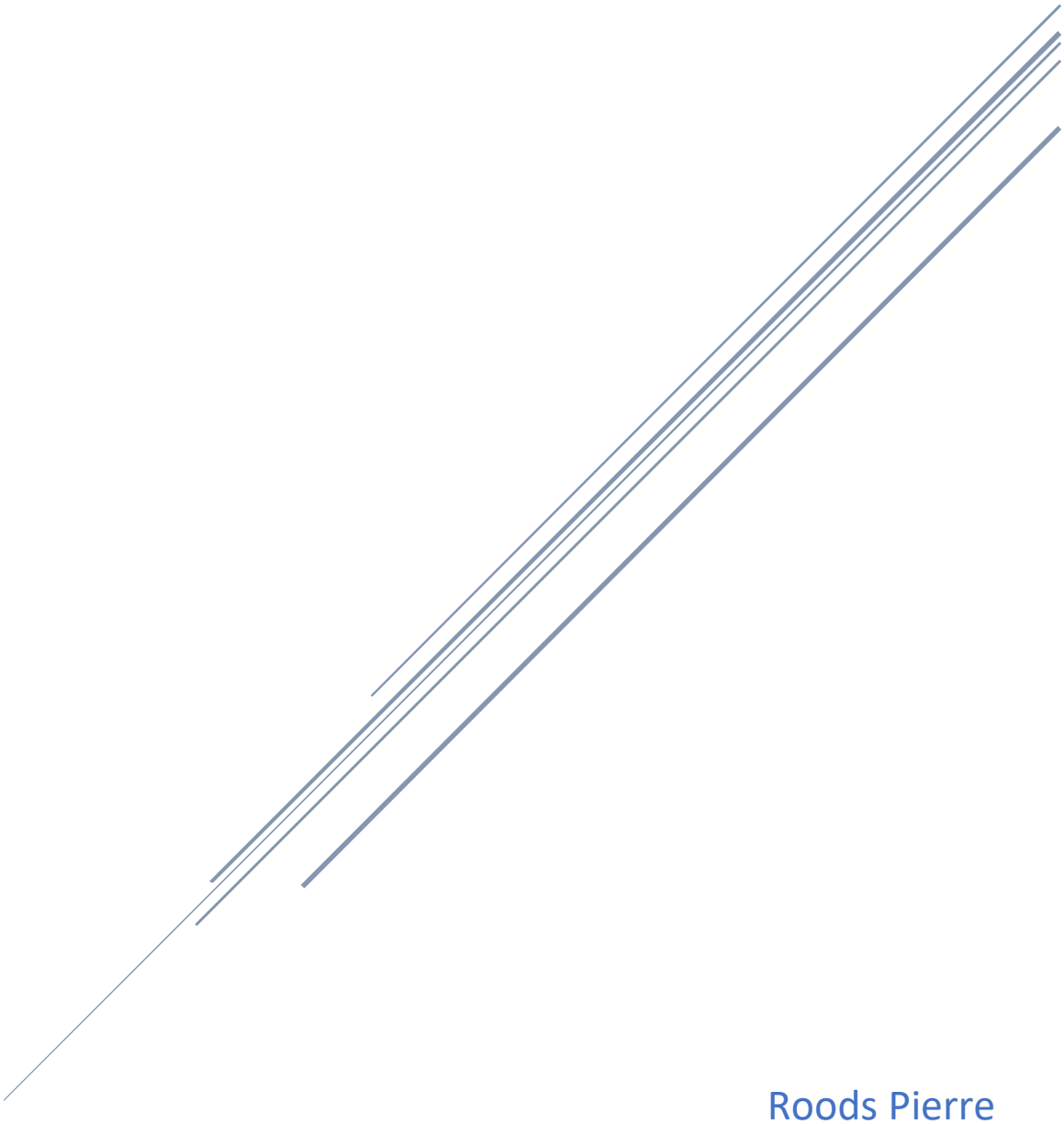
PHISHING EMAIL DETECTOR

Chrome Extension Using BERT

Project Proposal COMP 3260



Roods Pierre
T00653099

Chrome Extension for Phishing Email Detection Using BERT

Phishing attacks have emerged as one of the most prevalent cybersecurity threats, particularly in email communication. These attacks target users by mimicking legitimate entities and persuading them to disclose sensitive information, often leading to severe consequences like identity theft and financial losses. Otieno et al. (2023) state that "phishing messages can create a sense of urgency, curiosity or even fear in victims" and these threats are usually characterized in three ways:

- **The Hook**: A legitimate-looking tool (e.g., malicious website, email form, social media site) used by the attacker to collect the victim's sensitive information.
- **The Lure**: The motivation or incentive used to trick the victim, often through communication encouraging them to follow a spoofed hyperlink.
- **The Catch**: The use of the collected sensitive information to conduct illegal transactions or business.

Even though email providers have implemented various anti-phishing techniques such as spam filters and the like, the increasing sophistication of phishing strategies continues to bypass these safeguards. Current research emphasizes the need for more robust and automated detection mechanisms.

This project seeks to address this gap by developing a Chrome extension that integrates with Gmail to analyze emails in real-time using the power of the BERT (Bidirectional Encoder Representations from Transformers) model for phishing detection. Using pre-trained models like BERT for email content analysis, this tool will provide real-time feedback on potential phishing threats without requiring users to understand the complexities of machine learning or email forensics.

Objectives

The main goal of this project is to design and develop a Chrome extension capable of analyzing Gmail emails to detect potential phishing attempts using the BERT model. The extension will automatically scan emails' content and embedded links, providing users with an alert if a phishing threat is detected. This real-time detection aims to mitigate phishing attacks effectively.

Furthermore, the project will explore how the fine-tuning of pre-trained models like BERT can be applied to phishing detection without the need for extensive model training, focusing on practical implementation within a limited 7-week timeframe. The project's outcomes will contribute to the growing body of research on phishing detection mechanisms by assessing the efficacy of integrating BERT within a browser extension.

Literature Review

Phishing detection has been a focal point in cybersecurity research, with various methodologies proposed, including blacklist-based filtering, machine learning, and natural language processing techniques. Traditional phishing detection methods often rely on rule-based systems or blacklists, which are not always effective due to the evolving nature of phishing strategies and the limited ability to predict novel threats. Additionally, a quick review the literature reveals that although other approaches like ensemble learning, sentiment analysis, and topic modeling have been explored for phishing detection, these methods alone do not achieve the effectiveness seen with transformer-based models like BERT.

A 2023 study published in the IEEE Xplore by Otieno et al., “The Application of the BERT Transformer Model for Phishing Email Classification", outlines how BERT can be successfully applied to distinguish between phishing and non-phishing emails. This project will extend the work by focusing on practical applications within a Chrome extension.

According to Otieno et al. (2023), fine-tuning BERT for phishing email detection has achieved high accuracy (93%) and provides a promising solution to improve detection mechanisms. They showed that the model could capture the intricate textual patterns commonly found in phishing emails. Despite this level of accuracy, Otieno et al. note that BERT is an emerging model and cannot be fully relied on to classify phishing. However, further research conducted by Rifat et al. (2024) suggests that BERT can be a good baseline and have a great prospect in Spam and Intrusion Detection”. This project builds on their research by applying BERT in a practical, user-facing tool—a Chrome extension that will enhance Gmail's phishing detection capabilities.

Scope

The project focuses on the development of a Chrome extension that integrates with Gmail to detect phishing emails using a pre-trained BERT model. The extension will process the content and links within an email, classify it as either phishing or legitimate, and alert the user if a potential threat is identified. This implementation will primarily cover text-based phishing indicators and malicious link patterns, leaving out more complex analysis like attachment scanning.

Additionally, the project will not involve the training of a new model; instead, it will use an existing BERT model fine-tuned for phishing detection. This constraint keeps the scope manageable within the 7-week timeline while providing a proof of concept for the use of transformer models in practical cybersecurity applications. The extension that will be built in this project will not be created with the intention to function on any other platform than Gmail via the Google Chrome Browser. It should be noted that the project will not include extensive user interface development beyond basic warnings.

Methodology

The methodology for this project involves two key phases: the development of the Chrome Extension and the integration of the BERT model. The project will kick off with the development of the Chrome extension framework using JavaScript and the Chrome Extensions API. The extension will be designed to extract email content from Gmail and pass it to a backend Python-based server for analysis using a fine-tuned BERT model. BERT, known for its bidirectional contextual understanding, will serve as the core of the phishing detection mechanism. The project will use the pre-trained BERT model available through libraries like Hugging Face Transformers and fine-tune it on a phishing dataset similar to the one described in Otieno et al.'s 2023 study.

At the core of this project is the use of a pre-trained BERT model, which is particularly suitable for phishing email classification due to its ability to understand the context and semantics of email content. The fine-tuning process involves adapting the model to recognize phishing-specific textual patterns, including psychological manipulation techniques and phishing hooks embedded within email content. The steps in this process include tokenizing email text, encoding it into input vectors compatible with BERT, and then passing it through the model to classify the content. The methodology further includes implementing algorithms for link analysis to detect potentially malicious URLs. The model will be implemented using Python libraries like *transformers*, which facilitate easy model loading and fine-tuning for specific tasks. The extension will then provide users with a simple warning system, indicating whether an email is likely phishing.

The backend Python application will use Flask to handle requests from the Chrome extension, running the BERT model in real-time as email content is processed. Given the project’s time constraint, using an existing fine-tuned BERT model allows for rapid deployment while maintaining a high level of classification accuracy. The evaluation of the extension's effectiveness will be based on standard metrics such as precision, recall, F1-score, and accuracy, aiming to meet or exceed the performance demonstrated in the referenced study.

In summary, the key step in the development of this project will include:

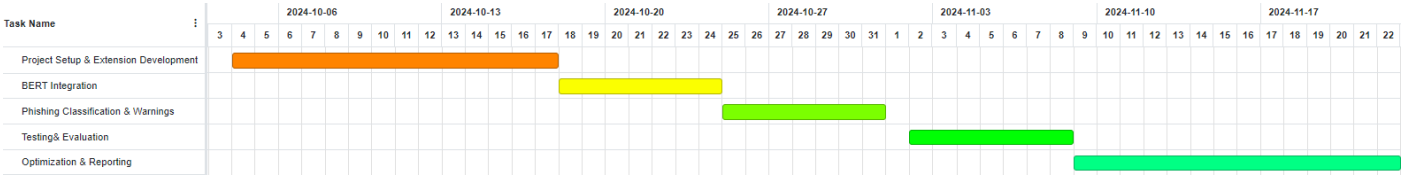
1. Extract email content from Gmail using the Chrome extension.
2. Pre-process the email data (cleaning, tokenization, etc.).
3. Run the content through the BERT model, which has been fine-tuned for phishing detection.
4. Analyze links within emails for common phishing patterns.
5. Return a verdict (Phishing/Non-Phishing) and display a warning message for the user.

Deliverables

The project will culminate in a comprehensive report and presentation that will document and showcase the research findings, methodology, implementation details, and an evaluation of the extension's performance against benchmark datasets. This report will contribute to the existing literature by demonstrating a practical application of BERT in the domain of phishing detection, highlighting its strengths and limitations in real-world usage. The following tangible component will be included:

1. A fully functioning Chrome extension that integrates with Gmail.
2. Source code with documentation.
3. Testing and evaluation report on phishing email detection accuracy.
4. Project Presentation

In order to deliver on its promise, the project will follow a structured timeline over a 7-week period:



Week 1-2
Conduct a detailed review of related research, finalize requirements, and set up the development environment. Begin developing the Chrome extension framework, focusing on interfacing with Gmail.

Week 3
Integrate the pre-trained BERT model into the backend system using Python. Develop functions for email content extraction and pre-processing within the extension.

Week 4
Implement the phishing detection logic, including text classification with BERT and link analysis. Design the user alert system to provide real-time feedback to the user within the Chrome extension.

Week 5
Conduct extensive testing and evaluation of the extension using a diverse set of phishing and non-phishing emails. Measure the accuracy, precision, recall, and F1-score to compare with benchmarks in existing research.

Week 6-7
Refine the extension based on testing feedback, debug issues, and optimize the performance. Finalize project documentation, prepare the report, and deliver the extension.

Conclusion

The significance of this project lies in its practical approach to addressing a real-world cybersecurity issue by developing a practical tool that leverages natural language processing techniques. The implementation of a BERT-based classification mechanism within a Chrome extension will provide users with an accessible, real-time phishing detection tool. This work builds upon recent research demonstrating the effectiveness of transformer models in phishing detection and seeks to translate these findings into a practical application that can enhance cybersecurity for a broad user base.

Moreover, the use of BERT allows for an advanced level of language analysis, enabling the system to detect subtle phishing cues that other models may overlook. The expected outcome is an efficient, easy-to-use tool that empowers users to identify phishing attempts and thereby enhance their personal cybersecurity. The project's success will not only contribute to the academic field but also offer a tangible solution to an ongoing cybersecurity challenge.

References

D. O. Otieno, A. Siami Namin and K. S. Jones, "The Application of the BERT Transformer Model for Phishing Email Classification," 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 2023, pp. 1303-1310, doi: 10.1109/COMPSAC57700.2023.00198.

N. Rifat, M. Ahsan, M. Chowdhury and R. Gomes, "BERT Against Social Engineering Attack: Phishing Text Detection," 2022 IEEE International Conference on Electro Information Technology (eIT), Mankato, MN, USA, 2022, pp. 1-6, doi: 10.1109/eIT53891.2022.9813922.