ECE 591: Phishing Data Analysis for Machine And Deep Learning

Date: April 18th, 2024

Payton Murdoch, V00904677 (MTIS)

Dhurvit Boricha, V01047102 (MTIS)

Raunak Ghetla, V01047108 (MTIS)

Oche Eko, V01033252 (MTIS)

Joe Nikesh Puthota John, V01033708 (MADS)

Koushik Sivarama Krishnan, V01031395 (MADS)

Swathi Gnanasekar, V01033644 (MADS)

Table Of Contents

Abstract

With emails being the most common means of communication in businesses, it is immensely important to guarantee that valid traffic remains in circulation. At the same time, we inhibit the circulation of fraudulent and malicious content. With this in mind, we will prioritize parsing socially engineered phishing emails and explore how Machine and Deep learning can be utilized in conjunction with the data analysis to create possible solutions for the ever-increasing circulation of fraudulent emails. With the assistance of the UVic systems team, we have been given access to a series of phishing emails, which we can parse and analyze to assess which aspects would have a strong influence on classifying fraudulent or valid data.

Introduction

In recent years, there has been a massive spike in the circulation of fraudulent emails designed to circumvent security and obtain valuable information from a susceptible audience. Through clever schemes, attackers have been able to embed emails with compelling details explicitly intended to trick the user. These can be faulty password reset requests with the authentic logo and formatting, a similarly authentic-looking document from what appears to be a trusted source, simple links within an email designed to grab the user's attention and more. The depth of complexity concerning email attacks cannot be understated, and its impact is immense. Taken from [1], the company Cloudflare alone has protected its customers from an estimated 250 million emails which could contain malicious content. Considering companies whose online influence spans a wider domain, we can consider this number within the billions. In [2], the author denotes that it can be estimated that Google blocks over 100 million malicious emails daily. In a report [3], the author states that within 2022 alone, Microsoft protected its users against 70 billion email-originating attacks.

Unfortunately, the UVIC domain is not immune to these attacks either. Recently, students have received many falsified emails, such as those with fake job offers embedded with links and official-looking Uvic logos, to attempt to undermine the student body into leaking confidential information. With these kinds of social engineering attacks at an all-time high, we must consider how we can utilize datasets derived from such attacks to increase the student body's and businesses' overall security. In the following paragraphs, we will establish a foundation of knowledge about phishing attacks and discuss how machine and deep learning have revolutionized phishing detection. Considering phishing detection, we will then parse and analyze the dataset of phishing attacks obtained by the UVic systems team in hopes of discerning pertinent information that can be utilized in an accurate Machine Learning or Deep Learning model.

Background

What is Phishing?

Following the rapid growth and evolution in the digital world, cybercriminals are relentlessly trying to illegally harvest digital assets, particularly personal information, from people for unethical reasons. This gave rise to the commonly encountered cybercrime known as identity theft**,** which is one of the most dangerous and has grown in popularity in recent years [4]. In the case of identity theft, one (the attacker) impersonates someone else's identity (the victim) to steal from them, the attacker's selfish benefit or further committing crimes under the victim's identity. The most effective means through which attackers achieve this identity theft crime is the technique known as **Phishing**. Phishing is a cyber-attack that exploits social engineering techniques to impersonate legitimate entities with the primary goal of personal information collection. However, phishing comprises technical and social engineering contributions for a successful attack. Hence, the phishing definition presented in a recent report describes it as an attack to pass a threat into the victim's system through tricks to convince them to take actions that benefit the attacker and circumvent security procedures in place.[5] Phishing, known as carding or brand spoofing, was initially used in 1996 when hackers generated random credit card numbers to obtain customers' passwords.[6] Then, by impersonating American online staff members, phishers utilize emails or instant chats to target individuals and persuade them to divulge their credentials. Phishing attacks began targeting major financial companies because attackers thought asking customers to update their accounts would be a good approach to revealing their sensitive information. The growth trend of phishing attacks in the last three years can be observed in Fig. 1

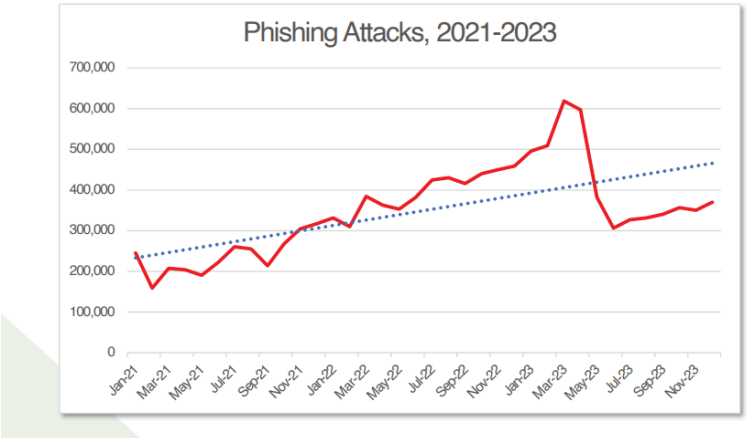

Fig. 1. The growth in phishing attacks 2021–2023 from [5].

Phishing Techniques

Phishing attacks are the most prominent cybersecurity attacks recorded in the Cybersecurity Breaches Survey 2020. [7] Phishing attacks are deployed in various forms and techniques, and they are highlighted:

- **Smishing**: This is a prominent form of phishing conducted via SMS and instant messaging platforms.
- **Email & Spam**: Here bulk spam emails are sent out without a directed target.
- **Vishing**: the phishing attack is conducted via phone calls from attackers to the victims.
- **Spear Phishing**: Similar to smishing and email, the attack is more streamlined to a specific target.
- **Pharming**: this involves creating a very close but fraudulent replica of an authentic web page and targeting it at untrained users to harvest login credentials and other private information
- **Domain spoofing** involves poisoning a legitimate website to redirect users to the malicious platform, which is usually a close replica of the original.

Impact

From an individual perspective, phishing attacks lead to financial losses, identity theft, loss of reputation, and, even worse, false criminal representation. Although the impact of phishing attacks is similar between individuals and organizations, the severity of this effect for organizations is much more significant. The burden on the organizations includes the cost of recovery from the attack and the potential loss of reputation to the public. They also incur fines from information laws/regulations, and these individual impacts tear down the organization's productivity.

Phishing Detection

Phishing detection remains a challenging problem. This is primarily because phishing is considered a semantics-based attack, which particularly exploits human vulnerabilities but not system vulnerabilities [8]. However, new and evolving technologies like machine learning, a subset of artificial intelligence, are being used to detect phishing emails. When a machine learning algorithm operates on a large-scale dataset representing phishing and legitimate emails, it analyzes the underlying pattern. The machine would derive the characteristics of phishing emails, ranging from the usage and frequency of certain words to the presence of malicious links in the body or the subject of the email, or advanced algorithms could also detect anomalies in the sender's email address [8]. This knowledge can be used by the machine to differentiate legitimate emails from phishing ones.

Deep neural networks can also help the machine understand the text semantics in the subject and body of these emails. These deep neural networks are sophisticated models comprising a network

of artificial neurons. The structure of these artificial neurons is layered in a way that could mimic the nature of the human brain. When these deep neural networks are deployed to distinguish these phishing emails, they can understand the semantics of the textual data in the subject and body of an email. It could increase the model's performance in classifying phishing emails as these DNNs could detect complex patterns in data and non-apparent cues in emails. [9]

Body

Pattern Recognition

Phishing email pattern identification is how individual senders identify particular tendencies that differentiate them from authentic communications. These include social engineering, spoofing sender information, attachment analysis, misleading content, and malicious URLs.
Following are a few ways how pattern recognition is applied in the context of phishing emails:

1. **Sender Information**: Phishing emails frequently pretend to be from banks, governments, or other reputable companies. Pattern recognition entails looking for anomalies and discrepancies in the sender's email addresses, display names, and domain names to determine whether or not a given email is a phishing effort. Red flags might include, for instance, strange sender addresses or misspelled domain names. [10]
2. **Content Patterns:** Deceivers frequently employ certain content or patterns to fool individuals. Among these are requests for urgent action, requests for private information, and requests for anything that looks too good to be true. You may identify common phishing tactics and linguistic trends in email content using pattern recognition algorithms. [11]
3. **URL Analysis:** Phishing emails sometimes contain links to phony websites that collect user passwords or sensitive information. Here, pattern recognition means looking at the structure of URLs and other characteristics found in email messages to identify odd patterns, such as misspellings, uncommon domains, or URL redirection. [12]
4. **Social Engineering:** Because phishing emails often follow certain patterns of attack, such as instilling a sense of urgency, fear, or curiosity to trick the target into clicking on links or divulging sensitive information, phishing emails typically use psychological tactics to make people feel as though they must take action. [13]

The Python notebook available through Jupyter Services is a useful tool which has allowed us to perform data analysis with the testing dataset. [14] In this way, emails' "Subject" and "Body" fields are tokenized, which enables the system to identify frequently used terms and cluster similar emails together using clustering analysis. This makes it possible to determine the commonality of phishing tactics, such as spoofing senders or misleading content. Since we have classified phishing attacks under the significance of finding trends in cyber security, this study advances attempts to combat them. By visualizing word clouds and examining cluster findings, the Python notebook enhances its capability to identify and classify phishing emails according to

their content and writing style. These techniques help us comprehend phishing emails' structure and deceptive tactics. The latter viewpoint suggests that detection capabilities could be enhanced while offering stronger protections against online attacks.

Machine Learning Model Data

Important email data components—primarily the 'Subject' and 'Body' text fields—must be analyzed and processed to create an effective ML model for identifying such phishing attempts. These components are enriched with possible phishing signs and, hence, critical in training algorithms to detect suspicious patterns. Below, we have discussed the methodologies using which we have standardized the sample data and also provided visualizations of the results obtained:

1. **Data Loading and Tokenization:** Moreover, the CSV file we examined has certainly already been tokenized for our ease of use as it has been separated into the ‘Subject’ and ‘Body.’ Tokenization is a straightforward technique that divides the text into smaller pieces, such as words, to assist in finding and analyzing patterns in text.
2. **Word Frequency Analysis:** It determines the frequency of words in email topics and bodies. Alternatively, this might prove advantageous in identifying often-used phrases like "call-to-action" or "urgent" that could suggest phishing or misleading information.
3. **Word Cloud Visualization:** Word clouds depict subjects and bodies based on the phrases they use most frequently. These graphic tools allow an analyst to identify phishing emails' most commonly used keywords quickly; ours is shown in Fig. 2.
4. **Clustering Analysis:** The email bodies are grouped based on similar information. This technique uses machine learning to group comparable data items (in this example, emails). Based on common words identified in those emails, each cluster would represent a distinct phishing strategy type, as shown in Fig. 3.

Fig. 2. Word cloud depiction of most common words utilized in phishing emails.

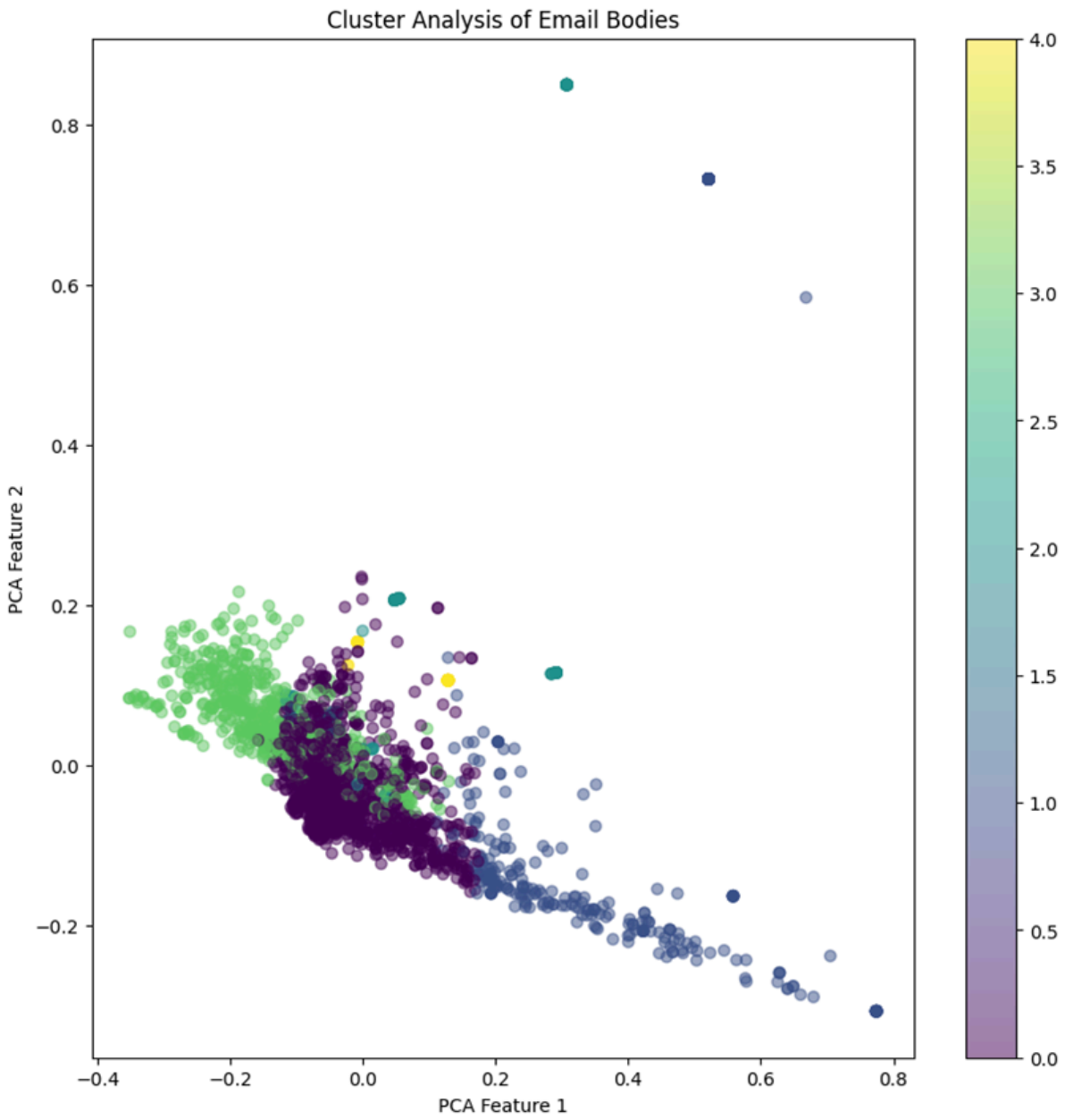

Fig. 3. Cluster Analysis of Emails.

When considering the abovementioned cluster analysis of the email bodies, we can observe that our code has grouped the emails into five distinct clusters. Therefore, we can provide some insights from this analysis:

- Cluster 0 (Purple) seems to be associated with emails containing terms like "information," "account," "details," and "contact." This could indicate that phishing emails imitate official communications and ask for personal details or account information.

- Cluster 1 (darker blue) features words like "notice," "cautious," "sensitive," and "links," which might be related to emails warning about security but contain malicious links.
- Cluster 2 (lighter blue) includes terms like "ups," "mailbox," "update," and "answered." These might be phishing attempts disguised as shipping notifications or email system updates.
- Cluster 3 (green) is characterized by terms such as "uvic," "helpdesk," "password," and "info," which suggests these emails may be trying to impersonate helpdesk or support services to solicit password information.
- Cluster 4 (yellow) has words like "upgrade," "accounts," "closure," and "inactive," which could relate to emails threatening account closure or requiring an account upgrade, a common phishing tactic.

The visualization created is a scatter plot of the clusters after reducing the data to two principal components. Each point represents an email, and the colours represent the different clusters. The distribution of emails across the clusters is as follows:

- Cluster 0 has the most emails, with 1,616.
- Cluster 3 contains 508 emails.
- Cluster 1 has 290 emails.
- Cluster 2 includes 121 emails.
- Cluster 4 has the fewest, with 41 emails.

These results provide a high-level view of the different types of phishing strategies that might be used in the dataset. The specific terms within each cluster give us a sense of the common themes within that group.

Possible Model Implementation

In today's digital world, the frequency of phishing attacks presents substantial issues for both individuals and enterprises. Phishing, a cybercrime in which bad actors seek to trick consumers into exposing sensitive information, is a constant issue in cybersecurity. As phishing assaults become more sophisticated, existing detection approaches frequently fail to identify and mitigate these threats adequately. To tackle this issue, we can leverage numerous computational methods and sophisticated pattern recognition techniques, including machine learning and deep learning, which can improve the accuracy and efficiency of detecting phishing attacks. We can use several machine learning models to implement these methods and train them on sample data of phishing attacks to identify other future phishing attacks. Scikit-learn is a popular library for creating typical machine learning models, whereas PyTorch is a popular library for deep neural networks.

Diving into the effective models and how they can be used:

Logistic Regression, a basic yet powerful linear model, can be a starting point strategy. Scikit-learn has an implementation of this model, which requires text input to be processed using numerical features obtained through techniques including TF-IDF or word embeddings. Another Scikit-learn choice, Random Forest, provides resilience against overfitting and noise. This ensemble learning approach generates numerous decision trees during the training phase and then returns the mode of classes as predictions. Like Logistic Regression, it requires text data preparation and hyperparameter tweaking, such as the number of trees and maximum tree depth. Although Random Forest is better at managing numerical and categorical information with feature significance ratings, it is more computationally costly than Logistic Regression. [15][16]

Convolutional Neural Networks (CNNs), explored alongside deep learning models here, are an excellent approach to capturing local patterns within text input. In PyTorch, CNNs consist of convolutional layers, pooling, and fully connected layers for classification purposes. Because of their capacity to automatically learn hierarchical representations, CNNs have proven particularly effective at recognizing smaller elements that hint at phishing emails. Furthermore, they are well-suited for parallel processing on GPUs, which results in faster training durations but require a big dataset for optimal performance and rigorous regularization to avoid overfitting. [17]

PyTorch-based Long Short-Term Memory (LSTM) Networks excel in detecting long-range dependencies in sequential data. The capacity of LSTMs to analyze diverse length sequences makes them suitable for assessing email content, which varies in length and structure. However, LSTMs offer significant benefits over CNNs regarding sequential data processing and coping with the vanishing gradient problem, which is common in classic recurrent neural networks. They still require careful hyperparameter adjustment and regularisation to avoid overfitting, much like CNNs. [18]

Bidirectional Encoder Representations from Transformers (BERT) is a tempting alternative for reaching cutting-edge performance. BERT fine-tuning improves text comprehension while capturing fine-grained word associations in email content through pre-trained contextual embeddings. PyTorch with Hugging Face Transformers has high computational needs and requires a long time to fine-tune, particularly for big models. Despite its high performance, overfitting must be avoided by applying good regularization algorithms and fine-tuning. [19]

To summarize, the choice of a model that can detect phishing attacks through email is influenced by several aspects, including dataset size, computing resources, and predicted performance. Logistic Regression (LR) or Random Forest (RF) is a simple, interpretable, and understandable approach, unlike more sophisticated techniques such as CNNs, LSTM networks, and BERT, which can capture complex email patterns.

Conclusion

To conclude, given the influx of phishing emails UVic and other major businesses have faced, it has become increasingly essential to increase the capacity of combative systems to help protect those dependent upon them. Phishing targets people directly by masquerading as normal traffic. It can trick a person into clicking a link, sharing private information and more so that the attacker can violate the user's privacy in some capacity. Recent developments in the field have developed pattern recognition methods with machine learning and deep learning. Machine learning considers datasets of previously established phishing methods. With the assistance of user input, these models essentially train their algorithm on these methods exclusively, thus rendering it immensely accurate against known phishing attacks but unable to detect zero-day attacks. On the other hand, deep learning eliminates the need for user assistance to parse the datasets as they can function like a human brain and infer information similarly. Using Jupyter Notebook, we have been able to parse the dataset provided by the UVic systems office to gain insight based on text analysis. Finally, we considered a series of machine learning and deep learning models that can account for the information we gathered and have traditionally worked well in phishing detection models.

References

[1] E. D. Cash et al., “Introducing Cloudflare’s 2023 phishing threats report,” The Cloudflare Blog, https://blog.cloudflare.com/2023-phishing-report (accessed Apr. 17, 2024).

[2] “The latest phishing statistics (updated April 2024): AAG IT support,” AAG IT Services, https://aag-it.com/the-latest-phishing-statistics/ (accessed Apr. 17, 2024).

[3] “State of Cybercrime: Microsoft security,” State of cybercrime | Microsoft Security, https://www.microsoft.com/en-ca/security/business/microsoft-digital-defense-report-2022-state-of-cybercrime (accessed Apr. 17, 2024).

[4] V. Ramanathan and H. Wechsler, “PhishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training,” *EURASIP Journal on Information Security*, vol. 2012, no. 1, Mar. 2012. doi:10.1186/1687-417x-2012-1.

[5] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, “Phishing attacks: A recent comprehensive study and a new anatomy,” *Frontiers in Computer Science*, vol. 3, Mar. 2021. doi:10.3389/fcomp.2021.563060.

[6] J. R. Talburt, “Principles of Information Quality,” *Entity Resolution and Information Quality*, pp. 39–62, 2011. doi:10.1016/b978-0-12-381972-7.00002-6

[7] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proc. 16th Int. Conf. World Wide Web (WWW '07)*, Banff, AB, Canada, 2007, pp. 649-656. DOI: 10.1145/1242572.1242660.

[8] M. Wu, R. C. Miller, and G. Little, “Web wallet,” *Proceedings of the second symposium on Usable privacy and security - SOUPS ’06*, 2006. doi:10.1145/1143120.1143133.

[9] N. Moradpoor, B. Clavie, and B. Buchanan, “Employing machine learning techniques for detection and classification of phishing emails,” *2017 Computing Conference*, Jul. 2017. doi:10.1109/sai.2017.8252096.

[10] “How to identify email spoofed phishing attacks - information security office - computing services - Carnegie Mellon University,” Carnegie Mellon University, https://www.cmu.edu/iso/news/2020/email-spoofing.html (accessed Apr. 17, 2024).

[11] “Phishing detection: Identifying phishing emails and websites,” Perception Point, https://perception-point.io/guides/phishing/phishing-detection-identifying-phishing-emails-and-websites/. (accessed Apr. 17, 2024).

[12] D. Pienica, “URL analysis 101: A beginner’s guide to phishing URLs,” Intezer, https://intezer.com/blog/incident-response/url-analysis-phishing-part-1/. (accessed Apr. 17, 2024).

[13] Google, https://developers.google.com/search/docs/monitor-debug/security/social-engineering (accessed Apr. 17, 2024).

[14] “Project jupyter,” Project Jupyter, https://jupyter.org/ (accessed Apr. 17, 2024).

[15] N. Sharma, “Spam detection with logistic regression,” Medium, https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522 (accessed Apr. 17, 2024).

[16] A. A. Akinyelu and A. O. Adewumi, “Classification of phishing email using Random Forest Machine Learning Technique,” Journal of Applied Mathematics, https://www.hindawi.com/journals/jam/2014/425731/ (accessed Apr. 17, 2024).

[17] C. McGinley and S. A. S. Monroy, "Convolutional Neural Network Optimization for Phishing Email Classification," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 5609-5613, doi: 10.1109/BigData52589.2021.9671531.

[18] V. S. Vinitha, D. K. Renuka and L. A. Kumar, "Long Short-Term Memory Networks for Email Spam Classification," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 176-180, doi: 10.1109/ICISCoIS56541.2023.10100445.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional Transformers for language understanding,” arXiv.org, https://doi.org/10.48550/arXiv.1810.04805 (accessed Apr. 17, 2024).