Modeling memory processes in phishing decision making using instance based learning and natural language processing

Tianhao Xu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Prashanth Rajivan, Chair
Linda Boyle
Shuai Huang

Program Authorized to Offer Degree:
Industrial and System Engineering

University of Washington

Abstract

Modeling memory processes in phishing decision making using instance based learning and natural language processing

Tianhao Xu

Chair of the Supervisory Committee:
Prashanth Rajivan
Department of Industrial and System Engineering

Phishing is a type of social engineering attack that uses psychological manipulations to influence people into revealing their personal information. Despite advancements in security technologies, phishing attacks continue to be rampant and successful because phishing attacks are primarily rare events, and discriminating phishing emails from legitimate emails continues to be challenging. Furthermore, attackers now exploit our personal information on the Internet to generate personalized phishing attacks, known as spear-phishing attacks. Current automation (e.g., spam filters) successfully filters most phishing emails but is poor at detecting spear-phishing attacks. Therefore, the onus is on the recipient to detect attacks that automation misses. However, in most instances, people who are the targets of spear-phishing attacks fall victim to them.

Past research has mostly blamed inattention for human susceptibility to phishing. But what is the underlying cognitive process that prevents people from paying attention to key indicators in phishing and spear-phishing messages? The answer to this question could be in the human working memory because our attention is inextricably linked with the contents in our memory and memory processes that govern our decision-making. Despite the large body of research on phishing attacks, there is a significant lack of work explaining the role of memory processes in end-user susceptibility to phishing and spear-phishing attacks.

This dissertation will address this research gap through laboratory experiments grounded in the principles of instance-based learning and the development of cognitive models of human decision-making. A novel multi-player, human-in-the-loop simulation environment called SpearSim was developed to study the human decision making to spear phishing attacks from both the attackers and end-users' perspectives. Results from the experiment show that access to more personal information about targets can enable attackers to produce spear-phishing attacks involving contextually meaningful impersonation and narratives, making end-users more vulnerable to spear-phishing attacks. Data from the experiment conducted using SpearSim was used to train Instance-Based Learning (IBL) models and natural language processing models (LSA, GloVe, and BERT) to predict and explain the role of working memory processes behind the human response to phishing and spear-phishing attacks. Results from my experiments with IBL models of phishing decision-making show that, compared to representations that only consider the semantic properties of emails, using representations that consider higher-order contextual meanings assigned by humans could enable IBL agents to predict human response with high accuracy. Furthermore, I found evidence that IBL models of phishing decision-making performed better in predicting responses in situations where participants made quick, system-1 like decisions, suggesting that instance-based learning satisfies the conditions for describing intuitive decision-making.

A follow-up study focused on end-user phishing decision-making to further test the insights obtained from the previous experiment and to understand how people encode emails to memory. The study involved the use of an eye tracker to monitor participants' eye movements when they processed the emails presented to them and to study how end-users' attention may influence their decision-making. Similar to previous experiments, data from the experiment was used to develop IBL models of phishing decision-making. I once again found that representations that consider higher-order contextual meanings assigned by humans enable IBL agents to predict human response more accurately than input representations that consider

human attention to words and phrases along with the semantic properties of the emails. I also found more evidence from eye-tracking data revealing that instance-based learning models effectively predict human responses in situations involving intuitive decision-making. This insight is crucial to cyber security defense because people are more likely to fall victim to phishing and spear-phishing attacks with more intuitive decision-making, and my work shows that models grounded in IBL can inform interventions to mitigate phishing threats.

Findings from this dissertation are expected to advance our understanding of the cognitive processes associated with detecting phishing attacks and could facilitate the development of personalized anti-phishing training solutions. The findings from this dissertation are also expected to contribute to our understanding of the cognitive models and how to apply them to analyze human decision-making in cyber security.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

NLP: Natural language processing

IBL: Instance-based learning

LIWC: Linguistic inquire word counts

LLM: Large language model

DEDICATION

To my mom and dad, thanks for the emotional and financial support along the way. You prepared me well for the mindset of facing uncertainties and complex problems.

This dissertation dedicates to my uncle-in-law, who unfortunately passed away on April 21st 2023. Fengguo Li, who has always been a brave man in his life, always encouraged me to be curious about the world and be exploratory.

ACKNOWLEDGMENTS

Thanks my advisor, Dr. Prashanth Rajivan for all the guidance and support along the way. There have been lots of inspiration and mind-blowing moment with you. You have always been a great visioner and mentor. I cannot get to this point without your guidance.

Thanks to my committee members, Dr. Linda Boyle, Dr. Shuai Huang, and Dr. Andrea Stocco for your valuable feedback and critiques.

Thanks to Dr. Kuldeep Singh for all the guidance and support with the experiment design and writing publications.

Thanks to Dr. Ashis Banerjee for the opportunity to work on a project in the early stage of my graduate study.

Thanks to my Behavioral Research in Computer Security lab colleagues for your help and great to be a part of the lab.

Thanks to my peers and friends, Dr. Zhanlin Liu and Dr. Tianchen Sun for the support and insightful discussion.

Last but not least, thanks to my girlfriend, Jiaxin Li has always been a reliable ally and thoughtful listener.

Chapter 1

INTRODUCTION

Phishing is a type of social engineering attack that uses psychological manipulations to trick individuals into revealing sensitive personal information. Unsuspecting victims fall for such manipulations because the emails appear legitimate and seemingly from a trustworthy entity such as a bank or government agency. In reality, they are email-based attacks carefully crafted by malicious actors to target people using protected computing systems instead of the system themselves. It only takes one email and one click to compromise the security of an entire organization.

A key characteristic of phishing emails is the use of social influence strategies, such as creating a sense of urgency or expressing authoritativeness that exploits human fears and emotions into taking an immediate actions. For example, an attacker may pretend to be an IRS (Internal Revenue Services) representative informing an issue with the recent tax filing and demanding immediate action. Recipients are usually persuaded into taking actions in the form of clicking a URL link within the email that will redirect people to spoofed web pages controlled by the attacker but could also include downloading an attachment containing malicious payloads. Falling victim to phishing emails can have damaging consequences, including identity theft, system compromise, financial loss, and data loss for individuals and organizations.

While regular phishing emails may cast a wide net by delivering them to many email addresses, spear-phishing emails target specific individuals or groups with an organization. Attackers collect information about their targets by scouring social media and personal websites to create phishing emails tailored to their targets. Furthermore, to make attacks trustworthy, attackers may take advantage of compromised email accounts to deliver their

attacks, known as lateral spear-phishing attacks. Phishing has been a problem for decades, and it is still a significant security threat. But, spear-phishing attacks have now become the major source of compromise for most organizations.

According to Federal Bureau of Investigation (FBI) reports, there was an 110.3% increase in reported phishing attacks from 2019 to 2020 [89]. There were 19,465 incidents in 2016 and 241,342 incidents in 2020. According to the data breach investigation report from Verizon in 2021, social engineering attacks are the primary cause of data breaches, causing a third of them [125]. These reports reflect the threats of phishing attacks to organizations. Even a single phishing attack can cause a data breach. The cost of such phishing attacks has increased dramatically over the past six years. Indeed, it has almost quadrupled, costing large U.S. companies an average of $14.8 million annually (or $1,500 per employee) [99].

Three main methods are used to manage phishing attacks: Security Protocols, Phishing Detection, and Human Training. Security protocols constitute a broad class of methods used to ensure that only emails from verifiable sources are delivered and that messages are not forged during transit. For example, email authentication protocols such as DMARC (Domain Message Authentication Reporting) are used to authenticate email servers and prevent unauthorized use for sending spam or malicious emails. Domain blacklisting is a related technique that filters emails from servers that have been flagged for sending spam and malicious emails. However, security protocols are often poorly configured [57]. So, preventing the delivery of messages purely based on email authentication schemes could result in the non-delivery of messages from legitimate sources and is counter intuitive to email providers' core business, which favors email delivery. Therefore, in addition to security protocols, systems to automatically detect Phishing attacks are used. They are typically bundled with SPAM filters and use machine learning and natural language processing algorithms to automatically differentiate phishing attacks from benign emails. Although SPAM filters successfully filter out most phishing emails (almost 99% of them), considering the volume of emails transferred on the Internet, the 1% that the algorithms miss still constitutes a large number of emails that can potentially harm its recipients. More importantly, machine learning algorithms are

poor at detecting spear-phishing attacks - phishing attacks of the personalized kind. Hence, humans become the last line of defense and are expected to detect attacks that security protocols and algorithms miss which tends to be novel and targeted and, therefore, difficult to detect [50].

Why is it challenging for people to detect phishing attacks? There are two major challenges. First, phishing attacks are largely rare events, and so, people tend trust the emails they receive. Second, phishing attacks are primarily deceptive communication that use impersonation to resemble truthful messages, applies emotional arguments to influence recipients, and could be tailored to exploit personal information. Most previous studies have blamed human susceptibility to phishing attacks on inattention. However, little has been found about the underlying cognitive process that prevents end-users from paying attention when processing phishing or spear-phishing messages. It is important to study human working memory because attention is inseparable from the contents of memory and the memory process that governs human decision-making. There is also a severe lack of models to adequately explain and predict the cognitive dynamics underlying end-user susceptibility to phishing emails.

1.1 Research Objectives

This dissertation addresses these research gaps in several ways. First, it develops a novel methodology to study human susceptibility to spear-phishing attacks in the laboratory. Since spear-phishing attacks involve the use of personal information, real-world datasets on spear-phishing attacks are challenging to acquire. Therefore, new experimental paradigms are necessary to collect and analyze data that measure susceptibility to spear-phishing attacks. There is a quote by Sun Tzu, '*If you know yourself and the enemy, you need not fear the result of a hundred battles*'. Unlike mass phishing attacks, studying adversarial behaviors is important in studies on spear-phishing attacks to understand how attackers would exploit the personal information of the targets and the strategies they would use to persuade end-users to respond to their phishing messages. So, studies on spear-phishing attacks need to move

beyond only studying end-users and investigating the attackers' behaviors. However, few studies have been conducted to study attackers' strategies and decision-making. I designed a multi-player synthetic task environment called SpearSim to address this critical gap. Using SpearSim, I conducted an experiment to understand how information exploitation in spear-phishing attacks influences end-user decision-making.

Second, this dissertation offers an understanding of the working memory processes associated with end-user susceptibility to phishing attacks. Figure 1.1 presents the overview of how the memory process involved in phishing decision making. A central hypothesis is that people make decisions on phishing messages based on past experiences by activating pertinent memories of decisions made in response to similar emails in the past. When facing incoming email instances, people retrieves their previous memory instance and compare the current instance with the previous ones. The feedback from the decisions stimulus the human reward system (dopamine) and drive future decisions. The experiences are consisting of a combination of environmental cues and these instances are stored in memory. With this system shown in Figure 1.1, it describes how the memory plays a vital role in the whole decision-making process. Instance-Based Learning Theory (IBLT) [42] was used as a backbone to formulate human decision making with multiple approaches of instances' representations, including natural language processing, human subject survey, and eye movement. Insights from the dissertation contribute to understanding the challenges from memory of processing emails.

1.2 Structure

In Chapter 2, I introduce previous studies related to these topics. Chapter 3 presents exploratory analysis investigating psycholinguistic factors associated with deception in phishing emails' suggested the important role linguistics play in determining end-user susceptibility to phishing. Chapter 4 discusses the experimental design called SpearSim and the findings of the experiment related to the information exploitation and strategy use in spear-phishing attacks. Chapter 5 presents a new modeling approach called SpearCog, which combines nat-

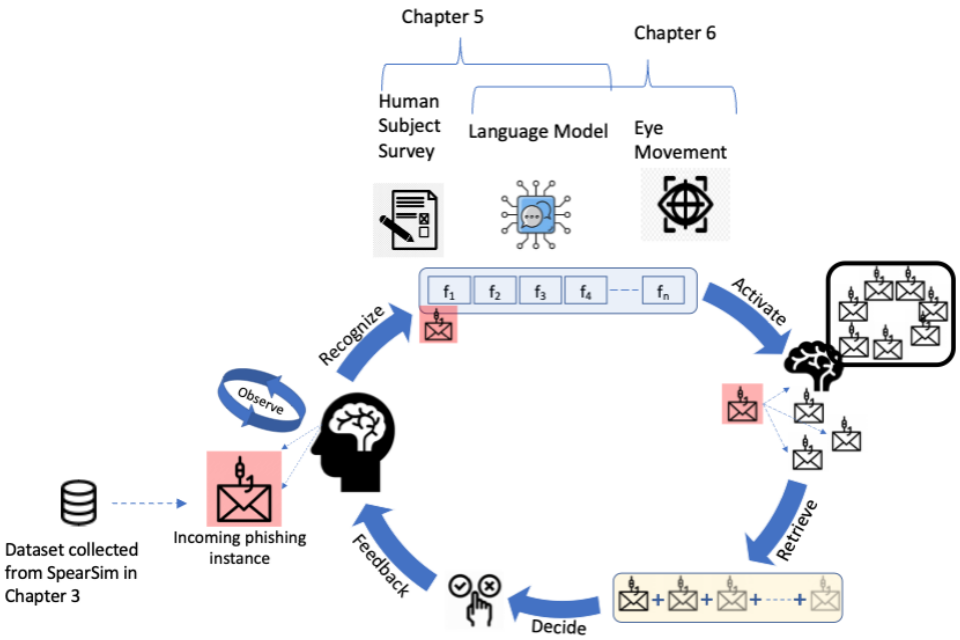


Figure 1.1: Overview of the dissertation

ural language processing and cognitive modeling (ACT-R and IBL) that predicts end-user decision-making related to phishing messages. In addition, further analysis of the general pattern of IBL was also demonstrated. Chapter 6 illustrates an eye-tracking experiment to study how attention and working memory would affect email decision-making. With such eye tracking dataset, the pattern discovered in Chapter 5 was validated.

Chapter 2
BACKGROUND AND RELATED WORK

2.1 Phishing

2.1.1 Deceptions in Phishing

Whether face-to-face or computer mediated, lying and deception is common in all forms of social interactions [108]. But, despite the prevalence of lying, people are poor at detecting them. A meta-analysis conducted by Bond and DePaulo show that people are only slightly better than chance in detecting a lie [11, 129].

Past research on the linguistic aspects of phishing has been focused on developing automation tools such as machine learning based spam classifiers to detect phishing attacks. Such machine learning classifiers are developed using insights from modeling and extracting common phishing topics found in known phishing attacks [44, 66]. Past research on classifiers have also developed new approaches to efficiently differentiate features present in phishing and ham emails [92]. For example, methods that integrate multiple feature extraction methods, including content-based, heuristic based, and fuzzy rule-based approaches [150].

These efforts have made significant improvements to spam classifier performance but the same cannot be said about the human performance. Despite text (subject and body of the email) being the primary stimuli in email messages, there is a severe lack of understanding about the linguistic features associated with deception in phishing emails and the effects of linguistic cues on end-user vulnerability to phishing, At present, the vast majority of theories used to explain why end-user fail to accurately detect phishing attacks are centered around human inattention and ineffective information processing (e.g., [29, 141, 30, 55, 48, 127, 106]). Other studies suggest social influence as a potential explanation for end-user vulnerability [93, 35, 119]. Among the past work on social influence in phishing, most have

attempted to test the influence of Cialdini's six generic persuasion principles in the context of phishing [18, 78, 94, 93, 35, 119]. However, so far, the results on the influence of these generic persuasion strategies have been largely inconclusive. Few studies suggest that social proof is the most effective strategy [119, 100], whereas other studies show that authority, consistency, and reciprocity were most successful and social proof and scarcity were the least effective strategies (e.g., [8, 62, 94, 36]).

I can, however, draw from decades of past research on deception and communication from fields such as psychology, communication, and linguistics. These works have been focused on identifying the underlying cognitive functions involved during deception and identifying valid indicators of deception by studying verbal and non-verbal cues of deception [33, 32]. However, the majority of this research has been centered around studying deception in face-to-face communication, particularly in law enforcement scenarios [38, 130, 129].

2.1.2 Spear Phishing

Despite efforts to train individuals to detect phishing emails, many nevertheless fall victims, particularly to phishing attacks of the targeted kind. Lack of human attention to important indicators (e.g., bad grammar, dubious claims, suspicious URL) in a phishing email is usually considered why end-users frequently fall prey to phishing attacks [106, 29]. However, unlike mass phishing attacks, spear-phishing attacks tend to be less conspicuous and lean on reliable cues that people can rely upon to identify spear-phishing attacks. So, people may fail to detect spear-phishing attacks, even after momentarily noticing the few indicators present in them [3, 94, 61, 136]. While inattention is certainly important, I posit there are deeper cognitive challenges when it comes to spear-phishing detection, such as targeted deception, information processing, and social influence [133, 43].

Human factors research on spear-phishing is limited and mostly focuses on testing if individual attributes such as age, personality, and education level play a role in human vulnerability to spear-phishing [39]. Few studies on the effect of age have found that, compared to older people, young people are more susceptible to phishing attacks [123, 109, 77, 140].

Other studies show that older people are more susceptible to phishing [90, 78]. However, a surprising lack of work has investigated the key aspect of spear-phishing attacks - the impact of personalization. There is a severe lack of work that have systematically studied how exploitation of personal information in spear phishing attacks impact end-user susceptibility. This research aims to address this critical knowledge gap.

Another critical attribute of a spear-phishing email is social influence through persuasion strategies [35, 148]. Phishing could be considered an act of persuasion used to convince others to give up their credentials or confidential information. Common examples of persuasion used in phishing attacks are urgency, such as a notification from the security office about multiple failed account logins; greed such as the offer for money in Nigerian 419 scams; and emotional appeal such as pretending to be a relief agency asking for money for a recent flood [55].

Among the past work on social influence in phishing, most have attempted to test the influence of Cialdini's six generic persuasion principles in the context of phishing [18, 78, 94, 93, 35, 119]. However, so far, the results on the influence of these generic persuasion strategies have been largely inconclusive. One study implies that social proof is the most effective strategy [119], whereas other studies show that authority, consistency, and reciprocity were most successful and social proof and scarcity were the least effective strategies (e.g., [94, 36]). The discrepancy in results, I suspect, is due to testing generic persuasion strategies in isolation without considering the context of the end-user and email topics. Also, these studies have predominantly focused on studying social influence in non-targeted, mass phishing messages [138]. Therefore, to a large extent, they have taken a static view of the system: non-adaptive adversaries and non-contextual social influence strategies.

2.2 Natural Language Processing

2.2.1 Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count (LIWC) toolkit [95] was used to analyze psycholinguistic characteristics of words that the participants chose to add and remove during each

attack trial. LIWC has been widely used in text analysis to categorize words within texts along several grammatical, psychological, and social dimensions [120]. LIWC analyzes the words used in a paragraph and provides the percentage of words that fall into different linguistic dimensions or categories identified by LIWC. LIWC features have also been widely used to analyze deception in communication [85, 75].

On the other side, there is some recent research on understanding deception in computer-mediated communication. For example, Ho has analyzed linguistic features associated with deception in an interactive chat environment [52]. They found that certain language-action cues such as cognitive load, affective process are often found in spontaneous online communication [51]. They used the Language Inquiry Word Count (LIWC) framework to extract and model linguistic and psycholinguistic features associated with deception in online communication. LIWC is a framework commonly used in analyzing natural language including text based communication. Specifically, LIWC is a framework and a software package that can analyze a given piece of text and calculates the percentage of word attributes (e.g., words describing negative emotions, social process, certainty) being used in a piece of text. LIWC features are also widely used in deception research (e.g., [85, 75]), LIWC has also been used as feature extraction method with machine learning to detect deception in phishing and spam messages by comparing the linguistic features in benign and phishing emails. [91, 80]. However, there is lack of previous research that have applied LIWC to analyze psycolinguistic features in phishing attacks and the effect of such word features on end-user vulnerability.

2.2.2 Word Embedding

As a baseline, I used *Latent semantic analysis* (LSA). It is a popular bag-of-words approach used to determine the similarity between two linguistic items (e.g., documents, emails) based on word frequencies. In LSA, linguistic items are organized into a word-frequency or a TF-IDF (term frequency-inverse document frequency) matrix to specify the number of times each word in a corpus appears within each document. Using the singular value decomposition method, this large dimensional matrix is factorized to represent the documents in a low di-

mension space and determine the latent factors (or topics) that may describe each document. The similarity between a pair of documents is calculated using the cosine distance between the low-dimension vector of latent factor values of each document. LSA effectively captures the similarity between documents based on word frequency that may suit information retrieval applications. However, they may not represent how humans process text. For example, two emails from the same bank could contain similar words (e.g., account, withdrawal) and branding but could be communicating two different things. A human would consider the two emails dissimilar to each other, but LSA is likely to consider them semantically similar because they contain words that belong to a common latent topic of banking.

2.2.3 GloVe

GloVe or Global Vectors algorithm was introduced to address some of the limitations in LSA and other algorithms (e.g., Word2Vec) used for learning word-level representations [96]. GloVe algorithm is used to produce a global vector representation of words based on their co-occurrence in a large corpus of text data. The algorithm generates a vector representation for each word in a given corpus, where words with similar vector representations are considered semantically similar. For example, let us consider two words, chase and account related in the banking context. The global vector representations of these two words are essentially the ratio of their co-occurrence probabilities with various probe words. When fine-tuned to the banking context, the algorithm is likely to generate similar vector representations for the words chase and account due to their co-occurrence in a banking email corpus. The algorithm generates a vector of fixed dimensions (usually 300) for each word in a corpus. These are machine-learned representations, learned by training the algorithm on a large text corpus to capture global co-occurrence statistics between words. Using the transfer learning approach, the global representations can be fine-tuned and applied to learn representations for words in a smaller dataset, such as a phishing dataset generated from the experiment described earlier. Although GloVe is significantly better at capturing global and contextual similarities between words, the sequential pattern of language and the context of a word

within a sentence is ignored.

2.2.4 Bert

Unlike LSA and GloVe, which represent natural language at a word level, BERT can be used to make inferences at a sentence level, allowing BERT and other related methods to achieve state-of-the-art performance on various natural language understanding tasks [28]. BERT is a bi-directional model because the algorithm considers the full context of a word in a sentence by processing words that come before and after it. The algorithm achieves bi-directional processing by considering all words in a sentence in parallel rather than one-by-one in a sequence, using a transformer-based self-attention mechanism originally introduced in the paper [124]. The architecture of BERT is complex to describe succinctly and is beyond the scope of this paper. However, in essence, the self-attention mechanism used in BERT takes inspiration from how humans pay visual attention to text stimuli such as words in a sentence - I fixate and associate relevant terms in a sentence and skip over irrelevant words. Similarly, the encoder module in BERT takes as input a long text sequence, encodes, and generates a word embedding for the input sequence, and using the self-attention mechanism, it recodes and weighs the relevance of words in the sequence to highlight the pertinent parts in the long text. For the full description of the BERT model, please refer to the original paper [124]. Like GloVe, I can use the transfer learning approach to fine-tune the BERT model for specific NLP tasks.

2.2.5 Semantic Similarity

With the word to vector approach, text is usually represented as matrix. I could calculate the matrix distance, such as cosine similarity, to represent the semantic similarity between text. Sentence-BERT is a modified version of the pertained BERT network that uses Siamese network structures to derive semantically meaningful sentence embedding that enables us to compute the cosine similarity between a pair of email texts [101].

2.3 Instance-Based Learning

IBL models use the formalization of the memory mechanisms from the adaptive control of thought-rational (ACT-R) cognitive architecture [5] and the decision process from Instance-Based Learning Theory (IBLT) [42]. An instance in the IBL model is a unit of experience, consisting of the state (attributes of task), the decision made in the current state, and the utility (the outcome of choosing an option in the current state). For each viable decision, the model computes an expected utility using the *blending* mechanism. The blended value is computed by averaging the past outcomes weighted by the probability of memory retrieval, which depends on the contextual similarity to past instances. It also take into account frequency and recency of the past experience instances. The decision is made with the highest expected utility. To calculate blended value $V_{k,t}$ for option k at trial t the following equation is used:

$$V_{k,t} = \sum_{i=1}^{n} P_{i,k,t} * X_{i,k,t} \tag{2.1}$$

Where $X_{i,k,t}$ represents the outcome of an instance i for option k at trial t and $P_{i,k,t}$ is the retrieval probability of an instance i for option k at trial t. The retrieval probability of an instance i is the ratio of activation of i_{th} instance corresponding to the activation of all instances (1, 2, ..., n; where n is total number of instances) created within the option k at trial t. The retrieval probability is defined as:

$$P_{i,k,t} = \frac{e^{A_{i,k,t}/\tau}}{\sum_{i=1}^{n} e^{A_{i,k,t}/\tau}} \tag{2.2}$$

Here, $\tau = \sigma * \sqrt{2}$ and τ is a free noise parameter. The noise parameter (τ) is used to capture the inaccuracy of remembering past experiences from memory. $A_{i,k,t}$ is the activation of an instance i on option k at trial t. It represents the linear aggregation of three cognitive elements: frequency and recency, similarity of an instance with past experiences, and noise. Based on the ACT-R theory of cognition [4], the activation value represents how readily

available an instance is in memory: the higher the activation, the easier and faster it would be to retrieve such an instance from memory. The Activation is computed as follows:

$$A_{i,k,t} = ln \sum_{t_i=1..t-1} (t - t_i)^{-d} + MP \sum_{k} Sim(v_k, c_k) + \sigma * ln \left(\frac{1 - \gamma_{i,k,t}}{\gamma_{i,k,t}} \right) \tag{2.3}$$

The term $(ln \sum_{t_i=1..t-1}(t - t_i)^{-d})$ reflects the power law of experience and forgetting, t_i represents all the previous trials where the instance i was either created or its activation was reinforced due to its recurrence. t_j is the time since the j^{th} occurrence of instance i and d is the decay (default value= 0.5) rate of each occurrence. The decay parameter accounts the the rate of forgetting the experienced events: higher the decay, faster the rate of forgetting of past events, which increase the reliance on recent events. The activation of an instance can increase with the frequency and recency of observing that outcome (i.e., by small differences in $t - t_i$).

$MP \sum_k Sim(v_k, c_k)$ represent a partial matching process which reflect the similarity between the current state (c_k) and the instances that are stored in memory(v_k), scaled by a mismatch penalty (set to 2.5). The similarity between numerical slot values are computed on a linear scale from distinct (0.0) to an exact match (1.0).

$\sigma * ln \left(\frac{1-\gamma_{i,k,t}}{\gamma_{i,k,t}} \right)$ represents the Gaussian noise mechanism for capturing the variability in individual choices. Where $\gamma_{i,k,t}$ is a random number drawn uniformly between 0 and 1. The σ i.e. the variance in the noise term is set to the default ACT-R value of 0.50.

2.3.1 SpeedyIBL: A faster version of IBL

However, original IBL model suffers from the curse of exponential growth of instance. As the number of observations increasing, the computation increase dramatically that leads to exponential slow down of the computational time. In addition, the dimension of a task could also significantly affect the amount of computation. Nguyen el at. implemented the *Speedy*IBL, which use matrix computation to speed up the computation [87].

$$S(D_t, f_k) = c \sum_{i=1}^{n} P_i \left(\frac{\partial Sim(f_k, v_{i,k})}{\partial f_k} - \sum_{j=1}^{n} P_j \frac{\partial Sim(f_k, v_{j,k})}{\partial f_k}\right) \tag{2.4}$$

2.3.2 Cognitive Salience

In addition, Somers et al. presented a modeling of cognitive salience that taking the derivative of the blending equation with respect to each feature as shown in equation 2.4 [115]. The salience is defined to be the influence of a factor on a decision. This is an extension of blending mechanism which exploits its analytical tractability to provide a closed form of the gradient-based salience of its representational features on it decisions. These two following works empowered the computation performance and post hoc analysis potential of the cognitive models.

Chapter 3

DETERMINING PSYCHOLINGUISTIC FEATURES OF DECEPTION IN PHISHING MESSAGES

Although phishing attacks are rampant on the Internet as a whole, thanks to automation, the likelihood of an individual encountering an attack on a daily or weekly basis is small compared to the large volume of legitimate emails received. The onus is placed on people to detect the rare attacks that automation misses [50, 88]. But, distinguishing phishing emails from legitimate emails continues to be a difficult task for most individuals. Decades of research on lying show that people demonstrate only near-chance accuracy at detecting lies despite the prevalence of lying in interpersonal communications [11, 27]. The reasons for poor accuracy at detecting lies include the use of invalid cues of lying when judging deception or due to the lack of cues that can reliably differentiate lying from truth telling. These reasons are also applicable to challenges people face with detecting phishing emails.

Unlike in face-to-face communication, where people have access to a rich set of cues, such as facial expressions and body language, computer-mediated communication through emails and text messages are largely '*cue lean*' [52]. With emails, people have access to only the information presented in the email clients. People are required to pay close attention to key phishing indicators (e.g., link in a phishing email that redirects people to a malicious website) to avoid falling victims to phishing attacks [29]. Nevertheless, most fail to do so. Lack of awareness and technical knowledge are considered important drivers for this individual inattention to security indicators [30, 141]. The rationale for this explanation is that when people don't possess the necessary knowledge, they would not know which indicators to attend and analyze. Consequently, anti-phishing training programs were developed to teach people to detect phishing attacks by presenting them guidelines on what indicators they

must attend to in a phishing message. However, to operationalize the training knowledge and actively search for indicators in a message, one must first become suspicious of it [135]. Without suspicion, people would fail to recognize phishing emails as a threat, even if they happen to momentarily notice the relevant indicators in the message [3, 94, 61]. Hence, it is important to understand the factors that prevent people from becoming suspicious of phishing messages [135].

The influence of structural elements of phishing emails (URL, reply-to address) on end-user vulnerability has been extensively studied [128, 58, 146, 111]. Past research has repeatedly found that when people pay attention to the source of the phishing email, they are less likely to fall victims to phishing attacks [25, 134, 128] because phishing messages typically contain illegitimate and suspicious-looking domain names and URLs. Although attention to such cues is essential, they are not sufficient to detect phishing attacks. When people rely disproportionately on structural elements of emails to detect phishing, they are likely to fall victims to attacks that use legitimate or legitimate-looking domain names [50]. Increasingly, attackers are adopting advanced spoofing methods to make the domain names, URLs, and email addresses in phishing emails seem legitimate [25]. Hence, it is important to train people to look beyond the structural elements of an email message to recognize increasingly sophisticated and targeted phishing emails. Note that this section is also published in [142]

3.1 Objective

Email text is one of the primary stimulus that people use to make judgments; whether the sender is trustworthy, whether the pretext is genuine, and, importantly, whether the actions requested in the message are benign or malicious. However, training people to use cues in the email text to detect phishing emails is challenging because it is currently unknown how to distinguish linguistic features used in regular communications from linguistic features (e.g., word categories that communicate threat, deadlines, reward) that attackers employ in a phishing email with the intention of deceiving and influencing the end-user into responding. Although there is a large body of work devoted to the development of algorithms and machine

learning models to discriminate phishing and legitimate emails using linguistic features [25], such features used by the models are often uninterpretable and unusable for people. There is a severe lack of fundamental work that has determined linguistic elements of deception in phishing messages and has analyzed their influence on end-user vulnerability. This section aims to fill this critical gap. Using a novel research methodology, I aim to reveal the linguistic items often used for deception in phishing email text and the effect of such linguistic items on peoples' judgment and decision making.

Deception is not a generic phenomenon and can vary widely based on situations and contexts [38]. Therefore, context is crucial in deception research, and findings about valid cues of deception from other contexts would not directly apply to phishing. Past research has also pointed to the importance of understanding risk communication and deception to create more efficient training protocols [26]. Therefore, we need new research methods to identify linguistic elements of deception specific to phishing, which is a type of asynchronous, non-iterative, computer-mediated communication. This work aims to fill this gap through a novel two-phase experiment design, discussed next.

3.2 Method

Data from this simulation study was specifically chosen because it was designed to understand how attackers constructed phishing emails, including the persuasion strategies they would use to influence end-users into responding. The dataset from this study contained specific words and sentences that participants chose to construct phishing emails during the study. Hence, this dataset was analyzed to determine the linguistic features used for deception in phishing emails and their effect on end-user vulnerability. It is not feasible to determine similar details, such as words and sentences that attackers chose to use beyond what is present in phishing templates over multiple attack attempts, from public available phishing datasets.

First, I will briefly describe the study procedures used to generate the phishing messages. Then, I will describe the methods used to validate the phishing data set generated from the

study, and finally, I will describe the procedure followed to extract and analyze the linguistic features in the messages. The protocols of the study were reviewed and approved by the Institutional Review Board(IRB) office of the Carnegie Mellon University.

3.2.1 Phase-1 of the Study

The phishing study was conducted in two phases. In phase-1 of the study, 105 participants from Amazon MTurk were recruited to play the role of a phisher. Participants in the study used a human-in-the-loop phishing simulation software which was specifically designed for the study to create and launch multiple phishing messages targeting fictional end-users. Their objective with each attack attempt ("trial") was twofold: (1) create messages that evade detection by fictional detection technology; and (2) create messages that persuade recipients of the message to respond. These were fake objectives, and no actual harm was caused. In reality, the players were only generating phishing *intent* messages without malicious payloads, targeting fictional end-user players. Also, the study procedures were designed such that participants playing the attacker role were not trained on the technical aspects involved with generation of phishing attacks (e.g., hosting spoofed webpages, typosquatting) and therefore, would not be able to transfer the experience from the experiment for malicious purposes.

The study simulated only the parts of the attack that involved phishing email message creation and did not involve the technical aspects of the attack, such as hosting spoofed websites or creating malicious payloads to be attached with the email, e.g., malware. Therefore, participants without any background knowledge in security could participate in the study. The rationale for such a simulation design is that social engineering is primarily human deception, and I need to analyze the human deception process to mitigate it.

To help participants assume the role of a phisher, they were provided detailed instructions about their role, tasks, and goals in the experiment. Participants also practiced the phishing email generation tasks before starting the main study. Like attackers in the real-world, participants in the study created the phishing emails using a randomly assigned phishing template at the start of the study. The templates were chosen randomly from a set of

ten actual phishing emails, each representing a known variant of phishing attack commonly encountered in the real world. For example, phishing templates for targeting consumer accounts (e.g., email communicating about a locked account); templates to target corporate accounts; tax refund scams; fake job requests scams; fake order placements; reward scams; and loan scams. The ten phishing templates used in the experiment did not differ significantly in their structure or word count (min = 95, mean = 111, max = 137, SD = 14).

Each participant generated a total of eight phishing messages. The messages generated did not include the sender address field. The focus of the study was to understand the effect of deception emerging from the body (text) of the email. In each trial, participants crafted a new phishing message by modifying the body of the phishing message created in the previous round. In the first trial, participants modified the body of the template randomly assigned to them.

Participants were encouraged and incentivized to create new phishing messages in each round. Participants were also told that it was imperative to modify the content of their phishing messages during each attempt to avoid detection and failure. Participants were also instructed to write strategic and influential messages that would persuade the recipients to respond immediately. After each attempt, participants received feedback on their success. Participants were rewarded, both for successfully evading detection and for successfully persuading end-users to respond.

The reward for evasion was directly proportional to the number of edits made during each round. The number of edits was calculated in real-time using the Levenshtein edit distance function - a function regularly used in text and speech processing. Participants could earn a maximum of 200 points as rewards for evasion, irrespective of the number of changes they made to the body of the message. Participants were allowed to make as many changes as they wanted. There were no time limits or word limits imposed.

The reward for persuading the end-user to respond was the result of a probabilistic function. The reward was a one-time, high-value payoff of 2000 points. Participants were eligible for earning this reward only once during the study, and their likelihood of winning

the reward increased with each attempt. Participants, however, were not aware of the fact that this was an on-time reward. The function was designed this way to keep the participant motivated and interested in the task. Also, the probabilistic nature of the reward is akin to the probability of success of mass phishing attacks in the real world.

3.2.2 Phase-2 of the study

In phase-2 of the study, a separate set of 340 participants were recruited from Amazon MTurk to play the role of an end-user performing a routine email management task on behalf of a fictional person named "Sally", a standard approach commonly used in phishing studies [110, 94]. Each participant in this phase received ten randomly selected phishing emails created by participants in phase-1 of the study. Additionally, participants also received ten regular emails. For each message, participants were presented with the subject and the body of the email, and they were asked to choose between five possible actions: 1 (respond immediately); 2 (flag and respond later); 3 (leave it in the mailbox); 4 (delete); 5 (delete and block the sender). End-users choosing actions 1 (respond immediately) and 2 (flag and respond later) in response to a phishing message are considered victims of that phishing message.

A randomization algorithm was used to distribute phishing emails to participants in phase-2 to ensure they responded to ten unique phishing emails generated by ten different attacker participants in phase-1. The randomization algorithm also ensured that each participant responded to messages belonging to all ten phishing varieties considered in phase-1 of the study. Such a randomization process was used to control for any learning effect on end-user response. The distribution was also designed such that each phishing message received responses from five different end-users. Table 3.1 presents an example of attack attempts made by a participant over two rounds. The table shows the textual changes that the participant made during each round. The participant in phase-1 deployed different impersonation strategies to influence recipients into responding to their messages. These two messages were distributed to two different groups of five end-user participants in phase-2.

Trial	**Email Body**	**Average Response Rating** ($1 \leq x \leq 5$)
1	Dear Sir or Madam: Per your instructions we have attempted to process your current payment. Because we were unable to process this payment you are in danger of having a late payment. In order for us to process your payment, please follow the link that we have provided for your ease of use. $< link >$To my account$< link >$ Thank you for banking with us. Sincerely, USAA Bank	2.75
2	Dear Sir or Madam: Because you have failed to update your account, we will find it necessary to close your account unless you immediately update it. Please promptly correct this situation by following the link at the bottom of this page. Failure to do so will result in the closure of your account. Click here Thank you for banking with us. Sincerely, USAA Bank	3.8

Table 3.1: Example of changes made to a message by a participant

The end-user participants in phase-2 provided a rating between 1 and 5 which indicated the action they would take in response to the message. For each message, responses from five end-user participants were averaged. The average response rating for each message could be anywhere between one and five where lower values indicate that most end-user participants who received the message chose to respond to the message. This work focuses on analyzing the message text data generated by participants in this study which has not been previously analyzed and reported.

3.2.3 Dataset

In this study, 674 phishing messages from a corpus of 840 messages were selected for analysis. The messages were selected based on the number of edits that participants made to the body of the message. If the message had at least 50 edits or more, it was selected for further analysis because it suggests that the participant had made at least a line or two worth of changes from their previous attempt.

As described earlier, for each message, responses from five different end-user participants were elicited. These five responses were averaged to measure the average performance of each phishing message - performance in deceiving the recipients. This response measure ranged from 1 (all five end-users chose to respond to the email) to 5 (all five end-users chose to delete the email and block the sender). Higher values indicate poor deception performance because most recipients chose to delete the message. In addition to the average response, the dataset also contained information about the participant who generated each message, the trial in which each message was generated, and the subject and body of the message. Thus, the dataset was suitable for studying the linguistic features used in a phishing email to deceive an end-user and the effect of such choices on end-user susceptibility.

However, these phishing messages were generated from a simulation experiment using participants who are not likely to be attackers in real life. Therefore, it was essential to evaluate the validity of the messages before further analysis.

3.3 Validation

Like real-world attackers, participants in the experiment created phishing emails using templates. Furthermore, participants playing the role of the attacker constructed only the text of the phishing emails which is primarily a deception/lying process and does not demand expertise in computer security. It was expected that participants would use their own experience with phishing emails to inform their choices during the study. Hence, it is reasonable to assume that phishing email text generated from the simulation experiment would resemble social engineering techniques used by attackers in real-world phishing emails. However, it was still imperative to test how well the phishing messages generated from the study were semantically similar to actual phishing emails generated by real-world attackers.

Manually validating text data generated from a laboratory study is laborious, challenging, and prone to errors, particularly when dealing with text data involving deception. Therefore, I used standard machine learning methods to evaluate linguistic similarity between actual phishing attacks and attacks generated during the experiment. Machine learning methods are used to build classifiers that discriminate phishing emails from benign emails [150, 103, 126]. Similarly, I developed supervised machine learning (ML) models to discriminate phishing emails from benign emails using the linguistic features present in the body of the messages. The model was developed using publicly available phishing datasets, for example, using data in the spam track of TREC dataset [20]. Next, I used the machine learning models to test whether the model trained on real-world phishing messages predicted the messages generated by participants during the study as phishing; to evaluate whether the messages generated by the participants were linguistically similar to actual phishing emails used to train the model.

I evaluated the data using representative machine learning models, including logistic regression(LR), XG boosting(XGB), Support Vector Machines (SVM), and Random Forest(RF). These are few of the standard methods in machine learning. Logistic regression is used to model the probability of discrete outcomes given a vector of input features. SVM is used to project input data into a multi-dimensional space and create hyperplanes to best

separate the space that discriminates the categories being predicted. Decision trees use information loss to identify the best features to build up if-else feature questions and to find the minimum number of features necessary to predict categories or classes. Generally, the decision tree approach has high variance and is considered as a weak classifier. Random Forest is based on decision trees but concatenates multiple decision trees together through an approach called bagging. Instead of the bagging method taking equal weights from each weak classifier, gradient boosting methods are used to iteratively optimize the ensemble of decision trees, which is XG-Boosting. I have used these methods to create our phishing email classifier for validation purposes. I used the spam track of TREC dataset to train the models. The TREC dataset contains 39,399 legitimate emails and 52,790 phishing and spam emails [20]. The models were trained using 5000 messages in the dataset (random selected 2500 legitimate and 2500 spam messages). The remaining messages in the dataset were set aside for model validation. The raw text data in the dataset were first preprocessed to remove links, stop words, and HTML formatting. This prepossessed data were used to train the models. I used the N-Gram and TF-IDF (Term Frequency-Inverse Document Frequency) methods to extract and tokenize the word features; a common approach used to represent text features for model training. The features extracted represented the words and topics present in the messages.

I validated the model's performance in classifying messages in the TREC dataset set aside for validation and messages in two other publicly available datasets containing phishing emails collected using a honeypot from 2005 to 2015 [84]. I will call these two datasets as Jose1 and Jose2. Table 6.5 presents the accuracy and precision of the different models on the different datasets used. As shown in the table, all four models performed well in recognizing phishing emails from legitimate emails. Finally, I tested the model's performance in predicting the messages generated from the laboratory study. I will call this the author's dataset. As shown in the table, I found that all four models accurately predicted all the messages generated from the study as phishing, suggesting the message generated by the participants in the laboratory study contained textual features akin to features present in

actual phishing messages.

Classifiers	TREC accuracy	TREC percision	TREC recall	author's dataset accuracy	Jose1 accuracy	Jose2 accuracy
Navie Bayes	95.12%	95.57%	94.97%	100.0%	97.72%	91.0%
LR	94.79%	98.95%	91.6%	100.0%	100.0%	99.94%
SVM	95.66%	98.95%	93.06%	100.0%	100.0%	94.68%
Random Forest	95.44%	98.73%	92.86%	100.0%	98.4%	99.01%

Table 3.2: Comparison of predictive accuracy between classifiers

(Jose1 contains emails before 2005 Jose2 contains emails between 2005 and 2015)

3.4 Linguistic Choices of Attackers

Next, I analyzed the linguistic choices that the participants playing the attacker role had made during each attack attempt in phase-1 of the study. Their linguistic choices were analyzed to understand the textual features used to manipulate and influence recipients into responding. Figure 3.1 is a word cloud of the words that the participants commonly added to their phishing messages. The bigger the word, the more frequent the word appeared in the messages. Although the word cloud shows the most frequently added words, it does not

reveal the underlying psychological manipulation and influence the participants intended.

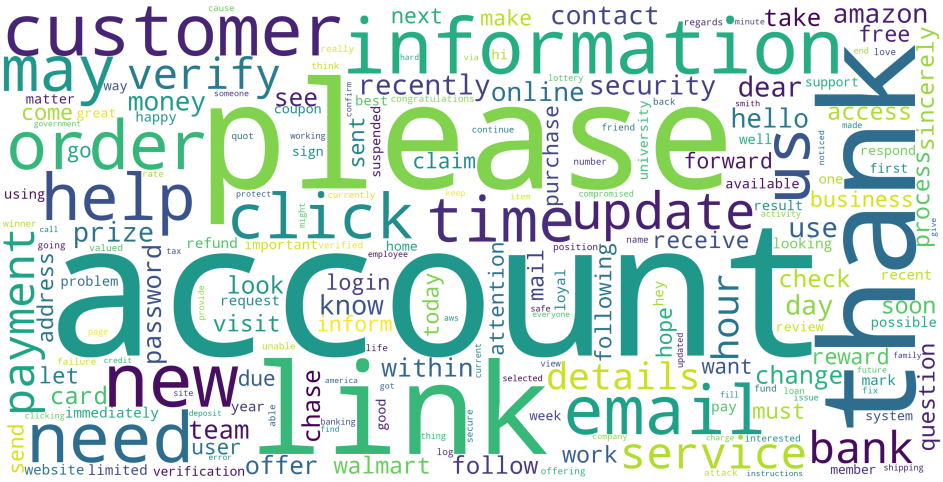


Figure 3.1: Word cloud of the most frequently added words in phishing messages

I used LIWC as a framework to do the feature extraction of phishing text. For reference, I include the LIWC categories relative to our study in Table 3.3. The table 3.3 provides examples of words that are categorized under each LIWC dimension or category. For example, words such as win, success, prize are categorized as words representing the achievement and reward categories. The table 3.3 also lists the mean percentage of each LIWC attribute found relative our dataset. For example, participants on average used more words reflecting 'cognitive process' while writing phishing emails (see the row for cognitive processes in table 3.3. Using the LIWC toolkit, I categorized words in the 674 messages along 85 linguistic dimensions present in the LIWC dictionary, including 21 grammatical dimensions (e.g., pronouns, verbs) and 41 psychological dimensions (e.g., cognition, affect). I observed large differences between participants and trials in the number of changes made during each attempt. Many participants in the study had made only a moderate amount of changes from one attack

attempt to the next. LIWC is not effective for analyzing individual words, short phrases, or short sentences because of the large number of dimensions used to categorize the input text. Therefore, instead of categorizing the text added by participants from one attack attempt to the next, I measured the relative difference in LIWC dimensions of phishing messages from one trial to the next.

For each participant and for each pair of consecutive trials, I calculated the difference in LIWC dimension between the two messages. For a pair of messages, LIWC dimensions with a positive difference will indicate that the participant added words belonging to those dimensions in the consecutive trial, and LIWC dimensions with a negative difference would indicate the opposite - the participant chose to remove words belonging to those dimensions. For example, if I observe an increase in the social dimension, that would indicate the participant chose to add words belonging to that dimension in the consecutive trial.

However, simply measuring whether an LIWC dimension increased or decreased between a pair of messages would be misleading because most dimensions tend to show marginal differences between a pair of text messages. Hence, I conducted further analysis to determine the dimensions that had a significant increase or decrease between attack attempts.

For each pair of consecutive messages, I identified LIWC dimensions with a difference of at least one standard deviation higher than the mean difference. Furthermore, taking a more conservative approach, I identified LIWC dimensions having one standard deviation difference in at least 50% of the messages. I identified 13 such LIWC dimensions, which include function words, pronouns, verbs, words describing cognitive process, social words, and words describing time and space relativity. These dimensions represent the words that participants most often used in their messages. Table 3.4 presents the LIWC dimensions that were used to modify each of the ten different phishing templates used in the study. The results show that for certain types of phishing templates, such as messages pretending to be coupons from Walmart, participants added or removed words belonging to one or two LIWC dimensions, whereas messages belonging to other kinds of templates involved the use of many more dimensions, such as messages pretending to be a security message from a bank.

How did such linguistic choices impact end-user susceptibility? To answer this question, I used penalized regression models for predicting participants' (playing the end-user role) responses to phishing emails in phase-2 of the study using the LIWC dimensions extracted from the messages created by participants (playing the attacker role) in phase-1 of the study, as described in the previous section. Since different end-user participants received messages created by the same attacker, I do not anticipate any learning effects. I discuss the results from this analysis in the next section.

3.5 Impact on End-user Susceptibility

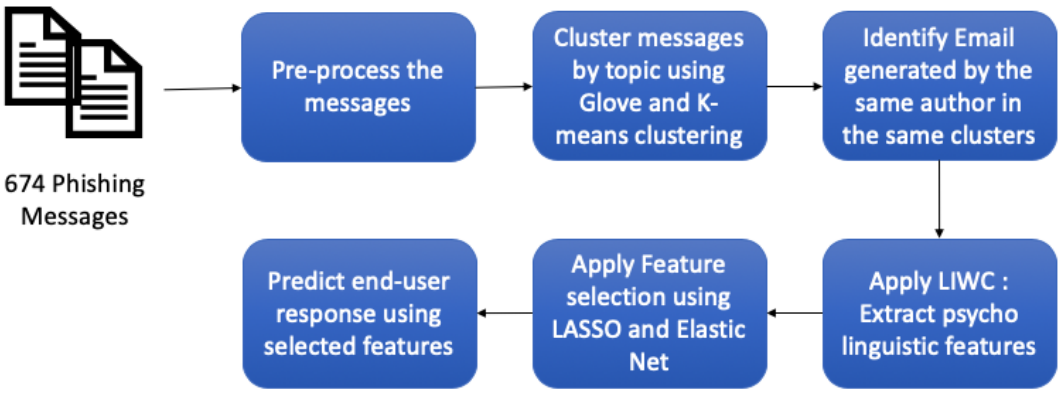


Figure 3.2: Summary of the analysis pipeline used for extracting and analyzing phishing messages

To analyze the linguistic features that the participants uniquely added and deleted during each round of the experiment, the text changes made between a pair of trials were extracted, and analyzed against the difference in phishing performance between those pairs of trials. For e.g., in trial 2 in Table 3.1, the participant added the lines *"it's very important that everyone.. ... to being written up"*. These changes, however, were not effective. Fewer people chose to respond to message in trial 2 compared to the earlier message. Average end-

user response lesser than 3 indicate that most participants who received that message chose to respond immediately or follow-up later. Average end-user response more than 3 indicate that most participants who received the message chose to ignore or delete the message. Since each message created by the attacker participants were received by 5 different end-user participants, I do not face the issue of collinearity while doing such analysis.

To extract the text changes made between a pair of trials, an analysis pipeline was employed as summarized in Figure 3.2. The 674 Phishing messages were first pre-processed to remove non-keywords and other content such as HTML formatting, numbers, and links. In the experiment, I observed that some participants playing the attacker role would work on a phishing topic for a few trials (e.g, issue with bank account), then moved to another phishing topic on their choices (e.g., email from FBI or IRS scams). Therefore, detecting changes made by attackers between all consecutive pairs of trials would not be accurate. So, after pre-processing, I applied topic modeling to cluster messages belonging to similar topics. To achieve this, the GloVe method was used as feature extraction method, which is a popular approach used for word embedding and document clustering [96]. Using the GloVe method, I transformed the phishing messages in such a way to capture the context for all the unique words present in the dataset. Towards this, I determined the number of unique words in the corpus to be 2798 and then using the GloVe method, I represented each of the unique words across the GloVe's pre-trained 300-dimension word vector. This produced a matrix of size 2798*300. I created another matrix (674*2798) that contained the number of times each of the 2798 unique words occurred in each of the 674 phishing messages. The two matrices were then multiplied to produce a final matrix (674*300) that contained the 300-dimensions word vector representation for each phishing message. This matrix was used as the input for clustering the phishing messages using K-means clustering [7]. Using the standard elbow method, the optimal number of clusters for this matrix was found to be 9. K-means clustering is a widely used unsupervised learning method that is used to group similar items in clusters. The K in k-means clustering represents the expected number of clusters. Elbow method is typically used to determine optimal values for K that leads to

minimum loss. In this work, the messages that share similar features would be clustered into one group. Using the results from K-means clustering, each phishing message in the dataset was identified with their corresponding cluster number. As a result, phishing messages sharing similar topic clustered together. Note, messages from one participant could appear in multiple clusters because a participant may have started with one topic at the beginning (e.g., bank account suspended scam). Once again LIWC features were extracted on the unique text added and deleted by each participant during each round of the experiment. After finding the pairs of phishing messages generated under the same topic by the same participant, I subtracted LIWC measurements between each pair of messages. Similarly, I also measured the change in end-user response by subtracting the average end-user response between each pair of message. I analyzed this LIWC transformed phishing dataset using two penalized regression methods (LASSO and Elastic Net), also known as regularization, because a general multivariate linear regression model would suffer from overfitting issues due to the large number of LIWC features (93) present in this dataset. LASSO [122] and elastic net [149] are the two most widely used approaches for generating a sparse model through feature selection based on linear dependencies between the input features (LIWC features in phishing message) and output values (average end-user response); by penalizing and forcing some of the input feature variables to zero, while retaining the remaining most informative features. Besides, I also analyzed the dataset using Random Forest and XGBoost [16, 76] to analyze the effects of LIWC features to average end-user response. I ran ten-fold cross validation 100 times to compare the performance of the four models on feature selection. LASSO ($R^2 = 0.233$; $MSE = 27.73$) and Elastic Net ($R^2 = 0.212$; $MSE = 26.74$). Table 3.5, presents the list of the most informative LIWC features in of text added by each participant and its effect on increasing or decreasing the vulnerability of end-user.

3.6 Discussion

Email messages produced from a phishing simulation study were analyzed using natural language processing methods and statistical models to determine the linguistic dimensions

that attackers may use while creating mass phishing messages. I also measured the impact of such linguistic choices on end-user susceptibility.

I found that more than 50% of participants while creating the phishing message often chose to use personal pronouns such as 'I', 'Me', 'She'; words conveying time-space characteristics such as 'under', 'in'; words presenting present time orientations such as 'today', 'now'; and words conveying human affiliations such as 'friend', 'social' etc. This shows that participants when asked to play the role of an attacker, they chose to deceive their end-user targets by pretending to be a familiar individual, and by presenting time pressure or deadlines. This result about the use of personal pronouns is somewhat contrary to past research which suggests that when people engage in deception, they try to distance themselves from the false narrative [85, 147, 46, 63]. The discrepancy in the results could be due to the difference in context and communication medium.

I also observed that participants regularly used words representing cognitive processing such as 'cause', 'know', and 'ought'. In a related study that analyzed LIWC dimensions in fraud messages, it was found that fraud messages have relatively higher cognitive complexity and use more WPS (words per sentence) [80]. Similarly, I found that attackers in the study used words representing cognitive complexity in 52.77% of trials. I also found that participants chose to add word relative to drives in 50.45% of trials. Drive words convey positive emotions and achievement, such as 'win', 'success'. This shows that participants often refrained from creating scam like phishing emails.

In Table 3.6, I present an example of changes that a participant made between a pair of trials, the most modified LIWC features and the average end-user response to each of these email messages. It can be observed that the participant first tried to deceive the end-user by pretending to be a Walmart customer care representative offering a reward which was largely unsuccessful as observed with a 3.2 average end-user response which meant most end-users who received this message chose to delete it. In the next trial, the attacker modifies the strategy from offering reward to one that informs an issue with the rewards account. This change leads to a higher success rate as observed with the 2.2 average end-user response

which meant at least some of the end-users who received this message respond to it.

Table 3.6 presents the list of the most influential LIWC features of the text added by participants in each round. When participants playing the attacker chose to use words corresponding to LIWC features such 'achieve', 'reward' and 'sad', the resultant phishing messages failed to deceive majority of the end-users. This is an intuitive result that most end-users were cognizant of scam like phishing emails and therefore, were less vulnerable to phishing messages containing such linguistic features. The example in Table 3.6 shows that removing words conveying reward and achieve increased the likelihood of response from the end-user.

I found that when participants chose to use words corresponding to features such as 'certain', 'work' 'differ', the resultant phishing messages were more successful in deceiving the majority of end-users compared to the earlier attempt. The LIWC features 'certain' and 'differ' are sub-features of the main LIWC feature called 'cognitive process' that represent words conveying cognitive processing such as know, inform, decision, etc. The specific feature 'certain' represent words conveying certainty such as 'always', 'never' etc. The feature 'differ' represents words conveying differentiation, such as 'haven't', 'but'. Past research have also found the use of words associated with cognitive processing is a strong indicator of computer-mediated deception [54, 53]. Furthermore, from Table 3.5 it can be found that words conveying personal concerns (e.g., work, death, home) and social affiliation (e.g., family, friends) are also significant predictors of end-user vulnerability to phishing. These features are also likely to be observed with phishing messages that are targeted in nature, and could be the topic for future work. The example shown in Table 3.7 shows that removing words conveying cognitive processing and differentiation decreased the likelihood of response from the end-users.

This study aimed at determining the linguistic elements of deception in phishing messages and its influence on end-user susceptibility to phishing. While the effect of structural elements of phishing emails (e.g., Source address, URLs) is certainly significant, this research describes the effect of email linguistics. This research shows that there could be classes of

linguistic features associated with deception in phishing emails that could be critical to the explanations for end-user susceptibility to phishing attacks. I need more work to clarify further the role of language in human susceptibility to phishing attacks.

The result from this work has multiple implications. Our result show that analyzing emails for psycholinguistic features associated with computer-mediated deception could be used to fine-tune spam and phishing detection technologies to become sensitive of linguistic features that end-users are most vulnerable. Furthermore, this work suggests that spam filters could also consider additional deception cues by including LIWC features inputs for model training. Results from this research can also be used for designing phishing training solutions. Past user research on phishing have led anti-phishing training solutions to focus the training around checking links and sender address. I recommend that future phishing training solutions could leverage results from our study and similar studies in future to train people to become aware of linguistic cues often used for deception in phishing attacks. Furthermore, I recommend future phishing training solutions to prioritize training people on phishing messages containing features people are most vulnerable to such as messages containing LIWC features such as 'certain', 'work' 'differ'.

3.6.1 Limitation and Future Work

This work has a few limitations which will be addressed in our future work. In the experiment, which is the source of the dataset, the participants playing the attacker role were told their goal was to persuade fictional end-users to respond but were not provided any specific attack goals or information about the targets. I are currently modifying this into an experiment on spear-phishing attacks where attackers will provided specific attack goals (e.g., steal bank account user name and password). Furthermore, the feedback that the attacker participants received in each trial were not directly based on actual end-user response. This was collected as part of phase-2. This may have affected the word choices that the attackers made in each trial. I are currently extending the experiment software with the capability of conducting live, multi-player experiments that allow participants to get feedback on their attacks directly

from the end-users, in real-time.

Another important limitation of the study is the use of MTurk participants as attackers, generating phishing messages. Although, I found strong semantic similarities between the email text generated by participants with text found within actual phishing emails, this does not discount the fact that it was created by regular people who may not be attackers in the real-world. Hence, future work needs to validate these findings with studies conducted with actual attackers for e.g., ethical hackers who conduct social engineering evaluations within organizations. Such studies with ethical hackers could also focus on other aspects of phishing generation such as use of phishing kits, malware deliver, domain hijacking and site hosting for data collection [68, 67]. Also, the study was focused on studying only the effect of linguistic features of phishing emails but there are other aspect of phishing emails (e.g., source address, signatures, URLs) that may interact and moderate the role of language in explaining end-user susceptibility.

The use of LIWC is a potential limitation because some of the categories present in LIWC may not be directly relevant to classifying words in phishing emails. In contrast, there is a lack of categories that would be relevant to analyzing words in phishing emails. For example, LIWC does not contain a specific category for words communicating fear and intimidation, e.g., terminated, canceled, forfeited, failure, removed, blocked, or suspended. Phishers frequently use such strategies to persuade end-user to fall victims to their attacks. So, future work could look at creating linguistic dictionaries like LIWC but for classifying and analyzing words specific to phishing attacks.

Finally, in this paper, I limited our analysis to word level analysis. I have not analyzed the phishing message at the sentence level to model the semantic relationships between sentences. Future work could focus on these factors.

Dimension	Abbreviation	Example	Mean(%)
Linguistic Dimensions			
Total function word	function	it, to, no, very	31.50
Total pronouns	pronoun	I, them, itself	9.55
Personal pronouns	ppron	I, them, her	7.39
1st pers singular	i	I, me, mine	1.35
2nd person	you	you, your	4.37
Articles	article	a, an, the	4.84
Prepositions	prep	to,with, above	8.40
Auxiliary verbs	auxverb	am, will, have	4.72
Conjunctions	conj	and, but, whereas	2.58
Negation	negate	no, not, never	0.41
Common verbs	verb	eat,come, carry	8.57
Common Adjective	adj	free, happy, long	2.75
Interrogatives	interrog	how, when, what	0.26
Numbers	number	second, thousand	2.68
Quantifiers	quant	few, many, much	0.91
Psychological Process			
Affective processes	affect	happy, cried	5.44
Negative emotion	negemo	hurt, ugly, nasty	0.50
Anger	anger	hatre, kill, annoiyed	0.07
Sadness	sad	crying, grief, sad	0.22
Social process	social	mate, talk, they	9.78
Friends	friend	buddy, neighbor	0.54
Cognitive processes	cogproc	cause, know, ought	6.10
Insight	insight	think, cause	1.81
Certainty	certain	always, never	0.55
Differentiation	differ	hasn't, but, else	1.30
See	see	view, saw, seen	0.55
Health	health	clinic, flu, pill	0.21
Drives	drives		8.14
Affiliation	affiliation	ally, friend, social	3.11
Achievement	achieve	win, success, better	1.52
Reward	reward	take, prize, benefit	1.32
Past focus	focuspast	ago, did, talked	1.23
Present focus	focuspresent	today, is, now	6.16
Future focus	focusfuture	may, will, soon	0.99
Relativity	relativ	area. bend, exit	7.91
Space	space	down, in, thin	3.60
Time	time	end, until, season	3.25
Work	work	job, majors, xerox	3.70
Home	home	kitchen, landlord	0.23
Money	money	audit, cash, owe	3.91
Assent	assent	agree, OK, yes	0.02

Table 3.3: Selected LIWC features

Template	Removed LIWC Dimensions(Percentage of Trials)	Added LIWC Dimensions(Percentage of Trials)
Walmart Coupons		social (51.02%)
Bank Security	number (61.29%)	relativ (56.45%), cogproc (54.84%), verb (54.84%), prep (54.84%), focuspresent (53.23%), function (53.23%), social (51.61%), pronoun (51.61%)
Account Limitation	money (55.1%), affiliation (51.02%)	relativ (55.1%), prep (51.02%)
Amazon account Suspension	auxverb (51.72%), affect (51.72%)	space (55.17%), ppron (55.17%), pronoun (55.17%)
Update Account		verb (53.4%), cogproc (52.43%), social (50.49%), prep (50.49%)
Tax Refund	time (64.52%), verb (51.61%)	you (51.61%)
Summer Internship	time (54.1%), money (52.46%)	cogproc (59.02%), space (55.74%), relativ (55.74%), drives (52.46%), conj (50.82%), auxverb (50.82%)
Help-Desk Support	time (54.55%)	cogproc (54.55%)
Incoming Payment		focuspresent (60.47%), you (58.14%), ppron (55.81%), pronoun (55.81%), function (53.49%)
Amazon Order	article (50.72%)	auxverb (52.17%), verb (50.72%), you (50.72%), function (50.72%)

Table 3.4: Most frequently added and removed LIWC dimensions across phishing templates.

Statistical Model	**Least Influential Dimensions**	**Most Influential Dimensions**
LASSO	achieve, reward, friend, adj, sad, social	certain, differ, work
Elastic Net	assent, achieve, friend, sad, reward, adj, home, negemo, see, health, social, prep	swear, anger, certain, work, focusfuture, negate, focuspast, interrog, differ, cogproc, i, quant, insight

Table 3.5: LIWC dimensions influential to end-user susceptibility extracted using penalized regression models

Email Body	**Achieve** (win, success)	**Reward** (prize, benefit)	**Average Response**
Dear Walmart shopper, I'm glad to inform you that you have have received a bonus in your Walmart Rewards account. To claim your bonus, please visit the link below by Thursday February 2 2017. Claim your bonus reward here: <link> Walmart Rewards < /link> Enjoy your reward bonus, and thanks for shopping at Walmart. Best wishes, Walmart Customer Service	9.52	9.52	3.2
Dear Walmart shopper, We have detected what we think to be an unauthorized use of your Walmart rewards card. To verify your identity, please click the link below before Thursday February 2 2017. <link> Verify my Walmart Rewards Account < /link> Thanks for looking after the security of your account. Best wishes, Eric Fraud Protection Services	4.0	4.0	2.2

Table 3.6: Example of phishing messages that demonstrate the change in LIWC features and average response corresponding to the changes made to the phishing text

Email Body	**Differ** (hasn't, but)	**Cogproc** (cause, know)	**Average Response**
Hello, I believe we spoke last week. I don 't mean to be a bother, but I haven't heard back from you. In case my email got lost or something, here is my resume again. Here is my <link> resume < /link> Thank you for your time. Best regards, Daniel Tyre	2.56	6.41	1.4
Hello, I believe we spoke last week. I haven't heard from your secretary. As requested, here is my <link> resume < /link> Hope I hear from you! Best regards, Daniel Esterson	0	3.45	2.0

Table 3.7: Another example of phishing messages that demonstrate the change in LIWC features and average response corresponding to the changes made to the phishing text

Chapter 4

QUANTIFYING SUSCEPTIBILITY TO INFORMATION EXPLOITATION IN SPEAR-PHISHING ATTACKS

Despite a substantial body of research on the human factors of phishing, there is surprisingly little understanding about spear-phishing attacks. For instance, why does the exploitation of personal information make people more vulnerable? How does it affect human suspicion and trust? What are the different ways an attacker could exploit personal information to create spear phishing attacks and how do they impact end-user susceptibility? These are some of the important questions about spear-phishing attacks that remain largely unanswered. Research on spear phishing attacks, thus far, has been centered around uncovering the technological characteristics of current or emerging threats [25], e.g., domain names, phishing kits, software vulnerabilities exploited, and payload information.

A primary reason for the lack of human studies on spear phishing is the lack of methods and data sets. To fill this gap, I have developed a synthetic task environment called *Spear-Sim* that is conducive to capturing both attacker and end-user behaviors. In this paper, I report findings from an exploratory experiment I conducted using the environment to understand how adversaries might exploit personal information while creating spear-phishing emails and the effect of such information exploitation on end-user decision making. This paper makes the following contributions: 1.Introduces a novel simulation methodology for studying the human factors of spear-phishing; 2.Discusses the challenges faced with conducting multi-player experiments on spear-phishing attacks, and 3.Discusses the results from the experiment that shows how people are vulnerable to certain kinds of phishing attacks and information exploitation.

4.1 Objective

To study susceptibility to spear-phishing attacks, I need to look beyond end-users. I must analyze susceptibility to spear-phishing attacks at a dyadic level to understand how deception, personalization, and persuasion initially manifest in targeted attacks and how they may influence end-users into falling victims to attacks. Researchers have analyzed phishing kits and domains hijacked by attackers to reveal technical behaviors associated with the design and delivery of phishing attacks, including countermeasure strategies (e.g., URL shortening and re-direction) that prevent the detection of attacks (add references). However, they are often analyzed independently of end-user behaviors and responses. In this study, I present a spear-phishing simulation system called SpearSim to simulate and study the interactions between attackers and end-users. Using SpearSim, I conducted an experiment to answer the following research questions:

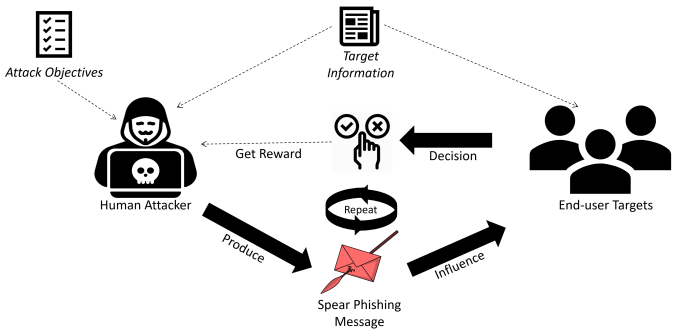


Figure 4.1: Overview of the SpearSim system design

- Research Question 1: Does availability of more information about targets enable attackers to create more convincing attacks? Specifically, how would attackers utilize targets' personal information to target individuals?

- Research Question 2: Previous work [100] indicates that the strategies used in phishing

emails may impact end-user susceptibility, but how do the persuasion and impersonation strategies implemented in spear-phishing emails impact end-user susceptibility?

- Research Question 3: From an end-user perspective, how effective are people at detecting spear-phishing attacks? How much more vulnerable are people to spear-phishing attacks compared to mass phishing attacks?

4.2 Method

4.2.1 Synthetic Task Environment

To study attacker and end-user behaviors associated with spear-phishing attacks in the laboratory, I designed a four-player synthetic task environment called SpearSim [143, 145]. Figure 4.2 presents the general structure of the SpearSim system. As shown in the Figure 4.2, each simulation run requires a group of four human subjects (players). Three players in each group are randomly chosen to play the role of an end-user, processing email massages during the study. Unbeknownst to them, the remaining participant in the group is chosen to play the role of attacker. The attacker's task is to target the end-user participants in the group by exploiting their personal information and launching spear-phishing attacks.

Participants playing the attacker role in the group are given three objectives to accomplish during the simulation: (1) steal bank account credentials from the target; (2) lure the target into downloading an attachment; and (3) steal work account credentials from the target. They are representative of the most common phishing objectives seen in the real-world [3]. The attackers are incentivized to persuade end-users to respond to their phishing emails. If the end-users choose to respond to a spear-phishing message, they are considered victims, and the attack trial is counted as a success for the attacker. For each successful attempt, the attacker in the group receives 1000 points. For each attack objective, attackers are provided with an initial phishing template, which participants modify and personalize using information available to them about their end-user targets. For example, for the objective of stealing work account credentials, the phishing template had a pretext of an email from

a business partner advertising their incoming products. No harm is caused because the attacker's task is to create phishing *intent* messages without any malicious payloads or links, and end-users are making decisions about whether they would respond to messages presented to them in this role-playing scenario.

Participants playing the end-user role are given a fictional persona to assume during the experiment. The persona included four types of personally identifiable information, as shown in Table 4.1. The persona design takes inspiration from the four privacy circles of individuals: the individual privacy layer, relational privacy layer, group privacy layer, and socio-political privacy layer [117]. During the experiment, the end-users performed a routine email management task as if they were that fictional persona - deciding whether to respond to the email messages received on behalf of the role they are assuming. To enable synchronized interactions between attackers and end-users, while end-user participants made repeated decisions about the benign email messages received on behalf of their fictional persona, the attacker launched spear-phishing messages targeting the end-users in the group. Role-playing is the standard approach used for conducting phishing experiments in the laboratory [94, 73]. Such personas and narratives are intended to simulate the target context, essential for studies on spear-phishing attacks.

Ham Email Corpus for End-User Profiles: To create a context for participants playing the end-user role in the group, I used ham (benign) emails addressed to actual people in publicly available Enron data set [64]. I randomly selected the names of three people and 70 emails from their inboxes from this dataset. During the experiment, the three end-users in each group assumed the names of the three people I extracted from the dataset and made decisions on the emails they received. These benign emails were intended to provide the necessary experience for the participant assuming the role. The ham emails were also essential for signal detection theory measurement [79] - to measure an individual's ability to discriminate spear-phishing from benign emails. In addition, end-user participants also received promotional emails and mass phishing emails, which had been randomly chosen from data sets used in past studies (references removed for anonymized review).

The rationale for such a simulation design is that social engineering is primarily a form of human deception and does not require extensive technical know-how, particularly in terms of message creation. Using a simulation with fictional phishing objectives and fictional target information enables researchers to conduct harmless and reasonably valid laboratory experiments.

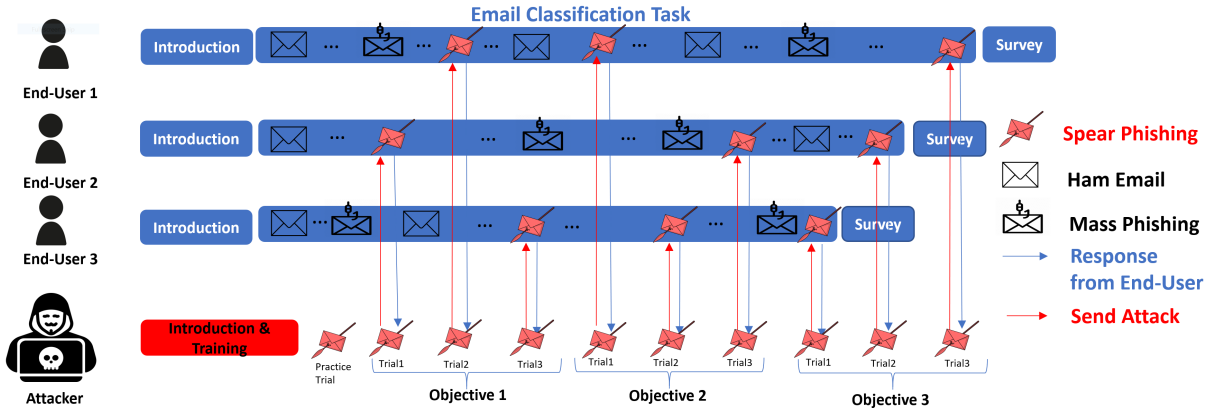


Figure 4.2: Experiment procedure overview

4.2.2 Experiment Design

Students from a large public university in the United States were recruited for the experiment. Twenty-eight groups of four persons each were recruited via email. The median age of the participants was 21 (SD = 3.6). Of the participants, 45.23% were juniors or seniors in college, and those remaining were pursuing graduate-level degrees. Due to Covid-19 restrictions, students participated in the experiment remotely from their homes. I developed SpearSim using PhP and MySQL, popular web application packages. Therefore, it was feasible to conduct experiments online without significant disruptions. On average, each participant received $13.50 for participating in the experiment.

Using the between-subjects design, I randomly assigned each participant group to one

of the two experiment conditions. Depending on the experiment condition assigned, the participant playing the role of attacker in the group received less or more information about the end-user targets in the group. In the first condition (low information), the attackers were given access to only primary information about their end-user targets (see Table 4.1). In the second condition (high information), the attackers were given access to primary, personal, and professional information about their targets. Attackers in both conditions were tasked with using the information provided to craft targeted phishing messages.

4.2.3 Procedure

All participants used an Internet browser of their choice (e.g., Google Chrome) on their personal laptop or desktop to access the SpearSim simulation system and participate in the experiment from their home. After providing informed consent, participants waited for others in the group to join the session. After successful grouping, the experiment was initiated. Three out of the four participants were randomly assigned the role of an end-user, and the remaining fourth participant was assigned the role of the phishing attacker. The experiment procedure consisted of three phases: Phase 1 - task introduction and training, phase 2 - the main experiment, and phase 3 - post-experiment survey and debriefing.

End-user: Participants playing the end-user role started the experiment at the same time as the attacker in the group. They were told that the goal of the study was to understand how people manage their emails because past research has shown that when participants know that the study is about detecting phishing emails, they tend to become cautious and biased towards reporting more emails as phishing [94]. Each end-user in a group was assigned a unique fictional persona. They also had access to full information about their persona, including their name, personal details such as location and birth date, details about the persona's work and personal interests (see Table 4.1). Following instructions, end-users in the group received a series of email messages that included benign emails addressed to the role they were assuming, a few promotional mails, and spear-phishing messages from the attacker in the group. For each email, participants playing the end-user role choose one of

five possible actions: (1) respond immediately; (2) flag the email for follow-up; (3) leave the email in the mailbox; (4) delete the email; or (5) delete the email and block the sender.

Initial	Category	Type of Information Included
Pri	Primary	Name, Gender, Location, Birth Date, Education
Per	Personal	Marital Status, Spouse, Children, Bank Accounts
Pro	Professional	Company, Position, Years of Service, Co-Workers
Int	Interests	Personal Hobbies, Recent Social Media Activity

Table 4.1: Overview of target information

Attacker: To help participants assume the role of an attacker, they were given detailed instructions about their role in the study. They also received training to practice the tasks they were expected to perform during the study. Muted videos were used to describe basic concepts about spear phishing attacks and how attackers typically exploit personal information to target people. The information they received was similar to what would also be included in an anti-phishing training program. To help participants with little to no background in cybersecurity, all the training material was presented without any jargon using visual aids and real-world examples. Participants were quizzed to help them recall the important information they received during training. The training procedure was improved through multiple rounds of pilot testing. Emphasis was given to make the training material concise, and jargonless. After training, the attacker gained hands-on experience in a short practice trial. Following the practice trial, the attacker performed the main tasks of the experiment.

During the main phase of the experiment, the attacker generated spear-phishing emails corresponding to three phishing objectives described earlier. The three objectives were assigned in a random order. During each trials, the attacker knew the trial's objective and had access to the information about the target. Depending on the information condition, the attacker in the group received either a low amount of information or a high amount of information about the target. Participants playing the attacker role customized the initial

phishing template using information available about the target and launched spear-phishing attacks targeting each end-user in the group. Attackers were told that sending the phishing template as-is was likely to be unsuccessful. The attackers conducted a total of nine spear-phishing attacks. During each trial, after crafting the spear-phishing email, the attacker self-reported the persuasion, impersonation, and emotional strategies they had employed in their attack. As shown in Figure 4.2, each spear-phishing email created by an attacker participant was sent to the intended end-user for their response decision. From the end-user perspective, the spear-phishing email appeared as another email stimulus about which they had to make a decision. The end-user's choice was sent as feedback to the attacker, who then moved on to the subsequent trial and targeted a different end-user in the group. This cycle was repeated until the attacker had completed all three phishing objectives. The experiment protocol was approved by the Institutional Review Board (IRB) at the authors' university.

Impersonation	**Keywords**
Pretended to be a government/law enforcement officer	Authority
Pretended to be a spouse/partner/ relative/friend/acquaintance	Friend
Pretended to be a workplace colleague/supervisor	Co-worker
Pretended to be an automated software reminder or notification	Notification
Pretended to be an IT/technology expert (e.g., email from the technology office at workplace)	Tech Expert
Pretended to be from a commercial organization (e.g., bank, store, shopping website)	Commercial Organization

Table 4.2: Impersonation choices provided to attackers in each trial

4.2.4 Pilot Test

Initially, a pilot experiment with five groups was conducted to evaluate the simulation software, experiment design, and procedure. The observations from the pilot experiment sug-

Social Influence	**Keywords**
Offering something (e.g., offering a reward)	Offer
Pretending to follow up on an earlier communication	Follow-up
Threatening unfavorable consequences (e.g., disclosing websites visited to police)	Threat
Providing information about a problem/failure and extending help (e.g., hacked account)	Failure
Appearing to be from a person or institution of authority (e.g., CEO, IRS, FBI)	Authority
Applying peer pressure by indicating that other people, often peers, had already taken this action (e.g., 80% of your friends have updated to this new version)	Inform
Providing information about an exclusive offer or resource or time (e.g., limited offer or deadline)	Time
Pretending to be a usual (personal or work-related) request	Usual Request
Trying to be familiar or desirable in terms of shared interest, beliefs, or background	Interest

Table 4.3: Social influence survey choices

gested there were issues with task synchronization. I observed that the end-users made decisions about the email stimuli at a brisk pace, while the attackers lagged behind due to their additional training and task requirements. To mitigate this problem, I streamlined the attacker training and introduced an additional secondary task for each end-user to perform. This secondary task was an image-processing task [47] in which each end-user was periodically (every five trials) presented with a simple puzzle that required counting the mathematical symbols shown in an image. Through follow-up tests, I found that these design changes were sufficient to resolve the synchronization challenge I originally faced.

Figure 4.3 presents the timing of the average progress made by the attacker and end-users. The x-axis represents the time (minutes) elapsed since the start of each experiment session. The y-axis on the left is used to represent the trials completed by the attacker. The blue smoothed line shows the average time taken by attacker participants to complete each

trial. Similarly, the y-axis on the right is used to represent the trials completed by end-users and the orange smoothed line shows the average time taken by end-users to complete each trial. From the slope of the lines, it can be observed that end-users completed their trials at a slightly faster pace than attackers. I found that the participants playing the attacker role, on average, took 12.2(std=3.8) minutes to read the instruction and to complete training. End-users, however, took only 3.8(std = 2.3) minutes to read the instructions related to their role in the experiment. I found that by the time end-users received their first spear-phishing email from the attacker in the group, they had processed, on average 25.26 ham emails related to their fictional role. This initial experience provides the necessary bootstrapping for the end-user participant to get familiar with their role and context. I found that attackers took an average of 6.78 min to complete each attack, whereas end users took an average of 0.68 minutes to make decisions on each email stimulus. I found that attackers took an average of 67.6 minutes(std= 16.9) to launch the last attack and the average total number of emails processed by end-users were 78.9. Due to the variance in the speed with which different end-users processed the emails, I observed large differences on when end-users received their last attack. This is reflected in the large variance observed in the top right corner of the graph in Figure 4.3.

4.2.5 Email Edits from Attacker

I measured the number of edits that was made by each attacker participant in each trial. The number of edits was measured to determine the effort participants exerted in drafting and customizing spear phishing emails during the trial. This was measured using JavaScript code that counted all key presses made by the attacker participant in each trial. Figure 4.4, shows the average number of edits made in each trial. Values shown in cyan belong to participants in the high information condition and the values shown in red belong to participants in the low information condition. In figure4.4, trial-1 to trial-3 represent the number of edits associated with attacks generated in response to the first objective, trial-4 to trial-6 represent the number of edits in response to the second objective and finally, trial-7 to trial-9 represent the

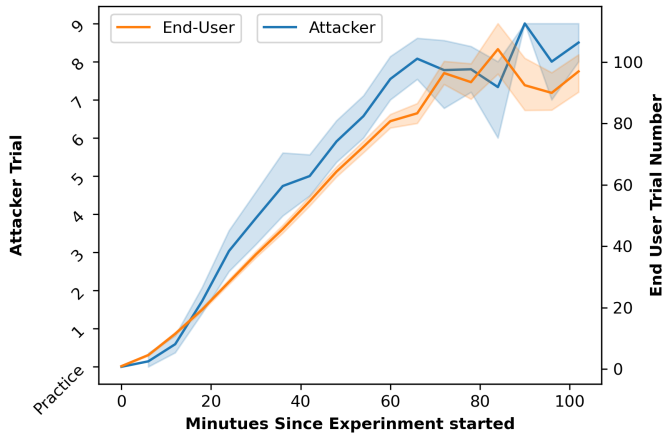


Figure 4.3: Attacker and End-user task time synchronization

number of edits made in response to the third objective. It can be observed that participants in the high information condition made significantly more edits compared to participants in the low information condition. I also observed that participants made a higher number of edits while drafting the first attack of each objective (see trial numbers 1, 4, and 7) compared to the consecutive trials for the same objective.

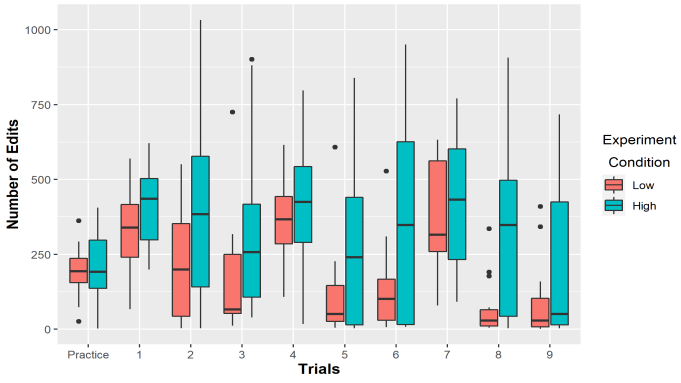


Figure 4.4: Number of edits made vs trials

Using a mixed effects ANOVA model, I analyzed the factors that would explain the variance observed in the number of edits made by the attacker. I tested the effect of experiment conditions, trial number, phishing objective, and the effect of outcome in the previous attack - whether the participant made more edits when their previous attempt was successful. Experiment conditions($F(1,26.14)=6.14$, $p=0.019$), trial number($F(7,183.06)=8.78$, p ¡0.001)) and response from previous trial($F(1,192.75) = 6.06$, $p=0.014$) were all found to have a significant effect on the number of edits.

A Tukey post hoc analysis revealed that attackers made 66.94 ($z=-2.463$, $p=0.0138$) more edits in the trial following a successful attack compared to trials following a negative response (i,e., failure). I also found that participants made fewer edits in the second and third trials of each objective. Finally, I found that participants in the high information condition made, on average, 167.1 ($z=2.48$, $p=0.0131$) more edits than the attackers in the low information condition. Participants appear to exert more effort in customizing their spear phishing emails with the availability of more information about the target.

4.3 Results

In this section, I report results pertaining to the evaluation of the simulation software I have developed and main results from the experiment conducted. I analyzed the data collected from the experiment to understand how the availability of personal information impacts adversarial behaviors and end-user susceptibility to spear-phishing attacks. Specifically, I analyzed the data set to understand how attackers exploited personal information about their targets when creating spear-phishing attacks, how attackers' performance and strategies differed under different information-availability conditions, and whether more personalized phishing attacks led to higher end-user susceptibility. Note that the pilot test data was excluded from the analysis. Valid end-user responses were critical for answering these questions. Therefore, I first analyzed responses to emails to test whether participants playing the end-user role had responded randomly. Secondly, I measured whether the experiment's design (i.e., role-specific narratives and emails used in the experiment) introduced spurious

differences in participants' susceptibility to phishing attacks.

On average, participants playing the end-user role responded to 79.9 emails. Figure 4.5 shows the distribution of end-user responses to different types of emails received during the experiment. To formally test if the participants had responded randomly, a chi-square goodness of fit test was performed. This test allowed us to determine whether the responses to different types of emails were likely to come from a theoretical uniform distribution. I found the distributions of responses to emails used in the experiment were significantly different from a uniform distribution (mass phishing : $\chi^2(4) = 80.77, p < 0.001$, primary email: $\chi^2(4) = 713.86, p < 0.001$, promotional email: $\chi^2(4) = 480.95, p < 0.001$ and spear-phishing email: $\chi^2(4) = 77.94, p < 0.001$). I also found that participants were more likely to reply to legitimate emails intended for their assigned fictional persona, whereas they usually deleted or left emails of a promotional variety in their mailboxes. For emails of the mass phishing variety, the participants often chose to either *respond immediately* or *delete the email and block the sender*. I found that participants did not fall victim to 55.66% of conspicuous mass phishing attacks, such as scams touting rewards or lotteries, whereas they fell victims to phishing attacks informing them about an issue with an online account. Importantly, I found that participants were most vulnerable to spear-phishing emails and chose to *respond immediately* or *follow up later*, as shown in the right-most graph in Figure 4.5. I compared the differences in participants' responses to mass phishing emails under the two different information conditions. I conducted a Welch t-test that showed no significant difference in responses to phishing emails between the two information conditions (t(df = 721.41) = -0.37,p = 0.7076). The non-significance further validates our design such that between-subject differences did not bias the participants' responses to mass phishing attacks.

Furthermore, I measured the consensus in responses between participants inhabiting the same persona. For example, I measured the consensus in responses to legitimate emails from all the participants inhabiting the Mark Taylor persona during the experiment. I conducted an inter-class correlation (ICC) analysis of the responses from each of the end-user personas following the guidelines suggested by Koo and Li [65]. Since the emails were randomly

assigned to participants during the experiment, the data fit the one-way random-effects ANOVA model. For each end-user persona and for each legitimate email received on behalf of that persona, I collected responses from 28 participants (i.e., raters). The ICC coefficients for the three personas, Mark, Joe, and Jeff, were 0.88 (95% CI: [0.84,0.92]), 0.89 (95%CI: [ICC 0.85,0.92]), and 0.86 (95%CI: [0.80,0.90]), respectively. Consensus is generally considered to be good if the ICC coefficient is between 0.75 and 0.9 [65]. Our results suggest that participants inhabiting the same persona in the experiment largely chose similar responses to the legitimate emails they received.

Additionally, I measured if there were significant differences in susceptibility to spear-phishing attacks among the three end-user personas used in the experiment; that is, whether the profile narrative and emails used for a certain persona caused participants to assume that persona was especially susceptible to phishing attacks. To test this, I performed a binary logistic regression analysis and categorized the options 'reply immediately' (RI) and 'flag and follow up later' (FF) as high-vulnerability. I merged the rest of the options, including 'leave in mailbox' (L), 'delete email' (D), and 'delete and block the sender' (DB), to form the low-vulnerability category. I found no significant differences in susceptibility to spear-phishing attacks among participants that had assumed the three personas ($\chi^2(df = 2) = 2.31, p = 0.31$). I likewise found no statistically significant evidence showing that the various personas affected the responses. Taken together, these analyses suggest that participants' responses to the emails are valid for analyzing how information availability, attack pretext, personalization, and attack strategies affect end-user susceptibility, as described in the subsequent section.

4.3.1 Impact of Information Availability and Pretext

To answer the first research question (i.e., Does more information available about targets increase attackers' ability to persuade end-users into responding and whether end-user susceptibility differ by attack pretext?), I used a binary logistic regression analysis to compare end-user responses to spear-phishing attacks between the two information-availability con-

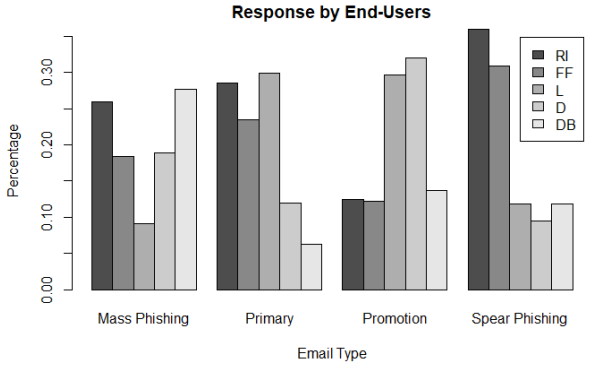


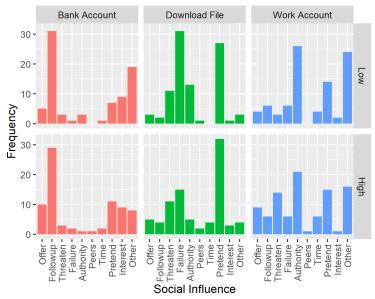
Figure 4.5: End-user responses across different types of emails

RI: Respond (reply or take action) immediately, FF: Flag the email and follow up later, L: Leave in the mailbox, D: Delete the email, DB: Delete the email and block the sender

ditions and three types of attack pretexts and goals, and attacker as the random effect. Five levels of responses were split into two categories: response ('respond immediately', 'flag and follow up later') and ignore ('leave in the mailbox', 'delete the email', 'delete the email and block the sender'). Table 4.4 reports the results from the binary logistic regression analysis. I found that both information availability and pretext had a significant impact on end-user susceptibility to spear-phishing attacks. The $\hat{\beta}$ coefficients for the experiment conditions reported in Table 4.4 indicate a significant negative association between information availability and end-user responses: end-user participants in the high information-availability condition were more like to respond to spear-phishing attacks than those in the low information-availability condition. The $\hat{\beta}$ coefficients provided in Table 4.4 are given in their log of odds ratio, which means that compared to end-user participants in the low information-availability condition, participants in the high information-availability condition were $e^{1.09} = 2.97$ times more likely to fall victim to spear-phishing attacks. Similarly, when responding to attacks developed from download attachment pretext(pretending to be about job applications mostly), participants were $e^{2.11} = 8.24$ times more susceptible than they were when responding to attacks developed from Bank Account pretext(pretending to be

(a) Responses across pretexts and experiment conditions

(b) Self-reported impersonations across pretexts and experiment conditions



(c) Self-reported social influence across pretexts and experiment conditions

about issues with a bank account) (see $\hat{\beta}$ coefficients for pretext (Bank Account—Download Attachment) in Table 4.4) Participants were also found no significant vulnerability (p = 0.6) between responding to attacks developed from work account pretext and work account pretext (see $\hat{\beta}$ coefficients for pretext (Work Account—Download Attachment) in Table 4.4).

I also found that attackers made a greater effort to generate spear-phishing emails under the high information-availability condition than under the low information-availability condition. I ran a logistic regression using the scaled number of edits as the independent factor and response as the dependent factor. I found that the more edits there were, the more vulnerable the spear-phishing email was to end-users (z=-2.95, p=0.003).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.31	0.36	0.87	0.39
Pretext(Download Attachment—Bank Account)	-2.11	0.45	-4.73	¡0.001
Pretext(Work Account—Bank Account)	-0.18	0.34	-0.52	0.60
Exp Condition(High—Low)	-1.09	0.45	-2.42	0.016

Table 4.4: Logistic Regression for pretext and experiment condition

4.3.2 Effect of Information Use

Why were participants in the high information-availability condition more susceptible to spear-phishing attacks than those in the low information-availability condition? One hypothesis is that attackers in the high information-availability condition were able to exploit the additional information available about their targets to create more persuasive attacks. To test this hypothesis, I analyzed the effect of using different types of target information in creating spear-phishing attacks on end-user susceptibility. During each trial, participants playing the attacker role in the experiment reported the type of information about the target (e.g., primary, personal, professional; see Table 4.1) they used to create the spear-phishing attack. Across the two conditions, 252 (28 groups * 9 spear-phishing attacks per group) spear-phishing attacks were created by the participants. Out of the 252 emails, 126 were created by participants in the low information-availability condition who had access to only primary information including name, gender, location, birth date, and education. The remaining 126 emails were created by participants in the high information-availability condition who had access to three categories of information: primary, personal, and professional (see Table 4.1 for details on what is included in these categories). To compare the effects of different types of information use, I excluded data from the low information-availability condition since participants in that condition had access to only one category of information. Focusing our analysis on data from the high information-availability condition, I found seven different ways participants in that condition used information to create attacks. In these attacks, the participants reported exploiting (1) only the primary information about the target (n=25); (2) only professional information about the target (n=6); (3) only personal information about the target (n=6); (4) both primary and professional information (n=45); (5) both primary and personal information (n=17); and (6) all three categories of information about the target (n=8). Using binary logistic regression analysis, I compared the difference in end-user susceptibility between attacks using only primary information (reference category) and the remaining six ways that information about the target was used to

create spear-phishing attacks during the experiment. Table 4.6 reports the results of this analysis. As shown in Table 4.6, I did not find a significant difference in end-user susceptibility to attacks using only primary information and attacks using other combinations of information about the target. This result suggests that spear-phishing attacks personalized with more information may not necessarily lead to higher end-user susceptibility. I further investigated how targets' personal information was used by attackers to deceive end-users. I found that while attackers used different information to deceive end-users, they mainly used two types of information: the target's name and the name of the organization where the target worked. Attackers personalized the attacks with their targets' names in 79.36% of the attacks, and I found that no significant difference between the end-user response distribution of spear-phishing attacks with and without targets' names with $\chi^2(1) = 0.55$.

In 46.03% of spear-phishing emails, attackers used the company name **Enron**. I observed that spear-phishing attacks including the company name 'Enron' received 15.9% more end-user responses than spear-phishing attacks that did not include it. In other words, attackers were more successful in deceiving end-users when they used the name of the organization where the given end-user was employed. Next, I analyze the effect of impersonation strategies.

4.3.3 Effect of Impersonation

Studies have shown that social influence strategies employed by attackers can significantly affect end-user susceptibility (references removed for double-blind review). For each trial in the experiment, participants playing the attacker role reported who they had impersonated (e.g., friend, co-worker, IT support person) and the social influence strategies (e.g., threaten, inform of failure, present a deadline) they had employed in the attack. Attackers were asked to choose the most representative impersonation from the available options (see Table 4.2) and select all applicable social influence strategies they had used in the given attack (see Table 4.3). The impersonation strategies and social influence strategies reported by the participants are shown in Figure 4.6b and Figure 4.6c, respectively. Notably, I found that

some of the strategies were more dominant in certain pretexts. For example, I found that among attacks using the pretext about an issue with a bank account, attackers frequently pretended to be following up on an earlier communication, whereas among attacks using the pretext about an issue with a work account, attackers frequently used an authoritative tone to persuade end-users into responding to their messages.

Importantly, as shown in Figures 4.6b and 4.6c, I found the distribution of social influence and impersonation strategy choices also varied between the two experiment conditions. For each pretext type, chi-square tests were used to compare social influence and impersonation strategy choice distributions between the two experiment conditions. The results from the chi-square tests are shown in Table 4.5. Unlike in the distribution of impersonation choices for the job application pretext, I found significant differences between the two conditions in the social influence and impersonation strategy choice distributions. As shown in Figure 4.6b, for attacks using the job application pretext, participants in both conditions most frequently reported impersonating a friend or a workplace colleague to persuade recipients to respond. However, I found differences in social influence and impersonation choice distributions in the other two pretexts. Generally, I observed that participants in the high information-availability condition explored more strategies in their attacks. For example, among attacks using the bank account pretext, I observed that participants in the low information-availability condition most often pretended to be from a commercial organization or automated software sending a notification about an issue with the target's account, whereas participants in the high information-availability condition reported using more diverse types of impersonation and social influence strategies. I observe similar differences in exploratory behavior in impersonation choices among attacks using the pretext of a work account in Figure 4.6b and in social influence strategy choices between the two conditions in Figure 4.6c.

Were certain impersonation strategies more successful than others? I analyzed the relationship between the impersonation strategies self-reported by attackers for each attack they generated and end-users' responses to the attack. As shown in Table 4.7, compared to at-

tacks impersonating a commercial brand, end-users were $e^{2.58}$ =13.2 times more vulnerable to attacks impersonating a friend, $e^{3.26}$ =26.0 times more vulnerable to attacks impersonating a co-worker, and $e^{2.15}$ =8.6 times more vulnerable to attacks that used other impersonation strategies. I analyzed the messages that were reported as using the *other* category and found impersonation of a job seeker, salesperson, and spouse.

Strategy	Pretext	χ^2	P value
Impersonation	Bank Account	78.9	< 0.0001
Impersonation	Download Attachment	10.42	0.660
Impersonation	Work Account	22.1	0.054
Social Influence	Bank Account	20.35	0.016
Social Influence	Download Attachment	33.62	0.0001
Social Influence	Work Account	17.1	0.047

Table 4.5: Results of chi-square test for strategy choices

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.88	0.65	-1.37	0.17
Pro—Pri	-0.85	0.95	-0.90	0.37
Per—Pri	1.85	1.40	1.32	0.19
Pri+Pro—Pri	-0.95	0.70	-1.36	0.17
Pri+Per—Pri	-1.32	0.93	-1.43	0.15
Pri+Per+Pro—Pri	-1.20	1.18	-1.02	0.31

Table 4.6: Binary Logistic Regression for analyzing impersonation

4.3.4 Who Is Best at Spear Phishing?

This section analyzes the spear-phishing performance at the attacker level. Specifically, what demographics of people are most adept at drafting spear-phishing emails? I use ANOVA to analyze the demographic factors driving participant-level performance. Attackers' performance is defined as the number of successful spear-phishing attacks, or attacks to which the target replied immediately or flagged and followed up on later. The performance ranges from 0 to 9. For example, if an attacker succeeded once, then the attacker's performance would be

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.23	0.42	0.56	0.58
High—Low	-0.71	0.51	-1.40	0.16
Friend—Commercial	-2.58	0.83	-3.11	0.00
Co-worker—Commercial	-3.26	0.88	-3.71	0.00
Notification—Commercial	0.11	0.60	0.19	0.85
Tech Expert—Commercial	0.33	0.45	0.73	0.46
Other—Commercial	-2.15	0.61	-3.50	0.00

Table 4.7: Binary logistic regression for analyzing impersonation

1. Age, gender, and primary language (binary) were used as independent variables and the experiment conditions were used as blocks. As Table 4.8 indicates, age and gender did not show a significant impact on spear-phishing performance. Primary language, however, did show a significant impact on spear-phishing performance. That is, participants who reported having English as their primary language performed better than other participants whose primary language was not English. Out of the 28 attacker participants, 15 had a primary language that was not English. Language fluency plays a vital role in crafting spear-phishing emails, which suggests why I found that primary language speakers succeed on average 1.76 (t =-2.422 ,p=0.237) more than non-native language speakers.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Experiment Condition	1	22.32	22.32	7.91	0.0099
Age	1	1.88	1.88	0.67	0.4230
Primary Language	1	17.77	17.77	6.30	0.0196
Gender	1	0.07	0.07	0.02	0.8770
Residuals	23	64.92	2.82		

Table 4.8: Number of successful spear-phishing emails vs. demographics

4.3.5 End-user Awareness

At the end of the experiment, end-user participants were asked to fill out a security awareness survey [31]. The survey consisted of 16 questions about security behavior, and the responses

were used to categorize end-users according to four dimensions, including password generation (e.g., creating strong passwords, changing passwords), device securement (e.g., using a PIN on a smartphone, locking a desktop screen when stepping away), proactive awareness (e.g., checking links before clicking them), and updating (e.g., applying software updates in a timely manner). One of the participants had a network issue and did not fill out the survey; this data was omitted. The security survey was included in the experiment to validate the experiment results by testing whether the participant pool in the two experiment conditions had the same level of security awareness. Hotelling's t-squared analysis was conducted to compare whether the two samples were from populations with the same multivariate mean [24]. I found that there was no significant difference in the security awareness of participants under the two conditions and summarized the result as ($T^2 = 0.71, F = 2.49, df = 4.78, p = 0.59$). The result indicates that end-user participants' vulnerability to spear-phishing emails was evaluated with the same baseline security awareness under the two conditions, which adds validity to our results.

I also tested another hypothesis about whether participant performance varied in response to mass phishing or spear phishing based on the participants' security awareness survey scores. I used password generation, device securement, proactive awareness, and updating as the independent variables. Mass phishing performance and spear-phishing performance were used separately as the dependent variables, and attacker ID was used as the random effects. A participant's performance in mass phishing and spear-phishing was determined by identifying the percentage of phishing emails that were rejected by the participant. Spear-phishing performance was defined as the frequency with which an end-user rejected spear-phishing attacks.

	Password	Securement	Awareness	Updating
Phishing Response	$F_{1,78} = 0.189, 0.66$	$F_{1,78} = 0.038, 0.846$	$F_{1,78} = 0.076, 0.783$	$F_{1,78} = 0.793, 0.376$
Spear-phishing Response	$F_{1,78} = 1.15, 0.289$	$F_{1,78} = 1.17, 0.2837$	$F_{1,78} = 0.314, 0.578$	$F_{1,78} = 1.07, 0.304$

Table 4.9: P-value of security awareness factors for phishing and spear-phishing performances (F value, P value)

No significant factors affecting end-users' phishing performance or spear-phishing performance were found. Specifically, the awareness category did not show significance for either phishing or spear-phishing performance. Although the survey estimated security awareness in multiple dimensions, I did not find survey performance to have any significant relationship with phishing susceptibility.

4.4 Discussion

Using *SpearSim*, I analyzed how access to personal information impacts attacker and end-user behaviors in spear-phishing attacks. I found that participants playing the end-user role in the high information-availability condition suspected only 19% of the spear-phishing emails they received and therefore were significantly more vulnerable. In comparison, end-users in the low information-availability condition were suspicious of 44% of the spear-phishing emails they received. I did not observe any significant differences in end-user security awareness or mass phishing detection capabilities between the two conditions. I also did not observe differences in susceptibility due to the differences in the fictional personas the participants assumed during the experiment.

Why were end-users in the high information-availability condition more susceptible to spear-phishing attacks? I found that when attackers had access to more information about their targets, they exerted more effort in exploring the available information to create spear-phishing attacks. I found that attackers in the high information-availability condition made, on average, 167 more edits than attackers in the low information-availability condition when crafting the attacks. However, contrary to general opinion, I did not find a relationship between the *amount* of personalization and end-user susceptibility. As shown in Table 4.6, I did not find any effect of customizing spear-phishing attacks with more units of personal information about the intended targets.

Instead, as shown in Figure 4.6b, our analysis indicates that attackers in the high information-availability condition with access to more personal information about their targets were able to explore more ways to impersonate and create convincing narratives that

led end-users to trust and respond to the attacks (see Table 4.5 and Table 4.7). For example, among attacks using the bank account pretext, I observed that attacker participants in the low information-availability condition most often pretended to be from a commercial organization or automated software sending a notification about an issue with the target's account. Participants in the high information-availability condition reported impersonating a friend or a workplace acquaintance, strategies that had a significant impact on end-user susceptibility (see Table 4.7). Similarly, I found that end-users across the two conditions were more susceptible to attacks created using the job application pretext template (see Table 4.4), which involved impersonating a friend or an acquaintance. Figure 4.7 shows an example of an attack created in the high information-availability condition. It can be seen from the example how the attacker impersonated a workplace colleague and exploited the fact that the target (the fictional Joe Parks, as played by an end-user in the group) had been working at Enron for more than five years to create a trustworthy and influential narrative. Ultimately, I theorize that access to more personal information may allow adversaries to create attacks containing novel and contextually meaningful narratives that end-users are not typically trained to suspect and detect.

Attackers' ability to create more trustworthy and believable spear-phishing messages in the high information-availability condition can be explained using theory of mind [74]. Theory of mind (ToM) refers to humans' ability to reason the mental state (i.e., thoughts, beliefs, and feelings) of others using shared world knowledge and social cues to predict their behavioral responses. ToM is integral to humans' ability to maintain social connections [98]. Decades of research on ToM have shown that when presented with short descriptions of scenarios, individuals are able to quickly infer the mental states of characters and predict characters' behaviors based on these mental states [104]. ToM provides individuals the ability to distinguish between their own mental states and the states of other through a self-other process [13]. I theorize that this self-other distinction process is crucial to successful deception and lying. With access to more information about their targets, adversaries can more easily access relevant episodic and emotional memory to represent, simulate, and anticipate

the behaviors of their targets [131]. This activation of working and long-term memory is necessary to create believable, albeit false, narratives within spear-phishing messages because false narratives are largely constructed by altering truthful episodic memories experienced personally by the adversary [132, 116, 82, 118]. More research in this area is necessary to fully understand the cognitive processes underlying adversarial decisions and actions. Paradigms like SpearSim provide the necessary simulation environment for this research in the context of spear-phishing attacks.

This work further demonstrates the need to develop anti-phishing training programs that help people become more sensitive to spear-phishing attacks. The results of our experiment show that end-users were 2.57 times more vulnerable to spear-phishing messages than they were to mass phishing messages and individual levels of security awareness did not have an impact. This raises important questions about the efficacy of existing anti-phishing training procedures in helping people detect spear-phishing attacks. Conventional phishing training programs such as embedded training [69] involve the use of serious games [137] to teach phishing concepts in a simulated game environment. These programs have been effective in providing just-in-time awareness of mass phishing messages that impersonate generic emails from popular brands [69]. However, it is unlikely such generic training methods would be effective against spear phishing. People would not be able to generalize training on mass phishing campaigns to spear-phishing attacks. To detect a phishing attack, people must first become suspicious of it. Only then can people use the knowledge gained from anti-phishing training to verify a phishing message's source and URL address. However, our results suggest that with access to more information, attackers can create attacks that do not raise suspicion but instead inhibit people's ability to notice discrepancies in the messages so that they ultimately fall victim to the attacks. Anti-phishing training for spear phishing, therefore, must be tailored to an individual's context and vulnerabilities.

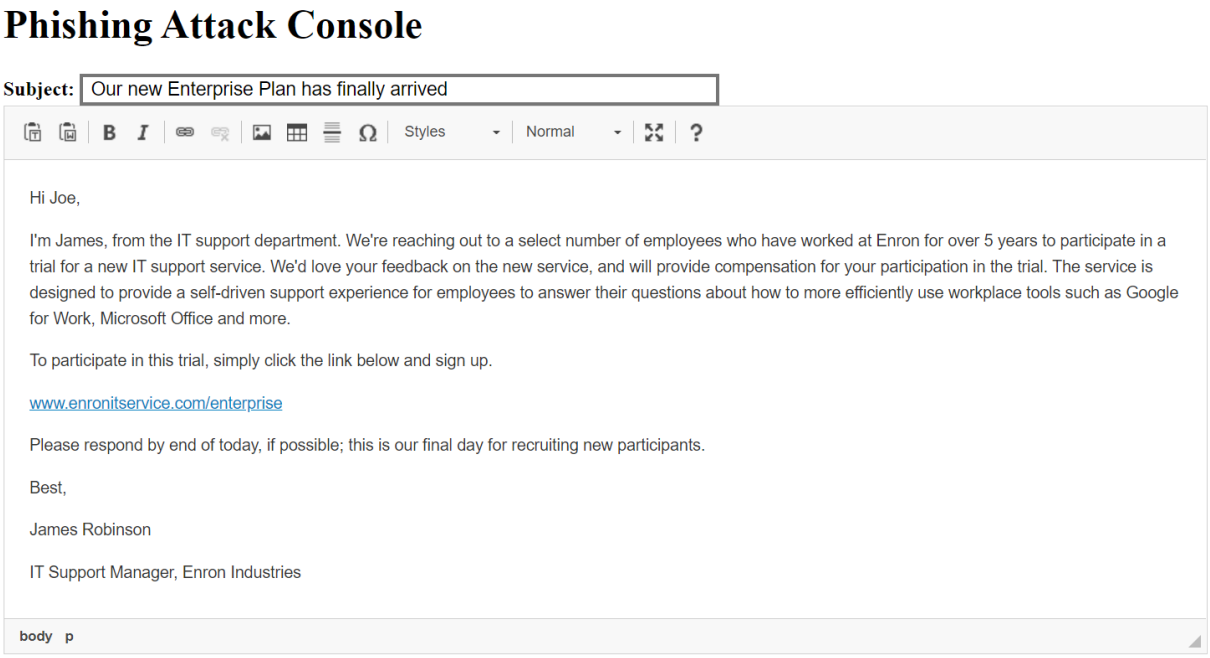


Figure 4.7: Sample spear-phishing email

4.4.1 Limitations and Future Work

Although I addressed some of the challenges I faced in running this multiplayer experiment in the pilot study, the current system is not without its limitations and challenges. Scheduling four-player groups and ensuring all players logged into the system at the expected time continued to be a significant challenge. For several reasons (e.g., internet connectivity and personal issues), I had several no-shows. Even among the groups that successfully formed, I had instances where participants dropped off during the experiment. To circumvent these challenges, I are testing cognitive models and deep learning models to replace end-user participants because end-users in the experiment primarily make classification decisions when they are given a stimulus [143].

In addition, since the focus of the study was on the deception side of spear phishing, the

screening process for participants was loose and the participants who participated as attackers were not cybersecurity experts or hackers. Therefore, the conclusions of this study cannot be generalized. I are planning to conduct a similar study with expert social engineers and white hackers to collect more realistic spear-phishing emails. Another limitation of the study is the limited data collected. The attackers created the spear-phishing emails, and there were three trials per pretext (i.e., each attacker created nine emails during the experiment). Therefore, I do not have sufficient samples to understand attackers' behavior in greater depth. For example, it is hard to decode the learning process of attackers and how they use personalization based on consecutive trials. Furthermore, our understanding of the exploration and exploitation behavior in attacker strategy is limited because of the limited number of trials. There is a dilemma of explore-exploit when choosing a strategy to generate a spear-phishing email. Attackers make decisions on whether to continually explore the previous strategy or try out another strategy. In complex environments such as spear phishing, understanding attackers' decision making related to explore-exploit is crucial for ensuring end-users' security. Psychologists have concluded that the trade-off between exploration-exploitation is driven by information seeking and behavioral variability [139]. When there is more information to process and greater uncertainty, people generally take more exploration actions [40]. Since our design is limited to three trials per pretext, I cannot answer these questions based on the current study and need to conduct an experiment with a longer trial.

Chapter 5

MODELING PHISHING DECISION MAKING USING INSTANCE-BASED LEARNING AND NATURAL LANGUAGE PROCESSING

Although phishing attacks are rampant on the internet, the likelihood of an individual encountering a phishing attack on a given day is small. Yet, people are expected to detect such a rare attack when they do experience one. Distinguishing phishing emails from legitimate emails remains a difficult task for a majority of people because phishing attacks are essentially *deceptive* messages that: a) are rare and constantly evolving; b) use impersonation to resemble truthful messages; c) applies emotional arguments to influence recipients; and d) could be tailored to exploit life context and recent world events [45, 113]. This section is published in [144].

5.1 Objective

Despite the large body of research on phishing attacks, there is a lack of models that explain the key cognitive processes governing end-user response to phishing attacks. Existing research on phishing has predominantly focused on: developing solutions to automatically detect phishing emails [2]; testing whether people pay attention to essential cues in a phishing email or website [29]; developing interventions to aid human attention [141]; and developing training programs to educate people about the concepts and strategies related to phishing (e.g., [71, 70]). Central to many of the past research is the aspect of human attention, or the lack thereof [106, 29]. An individual's lack of attention towards key indicators, such as the URL (Universal Resource Locator) of a website, and sender address in a phishing email, is widely considered why end-users fall prey to phishing attacks [29]. However, human atten-

tion is intimately linked to the contents of human memory [23, 112], and there is a severe lack of models that explain the role of such cognitive processes (e.g., memory activation dynamics) on end-user susceptibility to phishing emails. Our hypothesis is that people make decisions on phishing messages based on past experiences by activating pertinent memories of decisions made in response to similar emails in the past.

To test this hypothesis, I developed a cognitive model, named as SpearCog, based on Instance-Based Learning Theory (IBLT) of experiential-based decisions [42, 87]. The IBL cognitive model was developed in Python (PyIBL [83]). In the current research, I developed a model to understand how people may process phishing messages in memory and to determine the influence of past experience on end-user decision making. The objective here is to understand the cognitive processes driving end-user response to phishing attacks. Such cognitive models could be potentially used within email applications (e.g., Microsoft Outlook) to predict how people may respond to novel phishing samples which could inform embedded phishing training and phishing risk assessments. These models, however, may not be useful for discriminating phishing from legitimate emails. Secondly, what is the variance of IBL modeling performing on predicting the human decision making differentiated between heterogeneity groups? and what would be the factors affecting the IBL performance?

Survey	**Summary**
Request for action (Task assigned, click on a link, download attachment, etc)	Action
Request for information or opinion (send a reply message, contact info,send file, image, etc)	Information
Contains status update for an ongoing project or task	Project
Request for a meeting or other communication with you	Meeting
Contains reminder for a meeting, event, or upcoming deadline	Deadline
Spam or marketing or suspicious	Spam
Other	Other

Table 5.1: Survey questions presented to end-user during each trial

The end-users in the experiment were simply making decisions on whether they would

respond or not to any given message presented to them. They were rewarded based on their performance in the task. For each email, the end-user participants also responded to a small survey (see Table 5.1) about the email content. The questions in the survey were initially developed and improved in a previous study [100].

The participant playing the role of an attacker in the group was given specific goals in the experiment using the available information to them about the target end-user. The attacker objective was to steal bank credentials, work account credentials, and lure the target to download attachments. These were fake objectives, and no real harm were caused to any participants. If the attacker was able to deceive the end-user successfully in an email, i.e., end-user responded to that email then the attacker earned rewards. The overall goal for the adversary was to maximize their individual rewards in the experiment.

5.1.1 Dataset

To model phishing decisions, I leveraged the dataset generated from the experiment described in the previous section 4.2.2. This experiment was designed to understand the various factors involved in the spear-phishing attack. The dataset from this experiment contained responses from 84 participants. The dataset consists of a total of 529 unique emails with 6712 responses. Each participant responded to approximately 80 emails, including benign emails, phishing emails, and spear phishing emails. For full information about the study, please refer to this previous publication [143].

5.2 Methods

In a related work, Cranford et al. developed an instance-based learning model to predict end-user response to phishing emails [23]. LSA (Latent Semantic Analysis) and Wordnet were used to compute the semantic similarity between incoming emails and agent memories. However, using LSA to determine semantic similarities between two instances was raised as an important limitation of this work. They theorized that LSA was not effective at representing how people process email texts. Therefore, in this work, I investigate the effect

of using deep learning and attention-inspired natural language processing methods such as BERT, which has demonstrated impressive performance in other NLP applications.

I tested three different natural language processing methods (LSA [37], GloVe [96], and BERT [28]), often used in natural language understanding tasks, to represent and calculate the similarity between two email instances within the IBL model. I compared the performance of these three methods in predicting participants' responses to emails (ham, phishing, and spear-phishing) in the laboratory study described earlier. In the following section, I briefly describe each method and how these three methods were used to measure similarity between instances in the IBL model.

5.2.1 Partial Matching using NLP

The NLP methods described in the previous section were used in the IBL model to determine similarities between two email instances. Specifically, the NLP methods were used as sub-models for the similarity term $\sum_k Sim(v_k, c_k)$ that represent the partial matching process in the IBL model. The core idea is to use each of the three NLP methods to represent emails as fixed length numerical vectors. LSA and GloVe were used to produce vectors representing word-level embedding, whereas BERT was used to produce feature vectors representing sentence-level embedding. The similarity term in IBL is essentially defined as a cosine distance function measuring the distance between the feature vector representations of the current email instance (c_k) and email instances stored in memory (v_k).

LSA In the LSA method, all 529 email messages obtained from the study were mapped into a matrix with 529 columns and 6213 rows. The number of rows represents 6213 unique words present in the corpus. A truncated singular value decomposition method was applied to reduce the number of dimensions to 300. The number of dimensions was chosen to be 300 to match the vector sizes derived from GloVe and BERT. The final LSA matrix was of size $529 * 300$. Each row in the matrix represents the feature vectors for each message derived using the LSA approach. Similarity is calculated using a cosine distance between two email vectors from the LSA matrix, using the equation below.

$$similarity = \frac{v_k * c_k}{\|v_k\|\|c_k\|}$$

GloVe Using transfer learning, I fine-tuned and applied the GloVe matrix to represent each of the unique words across the GloVe's pre-trained 300-dimension word vector. This produced a matrix of size 6213*300, where 6213 is the number of unique words in the corpus. I created another matrix ($529 * 6213$) that contained the number of times each of the 6213 unique words occurred in each of the 529 messages. The two matrices were then multiplied to produce a final matrix (529*300) that contained the 300-dimensions word vector representation for each message derived using the GloVe method. Similarities between two emails were calculated using a cosine distance function between the corresponding two feature vectors in the GloVe matrix.

Transformers Transformers has been used to achieve high performance in many natural language processing tasks. I used nli-distilroberta-base-v2 model, a distill RoBERT model, fine-tuned on SNLI(Standard Natural Language Inference) corpus [12] with 84.38 performance on the STS benchmark dataset. The SNLI is a large collection of human written English sentence pairs labeled for semantic similarity. This Sentence-BERT(SBERT) [101] was applied to our dataset to derive the semantic similarity between every pair of emails in the dataset. Figure 5.1 shows the architecture of the SBERT. For each email pair (c and v), the SBERT was applied using two identical BERT/RoBERTa models to process each email separately. Use the pooling layer, I derive a fixed sized length embedding vector (a and b) representing the respective emails. Like other methods, the cosine similarity function was used to calculate the similarity between the two vector representations of emails.

User Perception The three NLP approaches (LSA, GloVe, BERT) described earlier enable us to capture semantic similarities between two email instances based on underlying statistical properties. However, these methods may not be representative of how humans actually process text, recall text from memory, and make decisions in the email management context. During the experiment, I had asked participants to self-report their opinion about each message presented to them along seven dimensions that described what was in the

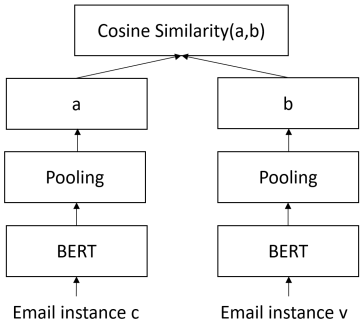


Figure 5.1: Sentence BERT architecture used to represent emails in IBL

message. For example, whether the message requests an action or whether the message contains a reminder for a meeting. Therefore, as an alternative approach, I represented each email along these 7 self-reported dimensions. The seven dimensions are listed in table 5.1. Emails with similar contents will have similar vector representations. These vectors represent how each individual participant perceived the contents of the message presented to them. Like with other methods, the similarities between two emails were calculated using a cosine distance function between the corresponding two self-reported feature vectors.

Perception Bert Practically, it is not possible to train cognitive agents based on user-reported email attributes. Therefore, I re-trained the SBERT model to learn similarities between emails based on user perceptions. For each pair of emails, I fine-tuned the SBERT model to predict the similarity score measured using the user-reported attributes. The similarity score was used as classification and it ranged from 0 to 1. 0 indicates emails that are least similar according to users, and 1 indicates identical emails.

There were 529 unique messages in the dataset. Therefore, there were 529*528/2= 139,656 similarity pairs. I randomly sampled 10000 pairs of similarities to fine-tune the downstream task in BERT. 10000 pairs of similarities were split and trained using the ratio 7:1:2. I used a package from [101] to manage the computation involving BERT, and all computation was conducted on a desktop with RTX 2060 graphics processor unit.

Heuristic features in phishing Building on the previous chapter, I consider incorpo-

rating heuristic features related to phishing into the representation of emails. These heuristic features include context-related information and psycho-linguistic characteristics.

In the phishing domain, context plays a crucial role, as attackers often exploit end-users' personal information to create persuasive stories that lure participants into falling victim to such attacks. Therefore, I include two heuristic factors related to personal information: (1) whether the recipient's name is included in the email and (2) whether the bank account of the recipient is included in the email.

In section**??**, I showed that the psycho-linguistic features significantly affect the decision-making process of end-user participants in the phishing context [142]. Language inquiry word count (LIWC) has been commonly used to analyze natural language in text-based communication. LIWC calculates the distribution of word attributes (e.g., words describing negative emotions, social process, certainty) in phrases of text. LIWC has been used in various types of research, particularly deception research and phishing research. Other work has used feature selection methods to discover that certain types of words are more likely to persuade end-users to fall victim to phishing emails.

Unlike transformer-based models producing attention-based features, LIWC provides insights into the psycho-linguistic features, some of which imply persuasive power in phishing contexts. Therefore, I incorporate LIWC as an amended feature set to represent the email instances.

To incorporate all of the heuristic features, I create a new feature set to represent the email instances. The language representation generated by the Bert encoder is combined with the other heuristic features. Figure 5.2 presents how I combined multiple data sources. The last layer is used as the last feed-forward layer of the deep learning model to fit the semantic similarity model.

5.2.2 Simulation Procedure

Dataset from a previous laboratory study was used to train the IBL agent and evaluate its performance in predicting human responses to phishing emails. The dataset includes

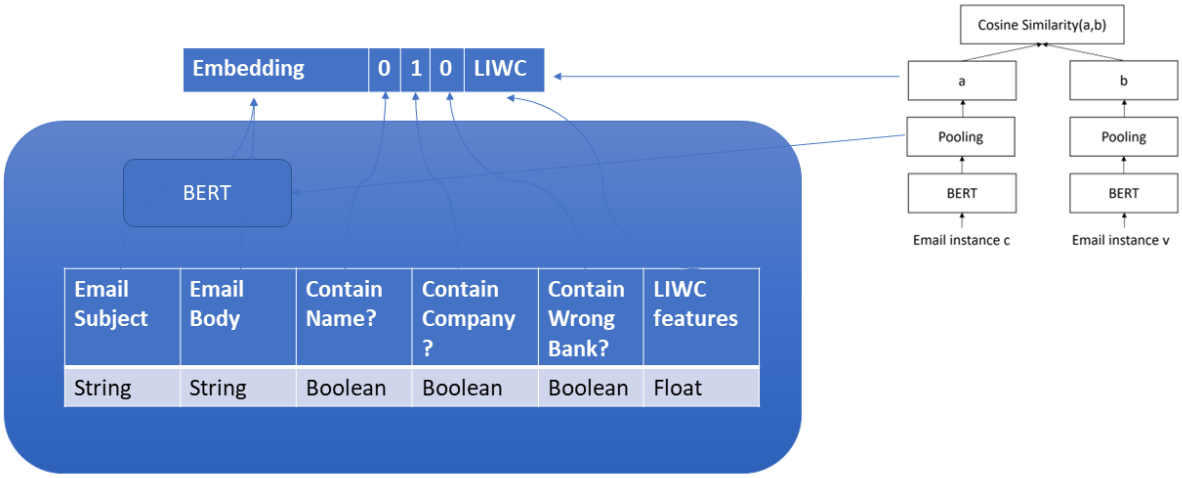


Figure 5.2: Representation structure for Perception Bert 2

529 unique email messages and contains email responses from 84 participants. Among the 529 emails, 210 were ham emails, 15 were mass phishing emails, 52 were promotion emails, and 252 were spear-phishing emails. To simplify the response from end-users, I encode the response (a) Respond Immediately Or (b) Flag the email for follow-up to ***response*** and (c) Leave the email in the inbox, (d) Delete the email, Or (e) Delete the email and block the sender to ***ignore***. I trained IBL agents to model and predict responses from each of the 84 human participants in the laboratory study. Each IBL agent represented a single human participant and was presented with the same emails experienced by its human counterpart. On average, participants in the experiment made decisions on 80 emails which included ham emails, mass phishing emails, and spear phishing emails. Similarly, the IBL agents representing each participant made decisions on the same number of emails experienced by its human counterpart. For each email presented, the model takes as input the context of the

email and generates an action (respond or not respond) by retrieving similar past instances. Typically, instances are encoded as chunks in an agent's declarative memory that represent the features of the decision: the context in which a decision is made, the action taken, and the outcome of that decision. In this work, the context was represented as a feature vector derived for each email message using the five similarity methods described in the previous section. This feature vector was provided as the input to the model. For each email feature vector presented, the model made a decision whether to respond or not. Each decision received outcome feedback: 1 point for correct response and -1 point for incorrect response. Except for the mismatch penalty parameter (see Equation 3), all other model parameters were set to their default values. For example, the decay parameter was set to the default value of 0.5. I experimented with three MP parameter values (1, 5, 30).

I also experimented with two split-ratios of training and testing: 50-50 and 80-20. With 50-50, the IBL agent was trained on 50% of randomly selected emails that its human counterpart experienced, for example, if a human participant made decisions on 100 emails during the study, the IBL agent representing the participant was trained on 50 randomly selected emails that the participant experienced. These emails served as the training instances for the agent and were encoded as instances in the declarative memory of the agent. The agent's decision performance was evaluated on the remaining 50% of the emails unseen by the agent during training. In the 80-20 split, the IBL agents were trained on 80% of randomly chosen emails the participant experienced and tested on the remaining 20% of emails. Since IBL is a stochastic model, the model was trained and evaluated 800 times to generate stable predictions of human behavior. Using the half-width approach [72], I estimated 800 model replications were necessary to achieve 95% confidence with our output estimates.

5.3 Results

I analyzed the performance of IBL agents in predicting human response to emails. I used model accuracy, hit rate, and correct rejection rate to measure model performance during the test phase. For each IBL agent representing a human participant, *accuracy* measured

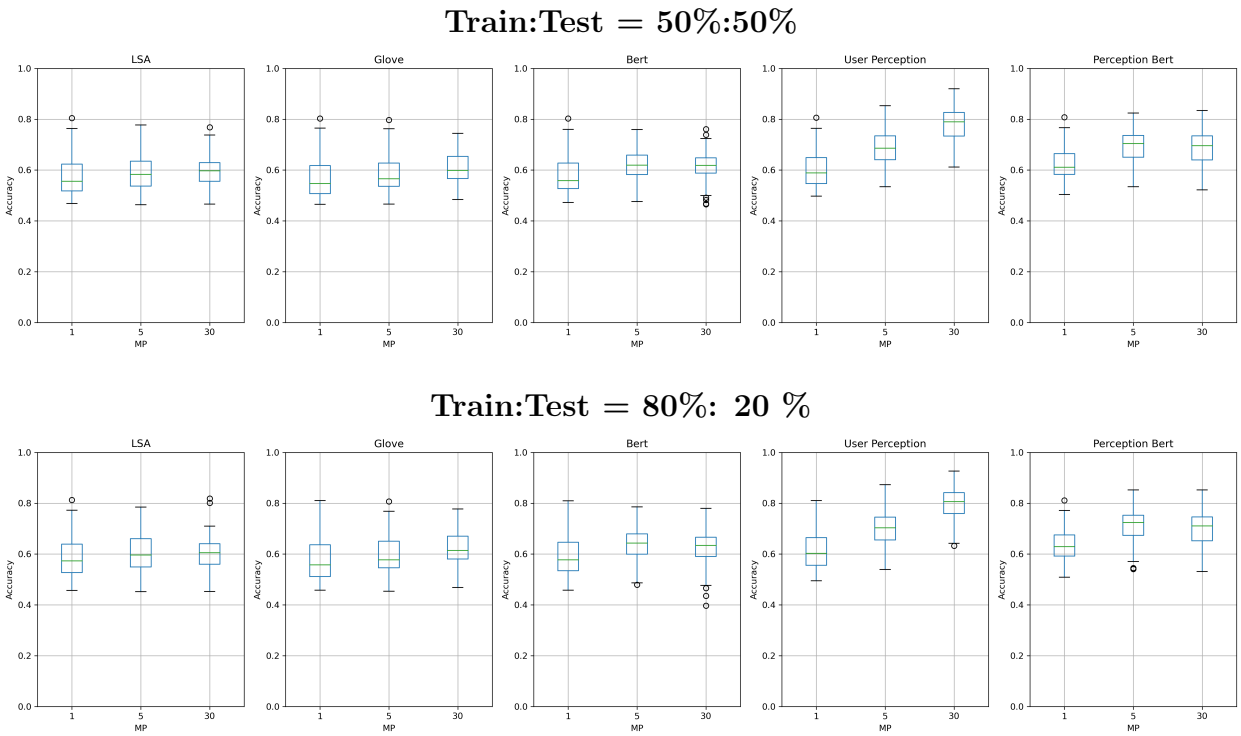


Figure 5.3: Accuracy for IBL agent across similarity approaches, MP values and split ratios

the percentage of emails for which the model decisions concurred with the human decisions. For example, a 70% accuracy would indicate that IBL agent accurately predicted the human response on seventy percent of emails presented during the test phase. In addition to the accuracy, the hit rate and correct rejection rates were also calculated. As shown in Table 5.2, for each IBL agent, the hit rate measured the proportion of emails for which the IBL agent accurately chose to respond to it, whereas correct rejection measured the proportion of emails for which the IBL agent accurately chose to ignore (not respond) the email.

Model accuracy, hit rate, and correct rejection rates were calculated by averaging the performance of 84 agents (representing 84 human participants) across 800 model runs. Figure 5.3 compares the accuracy of IBL agents in predicting human response across the five similarity methods (LSA, GloVe, BERT, user perception, perception bert), three mismatch penalty parameter values (1, 5, 30) and the two split-ratios used for training and testing the

		Human	
IBL		*Response*	*Ignore*
	Response	Hit	False Alarm
	Ignore	Miss	Correct Rejection

Table 5.2: Hit rate and Correct Rejection Rate

agents (50-50 vs 80-20). Figure 5.4 shows the distribution of hit rate and correct rejection rate. Each point in the graph represents the average hit rate and correct rejection rate of one IBL agent predicting the response of its human counterpart. The rates are color-coded to indicate the five different kinds of similarity approaches compared. The average performance with three mismatch penalty parameter values and two split ratios were also compared and presented in the Figure 5.4.

Using mixed-effects ANOVA, I tested the effect of different similarity approaches, mismatch penalty value, and split-ratio on all three performance measures. Tables 5.3, 5.4, and 5.5 presents the results from the ANOVA analysis for the three performance measures: accuracy, hit rate, and correct rejection rate, respectively. Similarity approach and mismatch penalty value were found to have a significant effect on all three measures, whereas the split-ratio had a significant effect on accuracy and correct rejection rates. I will discuss each of these results in more detail next.

	Df	**F value**	**Pr(>F)**
MP	2	184.78	0.0000
Approach	4	228.73	0.0000
Split Ratio	1	20.45	0.0000
Approach:Split Ratio	4	0.06	0.9934
MP:Split Ratio	2	0.23	0.7938
MP:Approach	8	39.66	0.0000

Table 5.3: Anova table for Accuracy

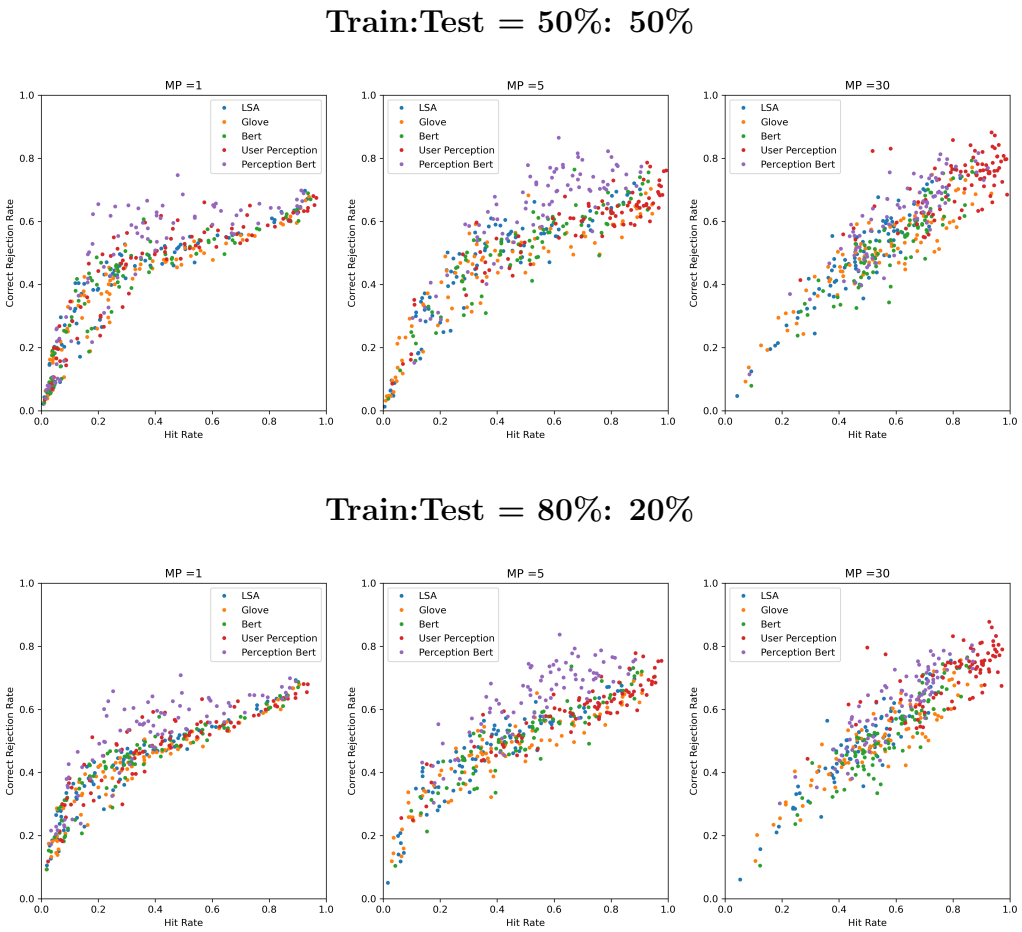


Figure 5.4: Correct Rejection Rate vs Hit Rate for IBL agent across similarity approaches, MP values and split ratios

5.3.1 Similarity Approach

Between the five similarity approaches I tested, I found that the approach that used participants' self-reports to represent emails within IBL model performed better than all other approaches (79.7% average accuracy). Following it closely, the second best performing model was the approach that used the perception Bert model, fine-tuned using participants self-reports. Although there is a significant difference in accuracy between the two approaches, I found that the perception Bert model performed, on average, only 2.23% ($p < 0.001$) lower

	Df	F value	Pr(>F)
MP	2	194.29	<0.001*
Approach	4	122.72	<0.001*
Split Ratio	1	4.48	0.0344*
Approach:Split Ratio	4	0.15	0.9627
MP:Split Ratio	2	5.29	0.0051*
MP:Approach	8	11.55	<0.001*

Table 5.4: Anova table for Correct Rejection Rate

than the model that directly used participants' self-reports to represent the emails. LSA, GloVe and canonical BERT, on average, achieved less than 60% accuracy in predicting human participant response. There was no significant difference in model accuracy between using LSA, GloVe, and BERTs. From Figure 5.4, it can be observed that the approach using participants' self-report to represent emails clustered at the top right corner indicating higher hit rates and higher correct rejection rates. The pattern is the same across all values of MP and split-ratio.

To further analyze the differences in performance, I compared the similarity scores (calculated using cosine distance function) between all pairs of emails derived using the GloVe approach against the similarity scores derived using participants' self-report. I compared these two approaches because the GloVe approach represents similarities calculated based on the statistical properties of text. In contrast, the user perception approach represents similarities calculated based on individuals' opinions about the email attributes. As shown in Figure 5.5, the GloVe approach considered the majority of the emails as similar to each other, whereas the user perception approach considered the majority of the emails to be dissimilar to each other. I calculated the correlation between the two matrices containing the similarity scores derived from the two approaches. The correlation between the two approaches was low, $r_{person} = 0.10, p < 0.001$. This indicates that for the same pair of emails, the similarity scores obtained using the two methods were different from each other. It is noteworthy that both approaches used the same cosine distance function to measure the similarity for any

given vector pair. This shows that pure semantic methods such as LSA, GloVe and canonical BERT were ineffective in capturing the deep contextual differences present in emails, which may have been considered as a significant difference from a human perspective.

	Df	**F value**	**Pr(>F)**
MP	2	235.46	<0.001*
Approach	4	84.17	<0.001*
Split Ratio	1	0.00	0.9591
Approach:Split Ratio	4	0.12	0.9760
MP:Split Ratio	2	0.95	0.3861
MP:Approach	8	7.17	<0.001*

Table 5.5: Anova table for Hit Rate

5.3.2 Mismatch Penalty

A large mismatch penalty(MP) value makes the IBL agent care more about the difference between the current instance and previous instances. Large MP values penalize agents more for incorrect mismatches between the incoming instance and the instance in memory. Overall, I found that the mismatch penalty has a significant effect on all three measures. A post-hoc analysis revealed that irrespective of the split ratio or similarity approach, there is a significant difference in model accuracy between MP = 1 and MP = 5 ($p < 0.001$). However, the difference is not significant for higher values of MP (MP = 30). The post-hoc analysis also showed that the increment in MP has a significant effect on hit rate and the correct rejection rates ($p < 0.001$). Hit rate increases by 13.85% ($p < 0.001$) from MP = 1 to MP = 5, and increases by 23.7% ($p < 0.001$) from MP = 1 to MP = 30. Similarly, the correct rejection rate increases by 9.22% ($p < 0.001$) for an increase in MP from 1 to 5 and a 13.4%($p<0.001$) improvement from MP = 1 to Mp = 30. For higher MP values, the improvement in hit rate is much more significant than the improvement in the correct rejection rate. As shown in Figure 5.4, I can observe a concave relationship between the hit rate and correct rejection rate for MP = 1, whereas this relationship becomes much more linear for MP = 30. There is

also a strong interaction effect between the similarity approach and MP values. As shown in Figure 5.3, the models using user perception and perception BERT approaches demonstrate an increase in accuracy for higher values of MP, but MP does not appear to impact the accuracy of models using LSA, GloVe, and canonical BERT models.

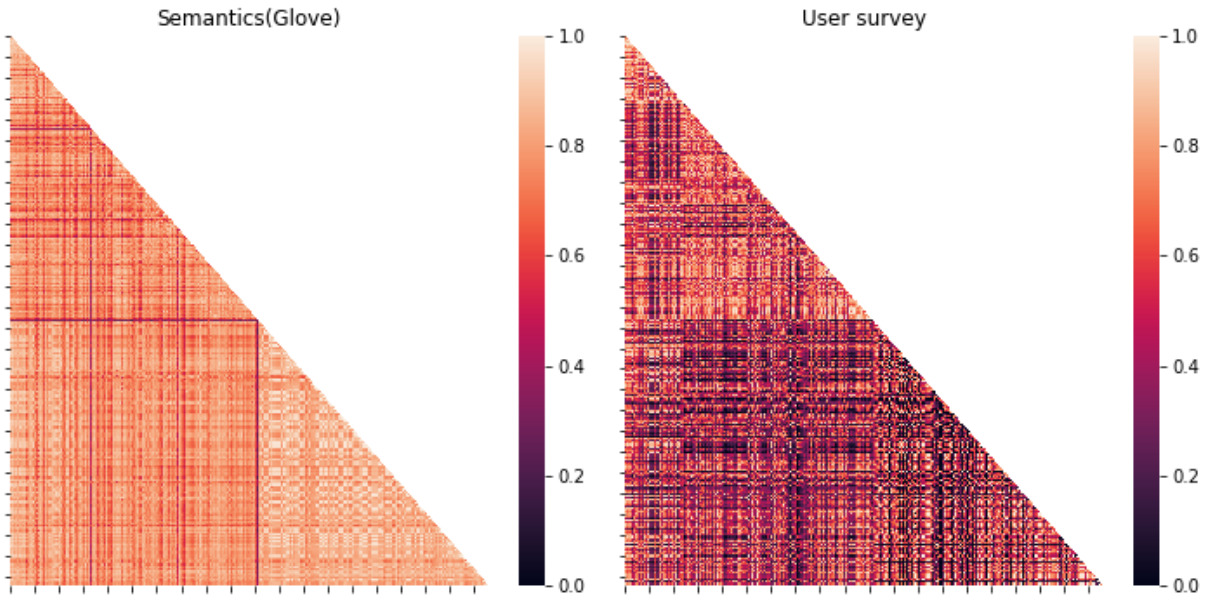


Figure 5.5: Visualizing similarity between every pair of emails. (Left) Similarity measured using GloVe. (Right) Similarity measured based on user perception

5.3.3 Split Ratio

Finally, I tested whether the IBL model performance depended on the amount of data used for training the agents. I tested two commonly used split ratios: 50-50 and 80-20. Although there was a statistically significant effect of the split ratio on the accuracy and correct rejection rate, as shown in Tables 5.3 and 5.4, the improvement was only marginal. On average, model accuracy increased by 1.4%($p<0.001$) from 50:50 to 80:20 split. The correct rejection rate increased by 1.2%($p = 0.037$). The split ratio did not have a significant

effect on the hit rate. There was also no interaction effect between split ratio and similarity approach or between split ratio and MP.

5.3.4 Heuristic information embedded in IBL

What is the effect of embedding additional heuristic information into the IBL model? The previous sections revealed that the performance of the IBL agent was not improved with the use of more advanced language models for semantic similarity prediction. Note that transformer-based models outperform the GloVe and LSA methods, but there is no improvement in predicting human responses to emails. This pattern suggests that the features captured by generic semantic similarity may not adequately reflect the perception humans have when they make email management decisions. Compared with the general semantic similarity, user perception obtained from survey responses provides insights into phishing decision-making. Therefore, I fine-tuned the NLP model to learn the similarity between emails based on user responses to the survey. In addition to using the perception Bert model, I included more heuristic features to represent the text, including LIWC features and personalized persona features. These features allow us to fine-tune the transformer model. Comparing the perception BERT model's results with those of the perception BERT II model shows that as more information was collected, the performance of predictive decision-making improved to 74.1%. The key factor in this improvement is the incorporation of more heuristic information related to contextual details and psycho-linguistic effects. By incorporating additional features that are relevant to human cognition and decision-making, the underlying IBL agents were more effective at predicting human responses to emails.

5.3.5 Heterogeneity groups of IBL models

I observed significant variance in IBL agents' performance in predicting end-user response to emails. What factors could explain this variance?

To answer this research question, the IBL agents were grouped based on performance percentile, yielding three groups: upper-performing (>75%), mid-performing (25% $\sim$ 75%),

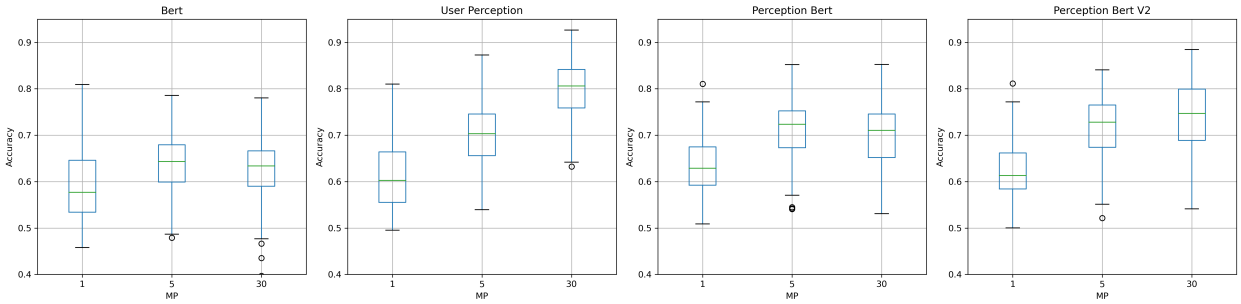


Figure 5.6: Accuracy for IBL agent across similarity approaches

and lower-performing groups (25%<). Logistic regression was then applied to each group to determine the coefficients and identify the factors that significantly influenced the decision-making processes.

One key assumption in logistic regression is the independence of the predictor variables. This assumption is made to avoid collinearity issues. To check for collinearity, I calculated the variance inflation factor (VIF). The VIF indicates the strength of the correlation between independent variables in regression analysis. A VIF value of 1 indicates there is no correlation between the independent factors. The VIF for our dataset was between 1-2, indicating that there are moderate correlations between a given predictor variable and other variables, but the correlation is often negligible for logistic regression.

To explain what factors drive decision-making in the email management context, we asked participants to respond to survey questions regarding the email content. Participants were allowed to select multiple responses for each email classification. The detailed choices are shown in Table 5.1. These choices were used as independent variables in the logistic regression, and their coefficients were calculated separately for each group. For the dependent variable, we used the decision-making of end-users, where the decision was to either respond or ignore. The same logistic regression formulation was applied separately to all the groups. Choices in Table 5.1 were found to significantly drive human decision-making, and coefficients are shown in Figure 5.7. Factors whose coefficients are positive made participants more

willing to reply to the email and vice versa. Participants were more inclined to reply to emails that they perceived as important to their profile. Emails that appeared to be spam or irrelevant were more likely to be ignored. Additionally, when an email specifically requested information, end-users were more likely to reply to the email.

The comparison of coefficients between different groups was conducted using the bootstrap resampling technique. By running the same logistic regression model multiple times with 90% of data points sampled from each group of email responses, I estimated the intervals for the coefficients and tested their significance. The coefficients and the interval is shown in figure 5.7. Generally, I observe a consistent trend in all groups on how factors affecting on decision making. If the coefficient is positive, then it means that the factor has more effect on the end-user responding to the email, and vest versa. I found that the upper group usually has higher absolute coefficient values which means they are more likely to choose decisions based on the corresponding factors. It could be posited that participants with higher coefficient values were more likely to make intuitive decision making than the participants who have less coefficient value.

5.3.6 Factors affecting the IBL agent performance

Although we observed differences between the three participant groups (grouped based on IBL prediction performance) in terms of the size of coefficient values, it is still unclear what factors explain IBL models' ability to better predict responses for some of the participants than others. To explore this further, I calculated the *entropy of email classifications* and *average email processing time* at the participant level to test if factors related to response time and decision complexity can explain the variance observed in IBL agent performance.

Entropy of email classifications: Information entropy is widely used to represent the uncertainty or randomness in a set of information. Entropy is generally calculated based on the probability of outcomes. In our case, if an email classification only contains one element, it could be inferred that participants experienced less complexity and uncertainty during decision-making and, therefore, lower entropy. On the other hand, if an email classification

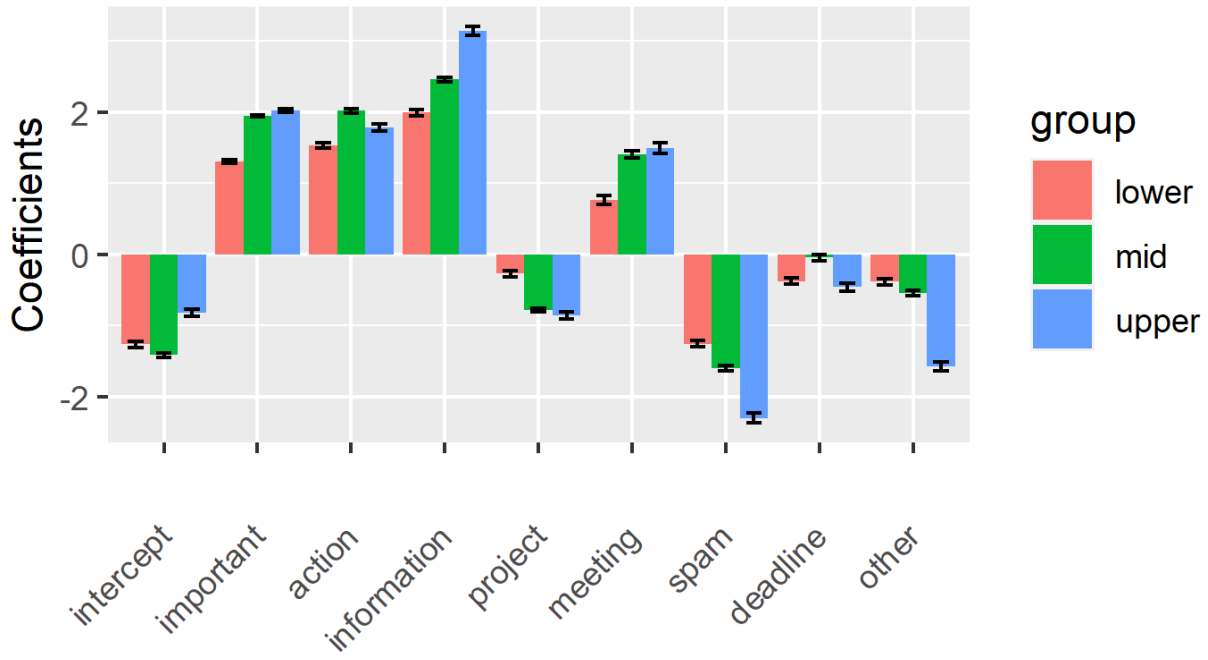


Figure 5.7: Coefficients across groups

contains multiple choices, it could be inferred that participants could have experienced more difficulty/uncertainty. Based on the quantile, I calculated the average entropy for each group. I found that the upper group has an average entropy of 0.2709, the middle group has an average entropy of 0.3426, and the lower group has an average entropy of 0.4092.

Email processing time: For each email classification, I have a record of how long the participants took to make a decision. The average time spent on email classification is 40.4 seconds (std = 34.9). The time a user spends on an email could have been affected by the length of the email, the context of the email, and the user's underlying cognitive state. As the emails have the same probability of distribution, most effects from the first two factors are negligible. Therefore, I calculated the average time spent on each email at the user level and evaluated if it able to explain the variance observed in IBL performance.

There is a moderate correlation ($r_{pearson}0.423$, $p < 0.001$) between the entropy of email classifications and the average email processing time at the participant level. When the

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	0.0525	0.179	0.294	0.770	-0.303	0.408
C(Enduser)[T.User2]	0.0263	0.252	0.104	0.917	-0.475	0.528
C(Enduser)[T.User3]	-0.1837	0.254	-0.724	0.471	-0.689	0.322
total_email_normalized	0.0215	0.148	0.145	0.885	-0.274	0.317
avg_process_time_normalized	-0.3843	0.147	-2.607	0.011	-0.678	-0.091

Table 5.6: Logistic mixed effect model on End-user level analysis

participants took less time processing emails, they were likely to experience less difficulty in classifying the emails (lower entropy). Using ANOVA I analyzed IBL performance in terms of profile assumed during the experiment, number of emails processed, and average time spent for each participant. The average time spent processing emails or the entropy of email classifications was identified as the only significant factor affecting the IBL agent performance with ($F(1, 78)=5.907$, $p = 0.017$). The persona or profile was used as a blocking factors. Table 5.6 presents the coefficients of the model.

Although in our earlier analyses we found that the split ratio significantly affected the performance which suggested that number of emails classified could be a significat predictor of IBL performance. However, I found that the total number of emails processed is not a significant factor when analyzed alongside average processing time. Figure 5.8 presents the relationship between the average time spent on email processing with IBL agent performance. The legend is the personas for end-users. It is evident that there is no effect of the profile that participants assumed during the experiment. Instead, I found a significant effect of average time spent on emails - IBL agents performed better in predicting the responses of participants who made quick and intuitive decisions.

The same pattern was also found in the email-level analysis. I trained 100 IBL agents to simulate each participant's decision-making on 80% of emails. The dependent variable was the probability of correct responses, with time as the fixed effect and email and participants as the random effects. Given that the residual did not follow a normal distribution, I established a threshold of 0.5 to categorize the numerical values into "high performance" and "low

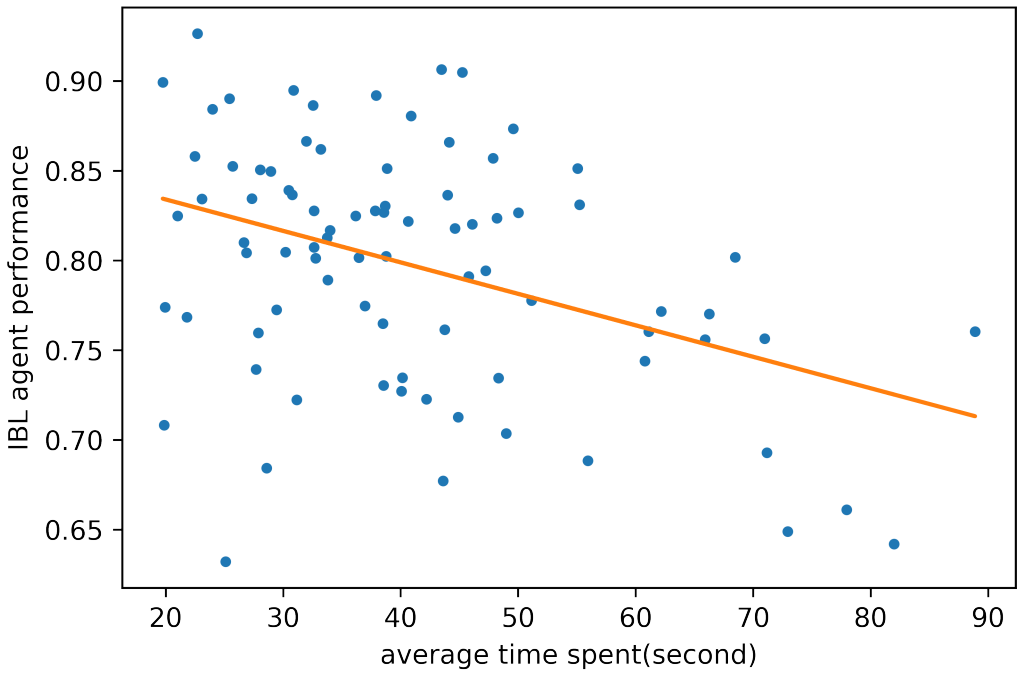


Figure 5.8: Average time spent on emails vs. IBL agent performance

performance" groups.

A mixed-effects logistic model was applied to this data set to determine the association between time spent on email classification and IBL agent performance. The results indicated a email processing time had a significant impact on IBL classification performance. The coefficient was -0.10 as shown in Table 5.7, which means that more time spent reading emails decrease the odds of IBL correctly classifying the email. In essence, the findings demonstrate that emails processed in less time have a higher probability of being correctly predicted by the IBL models.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.64	0.06	10.65	0.00
scale(time)	-0.10	0.03	-3.27	<0.001***

Table 5.7: Logistic mixed effect model on email-level

5.4 Discussion

Our work shows that the approach used to represent email text within IBL models can significantly affect agent performance in predicting human response. Furthermore, I found that using representations that could consider how humans perceive email messages can have a substantial impact on model performance. I predicted participants' responses to email messages in a laboratory study with 79.7% accuracy by training IBL agents with emails represented using attributes self-reported by participants in the study. In contrast, I achieved 71.7% accuracy by fine-tuning a BERT model to learn similarities between emails based on user-reported attributes. IBL agents using traditional NLP methods (LSA and GloVe) which are effective at capturing semantic similarities, performed better than chance at predicting human response, similar to the results obtained by Cranford and colleagues [23]. Overall, our results show that IBL models are able to adequately predict human response to phishing emails, providing evidence to our hypothesis that people make decisions on phishing messages based on past experiences by activating pertinent memories of decisions made in response to similar emails in the past.

Although methods like GloVe and BERT effectively highlight semantically important features in a piece of text, they may not be effective at highlighting features relevant to people in the email management context. Therefore, as shown in Figure 5.5, traditional NLP methods like GloVe may end up representing two emails as semantically similar to each other, whereas a human may think otherwise. For example, two emails may contain words with similar co-occurrence frequencies and may include similar latent topics. Yet, they may appear different to a human because they may be communicating two divergent messages.

Therefore, more work is necessary to understand how people encode email messages, what salient features of emails are encoded in the memory, and how the features encoded may vary by the type of email (ham vs. mass phishing vs. spear phishing). Understanding these issues could provide us important insights into how humans learn and make decisions on malicious deceptive signals such as phishing emails.

One important implication of this work is that IBL models, or cognitive models in general, could be potentially used in the future to develop personalized phishing probing and training solutions. For example, IBL models could be deployed in email management software such as Microsoft Outlook to actively learn phishing instances that an individual may be susceptible to and may benefit from receiving additional embedded phishing experiences. Furthermore, instead of probing every employee in an organization with generic phishing templates to learn who may be vulnerable to phishing, such human-centered, deep learning-based, and cognitive architecture inspired methods could be effective at quickly detecting human vulnerabilities within an organization.

5.4.1 Could IBL be a framework for intuitive decision making?

The relationship between IBL agent performance and average time spent on emails suggests that functions described in IBL framework could be representative of intuitive decision-making. There is rich research on System I and System II thinking [107, 60]. In most situations, people are likely to make decisions alike system 1 - making near-instantaneous decisions that happen automatically and intuitively. Specifically, System I decision-making is known for being fast, emotional, automatic, and independent of working memory. Quick responses and judgments are formed based on immediate associations and emotional cues [19, 34]. On the other hand, System II decisions are characterized as propositional, rational, and effortful, and involve decisions that are made more deliberately and slowly. System II decisions involve more logical and conscious thinking associated with working memory [19].

The modeling described in section 5.3.5 shows that time spent on email classification are good predictors of IBL performance at both the participant and email levels. The less

time spent on classification, the more accurate was IBL in predicting the human response. Our data indicate that instance-based learning is better at predicting decision-making that happens quickly and without much deliberation. There has been research arguing that working memory and short-term memory overlap or even may be the same thing [21]. The average decision making time of participants in email classifications are generally between 20 seconds to 90 seconds. Although there is no clear cut-off line on splitting the system I and system II thinking by seconds, I do found that the IBL model is better at predicting the participants' decision making and email classifications which take less time.

I infer that IBL models are good at predicting classifications with more System I thinking. These classifications are prone to be less cautious and more easily make errors. In the phishing domain, people who are less cautious and process emails carelessly are more likely to fall victim to phishing attacks [9]. In cyber security, there have been multiple studies that show that in instances when people are less cautious and intuitive while making decisions about phishing attacks, they are more likely to fall victims [14, 93]. My work shows that the IBL model is much better at predicting decisions driven by fast and intuitive thinking. Importantly, the IBL is the worst-case scenario modeling (big O) for phishing attack simulation.

This work, however, is not without limitations. The models were developed and validated using a small dataset containing human responses to emails (legitimate and phishing) collected from a laboratory experiment. The dataset may not truly reflect how people would respond to phishing emails in reality. Furthermore, there is a risk of overfitting from fine-tuning BERT using relatively small datasets. Finally, all the models in this work were built using cosine distance function, which may not represent how human partial match two given instances. Therefore, as part of our future work, I intend to test these models on large real-world email datasets and test its effectiveness in adaptive training and risk assessment.

Chapter 6

EYE TRACKING STUDY TO ANALYZE CONTEXT ENCODING DURING PHISHING DECISION MAKING

Previous chapters presented studies on spear phishing attacks that analyzed decision-making behaviors from both the adversarial and end-user perspectives. Cognitive models grounded in the theory of instance-based learning [42] were developed to understand and predict how end-users made phishing decisions. Multiple approaches to representing email instances within the cognitive models were developed and evaluated. Two fundamental research questions consistently emerged throughout all my previous investigations: One, how do people encode emails (including phishing emails) to memory, how should I construct such representations as inputs to the cognitive model, and how inefficiencies in the encoding process may affect end-user decision-making? Second, how robust are the results from the previous study that explained the variance observed in Spearcog predictions in terms of individual differences in decision-making, specifically system-1 decision making?

To answer these questions, this chapter describes a follow-up experiment I conducted to study how end-users made decisions on spear-phishing attacks. Unlike the previous study, this experiment was focused on end-user decision-making and did not include adversarial roles. Notably, in this experiment, an eye tracker was used to collect eye movement data while participants make decisions on a series of emails presented to them. The use of an eye-tracker was pivotal to answering the first question about the encoding process. Instance-based learning relies on the assumption that individuals make decisions based on past experiences by activating relevant memories. While it may not be possible to directly probe and understand the encoding mechanism during email processing tasks, past research has suggested that decision-making is intimately liked to human attention, as well as a re-

lationship between attention and memory [102, 17]. Therefore, in this study, I proposed to use eye-movement data to study the relationship between attention and decision-making in the context of phishing attacks and explore methods to use eye-tracking data to develop attention-based input representations for cognitive models.

Eye tracking has been used in the past to understand how people pay attention to different cues within phishing emails and to understand how attention to such cues impacts their susceptibility to phishing attacks. For example, in a recent study McAlaney and colleagues used an eye tracker to investigate the effect of phishing cues on end-user response [81]. They asked participants to assess the trustworthiness of the emails presented to them and used the eye tracker data to analyze their gaze. They found that participants were less likely to attend to phishing indicators, contrary to the expectation of the researchers [81]. Another study analyzed the attention people pay attention to in emails, including the attachment(s), email body, footer, header, and signature. They found that participants spending more time on email body and signature areas has a significant effect on their email classification tasks [97]. Eye movement has also been used to study the risk of phishing web pages. One study used the seconds spent in the areas of intersection, including web content, the address bar, and security icons, to access the end-user susceptibility to the phishing sites [6]. These past works collected and analyzed data on the attention paid to the non-textual regions of phishing emails and websites. However, to the best of my knowledge, there has not been any work that has analyzed end-user attention to the textual elements of phishing emails and has explored the use of eye-tracking data as input representations for cognitive models.

The experiment seeks to capture eye movement data during email processing to gain insights into the relationship between attention and decision-making in the phishing context. Moreover, it also aims to understand the approach of developing instance-based learning models for phishing tasks and to study how individuals' cognitive processes influence decision-making. The dataset collected is used to explore multiple representation approaches in predicting human decision-making using instance-based learning. Also, I utilize the eye tracking data to investigate the factors that contribute to individual differences in the instance-based

learning (IBL) predictions on individual decision making.

I hypothesize that the sentences, words, and other linguistic items in the email that participants fixate upon during the study are likely the items they store in their memory instances for decision-making. Furthermore, I hypothesize that the participants will "zoom in" on one word at a time when they encounter sentences relevant to decision-making [59] and analyzing the eye-tracking data collected from experiments conducted would reveal a collection of email features and linguistic items that people are theorized to be encoding to their memory. Toward this goal, I proposed, implemented, and tested multiple approaches to representing email instances with a combination of eye-tracking data and natural language processing models. For example, one way to represent the emails within IBL model would be in terms of the regions of interest that people paid the most attention to during decision making. Another way to represent the emails would be to represent the emails in terms of topics or linguistic items (words and sentences) that participants fixated the longest and recalled most. The prediction result was reported and compared with the transfer learning approach used in the previous chapter. These models are used to predict email decision-making and explore the general pattern of performance at an individual level.

In terms of the second question, I repeated the analysis from the previous study to explain the variance observed with the prediction performance of IBL models. However, in this work in addition to response time measures, I also used eye tracking measures such as gaze and fixation information to explain variance found in model performance. I found that the instance-based learning (IBL) model exhibits consistent patterns with the previous chapter, showing that IBL performs significantly better in predicting participants with a preference for quick and intuitive decision making which is representative of system-1 like thinking. Finally, results from this work show that language models trained on on "expert features" has the potential to leverage knowledge gained from a small dataset and contribute to unseen datasets effectively.

6.1 Research questions

To achieve the objective outlined in the previous section, the research is structured into the following separate five research questions.

- Research Question 1 (RQ1): What is the effect of fatigue and cognitive load on end-user attention during phishing decision-making?

  This question aims to investigate the impact of fatigue on the accuracy and efficiency of email response. By monitoring participants' performance over time and assessing their level of fatigue, the study aims to understand how fatigue influences decision-making in the context of email management.

- Research Question 2 (RQ2): How does eye movement affect human decision-making in an email context? This question explores the relationship between eye movements and decision-making processes during email classification tasks. By analyzing participants' eye movement metrics, I aim to identify specific eye movement patterns that are associated with effective or suboptimal decision-making outcomes.

- Research Question 3 (RQ3): Can I use eye movement as a feature and incorporate it with instance-based learning to predict human decision-making? This question focuses on leveraging eye movement data as an additional feature in the instance-based learning (IBL) model. By integrating eye tracking as representations, the section aims to assess whether the approach combined IBL and eye tracking representation improves the accuracy of predicting human decision-making in email classification tasks.

- Research Question 4 (RQ4): How do the instance-based learning (IBL) and perception Bert model demonstrate generalization and transfer learning abilities? This research question focuses on understanding how the instance-based learning (IBL) and perception Bert model exhibit generalization and transfer learning capabilities.

- RQ5: What is the general pattern from the instance-based learning variance on predicting individual decision-making? This question focuses on analyzing the variance in the instance-based learning (IBL) model's predictions. By examining the patterns and trends in the predictions, the section aims to gain insights into the cognitive factors (eye movements) contributing to individual differences in IBL prediction.

Addressing these research questions will provide valuable insights into the cognitive processes underlying email management and decision-making. The findings can inform the development of more effective cognitive models and personalized approaches to enhance end-user security and decision-making accuracy.

6.2 Eye tracking experiment

To answer the research questions described in the previous section, I designed and conducted a follow-up experiment using SpearSim but focused on end-user decision-making. In what follows, the detail of the experiment design and related statistical analysis is first presented. Then, the data pipeline of how I process the data is shown, and finally, I describe the modeling approach and procedure.

The objective of the experiment is to investigate words in the emails that participants fixated upon prior to decision making and how these words influence their final decision-making during email classification tasks. The study was conducted using a standard desktop computer equipped with a Tobii Nano pro eye tracker, which recorded participants' eye movement information as they made decisions about the presented email messages. Our hypothesis is that there is strong relationship between attention and memory recall processes on sentences and words that the participants find to be relevant to their decision-making.

Students from the University of Washington were recruited for the experiment. All the experiment sessions were held in person in the lab and were recruited via email. The median age of the participants was 22 (SD = 4.3). 29.1% were pursuing undergraduate, and those remaining were pursuing graduate-level degrees. 56.25% out of 48 participants' reported

their native language as English. 56% of participants identified themselves as Female, 37% as male, and 4.1% as non-binary. Participants were asked to do the email management task. They were told that the goal of the study was to understand how people manage their emails because past research has shown that when participants know that the study is about detecting phishing emails, they tend to become cautious and biased towards reporting more emails as phishing [94]. For each email, they were asked to choose one of these options: *'Response immediately', 'Response and follow up later', 'Leave in mailbox', 'Delete the email'*, and *'Delete and Block the Sender'*. Each participant processed a total of 50 emails.

The context for decision-making is crucial in phishing and spear phishing. Therefore, the participants were given a fictional role and were asked to make decisions on behalf of the role, a.k.a persona. The profiles utilized in my previous experiment were also retained for this current study 5. They also had access to full information about their persona, including their name, personal details such as location and birth date, details about the persona's work and personal interests. The persona provided contains individual information, family information, professional information, and social information. For more information about the persona, please refer back to the experiment design section in 4.

Email Corpus: The email content consisted of a mix of primary emails(27), general phishing emails(15), promotional emails(5), and spear phishing emails(3) randomly selected from a variety of resources including the Enron dataset, and spear-phishing emails from SpearSim. The whole corpus contains 481 emails, containing 186 phishing emails, 163 primary emails from Enron Dataset, 61 fraudulent emails from Nigerian fraudulent emails, 53 promotion emails, and 18 spear phishing emails obtained from past studies 4.

Mental fatigue Manipulation: Stress and mental fatigue have been identified as important indicators of human susceptibility to phishing attacks. For example, in a recent interview study, participants in the study who reported having been victims of phishing attacks also reported experiencing high levels of stress at the time of susceptibility. Fatigue is also associated with detriment in human attention [10]. So, in this study, participants conducted a task that was expected to induce significant workload and fatigue prior to processing emails

presented to them. To manipulate participants' fatigue levels, participants were asked to solve a series of puzzles [47] before making decisions on the emails. They were simple puzzles that required counting the mathematical symbols shown in an image. In the lower cognitive load condition, participants were required to solve only 2 such puzzles whereas participants in the high cognitive load group responded to 20 trials of puzzles. After solving the puzzles, participants in both condition were asked to filling a NASA TLX survey [49] to access their cognitive load.

After the image classification task, participants are shown surveys to evaluate the cognitive load taken during two conditions. It was found that the mental demand, temporal demand, and effort, frustration were significantly different between the two conditions. The participants in the high condition generally report significantly higher values than the low-condition group. In other words, they report a higher cognitive workload, which is as expected.

6.2.1 General pattern of eye tracking

During the experiment, I found a fatigue pattern in the participants. The anova table is shown in the table 6.1. As trial number increase, participants are spending less fixation on emails. I have found that the number of fixations is significantly affected by the trial number t(-6.35) = 2109.59, $p < 0.0001$. The experiment condition is shown a non-significant effect on the experiment condition(workload) on the number of fixation t(69.41)=1.82, p=0.07.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.27	0.11	69.49	2.43	0.02
Trail number	-0.01	0.00	2109.59	-6.35	0.00
ExpCondition(Low—High)	0.28	0.15	69.41	1.82	0.07

Table 6.1: Fixation across trials and experiment condition

6.2.2 Analysis from manipulating factors

Signal Detection Theory measures such as d-prime and response bias were used to analyze the effect of mental fatigue on participants' decisions to both the ham emails and phishing emails. The metric d-prime (d') serves as a measure of a participant's ability to discriminate between phishing and ham emails effectively. A larger value of d' indicates that the participant is effective in identifying phishing more often, whereas a lower d' value indicates that the participant could not discriminate phishing from ham emails. The metric response bias (β) reflects the overall bias of the participant to classify all emails as either phishing or ham. Both these measures are calculated using the hit rate and false alarm rate. The hit rate (HR) is a measure that quantifies the proportion of correctly identified phishing emails from the total number of phishing emails presented to participants (Signal). In other words, it measures how many of the phishing emails were correctly detected as phishing by the participants. On the other hand, the false alarm rate refers to the number of regular emails incorrectly classified as phishing emails. It is also known as false positives, in which regular emails are falsely identified as malicious.

$$HR = \frac{Hits}{Signal}$$

$$FAR = \frac{FalseAlarms}{Noise}$$

Using hit rate and false alarm rate, d-prime(d') and response bias c were calculated as defined in signal detection theory. When estimating d-prime and response bias for extreme proportions of stimulus, where either the hit rate (HR) or the false alarm rate (FAR) is 0 or 1, a correction is necessary to avoid division by zero. A common approach is to replace 0 with a small positive value(0.001) and 1 with a value slightly below(0.999) before calculating d-prime and response bias [114].

$$d' = z(HIT) - z(FA)$$

The sensitivity index d prime refers to how easy or difficult it is to detect that a signal exists in the presence of noise. The d prime is the distance between the means of signal and noise distributions. If a participant has a low d-prime, it means that the participant has difficulty distinguishing phishing emails from benign emails. The response bias c represents the number of standard deviations from the midpoint between signal and noise distributions. d prime and c were calculated using the following equations.

$$c = -0.5 * [z(HIT) + z(FA)]$$

A response bias greater than 0 indicates a bias toward classifying emails as ham and a response bias less than 0 indicates a bias toward classifying emails as phishing. To answer research question 1, I used a linear mixed effect model on both the d prime and response bias as the dependent variable. The independent variables were the experiment condition and whether the measurements were calculated in the first half or second half of the experiment. To reduce noise from individual differences and emails, the random effects in the logistic regression model include participants' ID and email ID. Since each participant may receive a random set of emails, these are considered crossed random effect factors.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Section(first half | second half)	0.18	0.18	1.00	46.00	0.31	0.5810
Exp Condition(High | Low)	0.58	0.58	1.00	46.00	1.01	0.3191
Section:Exp Condition	0.84	0.84	1.00	46.00	1.46	0.2327

Table 6.2: Mixed effect model predicting d-prime

For d', I found that participants demonstrated no difference between the first half and second half of the email classifications (F(1,46) = 0.31,p =0.58), and no significant difference due to experiment condition (high and low cognitive load (F(1,46) = 1.01, p =1.01)).

In terms of response bias, I found a marginal difference in participants' performance between the first half and second half of the experiment (F(1,46) =3.82, p = 0.0569) The average was 0.79 in the first half of the experiment and 0.9 in the second half of the exper-

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Section(first half | second half)	0.39	0.39	1.00	46.00	3.82	0.0569*
Exp Condition(High | Low)	0.12	0.12	1.00	46.00	1.16	0.2868
Section:Exp Condition	0.26	0.26	1.00	46.00	2.55	0.1168

Table 6.3: Mixed effect model predicting Response Bias

iment. As participants experienced more fatigue towards the end of the experiment, they were much more likely to choose to respond to the emails presented to them. Again, the experiment condition did not show a significant effect on the response bias ($F(1,46) = 3.82$, $p = 0.2868$).

From both tests, I did not observe effects from the experiment condition that would suggest that cognitive load and fatigue have a significant effect on susceptibility to phishing and spear-phishing emails. However, I found that the response bias has a marginal difference between the first half and second half of the experiment, which may suggest that fatigue may have an effect on response bias.

6.2.3 Eye tracking modeling result

After analyzing the impact of fatigue, I also analyzed the relationship between how eye movement affects decision-making. These analyses aims to answer research question 2: how would eye movement affect human decision-making? or what kind of eye metrics would result in decision-making in the email classification? The model initially consists of all eye tracking features obtained from Tobii Pro lab. The full list of features is listed in Appendixes A and B. Participants' response to each email was encoded as response vs non-response. The response label includes the '*response immediately*', and '*flag and follow up later*'. On the other hand, Non-response label includes the '*leave in the mailbox*', '*delete email*', '*delete the email and block the senders*'. To analyze the effect of eye-tracking metrics on decision making, I initially fit all the metrics to the model predicting the decision making and conducted feature selection to avoid correlating features. Eventually, table 6.4 shows the results from the final model. If

the coefficient is positive, then participants were more willing to respond to the email, and vice versa.

I found that the duration of the first whole fixation and the number of saccades were the significant features that positively affected the decision. If there are interesting elements in the email, then participants generally spent more time on the first fixation. A similar pattern could be found in the number of saccades. However, the average duration of whole fixations is found was found to be insignificant. On the other hand, the average whole fixation pupil diameter is the factor that negatively significantly affects decision-making. In other words, the more average whole fixation pupil diameter, the more likelihood participants not responding to the email. Note that it was found that the increasing mental fatigue could cause pupil constriction [56].

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.31	0.22	5.85	0.00
Average_duration_of_whole_fixations	-0.18	0.11	-1.69	0.09
Duration_of_first_whole_fixation	0.23	0.10	2.38	0.02*
Average_whole.fixation_pupil_diameter	-0.33	0.16	-2.08	0.04*
Number_of_saccades	0.63	0.11	5.74	0.00***
Maximum_peak_velocity_of_saccades	0.11	0.09	1.28	0.20

Table 6.4: Mixed effect logistic model predicting decision-making

6.3 Modeling protocol

6.3.1 Data Pipeline

To achieve the original goal of understanding participants' attention during email classification tasks, the appropriate cleaning process was conducted by matching pixel-level eye movement data with the areas of interest (AOIs) corresponding to words within each email. Figure 6.1 shows the pipeline of this data processing. The AOIs were generated using the optical character recognition (OCR) packages pytesseract and easyocr. These OCR tools were used to identify both the location of the text and its content within the emails. Although

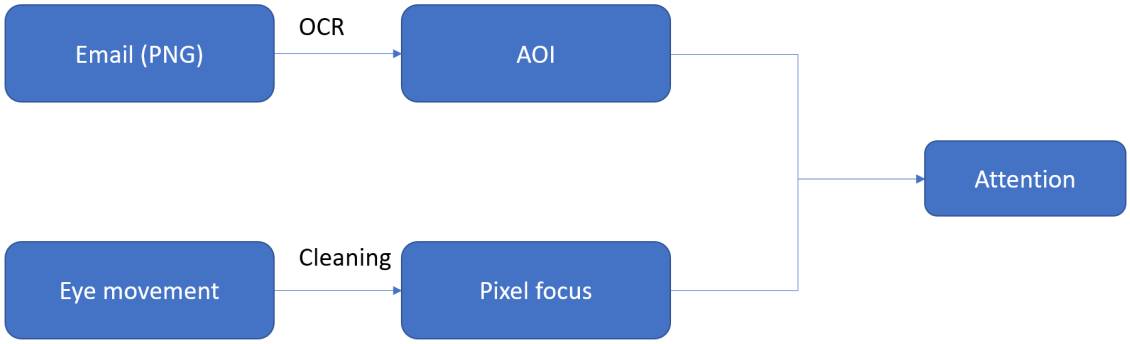


Figure 6.1: Data Pipeline for eye tracking study

OCR could be erroneous in recognizing words accurately, it still provided valuable information about the text's spatial arrangement and content. AOI and pixel-level focus were combined to generate attention weights, which served as indicators of how much attention each participant devoted to specific words within an email.

However, I found that eye tracker dataset suffers from systematic downshifting at the data collection time. Therefore, I implemented an algorithm to correct eye movement using heuristic methods. The correction algorithm considers the eye movement pixels that don't fall into the identified AOIs. The algorithm searched for the nearest AOI within desired pixel range. If no AOI was found within a desired distance, the pixel data was excluded from contributing to the weight given to the text. By implementing this heuristic correction algorithm, the study aimed to improve the accuracy and reliability of the eye movement data. This enabled a more precise representation of participants' attention distribution during email processing and subsequent decision making. Examples are given in table 6.5 and Figure 6.2.

The Table 6.5 presents the email processed by OCR, and attention-based eye tracking. When people read emails, they typically dont follow the word order as it is presented. Instead, people tend to skim first and scan for keywords or phrases to locate relevant information.

extraction type	emails
email	erwollam@hotmail.com man night again? Zero response from my last effort this matter: Are vouf emails working men? Zander called and suggested attend the CCA crawfish boil function held Thursday; May 2022. The reasonably priced event; beer; mud bugs and We can back and watch p1g Win raffle prizes I'm Send and receite Hotmail vouf mobile device: http: mobile msn com on we to be 2, sit in e. o
human reading	men? on effort on effort and called working men? working Zander and suggested and suggested and suggested and attend we sit back attend back sitk sit attend we p1g crawfish p1g CCA crawfish raffle boil raffle boil crawfish boil Win boil device: http:ḿobile msn com function raffle ion r function raffle function raffle function raffle function held be held be held May Thursday; May Thursday; reasonably event; priced reasonabevently priced event; beer; Im beer; Im in Im in Im in e. mudbugs e. mudbugs e. can Hotmail sit Hotmail on and on and p1g watch vouf watch Win ps e.1g raffle prizes e. Hotmail device:

Table 6.5: Email example and how human actually read email

The first row of table represents the text generated by OCR methods. There will be some word incorrectly identified. The second row is an example of how people actually reading the emails. People genernally read a paragraph of text back and forth, and the second row reflects on the pattern of reading.

Then, people read in chunks focusing on sections and paragraphs by words. People also read back and forth between paragraphs and words. If an email is like the first row in table 6.5, then the way people read it is like what is shown in the second row.

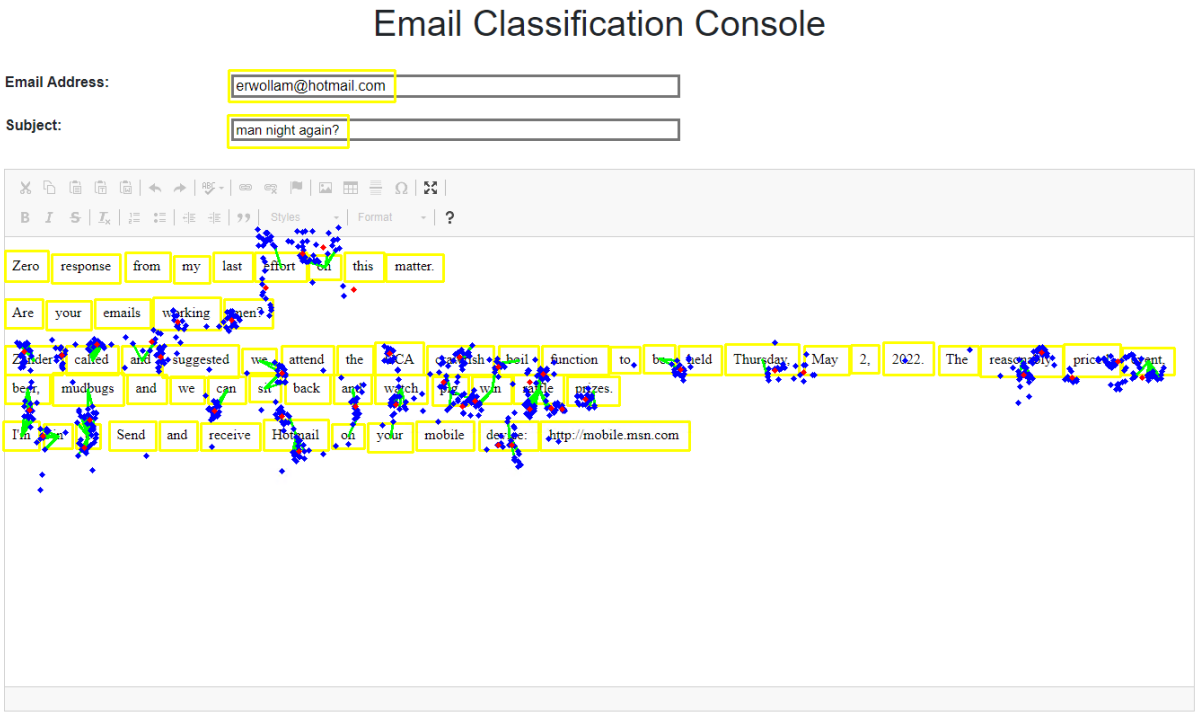


Figure 6.2: Eye movement example with AOIs, Fixation, Gaze, and corrections

The figure shows an email classification process, and how I process the data. The ocr was applied to extract the text in the email, information including words, location of the words. Other than the email information, there is also eye movement on the email. The blue points means gaze, and red dots are fixation. Correction was made to these gazes and fixation, and the green line connected the blue points, which are not inside the frames to the most adjacent aoi.

6.3.2 Prediction model based on Eye Tracking dataset

ACT-R models the cognitive process of decision-making, and human decision-making highly depends on attention and working memory. Eye movement is closely related to attention

and working memory, making it a valuable indicator of participants' cognitive processes during email classification tasks. The hypothesis of whether eye tracking features could be effectively encoded as input for the instance-based learning (IBL) model to predict human decision-making in email classification tasks is tested. By combining eye-tracking data with word embedding, it becomes possible to create a representation that incorporates both visual attention (captured through eye movements) and salient textual information (represented by word embedding). The eye movement data processing procedure described in the previous section involves segmenting words and augmenting them to fit into language models, which captures the attention and salience of each word. This approach provides a unique and comprehensive representation of participants' attention during email processing.

There are four models that are being tested to answer this research question.

- BERT + OCR text In this configuration, the BERT model is used, and the input text is obtained using OCR (Optical Character Recognition). The OCR extracts the text from the images of the emails, which is then fed into the BERT model for prediction. Since the other methods use the ocr generated text as input, this model also uses the OCR generated text as a baseline.

- BERT + eye tracking attention correction This configuration involves using the BERT model along with eye-tracking features. The eye tracking data is used to select and give weights to different words in the email. By incorporating eye tracking attention correction, the model aims to account for participants' actual attention patterns while processing the emails. Both gaze and fixation are considered in this model.

- BERT + eye tracking attention correction (fixation only) This configuration is similar to the previous one, but it focuses only on eye fixation data for attention correction. Eye fixations refer to the brief pauses the eyes make when processing information and the filtered rate is 100ms by default from the Tobii pro lab software. By considering only eye fixation data for attention correction, the model takes the pixels when participants

focused their attention on particular words.

- BERT +eye tracking attention correction (saccade only) Similar to the previous configuration, this setup also involves using eye tracking data for attention correction. However, in this case, the attention correction is based solely on saccades, which are rapid eye movements between fixations. By focusing on saccades, the model captures rapid shifts of attention between different words.

Each of these combinations is an approach to incorporating attention information along with instance/context represtations into the IBL model. Such a comparison between different approaches to incorporating attention in instance representation was expected to provide a deeper understanding of how attention influences instance encoding and therefore, decision-making.

6.3.3 Transfer learning from perception bert

In Chapter 5.2.1, the Perception Bert model was introduced and trained on the Spear-Sim dataset. This model demonstrated significant improvements in predicting participants' decision-making during email classification tasks. However, the generalization and transferability of the model to other datasets were not explored. To investigate transferability, I tested the perception bert with this dataset within the IBL model. This experiment aims to assess the model's performance and generalization capabilities when applied to new and unseen data.

6.3.4 Modeling procedure

For each participant, I randomized the order of emails to fit an IBL model on the train set, and predicted on the test set. The train test split ratio is 0.8. To get a reasonable representation of the performance, I fit IBL model 100 times for each participant and calculated the average accuracy and F1 score to evaluate the model performance.

6.4 Result

In this section, the major results of the experiment are presented, starting with exploratory analysis followed by modeling and statistical analysis.

6.4.1 IBL encoding with eye movement data

I hypothesized that analyzing eye movement data could inform how people process emails and how they may encode them to memory. The eyes are the window of our brain, and it was found that decision-making has significant associations with some eye movement metrics. To further understand this phenomenon, four approaches described in section 6.3.2 were tested and compared, following the procedure described in section 6.3.4.

The F1 score for the model is shown in Figure 6.3, and there is no significant difference that could be found between different input configurations.

6.4.2 IBL modeling with perception bert (transfer learning)

Chapter 5.2.1 presents multiple ways of representing text in the IBL model. Chapter 5.4 discussed the challenges in the perception Bert model and the potential risk of overfitting. To further validate the capability of the IBL model, the Perception Bert model is tested on the new dataset to assess whether the observed performance improvement remains consistent. This test will help determine whether the improvements observed in the previous dataset (SpearSim) generalize to the new dataset and provide insights into the model's robustness across different data. In addition, it is also unclear how much effort the attention term played versus the SpearCog model in chapter 5. To address this question, a 2 by 2 comparison is proposed. The comparison involves two factors: (1) fine-tuned versus not fine-tuned Perception Bert model, and (2) randomized versus non-randomized order of email instances.

Fine-tuned: I used the same perception bert model described in section 5.2.1 to predict the participants' responses in this study. The perception Bert model was trained on the SpearSim dataset, which was fine-tuned with the survey questions to force the large language

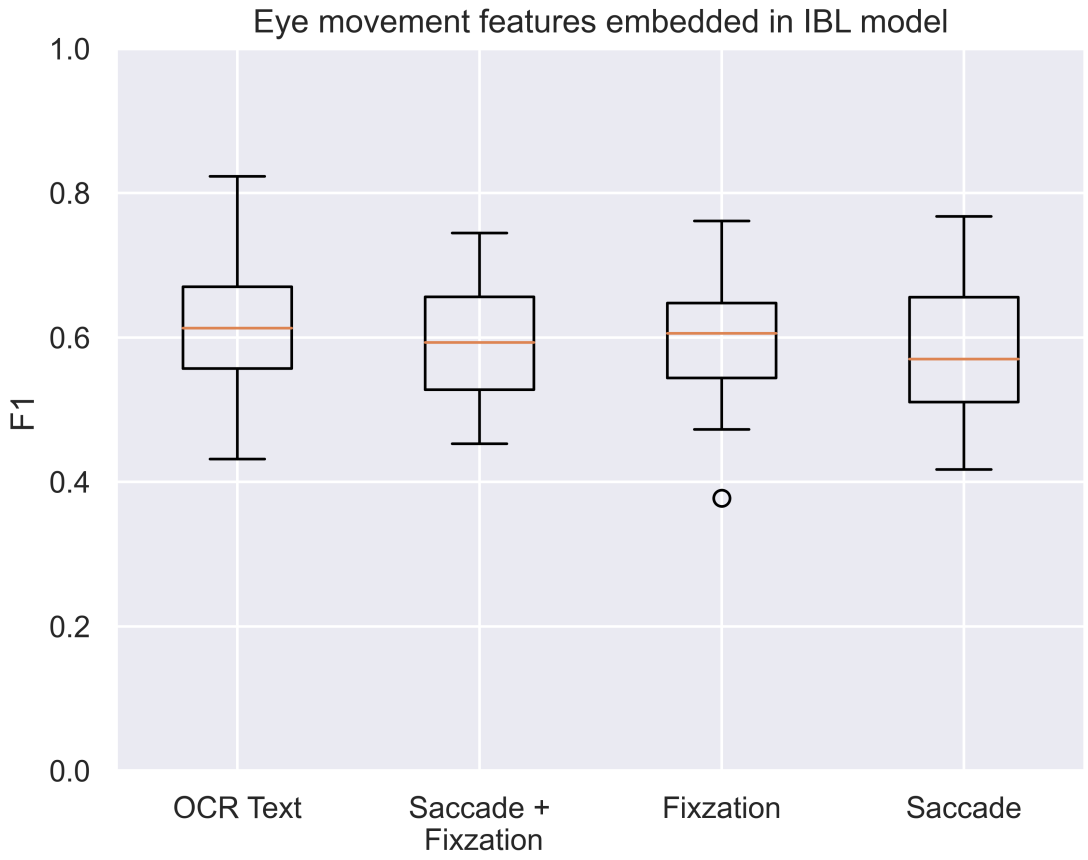


Figure 6.3: Eye movement features embedded in IBL model

model to learn how humans remembered the email content. The perception Bert model was able to achieve good performance.

Randomization: Randomization factor could be used to evaluate the contribution of the memory and decay process in the IBL model. For non-randomized IBL models, the IBL agents received email instances in exactly the same order as what the participant had experienced during the experiment, and the test set contains the last 20% of the emails that participants processed. On the other hand, for comparison, in the case of randomized IBL agent design, the agents' experience the email instances in random order.

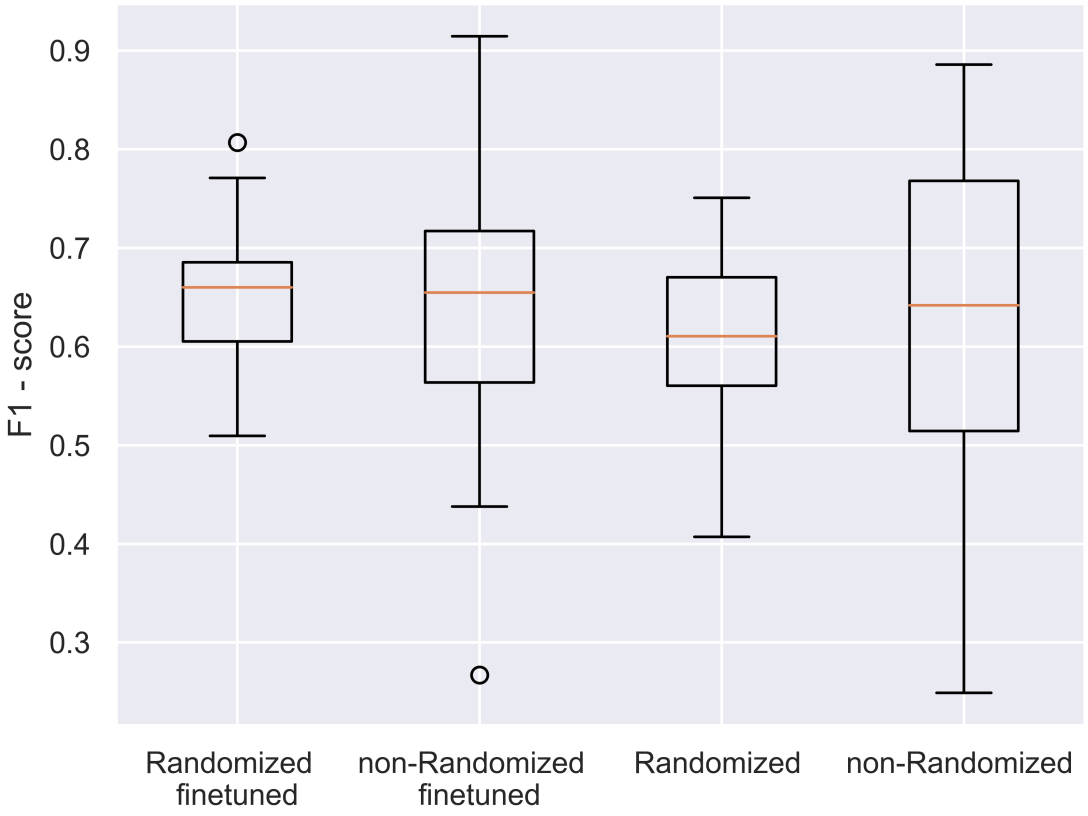


Figure 6.4: IBL modeling performance on email decision making across representation method and randomization

I tested the instance-based learning with the fine-tuned bert(perception bert) and the order of processing instances. Figure 6.4 and table 6.6 present the performance of the model under four scenarios. To assess the differences in the variance of F1 scores between the randomized group and non-randomized group within each fine-tuned factor, two separate F tests were conducted. Both tests show that the variance of the non-randomized group is significantly higher than the variance of the randomized group with $F_{47,47} = 0.2649 p < 0.0001$ and $F_{47,47} = 0.2856, p < 0.0001$.

Table 6.7 presents the result from the mixed effect model which shows that the fine-tuned model has a significantly better F1 score compared to the non-randomized model (2.93%). This suggests that the transfer learning ability of the Perception Bert model is retained across datasets, leading to improved performance on the new dataset.

finetuned	randomized	f1 mean	std
0	0	63.19	15.31
	1	61.19	8.18
1	0	65.05	12.49
	1	65.20	6.45

Table 6.6: Standard deviation of the performance

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	62.65	1.45	132.41	43.22	0.00
randomized	-0.92	1.40	142.00	-0.66	0.51
finetuned	2.93	1.40	142.00	2.09	0.04*

Table 6.7: Mixed effect result for IBL performance

6.4.3 Performance of the IBL vs time spent on emails and eye movement metrics

From the previous section, I observed variance in IBL performance between participants. Chapter 5.3.6 also discovered the pattern of IBL performance with the time spent on emails. Does this pattern hold for this dataset? Additionally, I aim to identify what eye-tracking features affect the IBL performance. The objective is to understand why the IBL agent performed well for some participants but not for others. Email level prediction was conducted and analyzed on what eye movement features would make the IBL agent easier or harder to predict participants' decisions. I utilized the result from the randomized baseline model in the previous section, and extract the IBL performance from the prediction result in the test set. For each classification, the percentage of IBL agent correct prediction of participants' responses across 100 simulation runs is calculated. Since the distribution of the IBL correct ratios is non-normalized, it poses challenges for using a mixed-effect linear model. The mixed-effect linear model requires the residual value to be a normal distribution. However, the residual value of the model appears to be non-normal distribution and violates the assumption. Therefore, I used the the binary logistic mixed effect model instead. I defined the dependent variable, IBL correct response category, as follow. If the IBL correct ratio is smaller than 0.5, then the label is created as classification cannot get good predictions from IBL, and vest versa. I then change the IBL correct prediction ratio to correct response category(1 as correct prediction ratio $> 50\%$ and 0 as correct prediction ratio $< 50\%$). Same as previous model, the participants and emails are used as the crossed random effects factors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.64	0.07	9.37	0.00
Average_duration_of_whole_fixations	0.08	0.06	1.45	0.15
Duration_of_first_whole_fixation	0.00	0.05	0.09	0.93
Average_whole.fixation_pupil_diameter	0.01	0.06	0.18	0.86
Number_of_whole_fixations	-0.13	0.05	-2.50	0.01

Table 6.8: Mixed effect logit model on fixation related features

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.65	0.07	9.25	0.00
Number_of_saccades	-0.15	0.05	-2.76	0.01
Average_amplitude_of_saccades	-0.06	0.06	-1.02	0.31
Time_to_first_saccade	0.09	0.07	1.43	0.15
Amplitude_of_first_saccade	-0.02	0.05	-0.42	0.67

Table 6.9: Mixed effect logit model on Saccade related features

To select appropriate feature sets from the lists of eye-tracking metrics in apendix A and B, the saccade and fixation features are usually tangled together, and therefore separate analyses were conducted on saccade and fixation features to predict the categories of IBL correct response. All features listed in Appendix were included in the mixed effect logistic model initially. During the model refinement process, factors that exhibited strong correlations with each other and features that were deemed irrelevant were gradually excluded from the model. Table 6.8 presents the result from fixation features on the IBL correct response category. The number of whole fixation shows a significant effect on the IBL prediction performance (z(-2.5) = 0.01). This result shows that a one unit increase in fixation during email processing, the odd of the IBL model correctly predicting decisions over 50% would decrease $e^{-0.13} = 0.87$ times than the IBL model correctly predicting decisions less than 50%. Table 6.9 presents the result from saccade features on the correct response category. The number of saccades shows a negative effect on the IBL prediction performance (z(-2.76)=0.01). A one unit increase in saccade during email processes, the odd of the IBL model correctly predicting decisions over 50% would decrease $e^{-0.15} = 0.86$ times than the IBL model correctly predicting decisions less than 50%.

Other fixation and saccade features do not show significant effects on predicting the IBL correct response category. It is interesting to note that the number of saccades and the number of fixations are highly correlated (correlation coefficient = 0.9768). The analysis results indicate that the more time spent on email classification, which is associated with more saccades and fixations, the more difficult it becomes for the IBL agent to make correct

predictions.

Furthermore, there is a strong correlation between email processing time with the number of saccades($r = 0.89$) and the number of fixations($r = 0.93$). Therefore, I separately tested the effect of email processing time on the IBL performance category. The email and participants are used as the random effects. The email processing time was found to be a significant factor and negatively affects the IBL performance(z=-2.4, p=0.02).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.82	0.10	7.87	0.00
time	-0.01	0.00	-2.40	0.02*

Table 6.10: Mixed effect logistic model on email processing time

6.5 Discussion

This chapter describes a human subject study I conducted to understand how people process and remember emails (including phishing emails) and how this affects their response decisions. Using eye movement data, I investigated the relationship between decision-making and attention to textual features within emails. I found that factors such as the duration of the first whole fixation and the number of saccades were significantly associated with a higher likelihood of response. This would suggest that a person's first glance at an email strongly affects their decision-making. Our findings also reveal that the smaller the average whole fixation pupil diameter, the more likely participants were to respond. When a person's mental fatigue increases, their pupils constrict [56]. This suggests that participants who were experiencing greater fatigue were more likely to respond to emails, which means they were also more likely to fall victim to phishing attacks.

Using the eye-tracking dataset, I experimented with different ways of representing email instances in models of instance-based learning. Although data from eye-tracking instruments reflect people's attention, it could be a noisy representation of how people process emails.

Also, embedding eye-tracking data of human attention toward text data is complicated to embed with cognitive models. Embedding attention information within cognitive models involves multiple steps, from processing eye pixels to language features and attention weights (a matrix). I experienced several challenges while processing the data. The augmentation approach used to represent how individuals read emails was not readily compatible with the large language model (LLM). In addition, there was also noise in the collected data. I observed a systematic downshifting in eye-tracking data. Although I developed an algorithm to rectify the downshifting issue, it may not have been entirely resolved and could have impacted the model performance.

I assessed the performance of the perception bert model developed in 5 on the data collected from this experiment. Remarkably, I found that the perception bert model's performance remained relatively high across datasets, showcasing its robustness and transferability. Note that the perception Bert model was fine-tuned based on participants responses to the survey related to user decision-making. This approach captures higher-level information about email decision-making behavior, which may make it more adaptable to models of instance-based learning. In contrast, eye-tracking data provides more insights into the fundamental attention processes during the processing of emails. Although eye-tracking data contains substantial information about how individuals process emails, it may not be effective in selecting the most relevant and salient features predictive of human decisions. Indeed, the user perception approach to representing features, as described in chapter 5, yields a comprehensive and high-level representation of features in the decision-making process. This pattern indicates that the perception bert model might be the correct model for bridging the gap between higher-level user-assigned features with low-level language features. Comparing multiple ways of presenting representations highlighted the significance of capturing the most relevant and informative features in decision-making.

Furthermore, to understand the contributions of memory effect vs. representation to the prediction, I manipulated the order of instances feeding into the IBL model. By training the IBL model with the same order of instances that participants experienced, I aimed to

assess the impact of memory on decision-making predictions. When the instance was not randomized in the training set, the variance of the IBL prediction result increased dramatically. The root cause of this could be a lack of generalization in the email instances in the training set. In such cases, the IBL model may have struggled to handle instances that were dissimilar to those encountered during training. Although the mean F1 scores do not show a significant difference, I observed that the standard deviation almost doubled from the approach involving randomized training to a non-randomized one. Further investigating this observation, I conducted an F test and found that the F1 scores trained with the randomized method had a much larger standard deviation than the non-randomized ones. This pattern could be attributed to the lack of variety in instances when the order is non-randomized. Some of the IBL agents achieved higher performance, such as 0.9. This demonstrates that some of the IBL agents trained with the original order could do better in predicting human decision-making due to representative memory activation processes. In addition, when applying the fine-tuned perception bert model to represent text, I found that the variance in the performance shrunk from 15.31 to 12.49, which means the robustness was improved. This result highlights the transferability of the Perception Bert model in predicting human decision-making.

I also utilized eye movement features to analyze variance observed with IBL performance. What type of eye-tracking metrics contributed to the observed variance in IBL model performance? The number of saccades and the number of fixations emerged as the primary eye-tracking metrics that could explain the variance in the IBL model's performance. Notably, a lower number of saccades or fixations was associated with higher accuracy in the IBL model's predictions. This suggests that participants who exhibited fewer eye movements while classifying emails tended to have more predictable decision-making patterns as governed by IBL functions. Meanwhile, the time spent on email classification was highly correlated with these factors and could not be separated out. Interestingly, this pattern of email processing time negatively affecting IBL performance aligns with the findings regardging intuitive decision-making presented in previous chapters.

In sum, this study employed eye movement data and statistical analysis to explain how individuals make decisions regarding emails. Notably, it pioneers the incorporation of lower-level data input into instance-based learning. The analysis has valuable implications for researchers and practitioners seeking to develop cognitive models for phishing decision making.

6.5.1 Lessons Learned About IBL

IBL was developed about 20 years ago and has been proven to be a valuable framework for decision-making in a variety of domains, including cybersecurity, aviation, and many others [42, 41, 23, 22]. New variants of IBL models have been developed to improve IBL from multiple aspects, including computational performance [87]. Recent research has also shown that an IBL model combined with q-learning can outperform traditional q-learning in the initial phases [86]. These improvements show the new possibility of IBL adopted into the big data era.

However, there is a growing demand for more advanced and adaptive representation methods to capture key characteristics of instances. IBL's setup is well-suited for representing memory activation and decision-making processes, but it is limited in its ability to select and encode features to represent the context. Deep learning techniques hold great promise for addressing this challenge. By leveraging deep learning methods for feature representation, I show that the IBL model's ability to capture and process complex patterns and interactions in the data is enhanced. In the big data era, a variety of behavioral data streams are available for modeling. For example, the IBL model has the potential to be a good backbone for human-centered digital twin models. These models can better understand and predict how humans process information, make decisions, and react to different situations. The integration of deep learning techniques allows for more sophisticated feature extraction and representation, enabling digital twin models to better mimic human behavior and decision-making processes. Furthermore, optimizing the feature selection process through transfer learning and other advanced techniques can enhance the IBL model's adaptability

and performance. The approach shown in section 5.1.1 use the transfer learning approach to improve IBL performance.

6.5.2 Limitations and future directions

I have identified several limitations that need to be addressed in future work. First, for the experiment design, the profile setup could be changed to contain email inboxes of familiar profiles (e.g., student inbox if using students as participants). The current email dataset represents communication that happens in a corporate environment, which may not have been familiar to the students who were the participants in the study. However, such a dataset is limited. This difference in the nature of communication may have affected participants' responses. It is important to further investigate this.

Second, the way of processing the data could be improved. To create representation for IBL models, I used a data augmentation technique to represent how the individuals actually read emails. As shown in Figure 6.5, the email text has many differences from how humans actually read email. Since this dataset is a small dataset, and it was limited to training our own LLM models, simple data augmentation was the most feasible approach to transfer the eye-tracking attention text to vector format. However, there could be other representation methods to explore in the future.

Related to this, the approach I have proposed is a two-step modeling process: 1)the IBL model to capture the long-term memory activation processes ands 2) deep learning methods for representing the instances within the IBL model (the similarity portion of IBL modeling). Deep learning methods are effective when working with large-scale datasets and representing language features. A more fundamental question that could be asked here is whether I can reformulate the IBL model to embed optimization methods into it. More work could be done to explore this area.

Third, another aspect of processing data is eye movement accuracy. Eye tracking suffers from downshifting, and it was challenging to accurately reflect the regions that participants attended to during the study. I created a heuristic-based algorithm aimed at mitigating the

impact of downshifting, but I have reservations about its ability to fully address the problem.

In addition, the modeling approach is modeled on the phishing-related task, and the pattern I have found related to the characteristics of IBL needs to be validated in other domains.

Chapter 7

GENERAL CONCLUSION

In this dissertation, I explored cognitive modeling, decision-making, and natural language processing related to phishing attacks from both the adversarial and end-user sides. I used human subject studies, simulation modeling, and machine learning to model human decision-making related to phishing and spear phishing emails.

Chapter 3 presented a study about text analysis with language inquiry word count to predict end-user vulnerabilities. The chapter revealed how psycho-linguistic features affect end-user susceptibility in mass phishing attacks.

Chapter 4 shifted the focus to spear phishing attacks, a customized version of phishing attacks. I designed an adversarial human subject study as a test bed to simulate behaviors from the perspectives of both attackers and end-users. This is the first test bed to incorporate live interaction and feedback between attackers and end-users in spear phishing attacks. Using the test bed, I compared the emails created by attackers with different amounts of available end-user information. Attackers generally create more convincing narrative stories in spear phishing emails when they have access to more of the target's personal information.

With such a dataset, Chapter 5 focused on developing end-user models to predict decision-making in response to emails using a combination of natural language processing and instance-based learning. I tested multiple approaches for representing email instances within the model. Notably, one model developed in the chapter, perception BERT, leveraged the learning from a small human subject study incorporating user-defined features to an NLP model embedded in IBL. I also analyzed the factors affecting IBL agent performance and found that the IBL model can much more easily predict the responses of individuals who tend to make decisions intuitively and quickly than in instances where participants may have made

decisions based on a more elaborate and analytical thinking process.

Lastly, Chapter 6 presented a human subject study and modeling experiments I conducted to investigate individuals' attention processes while processing emails and its effect on decision making. The chapter also described experiments that deployed eye-tracking data with IBL to predict end-user decision-making and discussed the implications. The chapter also validated the robustness of the IBL model and some general patterns developed in the previous chapter.

The dissertation focused on the effect of memory to phishing susceptibility. However, there are other aspects that have been shown effective in past research and could not be overlooked, such as end-user meta-cognition [15, 105] and social factors (e.g. culture and language) [121, 1]. In future work, these factors could be potentially explored to evaluate their effects on phishing susceptibilities.

The implications of this dissertation are threefold. First, I designed a test bed for simulating spear phishing attacks. This test bed could be utilized to conduct studies analyzing emerging questions from the attacker's side. Second, I modeled end-user decision-making with a combination of instance-based learning and natural language processing. This model could be used as a test bed to improve training and simulation attacks, which might in turn improve security awareness training efficiency. Third, the model that uses IBL as its foundation provides insight into how I could apply and improve the IBL model in the future.

BIBLIOGRAPHY

[1] Ahmed Aleroud, Emad Abu-Shanab, Ahmad Al-Aiad, and Yazan Alshboul. An examination of susceptibility to spear phishing cyber attacks in non-english speaking communities. *Journal of Information Security and Applications*, 55:102614, 2020.

[2] Ahmed Aleroud and Lina Zhou. Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68:160–196, 2017.

[3] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82:69–82, 2015.

[4] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.

[5] JRAL Anderson and C Lebiere. The atomic components of thought lawrence erlbaum. *Mathway, NJ*, 1998.

[6] Sabri Arik, Tingwen Huang, Weng Kin Lai, and Qingshan Liu, editors. *Neural Information Processing*, volume 9491 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2015.

[7] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[8] Piers Bayl-Smith, Ronnie Taib, Kun Yu, and Mark Wiggins. Response to a phishing attack: persuasion and protection motivation in an organizational context. *Information & Computer Security*, 2021.

[9] Minakshi Bhardwaj and GP Singh. Types of hacking attack and their countermeasure. *Int. J. Educ. Plann. Admin*, 1(1):43–53, 2011.

[10] Maarten AS Boksem, Theo F Meijman, and Monicque M Lorist. Effects of mental fatigue on attention: an erp study. *Cognitive brain research*, 25(1):107–116, 2005.

[11] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

[12] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[13] Marcel Brass, Perrine Ruby, and Stephanie Spengler. Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2359–2367, 2009.

[14] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. Quantifying phishing susceptibility for detection and behavior decisions. *Human factors*, 58(8):1158–1172, 2016.

[15] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. Better beware: comparing metacognition for phishing and legitimate emails. *Metacognition and Learning*, 14(3):343–362, 2019.

[16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[17] Marvin M Chun and Nicholas B Turk-Browne. Interactions between attention and memory. *Current opinion in neurobiology*, 17(2):177–184, 2007.

[18] Robert B Cialdini and Lloyd James. *Influence: Science and practice*, volume 4. Pearson education Boston, 2009.

[19] Brendan Conway-Smith and Robert L West. System-1 and system-2 realized within the common model of cognition. In *AAAI 2022 Fall Symposium*, 2022.

[20] Gordon V Cormack and Thomas R Lynam. Spam corpus creation for trec. In *CEAS*, 2005.

[21] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

[22] Edward A Cranford, Cleotilde Gonzalez, Palvi Aggarwal, Sarah Cooney, Milind Tambe, and Christian Lebiere. Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, 12(3):992–1011, 2020.

[23] Edward A Cranford, Christian Lebiere, Prashanth Rajivan, Palvi Aggarwal, and Cleotilde Gonzalez. Modeling cognitive dynamics in (end)-user response to phishing emails. *Proceedings of the 17th ICCM*, 2019.

[24] James M Curran. Hotelling: Hotelling's t-squared test and variants. *R. package version*, pages 1–0, 2013.

[25] Avisha Das, Shahryar Baki, Ayman El Aassal, Rakesh Verma, and Arthur Dunbar. Sok: a comprehensive reexamination of phishing research from the security perspective. *IEEE Communications Surveys & Tutorials*, 22(1):671–708, 2019.

[26] Sanchari Das, Christena Nippert-Eng, and L Jean Camp. Evaluating user susceptibility to phishing attacks. *Information & Computer Security*, 2022.

[27] Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979, 1996.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.

[30] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1065–1074, 2008.

[31] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2873–2882, 2015.

[32] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009.

[33] Paul Ekman and Maureen O'Sullivan. Who can catch a liar? *American psychologist*, 46(9):913, 1991.

[34] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

[35] Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. Principles of persuasion in social engineering and their use in phishing. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 36–47. Springer, 2015.

[36] Ana Ferreira and Soraia Teles. Persuasion: How phishing emails can influence users and bypass security measures. *International Journal of Human-Computer Studies*, 125:19–31, 2019.

[37] Peter W Foltz. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, 28(2):197–202, 1996.

[38] Matthias Gamer and Wolfgang Ambach. Deception research today. *Frontiers in psychology*, 5:256, 2014.

[39] Yan Ge, Li Lu, Xinyue Cui, Zhe Chen, and Weina Qu. How personal characteristics impact phishing susceptibility: The mediating role of mail processing. *Applied Ergonomics*, 97:103526, 2021.

[40] Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

[41] Cleotilde Gonzalez and Varun Dutt. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological review*, 118(4):523, 2011.

[42] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003.

[43] Stefano Grazioli. Where did they go wrong? an analysis of the failure of knowledgeable internet consumers to detect deception over the internet. *Group Decision and Negotiation*, 13(2):149–172, 2004.

[44] Thiago S Guzella and Walmir M Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222, 2009.

[45] Ziad M Hakim, Natalie C Ebner, Daniela S Oliveira, Sarah J Getz, Bonnie E Levin, Tian Lin, Kaitlin Lloyd, Vicky T Lai, Matthew D Grilli, and Robert C Wilson. The phishing email suspicion test (pest) a lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behavior Research Methods*, pages 1–11, 2020.

[46] Jeffrey T Hancock, Catalina Toma, and Nicole Ellison. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449–452, 2007.

[47] Uriel Haran, Ilana Ritov, and Barbara A Mellers. The role of actively open-minded thinking in information acquisition, accuracy, and calibration. 2013.

[48] Brynne Harrison, Elena Svetieva, and Arun Vishwanath. Individual processing of phishing emails: How attention and elaboration protect against phishing. *Online Information Review*, 2016.

[49] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[50] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. Detecting and characterizing lateral phishing at scale. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1273–1290, 2019.

[51] Shuyuan Mary Ho and Jeffrey T Hancock. Computer-mediated deception: Collective language-action cues as stigmergic signals for computational intelligence. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[52] Shuyuan Mary Ho and Jeffrey T Hancock. Context in a bottle: Language-action cues in spontaneous computer-mediated deception. *Computers in Human Behavior*, 91:33–41, 2019.

[53] Shuyuan Mary Ho, Jeffrey T Hancock, Cheryl Booth, and Xiuwen Liu. Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication. *Journal of Management Information Systems*, 33(2):393–420, 2016.

[54] Shuyuan Mary Ho, Jeffrey T Hancock, Cheryl Booth, Xiuwen Liu, Muye Liu, Shashank S Timmarajus, and Mike Burmester. Real or spiel? a decision tree approach for automated detection of deceptive language-action cues. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 3706–3715. IEEE, 2016.

[55] Jason Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[56] Jesper F Hopstaken, Dimitri Van Der Linden, Arnold B Bakker, and Michiel AJ Kompier. The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biological psychology*, 110:100–106, 2015.

[57] Hang Hu and Gang Wang. {End-to-End} measurements of email spoofing attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1095–1112, 2018.

[58] Mohammad S Jalali, Maike Bruckes, Daniel Westmattelmann, and Gerhard Schewe. Why employees (still) click on phishing links: investigation in hospitals. *Journal of medical Internet research*, 22(1):e16775, 2020.

[59] Johanna K Kaakinen and Jukka Hyönä. Task relevance induces momentary changes in the functional visual field during reading. *Psychological Science*, 25(2):626–632, 2014.

[60] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[61] Timothy Kelley, Mary J Amon, and Bennett I Bertenthal. Statistical models for predicting threat detection from human behavior. *Frontiers in psychology*, 9:466, 2018.

[62] Daejoong Kim and Jang Hyun Kim. Understanding persuasive elements in phishing e-mails: A categorical content and semantic network analysis. *Online Information Review*, 2013.

[63] Bennett Kleinberg, Yaloe Van Der Toolen, Aldert Vrij, Arnoud Arntz, and Bruno Verschuere. Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied cognitive psychology*, 32(3):354–366, 2018.

[64] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

[65] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[66] Michael C Kotson and Alexia Schulz. Characterizing phishing threats with natural language processing. In *2015 IEEE Conference on Communications and Network Security (CNS)*, pages 308–316. IEEE, 2015.

[67] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. Social engineering attacks on the knowledge worker. In *Proceedings of the 6th International Conference on Security of Information and Networks*, pages 28–35, 2013.

[68] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. Advanced social engineering attacks. *Journal of Information Security and applications*, 22:113–122, 2015.

[69] Ponnurangam Kumaraguru. *Phishguru: a system for educating users about semantic attacks*. Carnegie Mellon University, 2009.

[70] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[71] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 905–914, 2007.

[72] Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*, volume 3. McGraw-Hill New York, 2000.

[73] Patrick Lawson, Carl J Pearson, Aaron Crowson, and Christopher B Mayhorn. Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy. *Applied ergonomics*, 86:103084, 2020.

[74] Alan M Leslie, Ori Friedman, and Tim P German. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12):528–533, 2004.

[75] Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, 2018.

[76] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[77] Peter Alexander Lichtenberg, Michael A Sugarman, Daniel Paulson, Lisa J Ficker, and Annalise Rahman-Filipiak. Psychological and functional vulnerability predicts fraud cases in older adults: Results of a longitudinal study. *Clinical Gerontologist*, 39(1):48–63, 2016.

[78] Tian Lin, Daniel E Capecci, Donovan M Ellis, Harold A Rocha, Sandeep Dommaraju, Daniela S Oliveira, and Natalie C Ebner. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–28, 2019.

[79] Jaclyn Martin, Chad Dubé, and Michael D Coovert. Signal detection theory (sdt) is effective for modeling user behavior toward phishing and spear-phishing attacks. *Human factors*, 60(8):1179–1191, 2018.

[80] A Mbaziira and J Jones. A text-based deception detection model for cybercrime. In *Int. Conf. Technol. Manag*, 2016.

[81] John McAlaney and Peter J. Hills. Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking. *Frontiers in Psychology*, 11, 7 2020.

[82] Steven A McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M Wisner, and Xun Zhu. Information manipulation theory 2: A propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377, 2014.

[83] Dan Morrison and Cleotilde Gonzalez. Pyibl python implementation of ibl.

[84] Jose Nazario. Phishing corpus. 2016. `https://monkey.org/~jose/phishing/`.

[85] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[86] Thuy Ngoc Nguyen, Chase McDonald, and Cleotilde Gonzalez. Credit assignment: Challenges and opportunities in developing human-like ai agents. *arXiv preprint arXiv:2307.08171*, 2023.

[87] Thuy Ngoc Nguyen, Duy Nhat Phan, and Cleotilde Gonzalez. Speedyibl: A solution to the curse of exponential growth in instance-based learning models of decisions from experience. *arXiv preprint arXiv:2111.10268*, 2021.

[88] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis. In *2018 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–12. IEEE, 2018.

[89] Federal Bureau of Investigation. Internet crime report. `https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf`, 2020 [Online].

[90] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of*

the 2017 CHI Conference on Human Factors in Computing Systems, pages 6412–6424. ACM, 2017.

[91] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[92] Gilchan Park and Julia M Taylor. Using syntactic features for phishing detection. *arXiv preprint arXiv:1506.00037*, 2015.

[93] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, and Meredith Lillie. Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, 128:17–26, 2019.

[94] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. The design of phishing studies: Challenges for researchers. *Computers & Security*, 52:194–206, 2015.

[95] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[96] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[97] Kevin Pfeffel, Philipp Ulsamer, and Nicholas H Müller. Where the user does look when reading phishing mails-An eye-tracking study. Technical report.

[98] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[99] Proofpoint. The ponemon 2021 cost of phishing study, 2021 [Online].

[100] Prashanth Rajivan and Cleotilde Gonzalez. Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks. *Frontiers in psychology*, 9:135, 2018.

[101] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[102] Jennifer D. Ryan and Kelly Shen. The eyes are a window into memory, 4 2020.

[103] Said Salloum, Tarek Gaber, Sunil Vadera, and Khaled Shaalan. Phishing email detection using natural language processing techniques: a literature survey. *Procedia Computer Science*, 189:19–28, 2021.

[104] Idalmis Santiesteban, Sarah White, Jennifer Cook, Sam J Gilbert, Cecilia Heyes, and Geoffrey Bird. Training social cognition: from imitation to theory of mind. *Cognition*, 122(2):228–235, 2012.

[105] Dawn M Sarno and Mark B Neider. So many phish, so little time: Exploring email task factors and phishing susceptibility. *Human Factors*, 64(8):1379–1403, 2022.

[106] Ben D Sawyer and Peter A Hancock. Hacking the human: the prevalence paradox in cybersecurity. *Human factors*, 60(5):597–609, 2018.

[107] Marten Scheffer, Jordi Bascompte, Tone K Bjordam, Stephen R Carpenter, Laurie B Clarke, Carl Folke, Pablo Marquet, Nestor Mazzeo, Mariana Meerhoff, Osvaldo Sala, et al. Dual thinking for scientists. *Ecology and Society*, 20(2), 2015.

[108] Kim B Serota, Timothy R Levine, and Franklin J Boster. The prevalence of lying in america: Three studies of self-reported lies. *Human Communication Research*, 36(1):2–25, 2010.

[109] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373–382. ACM, 2010.

[110] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 88–99, 2007.

[111] V Shreeram, M Suban, P Shanthi, and K Manjula. Anti-phishing detection of phishing attacks using genetic algorithm. In *2010 International Conference on Communication Control and Computing Technologies*, pages 447–450. IEEE, 2010.

[112] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to detect phishing emails: Effects of the frequency of experienced phishing emails. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 453–457. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[113] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. What makes phishing emails hard for humans to detect? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 431–435. SAGE Publications Sage CA: Los Angeles, CA, 2020.

[114] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Cognitive elements of learning and discriminability in anti-phishing training. *Computers & Security*, 127:103105, 2023.

[115] Sterling Somers, K Mitsopoulos, Christian Lebiere, and Robert Thomson. Cognitive-level salience for explainable artificial intelligence. In *Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modeling*, 2019.

[116] Siegfried Ludwig Sporer and Barbara Schwandt. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 20(4):421–446, 2006.

[117] Stefan Strauß. Privacy analysis–privacy impact assessment. *The Ethics of Technology: Methods and Approaches*, pages 143–156, 2017.

[118] Leif A Strömwall and Rebecca M Willén. Inside criminal minds: Offenders' strategies when lying. *Journal of Investigative Psychology and Offender Profiling*, 8(3):271–281, 2011.

[119] Ronnie Taib, Kun Yu, Shlomo Berkovsky, Mark Wiggins, and Piers Bayl-Smith. Social engineering and organisational dependencies in phishing attacks. In *IFIP Conference on Human-Computer Interaction*, pages 564–584. Springer, 2019.

[120] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[121] Rucha Tembe, Olga Zielinska, Yuqi Liu, Kyung Wha Hong, Emerson Murphy-Hill, Chris Mayhorn, and Xi Ge. Phishing in international waters: Exploring cross-national differences in phishing conceptualizations between chinese, indian and american samples. In *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*, pages 1–7, 2014.

[122] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[123] P. Unchit, S. Das, A. Kim, and L. J. Camp. Quantifying susceptibility to spear phishing in a high school environment using signal detection theory. In Nathan Clarke and Steven Furnell, editors, *Human Aspects of Information Security and Assurance*, pages 109–120, Cham, 2020. Springer International Publishing.

[124] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[125] Verizon. 2021 data breach investigation, 2021 [Online].

[126] Rakesh Verma, Narasimha Shashidhar, and Nabil Hossain. Detecting phishing emails the natural language way. In *European Symposium on Research in Computer Security*, pages 824–841. Springer, 2012.

[127] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, 45(8):1146–1166, 2018.

[128] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011.

[129] Aldert Vrij, Maria Hartwig, and Pär Anders Granhag. Reading lies: Nonverbal communication and deception. *Annual review of psychology*, 70:295–317, 2019.

[130] Aldert Vrij, Samantha Mann, Susanne Kristen, and Ronald P Fisher. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499–518, 2007.

[131] Jeffrey J Walczyk, Laura L Harris, Terri K Duck, and Devyani Mulay. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, 34:22–36, 2014.

[132] Jeffrey J Walczyk, Karen S Roper, Eric Seemann, and Angela M Humphrey. Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7):755–774, 2003.

[133] Jingguo Wang, Tejaswini Herath, Rui Chen, Arun Vishwanath, and H. Raghav Rao. Research article phishing susceptibility: An investigation into the processing of a

targeted spear phishing email. *IEEE Transactions on Professional Communication*, 55(4):345–362, 2012.

[134] Jingguo Wang, Tejaswini Herath, Rui Chen, Arun Vishwanath, and H Raghav Rao. Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE transactions on professional communication*, 55(4):345–362, 2012.

[135] Rick Wash. How experts detect phishing scam emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[136] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–12, 2018.

[137] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[138] Emma J Williams and Danielle Polage. How persuasive is phishing email? the role of authentic design, influence and current events in email judgements. *Behaviour & Information Technology*, 38(2):184–197, 2019.

[139] Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38:49–56, 2021.

[140] Stacey A. Wood, Pi-Ju Liu, Yaniv Hanoch, and Sara Estevez-Cores. Importance of Numeracy as a Risk Factor for Elder Financial Exploitation in a Community Sample. *The Journals of Gerontology: Series B*, 71(6):978–986, 07 2015.

[141] Min Wu, Robert C Miller, and Simson L Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 601–610, 2006.

[142] Tianhao Xu and Prashanth Rajivan. Determining psycholinguistic features of deception in phishing messages. *Information & Computer Security*, 31(2):199–220, 2023.

[143] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. Spearsim: Design and evaluation of synthetic task environment for studies on spear phishing attacks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 1500–1504. SAGE Publications Sage CA: Los Angeles, CA, 2021.

[144] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. Modeling phishing decision using instance based learning and natural language processing. In *HICSS*, pages 1–10, 2022.

[145] Tianhao Xu, Kuldeep Singh, and Prashanth Rajivan. Personalized persuasion: Quantifying susceptibility to information exploitation in spear-phishing attacks. *Applied Ergonomics*, 108:103908, 2023.

[146] Huaping Yuan, Xu Chen, Yukun Li, Zbenguo Yang, and Wenyin Liu. Detecting phishing websites and targets based on urls and webpage links. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3669–3674. IEEE, 2018.

[147] Lina Zhou, Judee K Burgoon, Douglas P Twitchell, Tiantian Qin, and Jay F Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166, 2004.

[148] Olga A. Zielinska, Allaire K. Welk, Christopher B. Mayhorn, and Emerson Murphy-Hill. A temporal analysis of persuasion principles in phishing emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):765–769, 2016.

[149] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[150] AlMaha Abu Zuraiq and Mouhammd Alkasassbeh. Phishing detection approaches. In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pages 1–6. IEEE, 2019.

Appendix A

TABLE FOR EYE MOVEMENT METRICS RELATED TO FIXATIONS

Metric name	Description	Format
Total duration of whole fixations	The total duration of the fixations during an interval.	Milliseconds
Average duration of whole fixations	The average duration of the fixations during an interval.	Milliseconds
Number of whole fixations	The number of whole fixations occurring during an interval.	Count
Duration of first whole fixation	The duration of the first fixation during an interval.	Milliseconds
Average whole fixation pupil dial meter	The average pupil diameter of all whole fixation samples in this interval. Calculated using the resulting pupil diameter after applying pupil diameter filter.	Millimeters
Average whole fixation eye openness	The average eye openness of all whole fixation samples in this interval. Calculated using the resulting eye openness after applying eye openness filter.	Millimeters
Number of whole fixation starts	The number of whole fixations that starts during a bin.	Count

Average whole fixation pupil dia- meter	The average pupil diameter of all whole fixation samples in this bin. Calculated using the resulting pupil diameter after applying pupil diameter filter.	Millimeters
Average whole fixation eye openness	The average eye openness of all whole fixation samples in this bin. Calculated using the resulting eye openness after applying eye openness filter.	Millimeters

Appendix B

TABLE FOR EYE MOVEMENT METRICS RELATED TO SACCADES

Metric name	Description	Format
Number of saccades	The number of saccades occurring during an interval.	Count
Average peak velocity of saccades	The average peak velocity of all saccades in this interval.	Degrees/second
Minimum peak velocity of saccades	The peak velocity of the saccade with the lowest peak velocity in this interval.	Degrees/second
Maximum peak velocity of saccades	The peak velocity of the saccade with the highest peak velocity in this interval.	Degrees/second
Standard deviation of peak velocity of saccades	The standard deviation of all peak velocities of the saccades in this interval.	Degrees/second
Average amplitude of saccades	The average amplitude of all saccades in this interval.	Degrees
Minimum amplitude of saccades	The amplitude of the saccade with the lowest amplitude in this interval.	Degrees

Maximum amplitude of saccades	The amplitude of the saccade with the highest amplitude in this interval.	Degrees
Total amplitude of saccades	The total amplitude of all saccades in this interval.	Degrees
Time to first saccade	The time to the first saccade during an interval.	Milliseconds
Direction of first saccade	The direction of the first saccade in the interval.	Degrees
Peak velocity of first saccade	The peak velocity of the first saccade in the interval.	Degrees/second
Average velocity of first saccade	The average velocity of the first saccade in the interval.	Degrees/second
Amplitude of first saccade	The amplitude of the first saccade in the interval.	Degrees