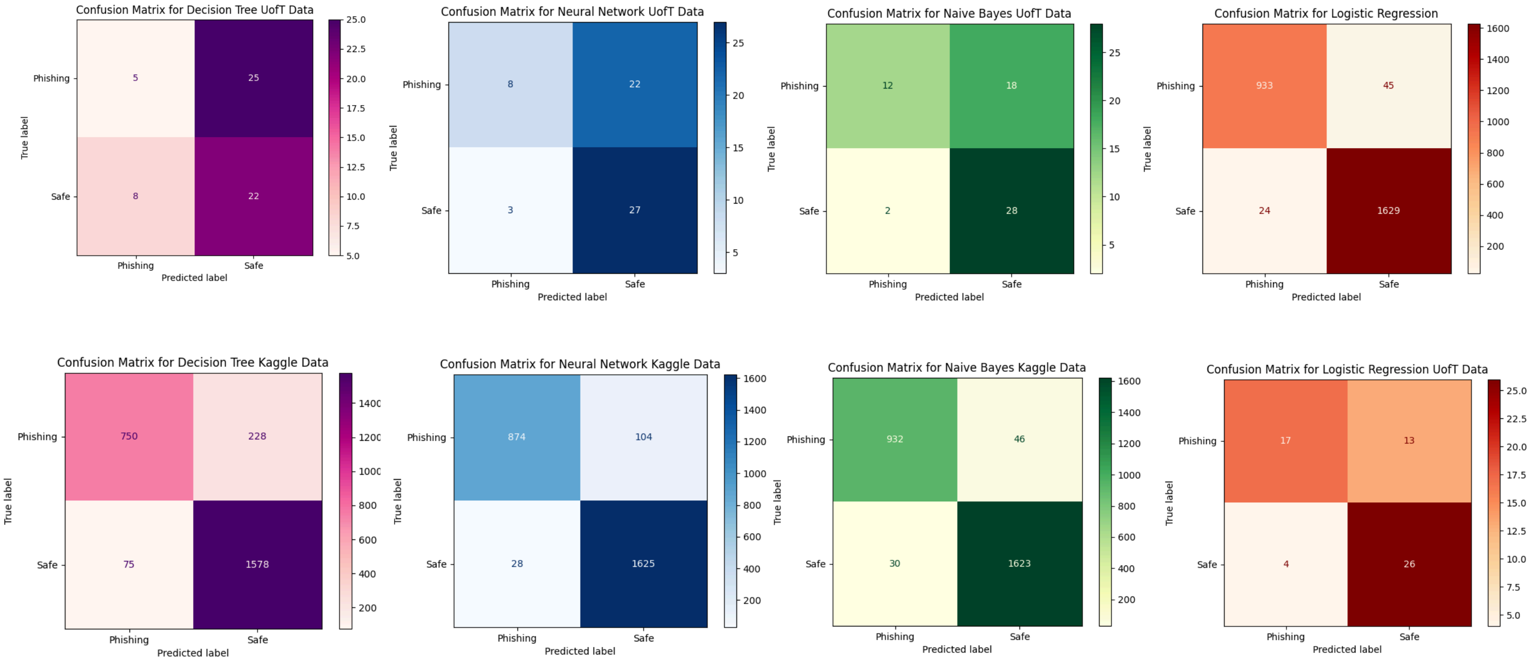
Perhaps unsurprisingly, the accuracy of the models using the UofT test set is significantly lower than that of the non-UofT (*Kaggl*e) test set. As we suspected, since UofT has a specific “brand” of spam emails that targets students, a general model has difficulty detecting them. More specifically, UofT phishing emails are typically related to scam job postings or UofT services, whereas the non-UofT (*Kaggle*) dataset includes a wide variety of general phishing emails. The word clouds below highlight the key differences between the two datasets, showcasing words that are frequently found in phishing emails for their respective datasets:

Figure: common spam keywords in the two datasets



Additionally, we noticed an interesting detail in the *type* of inaccuracies caused by the models. We will describe a “positive” result as an email flagged as phishing, “negative” otherwise. Analyzing the figures below, we notice a higher rate of false negatives (i.e. phishing email labeled as safe) than the false positive rate (i.e. safe email labeled as phishing).

Figure: Confusion Matrices For Each Test Set



These types of inaccuracies were expected; even as humans it's easier to fall for an actual phishing email than it is to suspect a safe email is spam. Additionally, we notice that the overwhelming majority of the UofT test set was classified as safe, regardless of the true label. As stated previously, this likely just due to the models not being trained of UofT-specific emails, and therefore it won't flag the keywords as spam.

Another result we found particularly interesting was the overall performance of the models. Initially, we assumed the ensemble method would yield the highest accuracy. This was due to the assumption that an ensemble would capture the best of the four models and smooth out any inaccuracies. That being said, the Logistic Regression model had the best accuracy for all of the metrics– including generalizing the best to UofT-specific phishing emails without being trained to recognize them.

Finally, the limitations of this model stem from the amount of training data we had available for UofT-specific spam emails. Initially, we contacted the UofT IT Help desk to inquire if they had any data available for us to use. Unfortunately, we did not receive a response, so we resorted to aggregating spam emails we found in our personal inboxes. As a result of only having ~60 samples, we were unable to train on this data without potentially overfitting our models.

Conclusion

Overall, every model performed significantly better on the non-UofT (Kaggle) dataset compared to the UofT-specific dataset. This further proves that models trained on general phishing email data cannot be used to flag UofT-specific spam emails due to their unique characteristics.

Despite their simplicity, Naive Bayes and Logistic Regression outperformed more complex models like Decision Trees and Neural Networks in terms of accuracy on both datasets. Surprisingly, the ensemble method did not perform better than Naive Bayes or Logistic Regression, indicating that combining models with voting may not always result in superior accuracy.

An interesting observation is the proportionally high rates of false negatives — where phishing emails are falsely marked as safe — in the UofT dataset. This raises a concern: individuals who are less familiar with specific UofT phishing tactics, like first-year students, are higher at risk. This may imply that general phishing detection models will not be very effective against the unique aspects of UofT-specific phishing emails.

In the future, it would be interesting to expand the UofT-specific dataset that we are able to train on extensively and potentially improve model accuracy. A potential strategy we could use to improve the model accuracy is utilizing natural language processing (NLP), since we noticed specific common keywords and phrases that are found in phishing emails that we could possibly identify using NLP. Moreover, our research proves that institutions like UofT with specifically-styled phishing email scams may need tailored phishing awareness training and uniquely trained machine learning models in order to successfully combat this problem.

Works Cited

Burns, L. (2021, November 15). *40,000 U of T student emails targeted by phishing attempt*. The Varsity. https://thevarsity.ca/2021/11/14/u-of-t-scam-emails-covid-19-support-team/

Bagui, Sikha, et al. “Classifying Phishing Email Using Machine Learning and Deep Learning.” *IEEE Xplore*, 1 June 2019, ieeexplore.ieee.org/document/8885143.

Griffiths, C. (2024, June 26). The Latest Phishing Statistics | AAG IT Support. *AAG IT Services*. https://aag-it.com/the-latest-phishing-statistics/

Kulkarni, A. D., & Brown, L. L., III. (2019, August 1). *Phishing Websites Detection using Machine Learning*. Scholar Works at UT Tyler. https://scholarworks.uttyler.edu/compsci_fac/20/

National Cybersecurity Alliance. (2023, June 2). *STUDY: Millennials and Gen Z Say They are Bigger Victims of Cybercrime*. https://staysafeonline.org/news-press/study-millennials-and-gen-z-say-they-are-bigger-victims-of-cybercrime/

Raza, M., Jayasinghe, N. D., & Muslam, M. M. A. (2021, January 13). *A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9334020

Appendix

GitHub Repository: https://github.com/UofT-ScamBusters/UofT-Email-Fraud-Detection

Survey Results: UofT Phishing Email Survey Responses