ISSN: 2182-2069 / E-ISSN: 2182-2077

Analyzing Social Engineering Attack Patterns Using Behavioral Psychology and AI-Driven Defense Mechanisms

Akash Parasumanna Sridhar[1*]

[1*]IT Cybersecurity Analyst, Campbell Clinic, United States of America.
akash2kparas@gmail.com, https://orcid.org/0009-0005-3917-458X

Received: December 27, 2024; Revised: January 30, 2025; Accepted: February 10, 2025; Published: February 28, 2025

Abstract

Complex Social Engineering (SE) attacks by manipulating people psychologically to mislead them and to weaken the security systems have become quite common. This research investigates the patterns of SE attacks through concepts of behavioral psychology integrated with protection frameworks driven by AI (Artificial Intelligence). This study evaluates significant weaknesses in the behaviour of human beings which hackers use against them by analysing the strategies of attackers such as gathering data, phishing techniques and exploitation of trust. We provide an enhanced protection architecture that includes BERT (Bidirectional Encoder Representations from Transformers) to deal with such attacks. Through the evaluation of text semantics in messages, emails, and websites, BERT's deep learning (DL) ability helps them detect phishing content, suspicious patterns of language, and fraudulent messages. Our approach decreases false positives and enhances contextual knowledge thereby improving traditional models used for detection. Experimental outcomes prove that BERT is more accurate in identifying harmful content than conventional Machine Learning (ML) approaches. This research focuses on the fact that the knowledge of behavioral psychology and AI-driven approaches can make cybersecurity systems more efficient and decrease the risk of SE attacks.

Keywords: Engineering, Fraudulent, Detection, Cybersecurity, Cyber-attacks, Behavioural Psychology, AI.

1 Introduction

An increase in the emergence of SE attacks based on behavioral psychology through AI-powered techniques has demanded the need to defend against such attacks. Industries can enhance their cybersecurity frameworks by comprehending similar patterns of attacks and then using AI to identify and reduce those patterns (Caviglione et al., 2021). When it comes to SE, the main goal of attackers is to manipulate the behavior of humans to access their private data in an unauthorized manner (Meurs et al., 2022). This technique manipulates the staff and clients by threatening them with their private data and using their behavior and flaws against them. The attackers utilize this information to access accounts or networks based on essential human characteristics like trust and cooperation. Similar to human flaws, the term human vulnerability can cause harm to an individual or an industry. A weak point can result in the collapse of a whole system, team, or program's integrity (Do et al., 2019). SE is still a weak point in cyber security, offering cyber-attackers a chance to use people's flaws against them to access accounts unauthorizedly. In the domains that involve cybersecurity and data security, the methodology of

Journal of Internet Services and Information Security (JISIS), volume: 15, number: 1 (February), pp. 502-519.
DOI: *10.58346/JISIS.2025.I1.033*

*Corresponding author: IT Cybersecurity Analyst, Campbell Clinic, United States of America.

"Weakest link" is critical. Even security systems that are well-defended and highly intricate have the possibility of getting breached by intelligent malware and trained hackers. This is achieved by spotting and using the weakest points as a defense. SE attacks can also occur disguised, where they pretend to make genuine access, and even intricate security systems often find it difficult to protect themselves against such attacks. In the digital environment, individuals are more likely to get attacked as their online activities on social media become an easy target to exploit. Cybercriminals often lead consumers to fraudulent web pages or accounts to make them use their accounts or networks and then take advantage of them (Gorment et al., 2023).

"SE" is a term used in the field of cybersecurity. It means manipulating individuals into exposing their private data or doing harmful activities like data breaches or illegally accessing accounts (Sood et al., 2017). SE is a concept that has become an integral part of cybersecurity. These attacks use the weaknesses of individuals more than the vulnerabilities in a system, using psychological and behavioral characteristics as a tool for manipulation. The concepts and tools used in SE attacks have become more intricated in nature over the past years, competing with the complexities of our digital architecture (Gao et al., 2016). Malefactors often find it easy to target human characteristics who are prone to manipulation through methods like intimidation, deceit, and impersonation. This is because networks and digital frameworks are growing to be durable and highly secure during attacks (Fung, 2011). Modern SE approaches have developed from traditional phishing methods to highly complex schemes like voice impersonation, spear phishing, and baiting, which use social media and virtual meeting applications (Thanh & Zelinka, 2019). The world of digital technology is expanding at an exponential rate. The Internet is becoming increasingly popular as a new medium for communication, commerce, information, and entertainment. On the other hand, this paradigm shift is giving rise to significant worries over users' privacy and security when they are online. Despite stringent security measures, many attacks on the internet are carried out by taking advantage of flaws in application design, engaging in fraudulent activities, or employing sophisticated technical approaches (Udayakumar et al., 2023; Siddiqi et al., 2016; Abroshan et al., 2021). Scamming is one of these assault strategies that has been around for a long time, even before things like computers and the internet came into existence. Both phishing and scamming fall under the category of social engineering (SE) assaults, which are classified in cybersecurity. When it comes to attacks, those that include taking advantage of human vulnerabilities are referred to as SE attacks (Wang et al., 2021). SE attacks exploit human vulnerabilities, such as deceit, persuasion, manipulation, or influence (Albladi & Weir, 2020). Social media and smishing attacks have been the most prominent forms of attack utilized for social engineering attacks during the past two years (Abdulrahman et al., 2023). Every single one of the cyberattacks that are based on SE is dependent on the actual interactions that take place between the attacker and the victim. In some instances, an individual may impersonate an employee to obtain information, such as a password or a PIN code, using a simple phone call. This type of attack is known as a social engineering attack (Truecaller, 2021). In 2020, the United States of America lost around USD 29.8 billion due to phone scams.

Social Engineering Attacks

Instead of a complex hacking technique, social engineering is a procedure that uses people's psychology (Govindankutty, 2021). A critical tool for attackers, SE is becoming more important as our reliance on technology grows. Conversely, technological countermeasures are improving as the number of cyberattacks rises (Siddiqi & Pak, 2020). Technical attacks are becoming more challenging due to the constant enhancement of security systems. However, SE is producing excellent results when used to launch cyberattacks. Attackers can infiltrate organizational networks, bypass firewalls, infect computers

with malware, open back doors in the organization network, and more using cyberattacks based on SE. Cyberattacks using SE can exploit or influence various forms of human behavior. For example, an attack can be launched due to human error (Raj & Dharmaraj, 2024). By interfering with a person's ability to make a decision, attackers can push them to make a mistake. The report delves into the specifics of the factors that impact decision-making later on. Cyberattacks based on SE can also take advantage of a number of other human traits (Human Cyber Risk—The First Line of Defense, 2021). The most prevalent SE-based cyberattacks are covered in this section. Figure 1 illustrates the classification of social engineering attacks.

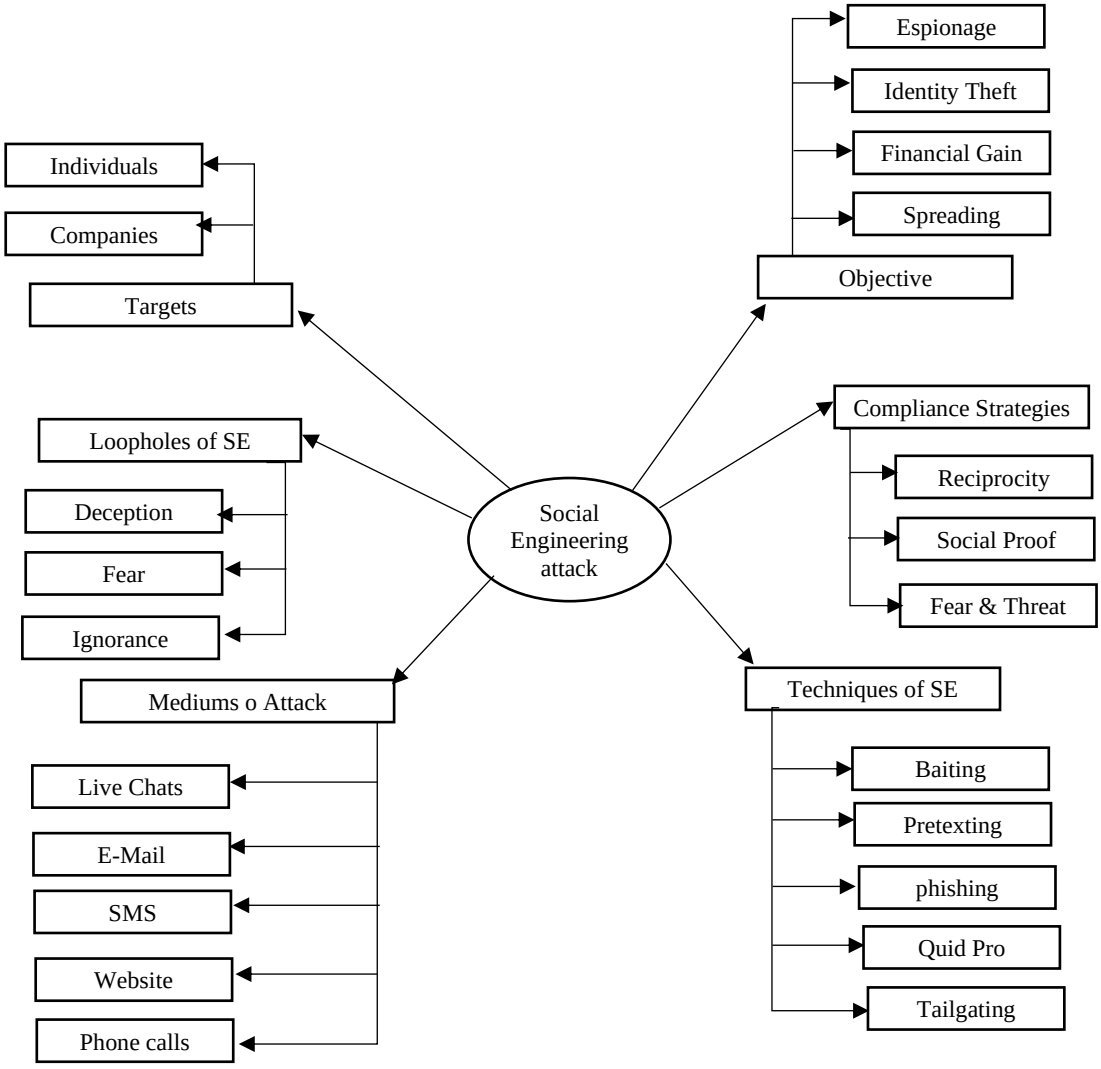

Figure 1: Classification of Social Engineering Attack

Objectives of the Study

- To study the effects of social engineering and other forms of cyberattack on human decision-making and behavior.
- To learn more about the social and psychological elements that put people and businesses at risk of falling for social engineering and phishing schemes.

- To assess how effective cybersecurity education and awareness programs have been in reducing human error and strengthening defenses against cyberattacks.
- To learn how social engineering and attack techniques have evolved and what effect they have had on cybersecurity throughout time.
- To help individuals and organizations decrease the success rates of assaults and social engineering by suggesting behavior-focused solutions and best practices.

2 Literature Survey

Cybercriminals use social engineering attack techniques, which are complex tactics, to breach targeted accounts by taking advantage of human weaknesses instead of technological ones. These approaches are highly efficient as they deceive people by offering too much information or enabling hackers to access their networks. The attackers may disguise themselves as trustable organizations or use psychological techniques like threatening to deceive them. SE is one of the most harmful approaches in the field of cybersecurity. It uses human emotions for manipulation and can be dangerous for industries and individuals. An approach used in SE attacks is phishing, where cyber-attackers send deceiving messages or emails as if they are sent from trusted organizations such as service providers or banks. Abdulrhman et al. (2023) suggest that emails such as these trick people into offering crucial data like credit card details or passwords or make them access links that may lead them to dangerous websites or download harmful software. Phishing is used to acquire sensitive data or to unauthorizedly access users' accounts by disguising themselves as trusted organizations.

SE attempts to manipulate the behavior of individuals for breaching security rather than carrying out intricate hacking approaches. Technical cyberattacks have become more challenging to execute as the intricacy of such attacks is higher due to technological enhancements. However, SE utilizes the weaknesses of individuals and manages to get past highly intricate security mechanisms successfully. With this strategy, intruders can easily install malware and firewalls, access networks, and create backdoors. Social engineers utilize human flaws and cognitive biases to access individuals' private data rather than depending on system vulnerabilities. SE is a fundamental approach to acquiring information about a subject using human vulnerability in an organization. Hackers use this approach to deceive people into providing their sensitive data, which is used against them for security breaches (Yash) As a result of this, human vulnerability is considered to be a crucial risk in cybersecurity. Despite a bad reputation, SE is highly efficient as it makes use of individuals' weaknesses. Though security mechanisms enhance system defense, the weakest points are still considered human flaws. SE is a practical approach amidst existing cyberattacks, which deceives humans and provides access to unauthorized people.

Human vulnerabilities are growing to be the primary reason for cybersecurity issues. However, it is evident from research that human characteristics play a crucial role in cybersecurity issues, especially the ones that involve phishing and SE. Humans are the masterminds of system attacks, and when making decisions, they take cognitive biases into account (Hadnagy, 2010). We need to learn these patterns of behavior so we can build good defenses. Phishing is one of the most prevalent methods of taking advantage of people's weaknesses. Research has shown that phishing tricks people into divulging critical information by using psychological manipulation techniques such fear, authority, and urgency (Jampen et al., 2020) Phishing emails are effective when well-designed, sent at the right time, and the recipient believes the sender is legitimate (Arachchilage & Love, 2013) Vulnerability in humans is an integral part of cybersecurity since people are prone to making mistakes, being manipulative, or being exploited,

all of which can result in security breaches (Demertzi et al., 2023). Even if systems and technology are getting more complex and resilient, humans are still a weak point in protection against cyberattacks.

- One typical flaw is a lack of knowledge about cybersecurity basics, which leaves many people vulnerable to social engineering attacks like phishing emails and viruses.
- Confidence: Cyber-attackers disguise themselves as trusted entities like government organizations or banks to access private data.
- A sense of fear and urgency: Cyber-attackers deceive users into making careless decisions without checking their authenticity by threatening to deactivate their accounts or take legal action.
- Inquisitiveness: Suspicious attachments or URLs with captivating titles like "Breaking news" or "Confidential Information" acquire people's attention and mislead them.
- Fifthly, carelessness can allow attackers to get unauthorized access. Password sharing, unattended devices, and missing software upgrades are all instances of this.
- Account Reuse or Use of a Simple, Predictable Password: This makes brute force or credential-stuffing assaults much more manageable for attackers.
- Cognitive Biases are tendencies to make irrational decisions all the time. Simplifying complicated security information causes people to develop these biases, which in turn cause them to make mistakes in assessing threats, evaluating risks, and developing response tactics.

Currently, social engineering attacks rank high among cybersecurity concerns (Arana, 2017; Chargo, 2018; Libicki, 2018; Costantino et al., 2018; Pavković & Perkov, 2011; Breda et al., 2017). You can find them, but you can't stop them, say the writers of (Libicki, 2018). To gain sensitive information, social engineers exploit their victims. This information can be utilized for specific reasons or sold on the dark web and black market. With the rise of big data, criminals can use companies' valuable data for their gain (Atwell et al., 2016). As commodities in today's markets, they bundle massive amounts of data for sale in bulk (Mahmood & Afzal, 2013). While every social engineering attack is unique, they all follow a similar pattern and go through the same phases. The typical procedure consists of four steps: (1) research the target; (2) establish rapport with the target; (3) use the data to launch an attack; and (4) disappear without a trace (Mouton et al., 2016). During research, often known as information gathering, the assailant chooses a victim according to specific criteria. In the hook phase, the attacker initiates contact with the victim, either in person or via email, to build their trust. Within the play phase, the perpetrator uses emotional manipulation to coerce the victim into divulging critical information or making security-related errors. When an attacker exits the out phase, they do so without leaving any evidence behind.

3 Methodology

Problem Statement

This research takes consideration of the usual suspects in attacker-engineered user attacks the attack approach, strategy, detection, etc but ultimately, it is up to humans to identify these aspects (Nejad & Shahriary, 2017). Thus, the danger arises when these assaults are combined with human involvement during attack reinforcement. This study presents a novel quantitative way to evaluating each attack node statistically using the standard notation in Table 1. It incorporates human traits, which are ignored by existing social engineering attack defense models.

Table 1: Frequently Used Symbols

Notation	Definition
X	Game model players
W	Is a constant
A	The game model's attack actions
S	The game models state space
R	The game model attacker's reward
D	The game model Defensive actions
a	The strategy of optimal attack
R	The game model defender's reward
8	The successful attack probability
d'	The strategy for optimal defensive
R (s. a.d)	The immediate reward
U (a,d)	The game model utility function
Q(s,a,d)	The Q-learning algorithms Q function
E (U')	In the next state of expected utility
y	Discount factor
a	Rate of learning

Physical attributes and target attributes are the two main categories into which node attributes fall. The system's nodes' impact sizes are primary considerations in a physical analysis. The degree of importance and the number of connections within a node are two of its physical properties. The target, character, security awareness, and knowledge of the target attributes make up the bulk of the node's target attributes. The robustness of the security defense is proportional to these characteristics.

Definition 1: Level of importance and level of connections are the primary physical qualities of a node. One measure of a node's significance within a social engineering system is its importance level (IL). The validity of the information gathered, its influence on future attacks, and the degree to which trust is developed are the three main components of this significance level (Yang et al., 2022). The number of other nodes connected to a node and the stage of the social engineering model in which it is located determine its connection level (CL), representing the relevance of the node's associations in the system.

Definition 2: Defensive technology, defensive means, basic target information, level of security knowledge, personality type, security awareness registration, and cognitive routes are the primary components of target attributes (TA) (Liu et al., 2020). Increasing the value of the target's attributes enhances node security and the target's defensive capabilities. The assigned protection is boosted in proportion to the node's high physical attributes.

The system administrator sets the values of the physical and target attributes, which are then assigned to a vector F ∈ [1, 2, 3]. The SESM, which stands for "social engineering security metric based on attributes," is

$$SESM = w * \log_2(1 + \frac{ILGL}{TA})$$

There are three main categories into which the social engineering security system is rated: high, medium, and poor. As a result, F belongs to the set [1, 2, 3]. The SESM, which stands for "social engineering security metric based on attributes," is

A constant, w, is used in the above equation. In the context of the node's state and action spaces, this constant is defined as $w = A + D/S$. A more considerable SESM value indicates a more critical node,

a more significant attack loss, and a lesser defense capacity. Social engineering threats can be seen by looking at the SESM value, which also displays how each node affects itself. The social engineering attacker in this study uses a time-based function to launch attacks, which include consuming resources for attacks, defenses, and loss recovery, by mainly taking advantage of the node's weaknesses. The time and resources consumed by an attack during its lifecycle, which includes tasks such as planning, carrying out, gathering target information, scripting, and building confidence, are referred to as attack resource consumption (AR) (Mouton et al., 2014). Defense resource (DR) consumption: time spent detecting attacks, determining when an attack is an attempt to collect information, and resources used to resist social engineering attacks (Ghasempour, 2019). Loss recovery consumption (LR) refers to the time needed to get back on one's feet after an attack, whether that's resetting one's password or replacing a secret key (Mayfield et al., 2019).

Proposed Attack Model

The social engineering concept that relies on human vulnerabilities allows an intruder to gain sensitive information through social engineering attacks. A more considerable vulnerability means more damage and more social engineering gain for the attacker. Protecting yourself from social engineering entails paying close attention to the countermeasures in response to such attacks. It would be perfect if the target was secure and had no exploitable weaknesses. Attackers can always uncover new vulnerabilities to exploit, targets have different attributes, vulnerability performance can vary, and defenses can be slow or even fail during an attack. Like two players on opposite sides of a game, an attacker seeks to maximize reward while a defender aims to minimize loss. It is possible to examine the optimal defensive strategy for the defender by applying a stochastic game model to the two sides of the attacker and defender. This work presents a novel approach to reward quantification that takes into account target features and the interplay between vulnerability and assault to quantify the impact of vulnerability in the model. You can see the model of the social engineering attack in Figure 2 down below.

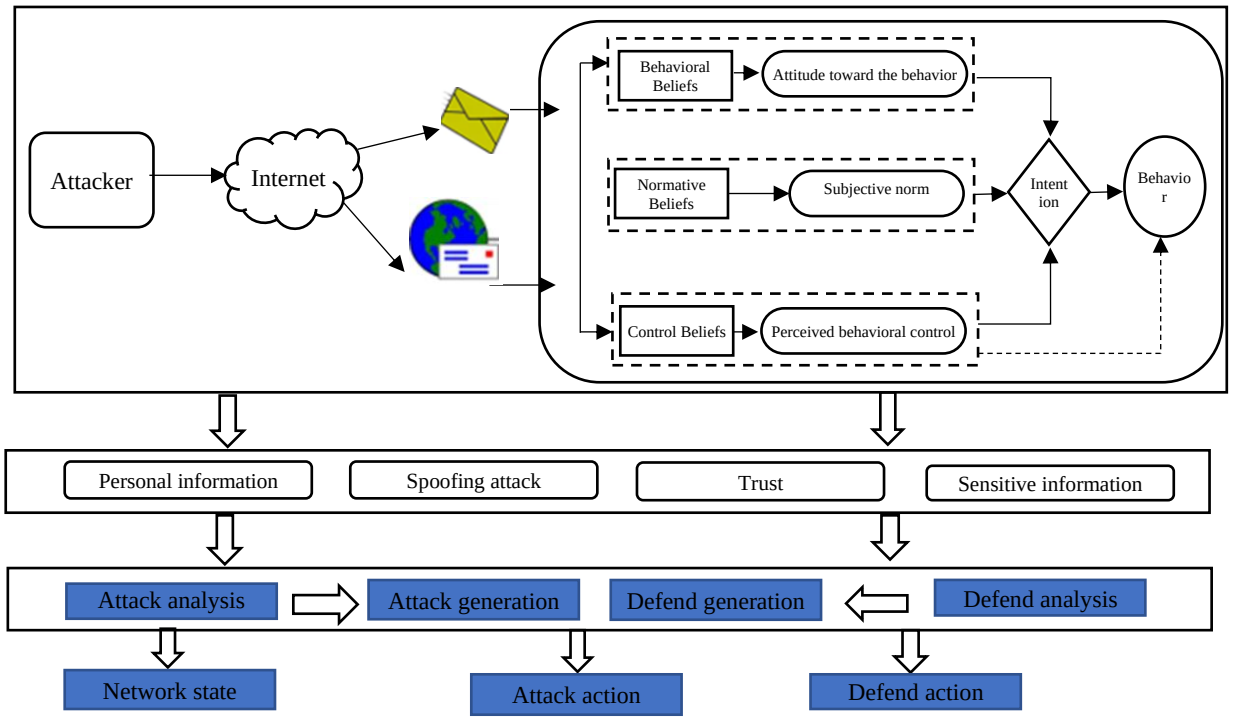

Figure 2: Social Engineering Attack Model Diagram

One paradigm for predicting people's actions is the theory of planned behavior (TPB). Figure 2 (Rishikesh et al., 2022) displays the TPB model. How a person feels about an action determines the variables that motivate them to perform in a certain way. For instance, a social media user may engage in an activity that puts their privacy at serious risk, even if they are unwilling to disclose personal information. The term "subjective norm" describes how a person's social group impacts their behavior. Belonging to a group can influence how an individual acts. A person's "perceived behavior control" reflects how much agency they feel over a given action. Based on the behavioral stimuli, the three branches of TPB might be utilized to induce the victim to divulge information. As illustrated in Figure 3, the social engineering system is primarily designed to mimic phishing attempts (Yang et al., 2022). The four phases of an assault are planning, preparing, executing, and gaining ground. An attack's preparation includes queries on information systems, collecting user data, and building scripts. Phishing websites, emails, and text messages are all examples of attack route nodes that must be located before an attack target can be selected. As part of the attack implementation stage, you will influence the target, execute psychological and script exploitation, evaluate participant behavior, and develop trust. Acquiring device permissions, sensitive information, and influencing target behavior are the main components of the attack gain stage.

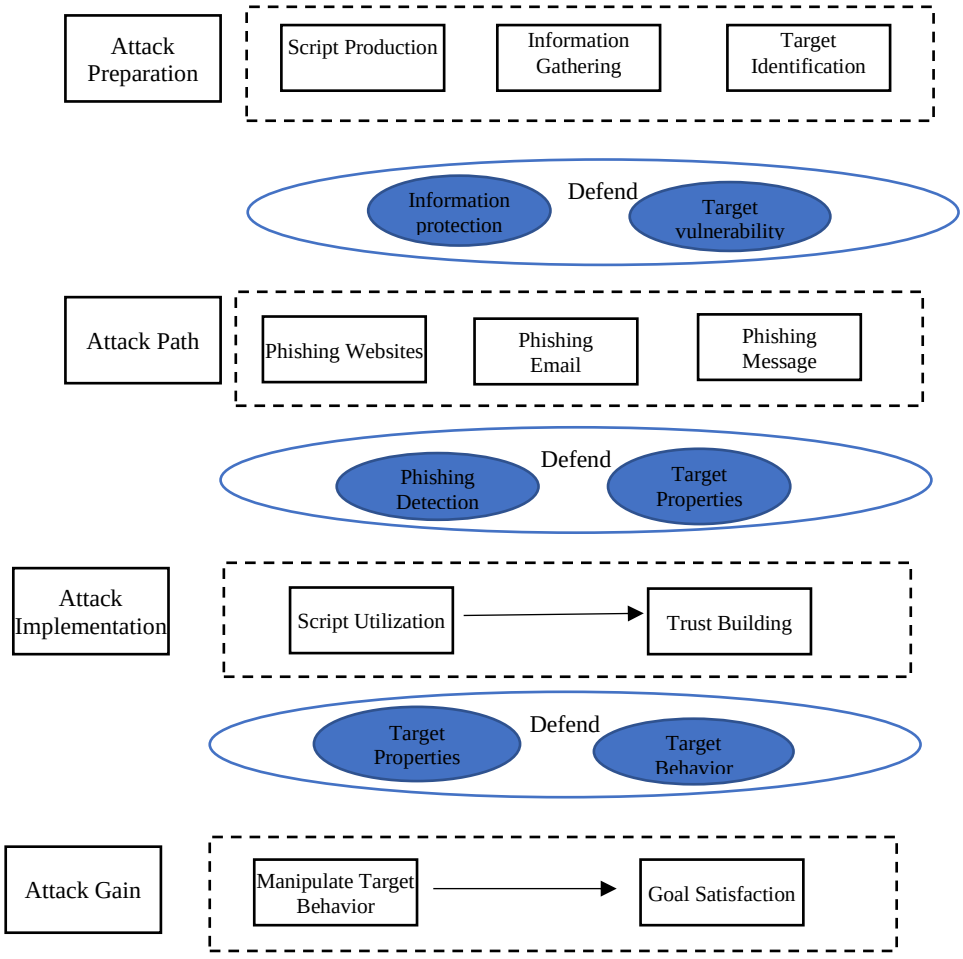

Figure 3: Attack Lifecycle and Defense Mechanisms

With its state-of-the-art performance, BERT has taken on numerous natural language processing tasks, including text classification, autocomplete/autosuggest, question answering, and more. Quicker

development, lower data requirements, and better results are further benefits of using BERT (Devlin et al., 2019). Here are the steps used to fine-tune BERT for text classification: When training the BERT model, it is necessary to provide data with a specific structure. The first step is to tokenize the data, which consists of review text. The BERT-based uncased model for tokenization is utilized in this research.

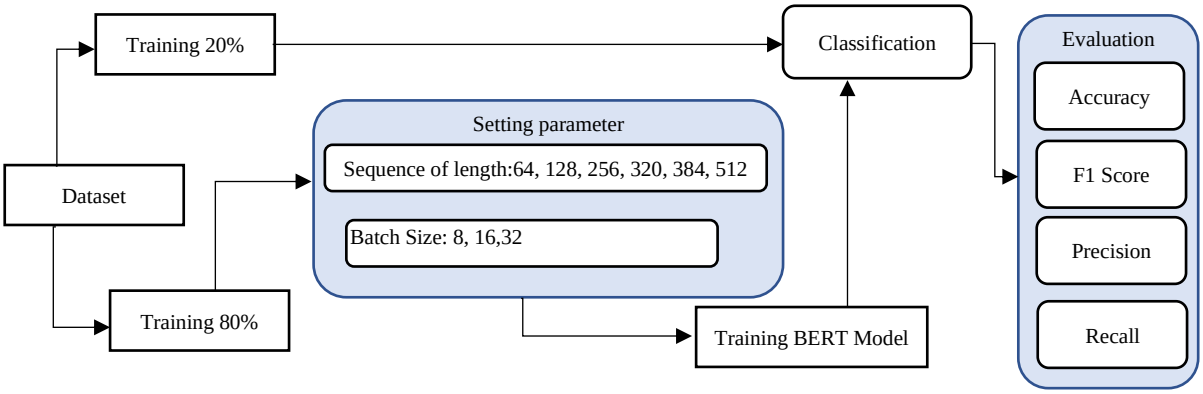

Figure 4: Hyperparameter Settings and Evaluation Metrics for Model Training

The Yelp dataset is split between training and testing, with 80% going to training and 20% to testing. Two thousand samples were used for testing, and 8000 samples were used for training. After that, we split the training dataset in half, using 7200 samples for training and 800 samples for testing, for a total of 90% and 10%, respectively. A total of 768 hidden sizes, 12 self-attention heads, and 12 transformer blocks make up the BERT model. Here, we use Google Colab to hone BERT's performance on categorization tasks. Sequences of 64, 128, 256, and 320 characters are used to construct BERT classification models, with 32 being the default for both the training and testing batches. Whether it's being used for training or testing, the BERT classification model is built with a batch size of 16 or 32 and a sequence of length 384. The BERT classification is also constructed with a 512-byte sequence and a batch size of 8 or 32 for testing and training. For BERT tuning, the researchers recommend 16 and 32 batch sizes (Devlin et al., 2019). The BERT model's suggested methodology is shown in Figure 4. To determine how effective each training iteration was, researchers used a validation split to calculate validation loss and accuracy. Using different sequence lengths, this work fine-tunes several classifiers.

4 Experimental Setup

The Python backend is responsible for developing the framework, which uses many Python libraries such as sci-kit, Sklearn, Numpy, Pandas, NLTK, and more. Using sci-kit learns train_test_split functionality; the dataset was partitioned into three parts: training, validation, and testing. Seventy percent of the data was used for training, fifteen percent for validation, and fifteen percent for testing. During data pre-processing, the NLTK tool was used to eliminate HTML tags, punctuation, multiple spaces, and other unnecessary elements. The fine-tuned embeddings on our dataset are obtained using the hugging face's distil-Bert-base-uncased model. For this data, we use an encoding scheme with a maximum length of 512, dimensions of 768, a dropout of 0.1, and 6 layers to obtain the input and mask identifiers. We have employed Keras's LSTM, Bidirectional, Conv1D, and Dense layers for classification purposes. Experiments informed the decision of how many units to use for each stratum. We tested the GloVe model with 100 and 300 vector dimensions and found that 100 produced the most

accurate results. The models were identical with respect to the loss function (cross-entropy) and the activation function (sigmoid).

Performance Metrics

The effectiveness of our model has been assessed using the following metrics: ROC, Accuracy, Precision, Recall, F1, and the Confusion matrix (Hossin & Sulaiman, 2015).

- An information-gathering tool for classifiers, a confusion matrix compares actual and expected classifications.
- The accuracy of a test is defined as the percentage of valid results relative to the total number of cases tested.
- Accuracy: Accuracy reveals the percentage of expected positive results that are positive.
- Memory: It reveals the percentage of true positives that are accurately rated.
- The F1 Score illustrates the symbiotic relationship between recall and precision.
- Receiver Operating Characteristic (ROC): ROC shows how effectively the positive and negative groups are separated by percentage.

Using these indicators, we were able to evaluate our model's outcomes. Table 2 shows the values of the following metrics: training, testing accuracies, validation, recall, precision, F1, and ROC.

Table 2: Results from Several Artificial Intelligence Models

	Models	Precision (%)	F1 (%)	Accuracy (%)	Recall (%)
Tokenizer	LSTM	85.1	86.9	86.6	88.7
	Bi-LSTM	84.9	85,5	85.4	86.2
	CNN	93.0	93.01	93.0	93.0
GloVe embeddings	LSTM	91.7	92.2	92.1	92.7
	Bi-LSTM	90.2	92.1	91.9	93.9
	CNN	91.6	90.9	91	90.2
GloVe embeddings and attention mechanism	LSTM	91.7	92.2	92.1	92.7
	Bi-LSTM	90.2	92.1	91.9	93.9
	CNN	91.6	90.9	91	90.2
BERT embeddings	LSTM	91.01	91.01	91.16	91.01
	Bi-LSTM	88.76	93.3	93.05	88.76
	CNN	94.11	95.31	95.32	94.11

5 Results and Discussion

Adding a new component called "stance" to the false news recognition algorithm is why we've suggested doing this in this effort. By analyzing the article's stance, we can learn how the title (the news headline) relates to the body (the news text). This characteristic, when combined with our content features derived from the pre-trained BERT model, gives us a deeper understanding of the piece. We have examined CNN, Bidirectional LSTM, ANN, and LSTM as AI models. We train and test these models using various vector representation techniques to get the results.

In some cases, we even add an attention layer to the mix. In Table 2 you can see the outcomes. Our proposed model's classification results are displayed in Table 3.

Table 3: Outcome of Proposed Classification Model

Models	Training accuracy (%)	Testing accuracy (%)	Validation accuracy (%)	Recall (%)	Precision (%)	ROC (%)	F1 (%)
LSTM	98.51	91.16	91.16	91.01	91.01	91.15	91.01
ANN	91.52	91.85	91.86	91.98	91.98	91.85	91.69
CNN	99.96	95.32	95.33	94.11	94.11	95.33	95.31
Bi-LSTM	98.48	93.05	93.06	88.76	88.76	93.47	93.3

As we used a pre-trained language model, we got the best results. Table 3 displays the accuracies with various parameters; for example, using GloVe Embeddings allows us to get a 92.1% LSTM model correctness. The accuracy improves by approximately 1% after these models incorporate the Attention layer. Our suggested model, which modifies the BERT embedding to extract article context, outperformed the competition with a 95.32% success rate. We conducted experiments using both the pre-trained and fine-tuned BERT models and received the findings. There is hardly any discernible difference in the outcomes from these two models. This could be because BERT was trained using an English language corpus, among other things. Our experimental dataset is similarly structured, features-rich, and likewise in English. We performed studies to demonstrate the impact of stance as a feature. We demonstrated that BERT encoding, which generates a representation of the news title and news body together with their similarity, outperforms encoding just the news title and news body. Table 4 displays the results, which show a significant improvement in testing accuracy.

Table 4: Classification of News Articles in Effectiveness of Stance Feature

Features	Models	Training accuracy (%)	Validation accuracy (%)	Testing accuracy (%)	Precision (%)	F1 (%)	ROC (%)	Recall (%)
News Title, News Body	Bi-LSTM	97.99	89.79	92.21	93.6	52.0	922:	93.6
	ANN	89.2	88.0	88.33	86.57	88.61	88.41	86.57
	LSTM	95.29	88.8	90.64	87.42	91.0	90.9	87.42
	CNN	99.1	93.68	93.90	91.3	94.0	94.0	91.3
News Title, News Body, Similarity between them (Stance)	Bi-LSTM	98.6	92.11	92.6	93.5	92,5	92 G	93.5
	ANN	89.31	89.37	89.38	56.4	89.8	89.4	86.4
	CNN	99.3	92.9	94.42	01.33	21.13	94.42	04.33
	LSTM	94.36	89.05	91.06	57.8	91.44	91.37	87.8

Due to the lack of available metadata (such as reposts, likes, shares, etc.) when a news piece is first published and not widely distributed, we have solely addressed the article's content. In that case, content alone can be used to identify false news. The outcomes shown in Figure 5(a), (b), (c), (d), (e), and (f) are clearly illustrated in the charts that follow. To assess our model and ensure that the outcomes are comparable to the other models on the identical dataset shown in Table 5, we have implemented a 5-fold Cross-validation resampling technique. While we did use stratified k-fold cross-validation, we did find several samples that were incorrectly classified. This is mainly because there is a lot of overlap between the two classes' attributes, making it hard to tell them apart.

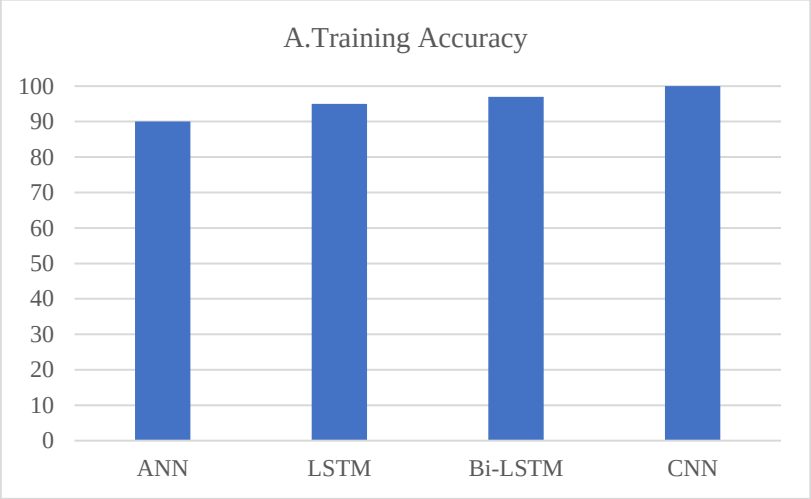

Figure 5(a): Training Accuracy

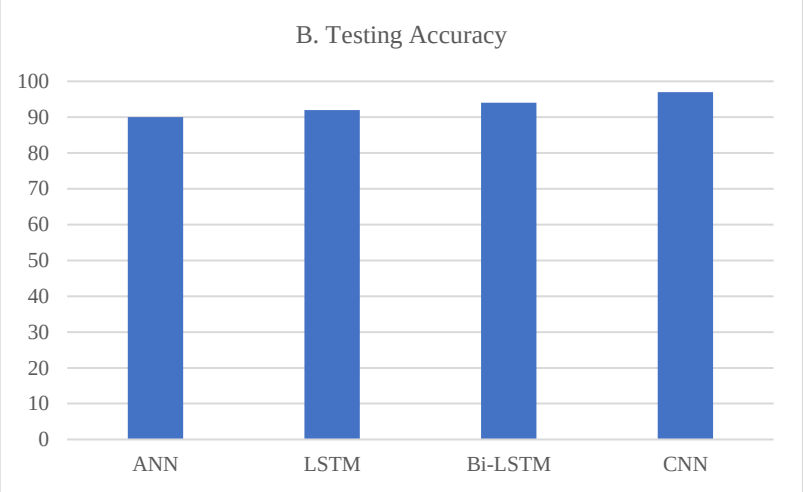

Figure 5(b): Testing Accuracy

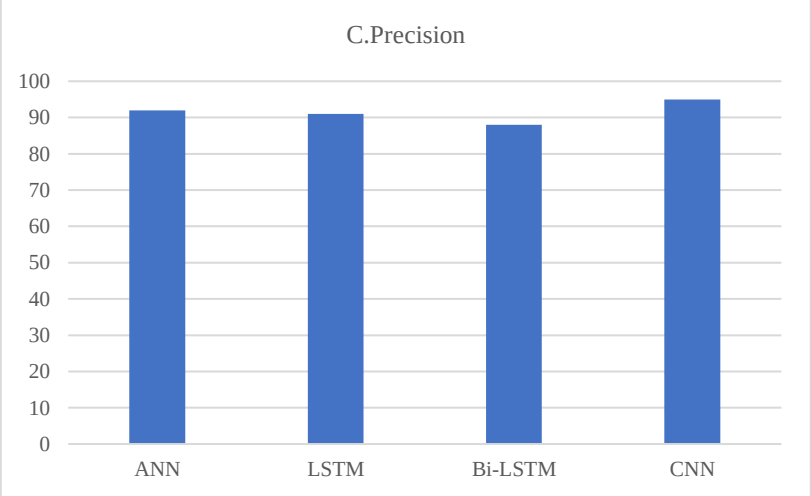

Figure 5(c): Precision

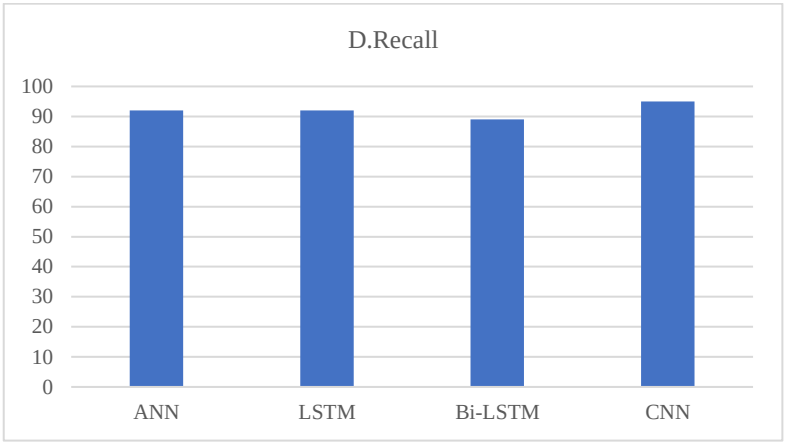

Figure 5(d): Recall

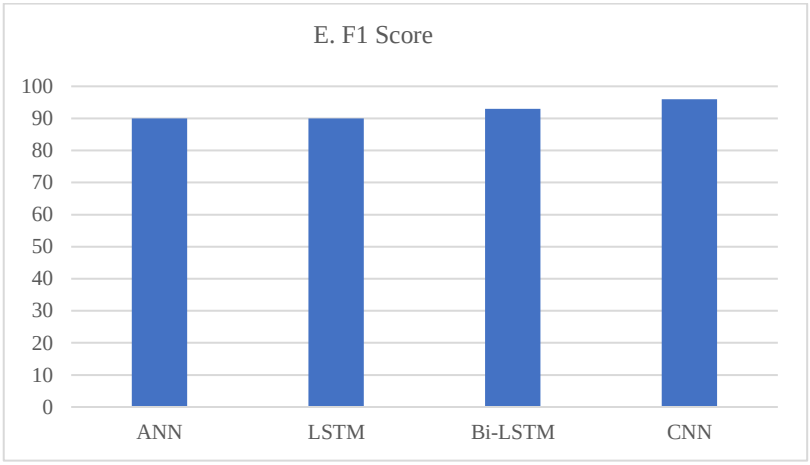

Figure 5(e): F1 Score

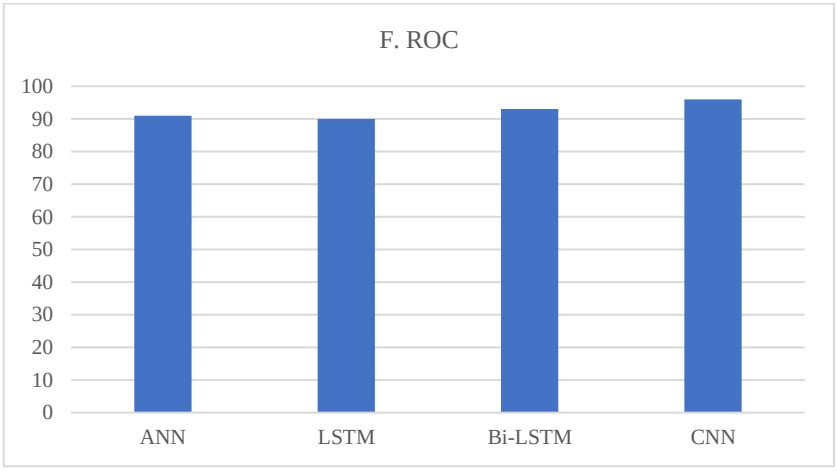

Figure 5(f): ROC

Figure 5: (a), (b), (c), (d), (e), (f) Different Evaluation Metrics are Applied to Our System

We contrasted our approach to other methods in the area using G. McIntire's Fake and Real News Dataset as a standard to guarantee its validity. Many Machine Learning models, including Boosting,

Naïve Bayes, Random Forest, Decision Tree, and Logistic Regression, as well as hand-crafted feature engineering, have been proposed as potential solutions to the problem of identifying false news. Combining stylometric and CBOW (Word2Vec) characteristics, (Reddy et al., 2020) demonstrated a 95% accuracy rate using a gradient boosting approach. Machine learning models were developed by (Bali et al., 2019) to distinguish between fake news based on variables such as sentiment polarity, readability, word count, word embedding, and cosine similarity. According to the evaluation, XGBoost achieved an accuracy of 87.3%, which is the highest. Extraction of features is executed by LSTM-encoder-decoder, which deals with the title, content, and a summary of the content (Esmaeilzadeh et al., 2019). Through his research, he showed that the generated summary acquired 93.1% accuracy. It initially had an accuracy of 87.7% and reached 93.1% through consistent improvements; the weights were modified with the aid of the SAHS (Self-Adaptive Harmony Search) algorithm. Some of the embedding techniques that were used were LSTM, LIWC CNN, depth LSTM, and also N-gram CNN (Huang & Chen, 2020). The N-gram GloVe embedding encoded features used in (Khan et al., 2019) research were word count, numbers, length of the article, the average length of the word, adjectives, numbers, count of exclamations, sentiment characteristics, etc. The Naïve Bayes Classifier acquired 90% precision in his research; on the other hand, 95% precision was acquired by the LSTM deep neural approach. Bhutani et al., (2019) acquired 84.3% accuracy by utilizing the Naïve Bayes approach and its features like sentiments, TF-IDF cores, and Cosine similarity scores. Using the Support Vector Machine, Gravanis et al., (2019) created a content-based model and boasted an accuracy rate of 89%. Convolutional Neural Networks (CNNs) were employed by (George et al., 2020) to handle linguistic information, while Multi-Headed Self-Attention was employed for contextual features. Machine learning models like Random Forest and Naïve Bayes classifier were able to attain accuracies of 83.9% and 84.3%, respectively. An impressive 94.3% accuracy rate was achieved with deep learning models such as LSTM and Fast Text embedding. Automatic feature extraction using Deep Learning algorithms has also been tested in response to disinformation's growing availability and production. Table 6 displays a comparison of our model with other works.

Table 5: 5-fold Cross-validation Results for the Proposed Model

Features	Models	Training accuracy (%)	Validation accuracy (%)	Testing accuracy (%)	Recall (%)	ROC (%)	Precision (%)	F1 (%)
News Title, News Body, Similarity between them (Stance)	ANN	91.31	88.72	90.73	87.35	91.17	87.35	91.14
	LSTM	97.08	87.61	88.60	84.98	89.40	84.98	89.29
	Bi-LSTM	99.23	92.72	93.24	92.03	93.44	92.03	93.34
	CNN	99.92	95.25	95.85	94.81	95.90	94.81	95.89

Table 6: Performance of Various Fake News Identification Models

Models	NB	RF	SVM	MLP	Boost	LSTM	Bi-LSTM	CNN
Reddy et al., (2020)	86	82.5			95			
Bali et al., (2019)	73.2	82.6	62	72.8	87.3			
Bhutani et al., (2019)	84.3	83.9						
Gravanis et al., (2019)	70		89					
Bharadwaj & Shao, (2019)	90.7	94.7				92.7		
George et al., (2020)	84.3	83.9				94.3		
Esmaeilzadeh et al., (2019)						92.1	93.1	
Huang & Chen, (2020)						84.9	87.7	91
Khan et al., (2019) Char-LSTM	90					95	85	86
Our Model						91.16	93.05	95.32

6 Conclusion

In this work, we provided an advanced AI-driven methodology for assessing social engineering attack patterns using state-of-art language models, particularly Bidirectional Encoder Representations from Transformers (BERT), utilizing behavioral psychology insights. Unlike conventional social engineering defense solutions that concentrate primarily on technological defenses and sometimes assume human behavior to be constant and predictable, our approach combines psychological elements influencing user decision-making and vulnerability. This all-encompassing approach helps create the best protection systems that fit different social engineering approaches. Our suggested structure presents a quantified tool to replicate the interaction between social engineering assaults and target vulnerabilities. We devised a dynamic defense strategy characterized as a two-sided stochastic game by considering several target features and behavioral patterns. This kind of strategic modeling helps defenders predict attacker actions and apply countermeasures that successfully lower security vulnerabilities. Our work mainly contributes to integrating BERT-based language embeddings, which significantly improved the detection of false information in social engineering assaults. Especially in phishing emails, bogus messaging, and manipulative social media posts, BERT's extensive contextual awareness enables exact recognition of harmful language patterns. We integrated a stance detection algorithm that assesses the alignment between article headlines and their related body material in order to raise detection accuracy. This extra tool helps the system spot information discrepancies, improving its capacity to separate between accurate and false material. Using Kaggle's open-source news article dataset, our evaluation showed that with 95% accuracy, the BERT embeddings-based CNN architecture performed remarkably. An increase in recall and precision will decrease false positives and help in efficient detection, which exceeds the efficiency of traditional ML approaches. The improved posture detection capability of the model improves its capacity to spot phishing attempts and false news before they become popular on social media. Our method has various restrictions even if it is really successful. The success of the model mainly depends on the presence of pertinent elements in the dataset, so inadequate or unclear data may lower its effectiveness.

Furthermore, the changing strategies applied in social engineering attacks need constant model updating to guarantee its robustness against new hazards. Future research will concentrate on enhancing the flexibility of our system to fit various datasets and surroundings. We also want to improve the model's capacity to spot recently developing social engineering trends so guaranteeing a proactive protection plan. We see a more complete and strong defense mechanism that significantly reduces the hazards presented by social engineering attacks in different digital contexts by combining deeper psychological insights with improving BERT's contextual awareness.

References

[1] Abdulrahman, L. M., Ahmed, S. H., Rashid, Z. N., Jghef, Y. S., Ghazi, T. M., & Jader, U. H. (2023). Web phishing detection using web crawling, cloud infrastructure and deep learning framework. *Journal of Applied Science and Technology Trends*, *4*(01), 54-71. https://doi.org/10.38094/jastt401144

[2] Abroshan, H., Devos, J., Poels, G., & Laermans, E. (2021). Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process. *IEEE Access*, *9*, 44928-44949. https://doi.org/10.1109/ACCESS.2021.3066383

[3] Albladi, S. M., & Weir, G. R. (2020). Predicting individuals' vulnerability to social engineering in social networks. *Cybersecurity*, *3*(1), 7. https://doi.org/10.1186/s42400-020-00047-5

[4] Arachchilage, N. A. G., & Love, S. (2013). A game design framework for avoiding phishing attacks. *Computers in Human Behavior*, *29*(3), 706-714. https://doi.org/10.1016/j.chb.2012.12.018

[5] Arana, M. (2017). How much does a cyberattack cost companies. *Open Data Security*, 1-4.

[6] Atwell, C., Blasi, T., & Hayajneh, T. (2016, April). Reverse TCP and social engineering attacks in the era of big data. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 90-95). IEEE. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.60

[7] Bali, A. P. S., Fernandes, M., Choubey, S., & Goel, M. (2019). Comparative performance of machine learning algorithms for fake news detection. In *Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3* (pp. 420-430). Springer Singapore. https://doi.org/10.1007/978-981-13-9942-8_40

[8] Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019, August). Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)* (pp. 1-5). IEEE. https://doi.org/10.1109/IC3.2019.8844880

[9] Breda, F., Barbosa, H., & Morais, T. (2017). Social engineering and cyber security. In *INTED2017 Proceedings* (pp. 4204-4211). IATED. https://doi.org/10.21125/inted.2017.1008

[10] Caviglione, L., Wendzel, S., Mileva, A., & Vrhovec, S. (2021). Guest Editorial: Multidisciplinary Solutions to Modern Cybersecurity Challenges. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *12*(4), 1-3.

[11] Chargo, M. A. (2018). You've Been Hacked: How to Better Incentivize Corporations to Protect Consumers' Data. *Transactions: Tenn. J. Bus. L.*, *20*, 115.

[12] Costantino, G., La Marra, A., Martinelli, F., & Matteucci, I. (2018, June). CANDY: A social engineering attack to leak information from infotainment system. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)* (pp. 1-5). IEEE. https://doi.org/10.1109/VTCSpring.2018.8417879

[13] Demertzi, V., Demertzis, S., & Demertzis, K. (2023). An overview of cyber threats, attacks and countermeasures on the primary domains of smart cities. *Applied Sciences*, *13*(2), 790. https://doi.org/10.3390/app13020790

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

[15] Do, P., Assaf, R., Scarf, P., & Iung, B. (2019). Modelling and application of condition-based maintenance for a two-component system with stochastic and economic dependencies. *Reliability Engineering & System Safety*, *182*, 86-97. https://doi.org/10.1016/j.ress.2018.10.007

[16] Esmaeilzadeh, S., Peh, G. X., & Xu, A. (2019). Neural abstractive text summarization and fake news detection. https://doi.org/10.48550/arXiv.1904.00788

[17] Fung, C. (2011). Collaborative Intrusion Detection Networks and Insider Attacks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *2*(1), 63-74.

[18] Gao, H., Yan, J., Cao, F., Zhang, Z., Lei, L., Tang, M., ... & Li, J. (2016, February). A Simple Generic Attack on Text Captchas. In *NDSS*.

[19] George, J., Skariah, S. M., & Xavier, T. A. (2020, February). Role of contextual features in fake news detection: a review. In *2020 international conference on innovative trends in information technology (ICITIIT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICITIIT49094.2020.9071524

[20] Ghasempour, A. (2019). Internet of things in smart grid: Architecture, applications, services, key technologies, and challenges. *Inventions*, *4*(1), 22. https://doi.org/10.3390/inventions4010022
[21] Gorment, N. Z., Selamat, A., Cheng, L. K., & Krejcar, O. (2023). Machine learning algorithm for malware detection: Taxonomy, current challenges, and future directions. *IEEE Access*, *11*, 141045-141089. https://doi.org/10.1109/ACCESS.2023.3256979
[22] Govindankutty, M. S. (2021). Is human error paving way to cyber security. *Int. Res. J. Eng. Technol*, *8*, 4174-4178.
[23] Gravanis, G., Vakali, A., Diamantaras, K., & Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, *128*, 201-213. https://doi.org/10.1016/j.eswa.2019.03.036
[24] Hadnagy, C. (2010). *Social engineering: The art of human hacking*. John Wiley & Sons.
[25] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1. https://doi.org/10.5121/ijdkp.2015.5201
[26] Huang, Y. F., & Chen, P. H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, *159*, 113584. https://doi.org/10.1016/j.eswa.2020.113584
[27] Human Cyber Risk—The First Line of Defense, (2021). https://www.aig.com/about-us/knowledge-insights/human-cyber-risk-the-first-line-of-defense
[28] Jampen, D., von Solms, R., & Kritzinger, E. (2020). An analysis of phishing attacks and their detection. *Journal of Cybersecurity*, *6*(1), 1-12.
[29] Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection.
[30] Libicki, M. (2018). Could the issue of DPRK hacking benefit from benign neglect?. *Geo. J. Int'l Aff.*, *19*, 83-89.
[31] Liu, C., Tang, F., Hu, Y., Li, K., Tang, Z., & Li, K. (2020). Distributed task migration optimization in MEC by extending multi-agent deep reinforcement learning approach. *IEEE Transactions on Parallel and Distributed Systems*, *32*(7), 1603-1614. https://doi.org/10.1109/TPDS.2020.3046737
[32] Mahmood, T., & Afzal, U. (2013, December). Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *2013 2nd national conference on Information assurance (ncia)* (pp. 129-134). IEEE. https://doi.org/10.1109/NCIA.2013.6725337
[33] Mayfield, K. P., Petty, M. D., Whitaker, T. S., Bland, J. A., & Cantrell, W. A. (2019, April). Component-based implementation of cyberattack simulation models. In *Proceedings of the 2019 ACM Southeast Conference* (pp. 64-71). https://doi.org/10.1145/3299815.3314435
[34] Meurs, T., Junger, M., Abhishta, A., Tews, E., & Ratia, E. (2022). Coordinate: A model to analyse the benefits and costs of coordinating cybercrime. *Journal of Internet Services and Information Security*, *12*(4), 1-22. https://doi.org/10.58346/JISIS.2022.I4.001
[35] Mouton, F., Leenen, L., & Venter, H. S. (2016). Social engineering attack examples, templates and scenarios. *Computers & Security*, *59*, 186-209. https://doi.org/10.1016/j.cose.2016.03.004
[36] Mouton, F., Malan, M. M., Leenen, L., & Venter, H. S. (2014, August). Social engineering attack framework. In *2014 Information Security for South Africa* (pp. 1-9). IEEE. https://doi.org/10.1109/ISSA.2014.6950510
[37] Nejad, M. B., & Shahriary, H. R. (2017). A Novel Approach for the Detection of Anomalous User Behavior in Web Applications Processes (Using Web Applications Firewall- WAF). *International Academic Journal of Science and Engineering*, *4*(2), 198–209.

[38] Pavković, N., & Perkov, L. (2011, May). Social Engineering Toolkit—A systematic approach to social engineering. In *2011 Proceedings of the 34th International Convention MIPRO* (pp. 1485-1489). IEEE.
[39] Raj, D. S., & Dharmaraj, A. (2024). Investment Pattern and Behaviour of Rural Households on Investment Avenues with Special Reference to Ernakulam District. *Indian Journal of Information Sources and Services*, *14*(4), 103–107. https://doi.org/10.51983/ijiss-2024.14.4.16
[40] Reddy, H., Raj, N., Gala, M., & Basava, A. (2020). Text-mining-based fake news detection using ensemble methods. *International journal of automation and computing*, *17*(2), 210-221. https://doi.org/10.1007/s11633-019-1216-5
[41] Rishikesh, Rupasri, Tamilselvan, Yoganarasimman, & Sujai, S. (2022). Intrusion of Attacks in Puppet and Zombie Attacking and Defence Model Using BW-DDOS. *International Academic Journal of Innovative Research, 9*(1), 13–19. https://doi.org/10.9756/IAJIR/V9I1/IAJIR0903
[42] Siddiqi, M. A., & Pak, W. (2020). Optimizing filter-based feature selection method flow for intrusion detection system. *Electronics*, *9*(12), 2114. https://doi.org/10.3390/electronics9122114
[43] Siddiqi, M. A., Mugheri, A., & Oad, K. (2016). Advance persistent threat defense techniques: A review. *Pakistan Journal of Computer and Information Systems, 1,* 53–65.
[44] Sood, A. K., Zeadally, S., & Bansal, R. (2017). Cybercrime at a scale: A practical study of deployments of HTTP-based botnet command and control panels. *IEEE Communications Magazine*, *55*(7), 22-28. https://doi.org/10.1109/MCOM.2017.1600969
[45] Thanh, C. T., & Zelinka, I. (2019, December). A survey on artificial intelligence in malware as next-generation threats. In *Mendel*, *25*(2), 27-34. https://doi.org/10.13164/mendel.2019.2.027
[46] Truecaller. (2021). 2021 U.S. spam & scam report. https://www.truecaller.com
[47] Udayakumar, R., Joshi, A., Boomiga, S. S., & Sugumar, R. (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification. *Journal of Internet Services and Information Security*, *13*(4), 138-157. https://doi.org/10.58346/JISIS.2023.I4.010
[48] Wang, Z., Zhu, H., & Sun, L. (2021). Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods. *Ieee Access*, *9*, 11895-11910. https://doi.org/10.1109/ACCESS.2021.3051633
[49] Yang, R., Zheng, K., Wu, B., Li, D., Wang, Z., & Wang, X. (2022). Predicting user susceptibility to phishing based on multidimensional features. *Computational Intelligence and Neuroscience*, *2022*(1), 7058972. https://doi.org/10.1155/2022/7058972
[50] Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2022). Prediction of phishing susceptibility based on a combination of static and dynamic features. *Mathematical Problems in Engineering*, *2022*(1), 2884769. https://doi.org/10.1155/2022/2884769
[51] Yash, R. Beyond the code: Exploring the human factor.

Authors Biography

Akash Parasumanna Sridhar is currently working as an IT Cybersecurity Analyst at Campbell Clinic Orthopaedics in Memphis, Tennessee, USA. In my role, I actively monitor and respond to real-time security threats, analyze logs from various security tools, and identify and report security incidents. Additionally, I utilize my red teaming expertise to conduct reconnaissance, exploitation, social engineering, and OSINT investigations on target systems and networks to strengthen security defenses. I hold a Master of Science in Cybersecurity from the University of Houston and am passionate about advancing AI-driven cybersecurity solutions to mitigate modern cyber threats. I have earned multiple industry-recognized cybersecurity certifications, including CEH v11, CHFI, eJPT v2, Pentest+, CCNA, SC-200, Security+, Cloud+, CySA+, CASP, SecurityX and Certified EC-Council Instructor, demonstrating my expertise across various domains of cybersecurity.