THE AI IMPACT IN DEFENSE MECHANISM OF SOCIAL ENGINEERING ATTACKS

Ivan Prole*,
[0009-0008-5391-0864]
Mladen Veinović
[0000-0001-6136-1895]

Singidunum University,
Belgrade, Serbia

Abstract:

This research paper explores the use of artificial intelligence in preventing social engineering attacks. Social engineering attacks are a significant cybersecurity threat that exploits individuals' emotions and psychological vulnerabilities. The paper examines various types of social engineering attacks and how AI can assist in thwarting the malicious intentions of attackers. It proposes a method for detecting the emotional state of potential attackers using AI technology. The research aims to identify the emotional state of potential attackers by analyzing their written communication. Identifying a person's emotional state is essential because it can help classify bad intentions and potential attackers, ultimately helping to prevent social engineering attacks. The methods used in this paper employ machine learning algorithms such as XGBoost, Naïve Bayes, KNN, and Random Forest to train the data. The experiment indicates that XGBoost, Naïve Bayes, and Random Forest have better accuracy rates, while KNN has a lower accuracy rate. The research results are based on a dataset. The paper demonstrates how identifying the emotional states of potential attackers can improve social engineering defense.

Keywords:

Social engineering, AI/ML, Supervised learning, Emotion detection.

INTRODUCTION

In today's world, companies face advanced and sophisticated attacks that exploit technical systems and human vulnerabilities. Artificial Intelligence (AI) and Machine Learning (ML) can be crucial to prevent these attacks. Social engineering attacks aim to steal sensitive information through technical and psychological manipulation. With emotion detection, we can identify negative emotions and build systems that can proactively defend potential targets. The latest developments involve using cognitive activities for emotion detection, providing an interdisciplinary approach for installing and deploying surveillance systems. It helps identify and record social engineering attacks easily. Most successful attacks that bypass security systems rely on human weaknesses. Social engineering is the art of manipulation to gain sensitive and confidential information. Social engineering attacks usually comprise technical, social, and physical elements, which are often interconnected. Developing appropriate models can help train protection systems to become more accurate in identifying and preventing attacks.

Correspondence:
Ivan Prole

e-mail:
prole.ivan@gmail.com

DOI: 10.15308/Sinteza-2024-18-24

AI can prevent attacks by analyzing the sentiment of text sequences. Human characteristics can be detected using the OCEAN model, based on the five major personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Cognitive science provides an interdisciplinary scientific approach that can provide a framework for the impact of AI in detecting and preventing social engineering attacks. It's important to note that AI can execute sophisticated cyber attacks. Therefore, the defender of the system needs to predict the attacker's behavior, enabling them to stay a few steps ahead of the attacker. In this paper, we explore how AI can help prevent social engineering attacks by incorporating human characteristics, focusing on detecting human emotions from the textual communication of potential attackers.

2. SOCIAL ENGINEERING ATTACKS OVERVIEW

As computer technology has developed, social engineering attacks have evolved from basic techniques like shoulder surfing, which involved secretly collecting data by looking at someone's screen, to more sophisticated methods. By implementing appropriate solutions to collect all data inputs, AI can assist in developing robust solutions to prevent social engineering attacks. [1] With the rise of social networks and new technology applications, identifying and preventing these attacks has become a significant challenge [2]. Social engineering attacks today include phishing, pretexting, road apple, tailgating, ransomware, fake software, reverse social engineering, robocalls, and help desk attacks.

Phishing attacks are the most common types of cyberattacks. The main objective of these attacks is to gain unauthorized access to sensitive personal information. Whistling is a form of attack that uses sound signals to achieve the goal. Phishing attacks can be categorized into different types, such as spear phishing, whaling phishing, vishing phishing, interactive voice response phishing, and email phishing. [3]

Pretexting is a phishing technique that relies on two-way communication. This kind of attack usually resides in mutual trust. Conversation can start by delivering a scam or message requesting the victim's confirmation, usually installing malware on the system. The final aim is to steal personal data that relies on trust between people. [4]

Tailgating is a method of bypassing security mechanisms without personal identification. In this method, the attacker follows a person who enters a secure area and gains access without proper authorization. To protect against this type of attack, staff should always ensure that access is granted only to those with proper identification cards and be vigilant to ensure that no one has entered the area without proper authorization.

Ransomware is a cyberattack that encrypts all the data on a computer system. The attackers then demand a ransom payment to release the locked data. This attack often preys on people's fear and urgency to regain access to their files. However, even if the victim pays the ransom, there is no guarantee that the attackers will return the data in a usable format. It is essential to consider the ethical implications of paying attackers before deciding.

Social engineers often combine social and technical tactics in their attacks. For instance, a baiting (road apple) attack might involve the attacker leaving a malware-infected disk in a location that is likely to be accessed. [2] This is a common technique in which storage media is labeled with Confidential, CEO, and Finance as teasers.

Robocall attacks are automated voice attacks that use pre-recorded messages to target unsuspecting individuals. These calls usually originate from unknown numbers and are sent to a list of phone numbers identified as potential targets. The primary goal of these attacks is to trick individuals into revealing sensitive personal information, such as PINs and Social Security Numbers. [5]

Reverse social engineering is a type of attack that involves three essential steps. First, the attacker creates a problem on the victim's network and then offers to fix it, claiming to be the only one capable of doing so. After fixing the problem, the attacker collects the desired information and leaves the network in the operating state [6].

Fake software attacks are often deployed on systems or websites, appearing to come from a known source. In these cases, attackers typically use malware to trick users into giving away their login credentials or other personal information, which can then be stolen. [3]. An example of this type of threat is online banking applications.

A help desk attack is when an attacker poses as an authority figure or an employee of a company and calls the company's help desk to request information or services [3] Companies are highly cautious of such attacks, particularly when they have many employees.

3. EXPLORING AI TECHNIQUES TO PREVENT SOCIAL ENGINEERING

Natural Language Processing (NLP) is a field that deals with the linguistic aspects of communication between computers and human language. It is a subfield of Artificial Intelligence (AI) that aims to connect human language understanding with computer information processing. [7]

Anomaly Detection is a crucial ability for analyzing data and identifying deviant behavior. It involves analyzing data points, models, signals, and patterns to spot a list of behaviors that can indicate suspicious activity [8]. This domain requires identifying abnormal personal patterns and detecting any malicious intentions.

Facial recognition technology uses neural networks to identify images, videos, or suspicious activity of attackers. This approach can solve various problems, including sentiment analysis from image and video sequences. [9] One practical implementation of this technology is detecting deepfake videos on social networks.

Graph Neural Networks (GNNs) can operate on graphs that have interconnected nodes through edges. In this context, nodes represent entities, while edges are the relations between these entities. GNNs can predict data on graph-structured data, and they can be practically implemented in several ways, such as node classification for a better understanding of attackers' interests and opinions, predicting relations between entities, or predicting a label for the entire graph when classifying the sequence of inputs into different entities. [10] To identify social engineering activities, GNNs can perform multi-modal detection analysis of devices, emails, and people.

Reinforcement learning is a type of machine learning that relies on trial and error. Unlike supervised learning, where all data is labeled, it depends on a system of rewards and penalties to guide the learning process. [11] An example is when an agent is placed in an AI-simulated environment of social engineering attacks and performs countermeasures to build a resilient defense.

Explainable AI is not a new field. It focuses on creating artificial intelligence models that humans can understand. It involves developing expert decision systems that model rules within a conceptual context. The goal is to create a mental model that leads to better performance and trust. [12] One example of this is helping to explain potential human attacks and evaluating ethical considerations.

4. EXPERIMENT DETECTING EMOTIONAL STATE

This experiment aims to explore the potential of artificial intelligence in identifying human emotions from text, which can be crucial in detecting potential attackers who may use emotional manipulation in social engineering attacks. To conduct this experiment, we utilized a standard dataset of Twitter communication posts, containing 416,809 different text sentences classified into five different emotional categories. The dataset was sourced from Kaggle, appropriately formatted to show the text of communication, the type of emotions, and corresponding labels that translate emotions into numerical values, thus enabling us to successfully perform the experiment.

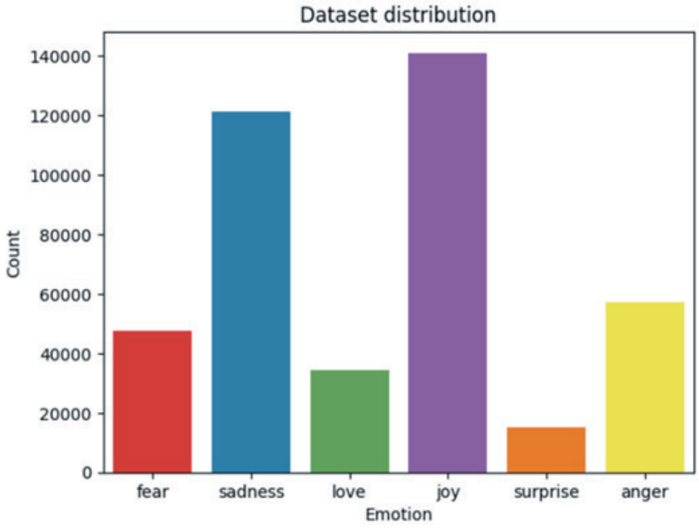

Figure 1. Emotion distribution.

The dataset's emotional distribution displays the frequency of various emotional states. We observed that the most common emotion is joy, represented by violet, in approximately 140,000 text sequences. The next most frequent emotion is sadness, marked in blue and present in around 120,000 text sequences. Anger is represented by the color yellow and appears in approximately 60,000 posts. Fear is marked in red and present in around 50,000 data rows, love in around 30,000 rows, and surprise in around 20,000 rows. We prepared the data using libraries in the Python environment, which helps analyze and manipulate data in datasets.

We used four machine learning algorithms to train our data: XGBoost, KNN, Naïve Bayes, and Random Forest. We used a 70/30 ratio with a test size of 0.3 to split our data into training and testing sets. Initially, we created a feature vector of 5,000 features and then transformed the train and test data. We used the random number generator seed with a parameter of 0 and a random state of 0. The XGBoost model instance was created with the parameter objective="multi: softmax" to predict more than two different classes, with five different classes predicted in this particular instance.

We used the same random state 42 to ensure repeatable decisions during the training and test data. The other algorithm train/test procedures followed the same steps. For the KNN algorithm, we set the parameter n_neighbors=5. The Naïve Bayes algorithm instance used the MultinomialNB() model. Finally, we implemented the Random Forest algorithm with n_estimators=100 and a reproducibility seed of 0.

5. RESULTS AND DISCUSSION

The XGBoost algorithm, also known as Extreme Gradient Boosting, is a machine learning algorithm that can easily handle classification and regression tasks. It has achieved an accuracy score of 89.32% in predicting values. One of its major advantages is its ability to handle large datasets effectively.

A confusion matrix diagram can be used to display the classification values of a model. The XGBoost model correctly predicted the following emotions: 32865 sadness, 37798 joy, 9738 love, 14919 anger, 12110 fear, and 4255 surprise.

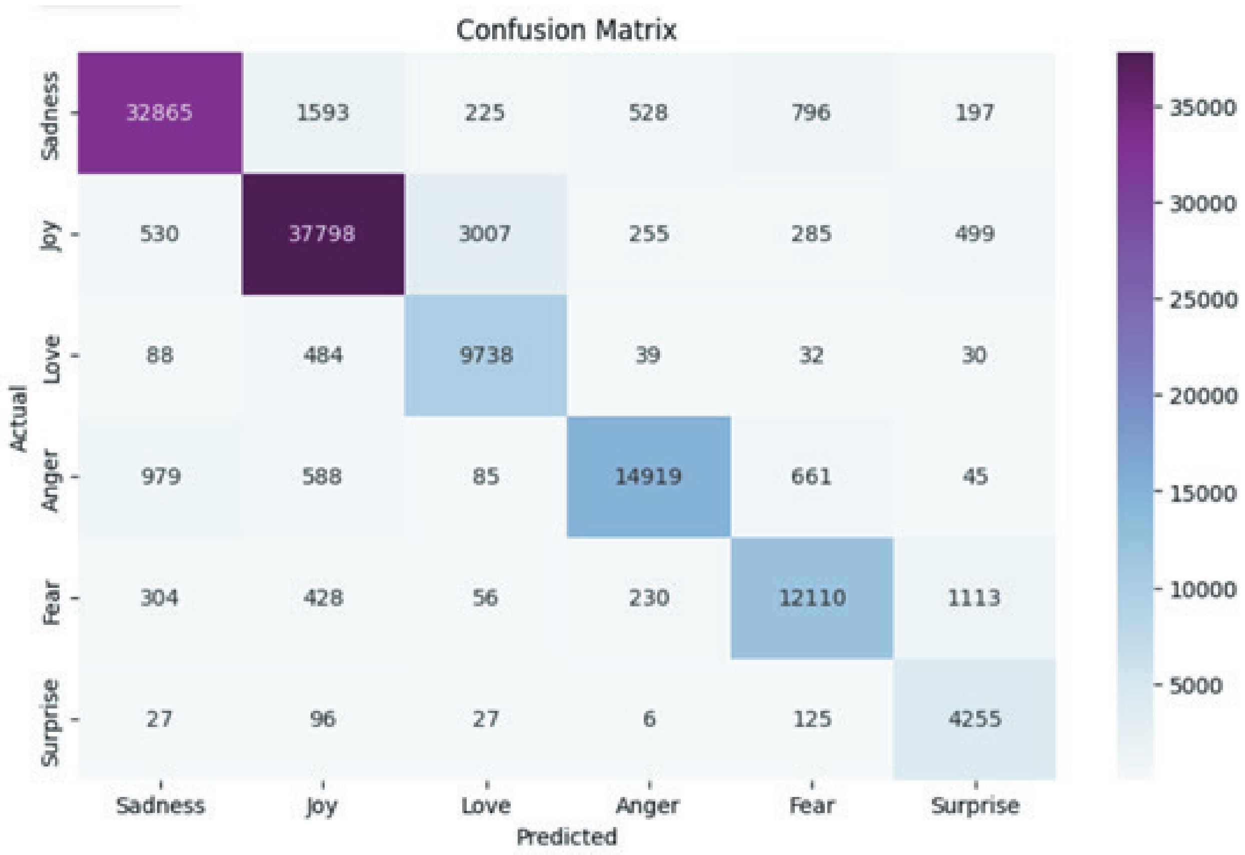

Figure 2. XGBoost algorithm prediction emotions.

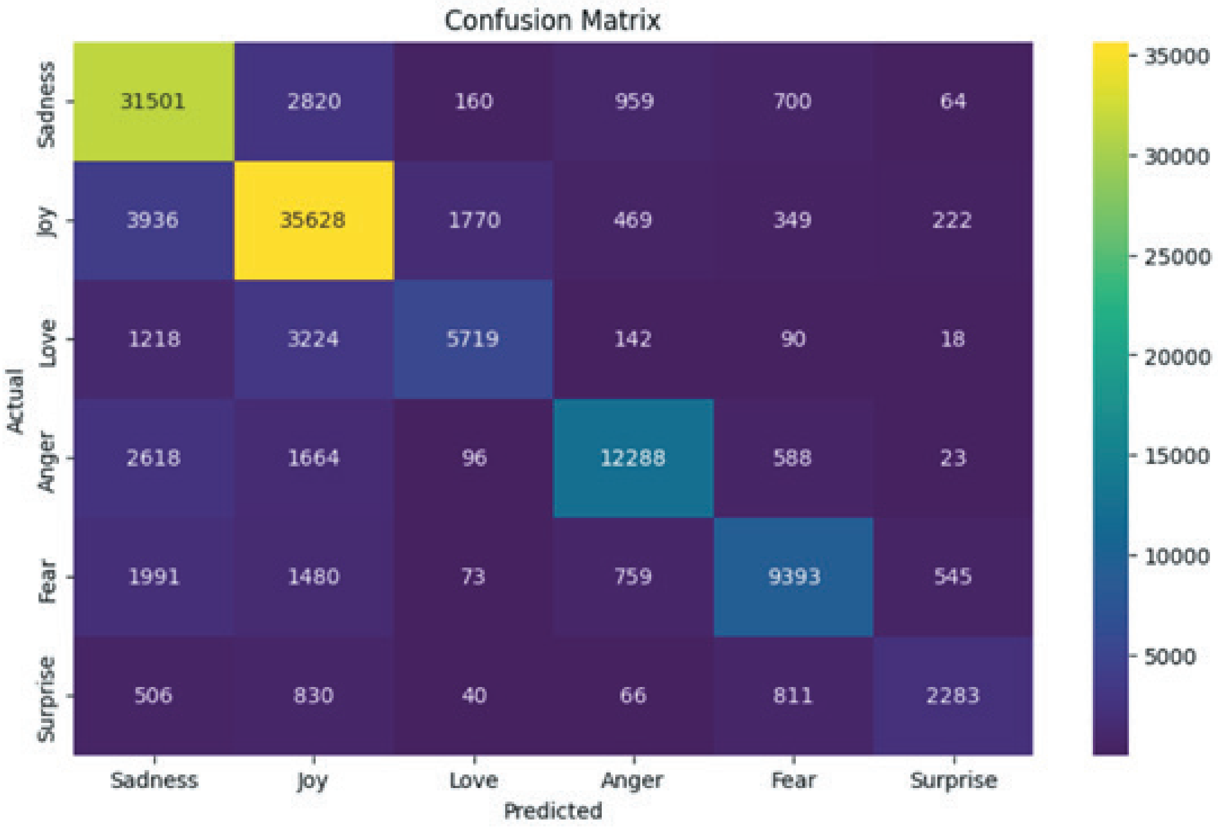

Figure 3. KNN algorithm prediction emotions.

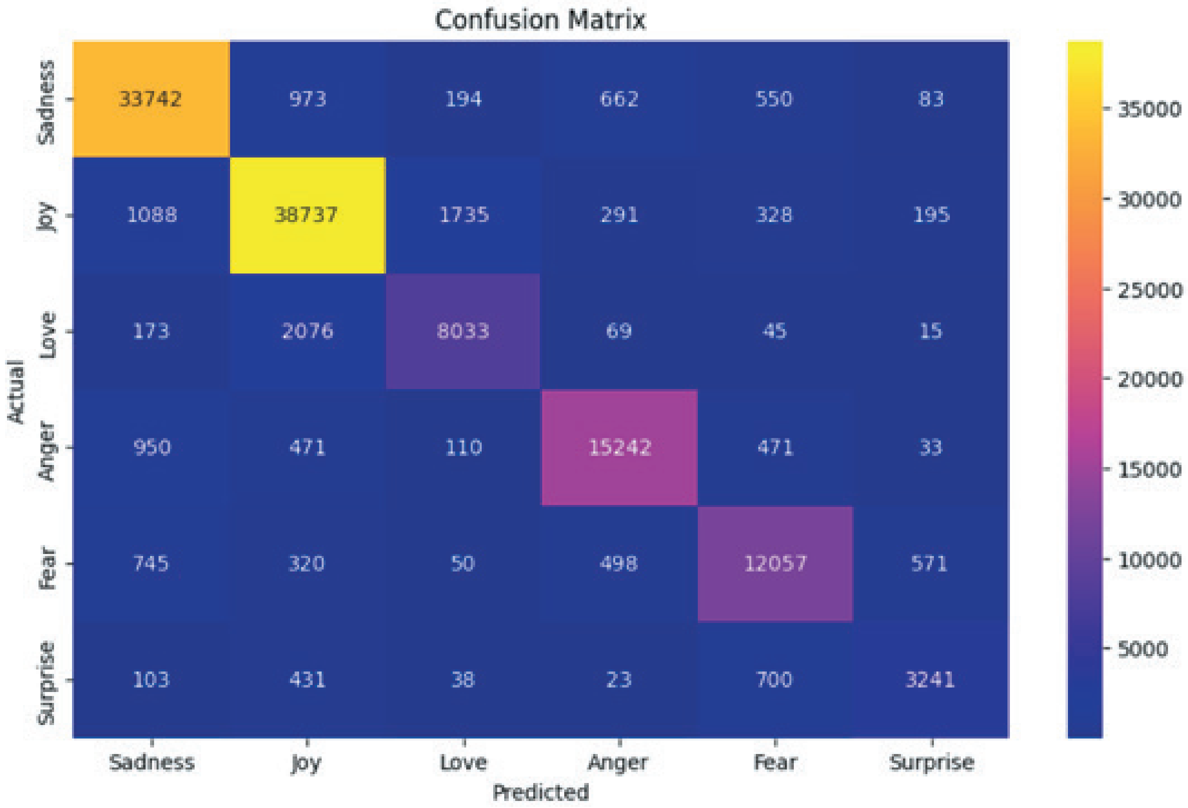

Figure 4. Naïve Bayes algorithm prediction emotions.

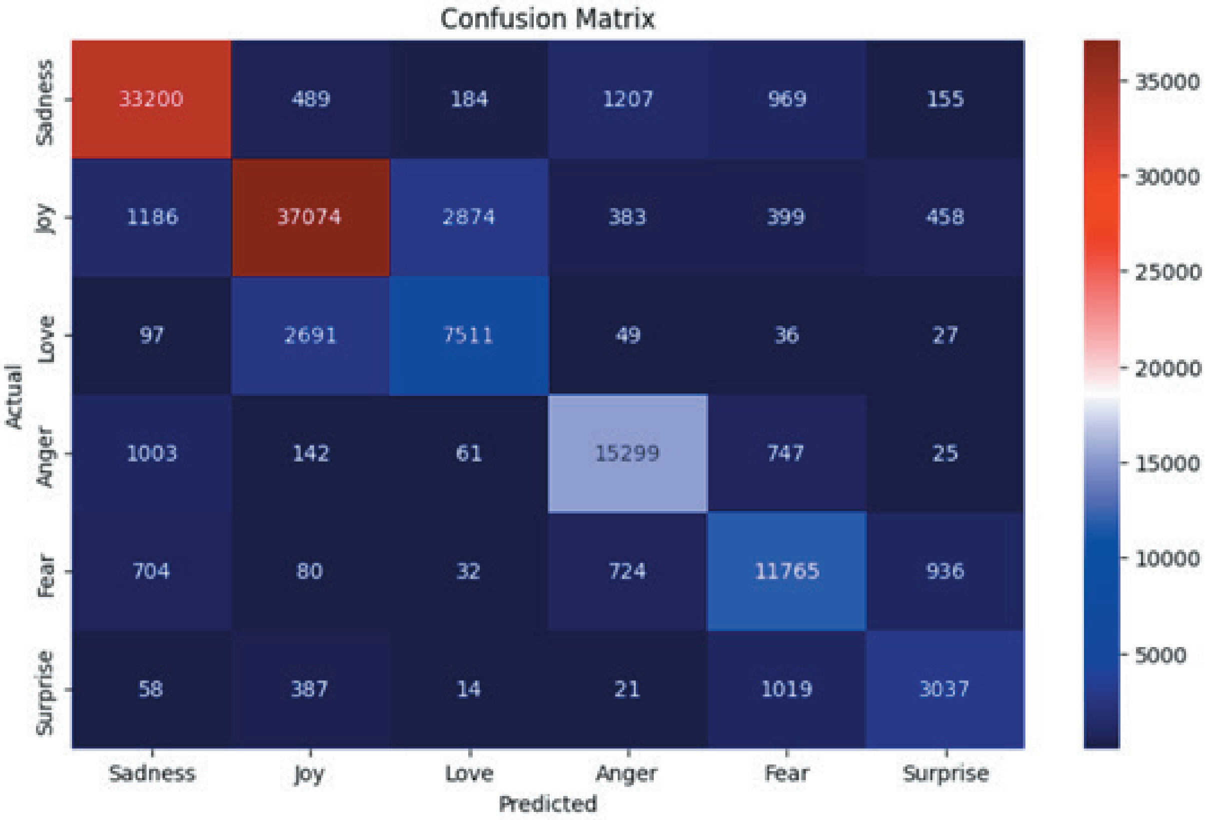

Figure 5. Random Forest algorithm prediction emotions.

The KNN algorithm is commonly used to identify the K nearest neighbors using the Euclidean distance. Analyzing the labels of the neighbors predicts the values or class classification. This algorithm has achieved an accuracy score of 77.42%, and its best advantage is simplicity and fast implementation. The confusion matrix classifies the model that predicted emotions of samples as follows: 31501 sadness, 35628 joy, 5719 love, 12288 anger, 9393 fear, and 2283 surprise.

The Naïve Bayes algorithm is widely used for classification tasks and has achieved an accuracy score of 88.81% in predicting values. The confusion matrix prediction reveals that the number of samples with correct predictions are as follows: 33742 for sadness, 38737 for joy, 8033 for love, 15242 for anger, 12057 for fear, and 3241 for surprise.

The Random Forest algorithm is used for both classification and regression tasks. It consists of a forest of multiple decision trees, where each tree is trained with random data and a random selection of features. The algorithm's predicted values have achieved an accuracy score of 86.28%. The confusion matrix shows that the following sets have been correctly predicted: 33200 sadness, 37074 joy, 7511 love, 15299 anger, 11765 fear, and 3037 surprise.

The current model has a limitation in that it can only take textual sequences as input, making it unable to predict emotions from image inputs. Therefore, implementing a multiple input data approach to overcome this limitation and involve artificial neural networks is considered. Future research can focus on sentiment detection from video/images or speech sequences. The OCEAN psychological model can represent personality traits based on the prominent five elements: openness, conscientiousness, extroversion, agreeableness, and neuroticism. We can improve their effectiveness by training AI/ML models to classify OCEAN personality traits based on psychological models. Moreover, future work could include a multi-modal approach to develop models considering social engineering behavioral aspects, including cognitive science.

6. CONCLUSION

Artificial Intelligence has become a crucial part of our daily lives, making social engineering attacks more sophisticated. The strength of any system is only as muscular as its weakest link, and humans are usually the most vulnerable link. There are differing opinions on the benefits and drawbacks of using AI, but it is neutral, like any other scientific approach. We have researched using AI and Machine Learning to develop models to prevent future attacks. We need to include cognitive processes and psychology to prevent manipulation. Limitations of the model are practical nature. How can we implement this solution if human communication needs to be observed in real time? Also, it is challenging how to predict an attack if it is not still happening. There is also tricky with user identification and privacy. Predicting based only on textual communication and emotional state can predict false potential attacks. Future work can involve other physical characteristics, such as image-face sequences, to identify emotional states from visual entities and predict potential attackers from the domain of provided image inputs. This approach can include emotion recognition from different sources, such as sequences of face images. It can be practically implemented as a Convolutional Neural Network (CNN) in Phyton language. From theoretical aspects, psychology can be used to gather five vital personal traits, including the ethical aspect of using emotion detection to prevent future social engineering attacks. As many inputs are involved in prediction, it could be a more precise real-life detection of social engineering attacks. Still, the critical question is how to implement the solution practically. Therefore, detecting human emotions can be crucial in identifying potential malicious intentions. Our research involved investigating a Twitter dataset to detect potential attackers' emotional states. The dataset included five different classes of emotions. We employed machine learning algorithms, such as XGBoost with an accuracy rate of 88.99%, Naïve Bayes with 88.81%, K-Nearest Neighbour with 77.42%, and Random Forest with an accuracy rate of 86.28%, to detect the emotional state of written communication. In the future, we can additionally include OCEAN psychology traits and cognitive science to develop Artificial Neural Networks and create a robust solution for future challenges of social engineering attacks.

7. REFERENCES

[1] M. Š. S. A. Aleksandar Jokić, "Next-generation Firewall and Artificial Intelligence," in *Sinteza 2023 - International Scientific Conference on Information Technology, Computer Science, and Data Science*, Belgrade, 2023.

[2] H. H. M. H. E. W. Katharina Krombholz, "Advanced Social Engineering Attacks," *Journal of Information Security and Applications*, vol. 22, pp. 113-122, 2015.

[3] N. K. Fatima Salahdine, "Social Engineering Attacks: A Survey," *Signal Processing for Next Generation Wireless Networks*, 2019.

[4] Z. S. U. A. M. R. S. M. A. I. Wenni Syafitri, "Social Engineering Attacks Prevention: A Systematic Literature Review," *IEEE Access*, vol. 10, 2022.

[5] A. D. Z. Z. G.-J. A. Huahong Tu, "SoK: Everyone Hates Robocalls: A Survey of Techniques Against Telephone Spam," in *IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, 2016.

[6] S. P. Kristian Beckers, "A Serious Game for Eliciting Social Engineering Security Requirements," in In *Proceedings of the International Requirements Engineering Conference*,pp 16-25, Beijing, China, 2016.

[7] Z. Feng, Formal Analysis for Natural Language Processing: A Handbook, Beijing: University of Science and Technology of China Press, 2023.

[8] Y. H. Liron Bergman, "Classification-Based Anomaly Detection for General Data," *arxiv.org*, 2020.

[9] A. P. Y. H. Liam Hiley, "Explainable deep learning for video recognition tasks: A framework & recommendations," *arxiv.org*, 2019.

[10] G. C. S. H. Z. Z. C. Y. Z. L. L. W. C. L. M. S. Jie Zhou, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57-81, 2020.

[11] Y. Li, "Deep Reinforcement Learning: An Overview," *arxiv.org*, 2017.

[12] H. U. Y. D. W. F. D. Z. a. J. Z. Feiyu Xu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *NLPCC 2019: Natural Language Processing and Chinese Computing pp 563–574*, China, 2019.