Phishing detection and reply system powered by Large Language Models

Aleksander Folfas
Department of Computer Science
Reykjavík University
Reykjavik, Iceland
aleksander23@ru.is

Hong Jing Toh
Department of Computer Science
Reykjavík University
Reykjavík, Iceland
hong23@ru.is

***Abstract*—This report delves into the utilization of advanced deep learning methodologies, specifically Bidirectional Encoder Representations from Transformers (BERT), in conjunction with Mistral 7B Instruct - a sophisticated Large Language Model (LLM), to detect and counter phishing emails. Our work encompasses data preprocessing, feature extraction, and model training, followed by the evaluation of the model's performance in classifying emails accurately. Furthermore, we showcase responses generated based on the provided input and prompt template. The results indicate that the BERT-based model achieved 99.3% accuracy in distinguishing between phishing and non-phishing emails.**

I. Introduction

Email communication is a cornerstone of modern life, playing a vital role in personal, corporate, and governmental interactions. The number of global email users is projected to reach about 4.38 billion in 2024, highlighting its widespread adoption. Professionals, on average, receive around 121 emails daily and send approximately 33 [1].

However, amidst the convenience lies a significant threat: phishing. This malicious tactic involves sending deceptive emails, posing as trustworthy sources, to trick individuals into divulging sensitive information like login credentials or financial details [2]. The consequences can range from financial loss to severe breaches of personal and organizational security.

Exploiting the cost-effectiveness of phishing, cybercriminals globally send out a staggering 3.4 billion phishing emails each day [5]. They don't rely on high success rates, even a small fraction of recipients falling prey is sufficient. Their strategy hinges on volume, targeting those most susceptible to manipulation.

To counter these threats, we propose a proactive approach. Our project utilizes advanced machine learning and natural language processing techniques to address the escalating phishing menace. We've developed a two-fold cybersecurity system: firstly, employing a deep learning model to classify emails as either phishing or non-phishing. Secondly, for identified phishing emails, we implement a large language model to craft intelligent, time-consuming responses. These responses are designed to engage with attackers, diverting their attention from genuine targets.

The remainder of this paper follows a structured approach. In Section 2, we provide a concise overview of the background. Following this, Section 3 reviews related work, while Section 4 outlines the email classification part of our system. Moving forward, Section 5 details the usage of Mistral 7B Instruct LLM to generate replies to emails classified as phishing. Afterward, Section 6 covers the system pipeline, and Section 7 describes future work. Finally, we conclude and summarize our contributions in Section 8.

II. Background

Phishing attacks have evolved significantly since their inception, leveraging sophisticated tactics and exploiting human psychology. Traditional defense mechanisms, such as rule-based filters and blacklists, are no longer sufficient due to their static nature and inability to adapt to new phishing strategies. Consequently, there is a pressing need for dynamic and intelligent systems capable of understanding the intricacies of human language and the ever-changing tactics of cybercriminals.

Deep learning, particularly models like BERT (Bidirectional Encoder Representations from Transformers), offers promising solutions due to its ability to understand context, nuances, and the structure of language. BERT's pre-trained models, fine-tuned on specific datasets, can capture the subtleties required to distinguish between legitimate and malicious emails effectively [7].

However, merely identifying and blocking phishing attempts is only a part of the solution. Engaging attackers in a way that wastes their time and resources presents a novel approach to cybersecurity. Here, large language models (LLMs), such as Mistral 7B, come into play [9]. These models can generate human-like text based on the content of phishing emails, creating responses that can hold the attackers' attention without risking real data. This strategy not only reduces the bandwidth of attackers to target potential victims but also serves as a unique form of active defense, potentially deterring future attacks.

III. Related work

The approach of engaging attackers by wasting their time and resources through responses generated by large language models is innovative in the cybersecurity landscape. This

strategy serves not only to engage attackers but also to potentially deter them from future attempts. It is a form of active defense that shifts the balance, requiring attacks to expand more resources and efforts, therefore reducing their overall effectiveness in targeting potential victims.

While there isn't direct research focusing on the use of LLMs for generating responses to phishing emails specifically, the broader context of using LLMs in cybersecurity provides a relevant backdrop. For instance, NVIDIA discusses the use of generative AI in cybersecurity, particularly for improving vulnerability defense and producing synthetic data to train models [3].

Moreover, recent advancements in cybersecurity have increasingly relied on artificial intelligence (AI) and machine learning (ML) to develop a more dynamic defense mechanism against phishing attacks. One of the innovative approaches that have emerged is the use of deep learning models to classify phishing emails [8]. This study explores various deep learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LTSM) networks, Recurrent Neural Networks (RNNS), and Bidirectional Encoder Representations from Transformers (BERT), for their ability to detect phishing emails. Their findings indicate that the integration of BERT and LTSM models yielded the highest accuracy, reaching up to 99.61%. This performance surpasses that of other state-of-the-art models, highlighting the capability of combined deep learning approaches to effectively understand and classify email contents.

IV. Email Classification

This section will discuss the implementation and findings regarding the classification model.

A. Preliminary data preprocessing

The dataset used for the training of the model is referred to as `combined_data.csv`, which consists of email attributes including sender, receiver, date, subject, body, URLs (indicating whether the email has a URL), label - a binary value (0 indicating not phishing and 1 phishing). The data has 145,369 rows and 7 columns. Some preprocessing steps were taken to ensure that the data was cleaned to be suitable for BERT processing:

1) Missing values in the 'subject', and 'body' were filled with whitespace, while missing values in the 'label' columns were filled with zeroes. This ensures that our model training and predictions are not hindered by missing information.
2) The presence of URLs in emails can be a significant indicator of phishing attempts. Hence, we transformed the 'urls' column into a binary flag 'urls_present', which indicates whether an email contains URLs. For those emails that are missing a value in their 'urls' column, a value of 0 is replaced to signify the absence of URLs.

B. Data Visualisation

Understanding the dataset is crucial for assessing the model's needs and potential biases. We plotted two different distributions to help us analyze the dataset:

1) A plot count of the 'label' column: This visualization helps us understand the balance between the phishing and non-phishing emails within our dataset. An uneven distribution could indicate a class imbalance, which might necessitate specific strategies during model training, such as weighted loss functions or oversampling, to ensure that the model does not become biased toward the more prevalent class. This chart shows a good mix of both phishing and non-phishing emails, which is indicative of a balanced dataset.

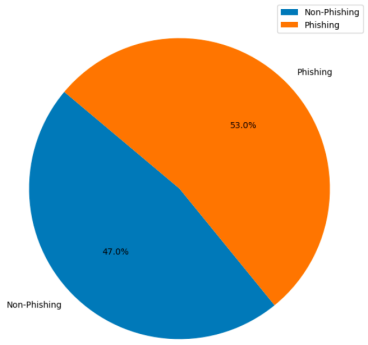

Fig. 1. Distribution of email labels

2) URL Presence by Category: This distribution shows how the presence of URLs varies between phishing and non-phishing emails. Since URLs are often used in phishing attempts to mislead recipients, a higher prevalence of URLs in one category could reinforce their importance as a feature in our classification model. This insight can guide us in feature selection and in designing more informed data preprocessing and feature engineering strategies. This chart shows that the majority of emails (phishing and non-phishing emails) in the dataset contain URLs, which further highlights the relevance of URLs in classifying whether an email belongs to the phishing class.

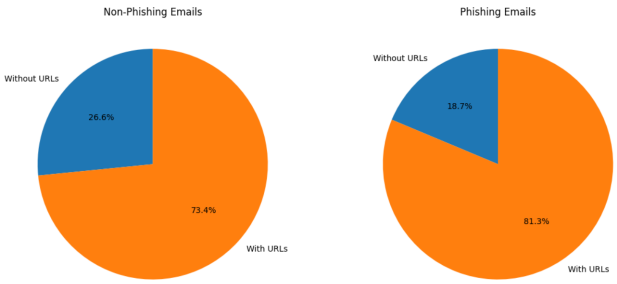

Fig. 2. Presence of URLs in Phishing vs. Non-Phishing Emails

C. BERT Data Preprocessing

In natural language processing (NLP) tasks, combining related textual fields can provide the model with a richer context. We concatenated the `subject` and `body` of emails using `[SEP]` as a delimiter. This method preserves the distinction between these elements while presenting them as a single piece of text, enabling the model to learn from both the subject line and the email body cohesively. We split the data into training and testing sets, with 80% allocated for training and 20% for testing. This ensures that the preprocessing is done independently on each dataset, maintaining a clear boundary between the training and testing environments. The `train_test_split` method from the Scikit-learn library facilitates this, ensuring a random and representative division of the data.

D. Model Development

We utilized TensorFlow Hub to integrate a BERT processor and encoder for this classification task. TensorFlow Hub is a repository of pre-trained machine learning models and components that facilitates the rapid and efficient integration of sophisticated machine learning capabilities without having the need to train the model from scratch.

The BERT layer from TensorFlow Hub serves as the foundation component of the model, responsible for transforming raw email texts into contextually enriched embeddings. This layer captures the intricate relationships and semantics within the email content.

Following the BERT layer, we included a Dropout layer with a rate of 0.1. This layer randomly sets input elements to zero during training at the specified rate, which helps prevent overfitting by ensuring that the model does not rely too heavily on any individual feature. This is crucial for maintaining the model's generalization capabilities, especially when dealing with noisy data, as is common in email classification.

The final layer in our model is a Dense layer with a single neuron and a sigmoid activation function. The sigmoid function outputs a probability score between 0 and 1, representing the likelihood of an email being a phishing email. This configuration is ideal for binary classification tasks like distinguishing between phishing and non-phishing emails.

We compiled the model using the Adam optimizer, with a set learning rate of 2×10^{-5}. Optimizers are algorithms used to change the attributes of the neural network such as weights to reduce the losses. The learning rate controls how much to change the model in response to the estimated error each time the model weights are updated. Alongside, we used binary cross-entropy as the loss function which is suitable for binary classification problems, and tracked accuracy as our primary performance metric.

E. Results

The BERT model did extremely well on the testing dataset, with a test accuracy of 0.993648, and a test loss of 0.02073. The confusion matrix in Figure 3 shows that there is a 99.5% probability that the model correctly classifies an email to be a phishing or non-phishing email and a 0.5% that it incorrectly classifies an email.

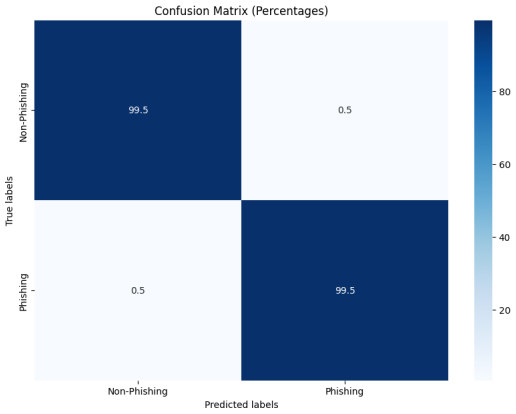

Fig. 3. Confusion Matrix for the email classifier model

V. Reply Generation

This section features information regarding the reply generation part of the project.

A. Model

For the reply generation component, we employed the LLM (Large Language Model) called Mistal-7B-instruct, developed by Mistral AI. This choice was based on its impressive performance in synthetic benchmarks, as illustrated in Figure 4.

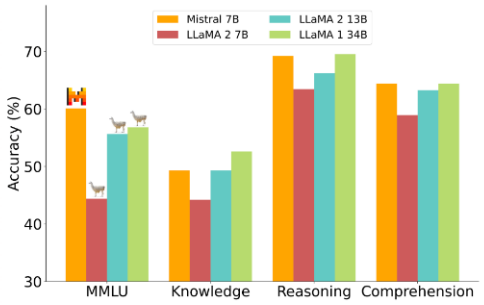

Fig. 4. Performance comparison between Mistral 7B and Llama model variants [4].

Additionally, to meet the demanding performance requirements for local model execution, we opted for a version with quantized parameters. Specifically, we selected a variant utilizing 5-bit quantization, as it was identified as a large model with minimal quality loss.

Running the LLM requires sufficient system resources, including 7.63 GB of available RAM and 5.13 GB of free disk space. The computations are executed on the CPU.

B. System used

The machine used for running the LLM had the following specifications: AMD Ryzen 7 4800H with Radeon Graphics, 2900 Mhz, 8 Core(s), 16 Logical Processor(s), installed with 32.0 GB of RAM and a 512 GB SSD.

C. Prompt

To use the model for reply generation purposes we needed to wrap the email into the following template:
user
You are receiving a scam email and want to reply to waste the scammer's time. Craft a response to the email below:
{*email*}
system
Feel free to engage, ask irrelevant questions, act confused, or provide false information to waste their time. Provide only the body of the message. Keep it short. Do not use Dear Scammer as a greeting.

Including the directive *Do not use Dear Scammer as a greeting* was essential, as certain responses tended to be initiated with this phrase. While amusing, this behavior was deemed counterproductive to the system's objectives.

D. Results

The results of the reply generation exhibit variability, implying that the response to the same email may differ each time it is generated. This dynamic behavior enables the system to cater to multiple users while ensuring that the generated messages remain fresh. This can be demonstrated in the following example: the initial phishing email can be seen in Figure 5, while the generated replies are showcased in Figure 6 and 7.

Threat detection

Our system indicates that there has been suspicious activity detected on your account and we require your immediate attention to verify your account information to prevent any unauthorized access. Please click on the following link to proceed with the verification process: Failure to verify your account within the next 24 hours may result in temporary suspension or permanent closure of your account.

Fig. 5. Phising email.

VI. Antiphishing system

The process of email classification and reply generation works as a unified pipeline, as depicted in Figure 8. Currently, the system operates with emails in JSON format as both input and output. These emails can be seamlessly integrated with email services using their respective APIs.

Upon classification, the classifier assigns a numerical value between 0 and 1 to each email, indicating the likelihood of it being a phishing attempt. To trigger the generation of a reply, we have set a threshold of 0.95. If an email receives a classification value equal to or greater than this threshold, the system initiates the reply generation process.

RE: Threat detection

I'm sorry, I don't have my verification code with me right now. Can you please send it again? Also, what type of suspicious activity was detected? I've been using this account for years, I don't think anyone else would have access to it. Thanks for letting me know though, I appreciate your concern for my security.

Fig. 6. Reply to the phishing email in the Figure 5.

RE: Threat detection

I'm sorry, I don't have my mouse with me right now. Could you please copy and paste the link into an email so I can verify my account from my phone? Also, what kind of information do you need me to provide for verification? I want to make sure I have all the necessary documents handy before proceeding. Thank you for your help!

Fig. 7. Different reply to the phishing email in the Figure 5.

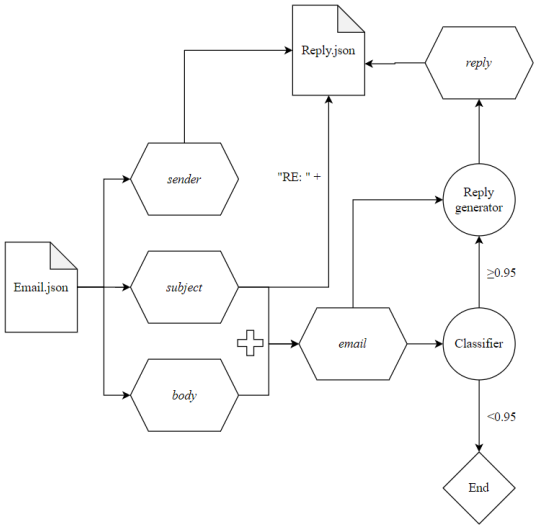

Fig. 8. Antiphishing system pipeline.

VII. Future work

Given more time we would focus our attention on the following areas:

1) Providing the reply generator with long-term conversation memory for better context.
2) Hosting the model with direct integration into mailbox services to observe its performance in real-life scenarios.

VIII. Conclusion

The utilization of advanced deep learning methodologies, particularly Bidirectional Encoder Representations from Transformers (BERT), in conjunction with Mistral 7B Instruct, a sophisticated Large Language Model (LLM), has been demonstrated in this report for the purpose of detecting and countering phishing emails. Through meticulous data preprocessing, feature extraction, and model training, followed by rigorous evaluation, the effectiveness of the proposed system has been showcased.

The results reveal that the BERT-based model achieved an outstanding accuracy of 99.3% in distinguishing between phishing and non-phishing emails. Furthermore, the integration of Mistral 7B Instruct LLM for generating responses to phishing emails exhibited promising capabilities, generating human-like responses to engage attackers and waste their time effectively.

The system's architecture, as illustrated in the antiphishing pipeline, demonstrates a comprehensive approach that connects email classification and reply generation. Moreover, the outlined future work provides insights into potential enhancements, such as incorporating long-term conversation memory and real-life scenario testing.

In summary, the proposed system offers a proactive approach to addressing the persistent threat of phishing attacks. With its high accuracy in email classification and intelligent response generation, it presents a viable solution for organizations and individuals seeking to bolster their cybersecurity defenses against phishing attempts.

References

[1] "Must-Know Email Statistics and Trends for 2024" Mailbutler, 2024. [Online]. Available: https://www.mailbutler.io/blog/email/email-statistics-trends/. [Accessed: 10-Apr-2024].

[2] "What is a Phishing Attack?" IBM, 2022. [Online]. Available: https://www.ibm.com/topics/phishing [Accessed: 10-Apr-2024].

[3] "Modernize Cybersecurity With AI" NVIDIA, 2024. [Online]. Available: https://www.nvidia.com/en-us/industries/cybersecurity/ [Accessed: 13-Apr-2024].

[4] "Mistral-7B-Instruct" Clarifai, 2024. [Online]. Available: https://clarifai.com/mistralai/completion/models/mistral-7B-Instruct [Accessed: 13-Apr-2024].

[5] "Top Phishing Statistics for 2024: Latest Figures and Trends" Stationx, 2024. [Online]. Available: https://www.stationx.net/phishing-statistics/ [Accessed: 13-Apr-2024].

[6] "Mistral 7B Instruct v0.2 - GGUF" TheBloke, 2024. [Online]. Available: https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF [Accessed: 13-Apr-2024].

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, vol. 1, 2019. doi: https://doi.org/10.18653/v1/n19-1423.

[8] S. Atawneh and H. Aljehani, "Phishing Email Detection Model Using Deep Learning," *Electronics*, vol. 12, no. 20, p. 4261, 2023. doi: https://doi.org/10.3390/electronics12204261.

[9] A. Q. Jiang et al., "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023. doi: https://doi.org/10.48550/arXiv.2310.06825.