Predicting the Threat: Investigating Insider Threat Psychological Factors With Advanced Natural Language Processing

Matthew Levy
U.S. Department of Defense
matthew.levy@niwc.navy.mil

Angela Horneman
Carnegie Mellon University
ahorneman@cert.org

Robert Dittmore
Carnegie Mellon University
rmdittmore@sei.cmu.edu

Craig Motell
U.S. Department of Defense
craig.e.motell.civ@us.navy.mil

Abstract

The term “insider threat” can take many forms, ranging from an information security risk to the threat of an active shooter. Accordingly, it is beneficial to researchers and practitioners to understand the relationship between psychological factors and the different types of threats an insider may pose to an organization. This research advances this understanding. Specifically, we investigate the three-way relationship between user-generated text, psychological factors espoused in insider threat literature, and risk indicator categories used by the U.S. Government. We employ advancements in machine learning and Natural Language Processing to investigate this relationship. Specifically, we use Bidirectional Encoder Representations from Transformers (BERT) for word embedding and vector space modeling. Our results indicate that there are indeed associations between established risk categories and the psychological factors seen as predictive of malicious insiders. Our exploratory research also reveals that further research is warranted to advance the predictive capability of insider threat modeling¹.

the Border Gateway Protocol (BGP) where attackers maliciously reroute Internet traffic can be quickly identified through response ecosystems such as the BGP Hijacking Observatory [2]. While not perfect, these response ecosystems amount to considerable efforts by multiple organizations and communities to strengthen the computing assets vulnerable to external threats.

On the other hand, insider threats require different types of response ecosystems. An insider is any person who has or had authorized access to or knowledge of an organization’s resources, including personnel, facilities, information, equipment, networks, and systems [3]. An insider threat constitutes the potential for an insider to use their authorized access or understanding of an organization to harm that organization [3]. Thus, insider threats originate from insiders seeking to use authorized access, wittingly or unwittingly, to harm information, technology, physical infrastructure, people, organizations, or nation-states through forms of espionage, terrorism, unauthorized disclosure of national security information, or through the loss or degradation of an organization’s resources or capabilities. To confront insider threats, research and practice must develop response ecosystems based on modeling the behavior of insiders to analyze potential insider threat risk.

1. Introduction

Organizations public and private face considerable struggle identifying and predicting where the next cybersecurity threat will originate. What makes this landscape more complicated is that organizations face cybersecurity threats from both inside and outside. External threats are constantly emerging and constantly becoming more sophisticated given the proclivities and resources of state-sponsored threat actors. But, at the same time, the many response ecosystems that exist in the threat intelligence community can quickly identify attack sources and the vulnerabilities attackers have exploited [1]. For example, a hijack of

Prior research on insider threats has sought to understand the behaviors of end-users when they interact with networks and systems and whether these indicators tell a story of the psychological characteristics seen as concomitant with the behaviors of known malicious insiders – but has stopped short of associating the proclivities of the insider with the type of potential threat. Indeed, prior research has taken significant strides in enumerating the psychological factors of possible malicious insiders, but, given the broad definition of insider threat [3], has refrained from associating them with insider threat types. Research in this area has taken significant strides in elucidating these factors through case studies [4], expert interviews [5, 6], and through comparisons between the psychological

¹ DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

profiles of the general population with known malicious insiders [7, 8].

In this article, we contribute to the body of knowledge on insider threat risk analysis (ITRA) [4, 5, 7] to better understand the behaviors of the potential malicious insider and advance the science of predicting the type of threat an insider may pose. Specifically, in using Natural Language Processing (NLP) to analyze data generated by an individual and data reported about an individual to discover relationships between these data, psychological factors, and known insider threat risk categories. Moreover, we advance the science of ITRA by moving beyond classical text mining techniques [7, 5, 9] and employ advancements in machine learning and NLP. Lastly, our research responds to calls made in ITRA research to explore additional datasets, methods to extract psychosocial features from those datasets, and methods that identify the specific type of insider threat [10].

Our paper proceeds as follows: Section 2 provides a background of ITRA research. Section 3 delineates the research model. Section 4 describes the machine learning and NLP methods used in our analysis. Section 5 provides results obtained from the analysis. Section 6 provides a discussion about our results, provides calls for further research in this area, and discusses the limitations of our research. Lastly, section 7 concludes the article.

2. Background

ITRA has its underpinnings in threat modeling. Research in this area is commonly viewed from an asset-based perspective – one that focuses on modeling the risk pathways in and out of physical or virtual computing assets [11]. The goal of research in this area is to identify and prioritize potential threats, and security mitigations to protect valued resources, such as confidential data or intellectual property [12]. The asset-based approach has its origins in research on attack trees to model external threats. In attack trees, the root node of the tree represents the goal of an attack, such as an asset that could be infiltrated, while the nodes represent the multitude of known attack vectors [13]. Related but modeled differently in practice are attack graphs – formal models used to illustrate the security vulnerabilities of a system and all possible sequences of exploits that an intruder can use to achieve a specific goal [14]. Using either the attack tree or attack graph approach, a systems administrator can model the computing environment they are responsible for and use one of these approaches to appropriate the right security measures for their computing environment.

Accordingly, penetration testers can use attack trees or graphs to model the computing environment they seek to infiltrate [15]. Both attack trees and attack graphs can be useful to model the possible vulnerabilities in a computing environment in protecting against insider threats, but, are couched in an asset-based perspective that lends itself to protecting against external threats.

Key Challenge Graphs (KCG) [16] are closely related to attack trees and attack graphs but were specifically developed to model insider threats. A KCG is also based on asset modeling but also incorporates human actors. In a KCG, the “keys” represent information sources, and the vertices can represent hosts or human actors. A KCG is essentially a tuple: $KCG = (V, E; K, V_0, V_s, \pi, \delta)$, where V is the set of vertices, E is the set of edges, V_0 is the initial set of compromised vertices, V_s is the set of target vertices, $\pi : V \rightarrow K$ is a function that assigns keys to vertices, $\delta : E \rightarrow K \times N \times N$ is a function that assigns key challenges and costs to edges, and N is the set of natural numbers. For example, $\pi(v_1) = k_0$ means that the key k_0 can be obtained at vertex v_1 , $\delta(e_1) = (a, c_1, c_2)$ which implies an assignment of a key challenge to edge e_1 , which requires an attacker to produce the key a . If he cannot do so, then he incurs a cost c_1 , which could be significant depending on the access control mechanism; otherwise, he incurs a smaller cost c_2 [16]. The point in elucidating the research on threat modeling is that, through formal modeling, research has sought extensively to understand potential vulnerabilities, internal and external, from an asset-based perspective. And indeed, approaches such as these are widely seen in the defense and aerospace industries for their ability to model potential threats against tamper-resistant electronics systems such as avionics on military aircraft and the computer control systems relating to the electric power grid [17].

Complementing the asset-based perspective is the use of behavioral modeling to analyze user activity and insider risk. Seminal research in this area includes the insider threat critical pathway explicated in [18, 19]. In conducting archival analyses of known malicious insiders, [18, 19] elucidate how individual characteristics, personal and professional stressors, and interactions with others in the workplace could increase the risk of attack. Shaw’s critical pathway discusses five interrelated components: (a) An occurrence of significant personal or professional stressors within six months of an attack, (b) a maladaptive behavioral reaction to the stressor hypothesized to be, in part, a result of a pre-existing vulnerability to frustration, underlying anger at authority, or poor social judgment or skills, (c) an emotional reaction to the stressor, (d) behavioral or emotional reactions in the workplace

that were sufficient to gain official attention in the forms of disciplinary action, counseling, or other forms of intervention, and (e) a managerial intervention that was insufficient to divert the subject from the destructive pathway and may have even escalated the process. Accordingly, research commissioned by the Defense Personnel Security Research Center (DPSRC) [4] builds on [18, 19] to elucidate 10 cases of insider abuse involving offenders working in critical sectors of national infrastructure. Each case study discusses the offender's background, the organizational environment in which the offense took place, the details of each event, the offender's presumed motivations, the final legal and investigative actions that resulted from the attack, and the lessons learned from each incident having implications for corporate policy or national security. The cases detailed the events that led to the discovery of abuse. They also elucidated the efficacy of the critical pathway method and the psychological states of the abusers leading up to the events. These studies laid the groundwork for further psychosocial-behavioral modeling for ITRA [6, 5, 8, 7].

The addition of psychosocial-behavioral modeling to ITRA adds a predictive component. In employing a combination of data generated by the individual such as email, text, chat, web-search, or event logs, as well as data generated about an individual such as documented encounters with law enforcement, credit scores, or documented psychological conditions, it is possible to develop psychological profiles of users and whether they may pose a threat. For example, [20, 21] use log data to create user-based and role-based profiles and then employ Principal Components Analysis (PCA) to identify features that deviate from normalized thresholds. In experimenting with mock scenarios [22] show that when insiders are engaged to commit malicious acts, their language changes concerning work motivation and their attitudes toward coworkers and the organization. Furthering this idea is research that has elucidated several frameworks on ITRA psychological factors. [8, 7] delineate several personality disorders and mental health disorders found in Axis I and Axis II of the Diagnostic Statistical Manual (DSM) [23] such as avoidant personality disorders and disruptive mood dysregulations that can manifest in a malicious insider. [6, 5, 24, 25] developed the Sociotechnical and Organizational Factors for Insider Threat Ontology (SOFIT), that delineates Psychological Factors as Dynamic States (attitude and mental states) and Enduring Traits (personality dimensions less likely to change over time). The research discussed above has made significant strides in bringing the analysis of psychological factors into the fold of ITRA. More

specifically, in explicating that it is empirically valid to measure psychological characteristics using computer data from an end-user.

Missing from the above research is the validation and advancement of established psychological factors using ground truth operational data from an actual insider threat organization. Indeed, ITRA research must advance beyond the delineation of psychological factors using only public data about the most extreme malicious insiders [8, 7], and must move beyond research that delineates psychological factors using only the opinions of expert judges [6, 5, 24]. Moreover, ITRA research streams must now advance beyond user-profiling using Linguistic Inquiry and Word Count models [21, 20, 5] to NLP methods that can get at underlying meaning. In particular, to supply a more grounded psychological component to the asset-based perspective:

Figure 1. The Insider Threat Prediction Model [26]

Our research brings additional context, validation, and depth to the use of psychological factors for ITRA in the following ways: (1) We validate and advance the SOFIT Psychological Factors branch and Enduring Traits sub-branch that relies heavily on the constructs of Narcissism, Machiavellianism, and Psychopathy that comprise the Dark Triad [27, 28]. (2) We do this using data from an actual insider threat organization and use NLP to derive underlying meaning and derive correlation metrics that reflect the relationship between underlying meaning and target constructs representative of insider threat psychological factors.

3. Research Model

To advance the science of ITRA we expand upon the use of the Dark Triad found in the Psychological Factors/Enduring Traits branch of SOFIT. In appropriation of the Dark Triad, the individual and the artifacts they generate through computer systems use is the unit of analysis. The paragraphs below describe two contrasting triads: The Dark Triad and

Light Triad. We included the prosocial Light Triad to provide a more robust and explainable model beyond the traditional highlighting of negative traits. Indeed, research on the Dark Triad has discussed the importance of integrating malevolent and beneficent sides of human nature [29], and more generalizable research on prosocial frameworks discusses the interactivity between malevolent traits and beneficent traits that seem particularly salient for ITRA. We do not theorize that Dark Triad traits and Light Triad traits rest on opposite sides of a continuum, as certain prosocial traits may illuminate based on specific types of affordances [30]. Instead, we theorize that Light Triad traits could serve as a moderating factor to Dark Triad traits, and thus, we include it to contemplate these effects.

The Dark Triad [31, 27, 32, 33] constitutes the first dimension of our model. The Dark Triad addresses the shortcomings of the Big Five, which has been looked upon favorably in ITRA research [5, 6, 34], but does not account for many of the other behavioral indicators that are difficult to measure from user-generated data in ITRA.

The Light Triad [29] constitutes the second dimension of our model. When incorporated under the auspices of ITRA, the Light Triad helps theorize about the moderating effects of prosocial speech acts. In particular, about a person's connectedness or disconnectedness with humanity and how it may play a role in their proclivities to commit a malicious insider attack. Moreover, low or negative metrics on the components of the Light Triad (Kantianism, Humanism, Faith in Humanity) can provide indications about social-isolationism, avoidant personality, and other anti-social characteristics that have been explicated in insider threat research but are not easily measurable from data generated by an individual.

The Potential Risk Indicators (PRI) delineated by the U.S. Center for Development of Security Excellence (CDSE) [35] constitutes the third dimension of our model. All data provided to us from the insider threat organization is labeled with CDSE categories, and thus, represents a ground truth for analysis and cyber-analysts in the insider threat organization. CDSE has eight major PRI categories: (1) Professional and Lifecycle Performance, (2) Foreign Considerations, (3) Security and Compliance Incidents, (4) Technical Activity, (5) Criminal, Violent or Abusive Conduct, (6) Financial Considerations, (7) Substance Abuse and Addictive Behaviors, and (8) Judgement, Character, and Psychological Conditions. Each classification also has subcategories for a total of 162 different indicators. There is also a metadata category that pertains to access attributes that describe a person's access to sensitive

information.

The sections below describe the three dimensions of the Dark Triad: Machiavellianism, Narcissism, and Psychopathy, and the three dimensions of the Light Triad: Faith in Humanity, Humanism, and Kantianism.

3.1. Machiavellianism

Machiavellianism is considered antecedent to the behavior of malicious insiders [27]. Research characterizes Machiavellian behaviors by (a) a shallow or cold affect observed in interpersonal situations, (b) the manipulation and exploitation of others for self-beneficial goals, (c) the explicit negativity towards other people, social-consensual beliefs and institutions, and (d) the world in general. In turn, Machiavellian desires can be described as generally agentic, showing cold selfishness and instrumentality in pursuing goals such as status, money, power, and competitiveness. In contrast, relatively low priority is afforded to goals such as harmony, love, and family [36].

3.2. Narcissism

Liang and colleagues [7] enlist narcissistic personality disorder as a principal marker of malicious insiders. Research characterizes narcissism by (a) grandiosity, (b) a sense of entitlement, and (c) a belief that you have unparalleled skills or talent. Narcissism is closely associated with the fetishization of power, success, and attractiveness. It is also associated with a perception of the self as deserving special or preferential treatment. According to Liang and colleagues [7] narcissism manifests in textual sources as an insatiable craving for admiration and special attention that cannot be met where someone might seek validation and affirmation of their self-importance from other sources such as competitors or opponents [7].

3.3. Psychopathy

Numerous sources have demonstrated the association between psychopathy and computer criminal behavior [37, 28]. Research characterizes psychopathy as the inclusion of high impulsivity and thrill-seeking behavior along with low empathy and low anxiety. Three qualities delineate psychopathy: (a) an arrogant and deceitful interpersonal style, (b) a deficient affective experience, and (c) impulsive and irresponsible behavior [28].

3.4. Faith in Humanity

Faith in Humanity reflects a tendency to believe in the fundamental goodness of people. Research characterizes Faith in Humanity by individuals who score positively on this factor as (a) those that see the best in people, (b) those that believe people will see the best in them, and (c) inherently trust they will be dealt with fairly. Individuals who also score highly are generally those that believe most people are good in a moral sense, and as a result, are quick to forgive others who may have hurt them [29, 38].

3.5. Humanism

Humanism refers to the valuing of the dignity and worth of each individual. Research characterizes Humanism by individuals who (a) tend to admire others, and (b) are quick to applaud the success of other people. In addition, individuals who score highly on Humanism tend to value the treatment of others and enjoy listening to people from other walks of life [38].

3.6. Kantianism

Kantianism refers to Immanuel Kant's second formulation of his categorical imperative, "Act in such a way that you treat humanity, whether in your person or in the person of any other, never merely as a means to an end, but always at the same time as an end" [39]. Research characterizes individuals who score high on Kantianism as (a) those tending to prefer honesty over charm and (b) those who feel uncomfortable manipulating people to get what they want. In addition, individuals who score highly on Kantianism typically do not think about what they want from someone when talking to them and would risk damaging their reputation as a means to preserve authenticity.

4. Method

The methods employed in the current study include four overarching steps: First, using the lexicon from the definitions of the psychological factors discussed in Section 3, we constructed target clusters of words and n-grams based on the underlying meaning of the construct. Second, we performed a series of data preprocessing steps on both the target clusters and our data to ready it for use in a machine learning algorithm. Third, we performed feature engineering by embedding both the target clusters and data corpus into the same vector space. Fourth, we built our models and evaluated them using common data science metrics.

Figure 2. Model Development Workflow.

4.1. Data Collection and Sampling Criteria

Data collection took place at an insider threat organization within the U.S. Government created in 2015. The mission of this organization is to predict insider threats so that it does not have to react to a malicious insider. The data discussed in this research essay is actual data from employee background investigations – not from existing corpora such as the Enron corpus or CERT insider incident corpus. The researchers act as behavioral scientists in this program and were granted access to publish an analysis of these data by the insider threat's parent organization.

The data collected is in two general forms: event-driven and non-event-driven. Event-driven data consists of time-series-based alerts about user activity, such as a user inserting a non-authorized USB drive into their work computer at a time considered to be outside the scope of typical work hours. In contrast, non-event-driven data consists of data sources generated by the individual, such as email correspondence, chat correspondence, system event logs, or web search history, and data sources generated about an individual, such as background checks, credit reports, or periodic employee evaluation reports. Lastly, there is metadata about an employee, such as their access level, current work projects, and rank.

For the analysis in this research essay, we focus exclusively on continuous employee evaluation reports. We analyzed the data from approximately 72,000 employee evaluations taken from a much larger corpus in 2020. We estimate the population corpus to be around 200,000. We sampled evaluations from this larger corpus to focus on a subset of categories where we had enough labeled records. Each record contains information about rank, information access level (Unclassified, Secret, Top Secret), organizational association, and a text field that has the body of the report. The reports themselves are alert-based. Thus, they constitute reports of a person's financial considerations if it is not stable, documented criminal conduct that is violent or abusive, or documented substance abuse or addictive behaviors. Each report is labeled by one or more of these categories, which are equivalent to CDSE categories [35]. However, our data source only contained three categories from CDSE (Criminal, Violent, or Abusive Conduct, Financial

Considerations, and Substance Abuse and Addictive Behaviors).

The following is an anonymized text segment. The employee evaluation record containing this segment has the following labels: "Criminal, Violent, or Abusive Conduct" and "Substance Abuse and Addictive Behaviors."

"Subject was charged in the State of **** with the following: Offense Date: 0*/2*/20** Arrest Date: N/A Arresting Agency: N/A Charge at time of Arrest: N/A Name of Court: DISTRICT COURT FOR **** COUNTY 1. Charge at Disposition: CDS: POSS W/I DIST: NARC Disposition on Disposition Date: N/A on N/A 2. Charge at Disposition: FIREARM/DRUG TRAF CRIME Disposition on Disposition Date: N/A on N/A 3. Charge at Disposition: CDS: DISTR ETC. W/FIREARM Disposition on Disposition Date: N/A on N/A 4. Charge at Disposition: HANDGUN IN VEHICLE Disposition on Disposition Date: N/A on N/A 5. Charge at Disposition: LOADED HANDGUN IN VEHICLE Disposition on Disposition Date: N/A on N/A 6. Charge at Disposition: CDS: POSSESS-NOT MARIJUANA Disposition on Disposition Date: N/A on N/A 7. Charge at Disposition: CDS: POSS MARIJUANA 10 GM+ Disposition on Disposition Date: N/A on N/A"

4.2. Target Variable Construction

We constructed target variables as clusters of words and n-grams representative of the psychological factors in the Dark and Light Triads. We then augmented the words and n-grams using the gensim open-source library for unsupervised topic modeling [40]. Lastly, we used Bidirectional Encoder Representations from Transformers (BERT) [41] for word embedding (to get at the underlying meaning of gensim-augmented segments) and vectorization.

We constructed the initial target clusters manually using open coding [42] and researcher interaction to arrive at a list of agreed-upon words and n-grams representative of each psychological factor. Following this, we augmented the initial target clusters using a gensim function that finds the top-N most similar words or n-grams and manually analyzed the output using word clouds to ensure the target clusters were still representative of the underlying meaning of the psychological factor. Additionally, we did a manual investigation to ensure we avoided polysemy – the coexistence of many possible meanings for a word or phrase.

The gensim library and gensim-data corpora were instrumental in developing the target clusters

representative of psychological factors. Gensim can return the most similar words for any word in their vocabulary based on words or n-grams given as input, thus providing context interpretation of a word or n-gram. For example, the command: `nlp.most_similar(["obama"], topn = 3)` will return the words "barack", "mccain", and "bush" with similarity scores. This functionality also helped to ensure distinctiveness. Given the psychological factors in the Dark and Light Triad, we knew there would be overlap, but we needed to ensure elements in the target clusters would at least be distinctive enough to analyze similarity scores between vectorized features and vectorized target clusters – this was where t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction was instrumental. t-SNE is a technique for dimensionality reduction that is particularly suited for high dimensional datasets. In contrast to Principal Components Analysis, t-SNE is probabilistic – where given a set of N high-dimensional objects $x_1 \dots x_N$ t-SNE first computes probabilities p_{ij} proportional to the similarity of objects x_i and x_j as follows:

$$p(j | i) = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Last, we used BERT for embedding and vectorization of both target clusters and our feature variable. We chose BERT because of its advancements in the ability to provide context. Context-free models such as word2vec or GloVe generate single word embedding representations for each word in the vocabulary. In contrast, BERT accounts for the context of each occurrence of a word. For example, where the vector for "running" will have the same word2vec vector representation for both occurrences in the sentences "he is running a company" and "he is running a marathon", BERT will provide a contextualized embedding that will be different according to the sentence. BERT is pre-trained using a plain text corpus and fine-tuned using the pre-trained model as the basis of a new purpose-specific model. In our case, we used `bert – base – uncased` as the pre-trained model and applied BERT tokenization and embedding to our target variables and feature text.

4.3. Data Preprocessing

The next step was to preprocess our feature variable – the body text of the employee evaluations. We preprocessed our data using the following sequence: (1) We cleaned the feature variable by converting to lowercase, removing punctuation, removing punctuation characters, and (2) applying a lemmatization algorithm, which groups together the

inflected forms of a word as a single item in an attempt to determine the lemma of a word based on its intended meaning.

4.4. Feature Engineering

The next step was to perform feature engineering. In our case, we are using a single feature – the body text of the employee evaluation – as the single feature variable. As mentioned, we had already preprocessed the text. Thus, the next step was to perform tokenization and embedding on the feature variable to place it in the same vector space as the target clusters. After using BERT for tokenization and embedding we summarized entries based on labels into a mean vector. Likewise, we did the same with the target clusters, where each target is identified by the list of words from gensim-data which helps BERT understand the context within the cluster to generate vectors. In our code, this was actuated using a cluster mean vector as our dictionary label.

4.5. Model Building and Testing

The principal measure we used to test the association between the feature variable and the target clusters representing psychological factors was cosine similarity. Cosine similarity looks at the cosine of the angle between two vectors and returns a $n \times m$ vector where n equals the number of data points we have in the dataset, and m equals the number of target clusters.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A^i B^i}{\sqrt{\sum_{i=0}^n A_i^2} \sqrt{\sum_{i=0}^n B_i^2}}$$

We also calculated the evaluation metrics consonant with multilabel classification. These metrics include Exact Match Ratio (EMR), Hamming Loss (HL), Recall (Rec), Precision (Prec), F1-Score, and 0/1 Loss. This was performed to test the strength of our model to be able to classify the labeled feature variable against the CDSE categories. While cosine similarity was the measure used to investigate the association between the feature variable and target clusters, we still needed to know the association between the feature variable and CDSE categories was strong in order to analyze the possible links between feature variable, target clusters, and the type of threat an individual may pose.

5. Results

Our analysis of results proceeded as follows: (1) We trained a model for multilabel text classification

to evaluate the inferential link between the CDSE categories and the feature variable. This step establishes the linkage between data from the insider threat organization and the type of insider threat risk an individual may pose. (2) We analyzed the relationship between the target clusters, representative of psychological factors from the Dark and Light Triads, and the feature variable using cosine similarity. This step establishes the linkage between data from the insider threat organization and the psychological factors.

For the first step, we trained and evaluated a multilabel classification model. Multilabel, as opposed to multiclass, applies to our data since employee evaluations can be labeled as more than one CDSE category. For example, an employee evaluation that describes a reported DUI will typically be assigned Category 6 (Criminal, Violent, or Abusive Conduct) and Category 8 (Substance Abuse and Addictive Behaviors). Considering we are using multilabel classification model training as criteria to evaluate linkage and not necessarily to employ a model for large-scale inferencing, we used One vs. Rest (OvR) as our multilabel classification technique. The metrics from model training indicate strong support for the link between CDSE categories and our feature variable. We expected this to be extremely high given that the nomenclature used to write evaluation reports is supposed to match up with CDSE categories for the investigative organization that formalizes them. Nevertheless, we report these statistics to illustrate the strong linkage between evaluation reports and CDSE categories:

Accuracy, the ratio of correctly predicted observations to total observations, for each of the three CDSE categories we analyzed were: Criminal (0.959444), Financial (0.938889), Substance Abuse (0.910045). The Exact-Match Ratio ($EMR = 0.902$) indicates the percentage of samples where all labels are classified correctly. Conversely, Hamming Loss ($HL = 0.018$) measures the fraction of labels incorrectly predicted to the total number of labels. Recall and Precision were also significant. To calculate recall and precision on a multilabel classification model, we used micro-averaging. Using micro-averaging, all true positives, true negatives, false positives, and false negatives for each class were summed and averaged ($Rec = 0.977, Prec = 0.939$). In addition, the $F1 - Score = 0.951$, and $0/1Loss = 0.097$. The following figure illustrates the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) for each of the three categories:

In sum, model training to evaluate the linkage between CDSE categories and our feature variable was

Figure 3. AUC-ROC Curve

strong, indicating support to investigate the link between Dark and Light Triad psychological factors and the categories representing the type of potential insider threat.

For the second step, we investigated how well text from our feature variable activated the target clusters. We used cosine similarity to perform this evaluation. In the application of machine learning, cosine similarity is a measure based on the cosine of the angle between two non-zero vectors, which equals the inner product of the same vectors normalized to both have length 1. Cosine similarity was utilized in two ways: between the feature variable and the CDSE categories (as a sanity check) and between the feature variable and the target clusters. In both cases, we used the mean vectors.

When we computed the cosine similarity between the feature variable and the CDSE categories, the results (again) indicated a strong relationship. In contrast, the relationship between the feature variable and the target clusters was less strong but strong enough to indicate future research should consider investigating the applicability of developing measurements of psychological factors based on how well user-generated feature variables associate with representative target clusters. For example, there was a clear indication that the mean vector on our feature variable that had a positive label for Category 6 (Criminal, Violent, or Abusive Conduct) also had a higher probability score on the target cluster for psychopathy than other target clusters $\cos(\theta) = 0.2833$ which contains embedded terms such as "impulsive", "hyperactive", "loutish", and "destructive." There was also a smaller similarity between feature variables with positive labels for Category 7 (Financial Considerations) and the target cluster for narcissism $\cos(\theta) = 0.1750$, which contains embedded terms such as "materialism", "selfishness", and "obsession". We should also note the lower-to-negative scores on factors found in the Light Triad. This was more significantly seen in the low cosine

similarities between Kantianism and the mean vectors whose primary labels were Category 6 $\cos(\theta) = 0.017$ and Humanism $\cos(\theta) = 0.073$. Further investigation is warranted to investigate how the Light Triad would perform with more diverse data sources.

To summarize, our model for multilabel text classification to evaluate the link between CDSE categories and the text from the employee evaluation body provided a stable platform to investigate the linkage between our feature variable and target clusters representative of psychological factors in the Dark and Light Triads, and consequently, the link between psychological factors and the type of insider threat someone may pose.

6. Discussion

As previously discussed, the range of what constitutes an insider threat ranges from an insider who is a risk concerning information leakage to someone who is a risk of verbal or physical violence in the workplace. Also previously discussed is that extant research on the psychological factors that serve as potential risk indicators of an insider threat simply points out the psychological factors that may lead to the general notion of an insider threat while eschewing any attribution to the type of threat they may pose. We contend research in this area needs to move towards increased specificity – on the category of threat, an insider may pose – as logically, the psychological factors leading to violent acts in the workplace are likely far different from the psychological factors that may lead to someone compromising sensitive information.

Our research seeks to advance this notion of specificity. In particular, our data, labeled by CDSE categories, exhibits variance across target clusters representative of psychological factors. Thus, further research is warranted to understand the psychological factors that can indicate specific types of insider threats. In addition, we have also shown that publicly available NLP can be applied to simulate psychological factors as target clusters. This finding potentially represents a discovery as it relates to the science of psychometric measurement.

Future research must refine target clusters and expand the variety of data sources to include email, chat, web search records, and system activity logs. Future research can also develop industry standards on behaviorally-oriented ITRA risk scoring. For example, risk score classes can have multiple dimensions: the postulations about user behavior, the assets they access, the computing environment in which a user resides, and their access level. Thus, we consider the idea of future

research that investigates and validates the appropriate methods to weigh these dimensions.

Lastly, underpinning the idea of research in this area are ethical considerations. Indeed the role of "the machine" and how it acts as a proxy for human behavior should not be underestimated. Moreover, if technology manifests from this research area, careful thought must be given to the model training of such technology and continuous analysis of its performance. Does the model still measure according to its original intent? Does its use inextricably lead to bias based on protected categories or other group associations? Not only do these considerations apply to this research, but the myriad applications of machine learning in practitioner settings.

6.1. Limitations

This article represents the results of a research stream that seeks to validate the prospectus of psychological factors in ITRA, and in a practical sense, discover insider threat risk before it becomes a malicious insider act. Naturally, nascent research in this regard has limitations to be elucidated.

The first limitation has to do with data derived from a continuous employee evaluation process. While this data carries the benefit of already having a CDSE category assigned to it, it represents reported events and not organic discourse. Accordingly, the reported results are associative, not causative, and thus, associations with other data sources could be investigated. In addition, while we engaged in a rigorous selection of words and n-grams to develop target clusters, the data making up these target clusters was a subjective process based on what we interpreted as the words and phraseology that captured the underlying meaning of the psychological factors in the Dark and Light Triads. Subsequent research should validate our selections, and possibly, use alternative machine learning textual analysis methods to extract meaning from multiple authoritative sources to build alternative target clusters. To aid in this process, we are seeking approval to publish the dataset and Jupyter Notebook from this study. These artifacts have the techniques we used for multilabel classification, the techniques we used for BERT fine-tuning, and the process for target variable construction. Lastly, we chose BERT for our natural language processing tasks and did not evaluate other models. Future research could evaluate the efficacy of BERT against alternative NLP-oriented algorithms such as GPT-3, word2vec, Elmo, or GloVe.

7. Conclusion

This article contributes to the body of knowledge in ITRA by applying testing, attribution, and advanced techniques on what is considered the spectrum of psychological factors associated with insider threats. We advance this research in this area by analyzing the link between risk indicator categories and insider threat-associated psychological factors. Further, we used machine learning and NLP – techniques not previously used in this research area. Lastly, we discuss trajectories for future research and the inherent limitations.

Across the spectrum of applied psychology, computer science, information science, and information systems, we contend a paucity of research exists that analyzes the psychological factors of insider threats. We further contend research in this area offers significant value to both researchers and practitioners. Thus, we hope our research engenders new forms of argumentation in this area. Specifically, those that advance academically-oriented knowledge contributions, and practitioner applications that hedge an organization's exposure to malicious insiders.

References

- [1] B. J. Bakis and E. D. Wang, "Building a national cyber information-sharing ecosystem," tech. rep., Mitre Corporation, McLean VA, 2017.
- [2] P. Sermpezis, V. Kotronis, P. Gigis, X. Dimitropoulos, D. Cicalese, A. King, and A. Dainotti, "Artemis: Neutralizing bgp hijacking within a minute," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2471–2486, 2018.
- [3] CISA, "Defining insider threats," tech. rep., 2021.
- [4] E. D. Shaw and L. F. Fischer, "Ten tales of betrayal: The threat to corporate infrastructure by information technology insiders analysis and observations," tech. rep., Defense Personnel Security Research Center, Monterey, CA., 2005.
- [5] F. L. Greitzer, L. J. Kangas, C. F. Noonan, C. R. Brown, and T. Ferryman, "Psychosocial modeling of insider threat risk based on behavioral and word use analysis," *e-Service Journal: A Journal of Electronic Services in the Public and Private Sectors*, vol. 9, no. 1, pp. 106–138, 2013.
- [6] F. L. Greitzer and D. A. Frincke, "Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation," in *Insider threats in cyber security*, pp. 85–113, Springer, 2010.
- [7] N. Liang, D. P. Biros, and A. Luse, "An empirical validation of malicious insider characteristics," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 361–392, 2016.
- [8] N. Liang and D. Biros, "Identifying common characteristics of malicious insiders," 2015.

- [9] S. Debortoli, O. Müller, I. Junglas, and J. vom Brocke, "Text mining for information systems researchers: An annotated topic modeling tutorial," *Communications of the Association for Information Systems*, vol. 39, no. 1, p. 7, 2016.
- [10] L. L. Ko, D. M. Divakaran, Y. S. Liau, and V. L. Thing, "Insider threat detection and its future directions," *International Journal of Security and Networks*, vol. 12, no. 3, pp. 168–187, 2017.
- [11] A. Shostack, *Threat modeling: Designing for security*. John Wiley & Sons, 2014.
- [12] W. Xiong and R. Lagerström, "Threat modeling—a systematic literature review," *Computers & security*, vol. 84, pp. 53–69, 2019.
- [13] B. Schneier, "Attack trees," *Dr. Dobbs's journal*, vol. 24, no. 12, pp. 21–29, 1999.
- [14] O. Sheyner, J. Haines, S. Jha, R. Lippmann, and J. M. Wing, "Automated generation and analysis of attack graphs," in *Proceedings 2002 IEEE Symposium on Security and Privacy*, pp. 273–284, IEEE, 2002.
- [15] K. Ferguson-Walter, S. Fugate, J. Mauger, and M. Major, "Game theory for adaptive defensive cyber deception," in *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*, pp. 1–8, 2019.
- [16] R. Chinchani, A. Iyer, H. Q. Ngo, and S. Upadhyaya, "Towards a theory of insider threat assessment," in *2005 International Conference on Dependable Systems and Networks (DSN'05)*, pp. 108–117, IEEE, 2005.
- [17] C.-W. Ten, C.-C. Liu, and M. Govindarasu, "Vulnerability assessment of cybersecurity for scada systems using attack trees," in *2007 IEEE Power Engineering Society General Meeting*, pp. 1–8, IEEE, 2007.
- [18] E. D. Shaw, K. G. Ruby, and J. M. Post, "The insider threat to information systems," *Security Awareness Bulletin*, vol. 2, no. 98, pp. 1–10, 1998.
- [19] E. D. Shaw, "The role of behavioral research and profiling in malicious cyber insider investigations," *Digital investigation*, vol. 3, no. 1, pp. 20–31, 2006.
- [20] P. A. Legg, O. Buckley, M. Goldsmith, and S. Creese, "Automated insider threat detection system using user and role-based profile assessment," *IEEE Systems Journal*, vol. 11, no. 2, pp. 503–512, 2015.
- [21] P. A. Legg, "Human-machine decision support systems for insider threat detection," in *Data Analytics and Decision Support for Cybersecurity*, pp. 33–53, Springer, 2017.
- [22] P. J. Taylor, C. J. Dando, T. C. Ormerod, L. J. Ball, M. C. Jenkins, A. Sandham, and T. Menacere, "Detecting insider threats through language change," *Law and human behavior*, vol. 37, no. 4, p. 267, 2013.
- [23] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, 2013.
- [24] F. Greitzer, J. Purl, Y. M. Leong, and D. S. Becker, "Sofit: Sociotechnical and organizational factors for insider threat," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 197–206, IEEE, 2018.
- [25] F. L. Greitzer, J. Purl, P. J. Sticha, M. C. Yu, and J. Lee, "Use of expert judgments to inform bayesian models of insider threat risk,"
- [26] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and D. Gritzalis, "An insider threat prediction model," in *International conference on trust, privacy and security in digital business*, pp. 26–37, Springer, 2010.
- [27] M. Maasberg, C. Van Slyke, S. Ellis, and N. Beebe, "The dark triad and insider threats in cyber security," *Communications of the ACM*, vol. 63, no. 12, pp. 64–80, 2020.
- [28] M. Maasberg, J. Warren, and N. L. Beebe, "The dark side of the insider: detecting the insider threat through examination of dark triad personality traits," in *2015 48th Hawaii International Conference on System Sciences*, pp. 3518–3526, IEEE, 2015.
- [29] S. B. Kaufman, D. B. Yaden, E. Hyde, and E. Tsukayama, "The light vs. dark triad of personality: Contrasting two very different profiles of human nature," *Frontiers in psychology*, vol. 10, p. 467, 2019.
- [30] I. Thielmann, G. Spadaro, and D. Balliet, "Personality and prosocial behavior: A theoretical framework and meta-analysis," *Psychological Bulletin*, vol. 146, no. 1, p. 30, 2020.
- [31] K. Padayachee, "Joint effects of neutralisation techniques and the dark triad of personality traits on gender: An insider threat perspective," in *2021 Conference on Information Communications Technology and Society (ICTAS)*, pp. 40–45, IEEE, 2021.
- [32] A. Furnham, S. C. Richards, and D. L. Paulhus, "The dark triad of personality: A 10 year review," *Social and personality psychology compass*, vol. 7, no. 3, pp. 199–216, 2013.
- [33] P. A. Legg, N. Moffat, J. R. Nurse, J. Happa, I. Agraftiotis, M. Goldsmith, and S. Creese, "Towards a conceptual model and reasoning structure for insider threat detection," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 4, no. 4, pp. 20–37, 2013.
- [34] M. Warkentin, M. McBride, L. Carter, and A. Johnston, "The role of individual characteristics on insider abuse intentions," 2012.
- [35] CDSE, "Potential risk indicators: Insider threat - what is the "something" we should be looking for?," 2020.
- [36] D. N. Jones and D. L. Paulhus, "Machiavellianism," 2009.
- [37] K. C. Seigfried-Spellar, N. Villacić-Vukadinović, and D. R. Lynam, "Computer criminal behavior is related to psychopathy and other antisocial behavior," *Journal of Criminal Justice*, vol. 51, pp. 67–73, 2017.
- [38] R. Gerymski and D. Krok, "Psychometric properties and validation of the polish adaptation of the light triad scale," *Current Issues in Personality Psychology*, vol. 7, no. 4, pp. 341–354, 2019.
- [39] I. Kant, *Groundwork for the Metaphysics of Morals*. Yale University Press, 2008.
- [40] R. Řehůřek and P. Sojka, Sep 2021.
- [41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [42] J. Corbin and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.