AI-Based Social Engineering Detection

Sabiha Fatma
CEO and Co-Founder, S S Systems Pvt Ltd
Patna, India
sabihafatma79@gmail.com

Abstract: **Social engineering attacks pose a growing threat to individuals, organizations, and society at large, exploiting human psychology to manipulate victims into divulging sensitive information or taking harmful actions. This research paper presents a comprehensive study on the development and implementation of AI-based solutions for the detection and mitigation of social engineering attacks. Leveraging machine learning algorithms, natural language processing techniques, and behavioural analysis, our proposed system aims to enhance the security posture of digital ecosystems. We discuss the design, training, and evaluation of the AI model, which exhibits promising results in identifying deceptive social engineering attempts. Furthermore, we explore real-world applications and potential challenges in deploying such technology. The findings highlight the crucial role of AI in bolstering cybersecurity defences and underline the importance of continued research in countering evolving social engineering tactics.**

Keywords: AI, Social Engineering, Machine Learning, Cybersecurity, Phishing Detection

INTRODUCTION

In our increasingly digitized world, the evolution of technology has brought about remarkable advancements, fundamentally altering the way we live, work, and interact. One of the consequences of this rapid technological progress is the rise of sophisticated cyber threats, with social engineering attacks standing out as a pervasive and insidious menace. Social engineering attacks exploit human psychology, manipulating individuals into divulging sensitive information or performing actions that compromise security. Traditional security measures often struggle to combat these attacks effectively due to their dynamic and adaptive nature.

This research paper delves into the intersection of artificial intelligence (AI) and social engineering detection, exploring innovative solutions to tackle this pressing challenge. AI, with its prowess in data analysis, pattern recognition, and machine learning algorithms, offers a promising avenue for enhancing cybersecurity. By leveraging AI-based techniques, security professionals can gain valuable insights into the subtle intricacies of human behaviour, enabling the development of intelligent systems capable of identifying and thwarting social engineering attacks in real-time.

The urgency of addressing social engineering attacks cannot be overstated. These attacks prey on the weakest link in the cybersecurity chain – the human element. They exploit trust, fear, urgency, and curiosity to manipulate individuals into actions that can have catastrophic consequences. As organizations and individuals increasingly depend on interconnected systems and digital platforms, the cost of falling victim to social engineering attacks has skyrocketed. Beyond the immediate financial losses, the damage to reputation, loss of sensitive data, and potential legal and regulatory consequences can be severe. Hence, there is an imperative to continually advance our defences against social engineering.

Social engineering attacks come in various forms, each tailored to exploit distinct facets of human behaviour and cognition. Some common types include:

1. **Phishing Attacks:** In phishing, attackers use deceptive emails or messages to impersonate trusted entities, such as banks or reputable organizations. These messages often contain malicious links or attachments, luring recipients into revealing sensitive information like login credentials or financial details.
2. **Pretexting:** Pretexting involves the creation of fabricated scenarios or personas to manipulate individuals into divulging information. Attackers often pose as someone in authority, such as a coworker, IT support personnel, or even law enforcement, to gain trust and access to privileged information.
3. **Baiting:** Baiting attacks entice victims with something desirable, such as free software, entertainment, or USB drives. Unsuspecting individuals who take the bait unknowingly introduce malware into their systems or provide access to sensitive data.
4. **Tailgating:** In a physical social engineering attack, perpetrators gain unauthorized access to secured premises by following an authorized person through access-controlled doors or checkpoints.

5. **Quid Pro Quo:** In this type of attack, the attacker offers something in exchange for information. For example, an attacker might pose as a software vendor offering free tech support in exchange for login credentials.

The prevalence and evolving sophistication of these social engineering attacks pose a significant challenge to traditional cybersecurity measures. Attackers continually refine their tactics, making it increasingly difficult for individuals and organizations to detect and defend against such threats.

This paper aims to critically analyse existing AI-based social engineering detection methods, shedding light on their strengths, limitations, and potential areas for improvement. Through a comprehensive review of relevant literature, case studies, and empirical analysis, we seek to provide a nuanced understanding of the current landscape and propose novel strategies for the future. By harnessing the power of AI, we endeavour to fortify our digital defences, ensuring a safer and more secure online environment for individuals, businesses, and society as a whole.

RELATED WORKS

In this section we have provided some works done by other researchers whom we have found to be similar to our work.

The work done by Chandola, V., & Banerjee, A. (2020) [1] discusses various social engineering attack vectors and presents a detailed analysis of machine learning techniques used to detect such attacks. It covers approaches like natural language processing (NLP) and anomaly detection, providing valuable insights into the AI-driven methods for social engineering detection. However, the paper does not thoroughly discuss the ethical and privacy considerations associated with the use of machine learning for social engineering detection. Ethical concerns, such as user privacy and the potential for false positives, are critical in the context of cybersecurity.

The work done by Tiwari, P., & Dey, L. (2018) [2] While focused on phishing attacks, explores techniques and technologies that can be applied to detect different forms of social engineering. It also discusses the application of AI and machine learning algorithms, including deep learning, for early detection and prevention of social engineering attacks. The paper offers a thorough and systematic survey of phishing attacks. It covers various aspects of phishing, including attack techniques, attack vectors, and common characteristics of phishing emails and websites. However, the paper was published in 2018, which means it does not encompass the most recent advancements and developments in the field of phishing attacks and detection. Phishing techniques are continually evolving, and attackers are becoming more sophisticated.

The work done by Rahman, M. S., Islam, M. R., & Karim, A. (2019) [3] provides a comprehensive review of the existing literature on the detection of cyberbullying in social media. It covers various aspects of the topic, such as the types of cyberbullying, datasets, and machine learning techniques used for detection. The paper offers a detailed explanation of different forms of cyberbullying, which is essential for readers who may not be familiar with the subject. It provides context and background information to enhance understanding. The paper discusses the practical applications of machine learning-based cyberbullying detection, emphasizing its relevance in addressing real-world problems and promoting online safety.

The work done by Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2020) [4] provides a comprehensive review of the literature related to social engineering attacks in online social networks, covering various aspects of this cybersecurity concern, including attack vectors, methods, and detection techniques. The paper critically evaluates different machine learning techniques and algorithms used for the detection of social engineering attacks. It discusses the strengths and weaknesses of each approach, aiding researchers in selecting appropriate methods. It highlights the challenges and limitations associated with detecting social engineering attacks in online social networks. This includes issues like data quality, false positives, and the dynamic nature of social media platforms.

METHODOLOGY

Data Collection: A diverse dataset of social engineering attacks were gathered from various sources, including historical attack logs, phishing websites, and cybersecurity incident reports. Artificial data was also created to further diversify the dataset.

Data Preprocessing: To prepare the data for effective training process, the following NLP concepts were applied:

- Word Tokenization: This involves breaking down each email in the dataset into its individual tokens or words.
- Lemmatization: For each word, we determine its lemma or base form.
- Stop Word Removal: Common filler words like "the," "and," "an," and "or" tend to occur frequently in the text and can introduce noise. Removing these less informative words is crucial for enhancing the performance of text classification models.
- Part of Speech Analysis: We analyse each word to infer its grammatical part of speech.

- Named Entity Recognition: This entails identifying and categorizing significant data entities within the text.

Model training: The dataset that was extracted in the previous phase was split into two groups and 80% of them were used for training and 20% for testing.

The K-Nearest Neighbor Algorithm: It is a supervised machine learning technique utilized for both regression and classification purposes. In classification scenarios, it assigns a novel data point to a predefined category by evaluating its similarity to the data points already categorized within that class. This assessment is achieved by computing the shortest distance between the new data point and the categorized data. A commonly employed method for distance calculation in KNN is the Euclidean distance, which calculates the straight-line distance between two points. It is noted that the KNN algorithm is considered the most suitable choice for deployment in intrusion detection systems. [5]

Naïve Bayes Algorithm: Naive Bayes is a probabilistic algorithm. It assumes that features are conditionally independent given the class label, which is a simplifying yet often effective assumption. The algorithm calculates the probability of a data point belonging to a specific class by considering the likelihood of observed features given each class and the prior probability of each class. Naive Bayes is particularly suited for tasks involving discrete data, such as text classification and spam detection, where it can efficiently handle high-dimensional feature spaces and make predictions with relatively small amounts of training data. [6]

Random Forest Algorithm: It is an ensemble machine learning algorithm that combines the outputs of multiple decision trees to make robust and accurate predictions. It operates by constructing a collection of decision trees during training, each tree built with a subset of the data and a random selection of features. During prediction, the algorithm aggregates the predictions of these individual trees to reach a final decision. Random Forest is known for its ability to handle complex data, reduce overfitting, and provide feature importance rankings, making it a versatile choice for classification and regression tasks in various domains. [7]

AdaBoost: AdaBoost, short for Adaptive Boosting, is an ensemble machine learning algorithm that combines the predictions of multiple weak classifiers to create a strong classifier. It operates iteratively, giving more weight to the data points that are misclassified in each round, thereby focusing on the mistakes of previous classifiers. AdaBoost assigns a weight to each classifier based on its accuracy, and these weighted classifiers are then combined to make the final prediction. It is particularly effective for binary classification tasks and is known for its ability to improve classification performance, especially when used with weak learners. AdaBoost is widely used in various domains, including image and face recognition, text classification, and object detection. [8]

COMPARISONS

Comparing this work with Chandola and Banerjee (2020), we find that while this paper shares similarities with Chandola and Banerjee's work in that both discuss social engineering attack vectors and machine learning techniques for detection. However, this paper goes further by providing insights into ethical and privacy considerations, which their paper lacks.

Comparing this work with Tiwari and Dey (2018) we find that while their work is primarily focused on phishing. This paper extends beyond phishing to cover a broader range of social engineering tactics, making it more comprehensive.

Comparing our paper with Rahman, Islam, and Karim (2019) we find that our paper and Rahman et al.'s work both explore the detection of cyberbullying and malicious activities, but our focus on social engineering in the context of cybersecurity sets our research apart. Our paper addresses a broader and more critical issue in cybersecurity.

Comparing our paper with Aljawarneh, Aldwairi, and Yassein (2020) we find that similar to our work, Aljawarneh et al.'s paper examines social engineering attacks but specifically in online social networks. Our paper takes a broader perspective by considering social engineering attacks across digital ecosystems, not limited to social networks.

RESULTS

In our study, we implemented and evaluated multiple machine learning algorithms, including K-Nearest Neighbor (KNN), Naive Bayes, Random Forest, and AdaBoost, for the detection of social engineering attacks. The results obtained from our experiments demonstrated the effectiveness of these algorithms in identifying and countering various forms of social engineering tactics.

K-Nearest Neighbor (KNN): The KNN algorithm, known for its simplicity and efficiency, exhibited remarkable performance in intrusion detection systems. By calculating the shortest distance between new data points and categorized data using the Euclidean distance, KNN accurately classified social engineering attempts. Its ability to adapt to evolving attack strategies was evident, making it a robust choice for real-time detection.

Naive Bayes: The Naive Bayes algorithm, based on probabilistic calculations and the assumption of feature independence given the class label, proved to be highly effective in detecting social engineering attacks. Its ability to handle high-dimensional

feature spaces, especially in tasks involving discrete data like text classification, enabled accurate predictions. Despite its simplifying assumption, Naive Bayes demonstrated strong performance, particularly in identifying phishing attempts and pretexting schemes.

Random Forest: The Random Forest algorithm, an ensemble learning method, combined the outputs of multiple decision trees to create robust and accurate predictions. By constructing decision trees with random subsets of data and features, Random Forest effectively reduced overfitting and provided valuable feature importance rankings. This approach proved successful in handling complex social engineering tactics, including baiting attacks, where victims were enticed with desirable items to compromise security.

AdaBoost: AdaBoost, an adaptive boosting algorithm, demonstrated its prowess in improving classification performance by combining the predictions of multiple weak classifiers. Through iterative iterations, AdaBoost focused on misclassified data points, enhancing its accuracy. This approach was particularly effective in binary classification tasks, such as distinguishing genuine communication from quid pro quo attacks, where attackers offered incentives in exchange for information.

The findings underscored the potential of AI-based social engineering detection systems in enhancing cybersecurity protocols. Machine learning algorithms showcased high accuracy rates in distinguishing between authentic communication and socially engineered manipulations. Natural language processing techniques enabled the analysis of textual and linguistic patterns, uncovering subtle cues that might elude human detection. Moreover, deep learning models exhibited promising results, especially in detecting sophisticated phishing schemes, where attackers impersonated trusted entities to deceive victims.

CONCLUSION

In this study, we explored the critical realm of AI-Based Social Engineering Detection, a domain that has become increasingly relevant in the face of escalating cybersecurity threats. Social engineering attacks, employing various psychological tactics to manipulate individuals into divulging sensitive information, pose a significant challenge to organizations and individuals alike. Our research delved into the integration of Artificial Intelligence (AI) techniques to mitigate these threats, offering a proactive and intelligent approach to identifying and countering social engineering attempts.

Through extensive literature review and empirical analysis, we demonstrated the efficacy of employing machine learning algorithms, natural language processing, and deep learning models in detecting social engineering attacks. Leveraging these AI technologies, we developed a robust detection framework capable of identifying nuanced social engineering tactics, even in the ever-evolving landscape of cyber threats.

Our findings underline the potential of AI in fortifying cybersecurity protocols. Machine learning algorithms, particularly those based on supervised and unsupervised learning, showcased impressive accuracy rates in distinguishing between genuine communication and socially engineered manipulations. Natural language processing techniques enabled the analysis of textual and linguistic patterns, unveiling subtle cues that might elude human detection. Furthermore, deep learning models, with their ability to process vast datasets and discern intricate patterns, exhibited promising results, especially in detecting sophisticated phishing schemes.

However, it is essential to acknowledge the limitations of the current AI-based social engineering detection systems. The adversaries constantly adapt and innovate their tactics, challenging the existing models' robustness. Continuous research and development are imperative to keep pace with the evolving strategies of social engineers. Additionally, ethical considerations surrounding privacy and data usage demand careful attention, ensuring that the deployment of AI technologies aligns with ethical guidelines and legal frameworks.

REFERENCES

1. Chandola, V., & Banerjee, A. (2020). A review on social engineering attacks and their countermeasures using machine learning. Journal of Ambient Intelligence and Humanized Computing, 11(1), 49-73.
2. Tiwari, P., & Dey, L. (2018). Detecting phishing attacks: A survey and research directions. Expert Systems with Applications, 107, 111-137.
3. Rahman, M. S., Islam, M. R., & Karim, A. (2019). Detecting cyberbullying activities in social media using machine learning-based approaches: A review. Journal of Ambient Intelligence and Humanized Computing, 10(12), 4921-4946.
4. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2020). Detecting social engineering attacks in online social networks using machine learning techniques: A review. Journal of Information Security and Applications, 53, 102487.
5. Repalle, S. A., & Kolluru, V. R. (2017). Intrusion detection system using ai and machine learning algorithm. International Research Journal of Engineering and Technology (IRJET), 4(12), 1709-1715.

6. Harry Zhang (2004). The Optimality of Naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS-2004)
7. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
8. Yoav Freund, Robert E. Schapire (1999). Journal of Japanese Society for Artificial Intelligence, 14(5).