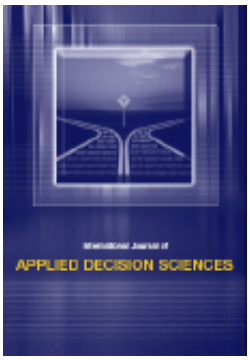


International Journal of Applied Decision Sciences

ISSN online: 1755-8085 - ISSN print: 1755-8077
https://www.inderscience.com/ijads

Human and machine partnership: natural language processing of army insider threat hub data

Saleem Ali, Hayden Deverill, Joseph Lindquist, Jonathan Roginski

DOI: 10.1504/IJADS.2025.10071558

Article History:

Received:	26 September 2024
Last revised:	04 April 2025
Accepted:	07 April 2025
Published online:	04 June 2025

Human and machine partnership: natural language processing of army insider threat hub data

Saleem Ali

School of Engineering,
Brown University,
Providence, RI 02912, USA
Email: saleem.ali@brown.edu

Hayden Deverill, Joseph Lindquist* and Jonathan Roginski

Department of Mathematical Sciences,
United States Military Academy,
West Point, NY 10996, USA
Email: hayden.deverill@westpoint.edu
Email: joseph.lindquist@westpoint.edu
Email: jonathan.roginski@westpoint.edu
*Corresponding author

Abstract: Threats to organisational efficacy and wellness may come from competitors (external threat) or from trusted agents (insider threat). Countering the insider threat is an imperative for the security of governments, the military, businesses, and all other organisations and institutions that employ people. This paper presents a case prioritisation system that utilises a deep learning classification model trained on expert evaluated insider threat cases to label cases as 'negligible', 'low', 'medium', or 'high' threat level. This classification model enables a partnership between machine and human that focuses human effort for the greatest impact. To evaluate the models created, the authors created a metric called 'detection accuracy rate' that measured correct prediction and over- estimations of threat, with the best model achieving a 96% detection accuracy rate.

Keywords: insider threat; natural language processing; NLP; classification; machine learning.

Reference to this paper should be made as follows: Ali, S., Deverill, H., Lindquist, J. and Roginski, J. (2025) 'Human and machine partnership: natural language processing of army insider threat hub data', *Int. J. Applied Decision Sciences*, Vol. 18, No. 7, pp.1–22.

Biographical notes: Saleem Ali is a graduate student studying Data-enabled Computational Engineering and Science at Brown University. He earned his BS in Mathematical Sciences from the USA Military Academy. His current research interests are hypersonic flow and neural operators.

Hayden Deverill is an active duty Army Officer currently serving as an Instructor in the Department of Mathematical Sciences at the USA Military Academy. He has a BS and MS in Systems Engineering from the

USA Military Academy and the University of Virginia, respectively. His research focuses on leveraging data science to generate insights and create practical solutions to complex, real-world problems.

Joseph Lindquist serves as Professor of Mathematics at the USA Military Academy at West Point where he teaches courses in numerical techniques and machine learning. He earned his PhD from the Naval Postgraduate School in Monterey, California and enjoys using analytical tools to inform decisions.

Jonathan Roginski serves as the Program Manager for the Insider Threat Research Program and Assistant Professor of Mathematics at the USA Military Academy, both located at West Point. He teaches courses in network science and graph theory and earned his PhD from the Naval Post-Graduate School in Monterey, California. Jon enjoys leveraging science against complex problems that inform decision making.

1 Introduction

1.1 Insider threat overview

Countering the insider threat enhances the security of governments, the military, businesses, and all other organisations and institutions that employ people (Kelly, 2018; Nurse et al., 2014; Inayat et al., 2024). Serious insider threat cases can result in great harm to not only an organisation (and its employees), but also the customers of that organisation and proximate organisations, including the loss of life, resources, and information security (Kelly, 2018; Kamatchi and Uma, 2025; Al-Shehari and Alsowail, 2021; Whitelaw et al., 2024). For instance, purposeful disclosure of military battle plans to enemy forces by an insider threat directly reduces the effectiveness of military operations and hinders the military's capability to protect the people it serves. Consequently, it is imperative that organisations are able to protect themselves from insider threats.

Many organisations establish counter-insider threat functions to protect against exactly such a threat. In response to mass shootings and the unauthorised release of sensitive information by DoD employees and contractors, the USA Department of Defense (DoD) established the DoD insider threat management analysis centre (DITMAC) to implement insider threat policy and analyse indicators of insider threat activity. The army insider threat hub (InT Hub) coordinates with the DITMAC to focus solely on insider threat within the Army. It functions as the central facility for the Army's insider threat analysis, reporting, and response efforts, equipped with the necessary tools and capacity to offer prompt information and risk-based analytical support.

1.2 Traditional insider threat detection

The purpose of the InT hub is to detect, deter, and mitigate insider threat across the entire army. They do this by assigning potential threat levels (negligible, low, medium, high) to individuals reported to the InT hub after conducting extensive research on the individual using data from various law enforcement agencies, human resources, open source information, and other authorised data sources. Since its establishment, the InT hub has

used various insider threat frameworks to guide its analysis of the information it collects (Shaw and Sellers, 2015; Lenzenweger and Shaw, 2022). Insider threat frameworks are models that attempt to illustrate the development of insider threat within organisations to the point of an incident. They are an analytical framework that counter insider threat organisations can use to streamline their identification process through identifying specific risk indicators, some of which can be measured empirically (Shaw and Sellers, 2015; Harms et al., 2022; Al-Mhiqani et al., 2020).

There are two categories of risk indicators: dispositional traits and situational traits. Dispositional traits refer to enduring aspects of an individual's personality, such as their overall stability or impulsiveness, which may predispose them to certain behaviours. In contrast, situational traits are influenced by external conditions or immediate circumstances that can provoke certain actions from individuals. For example, (Johnston et al., 2016) explored how personality meta-traits like Stability and Plasticity moderate the influence of situational perceptions on an individual's intent to violate information security policies, revealing how personal characteristics can amplify or mitigate the impact of situational triggers. Similarly, Harris and Teasdale (2017) found that both dispositional (such as personality traits) and situational factors (like marital status or drug use) significantly contribute to the likelihood of repeated violent behaviours among individuals with serious mental disorders, illustrating the complex interaction between inherent traits and external conditions (Harris and Teasdale, 2021).

In the context of mitigating insider threats, it is essential to consider how these traits interact. Dispositional traits can provide a stable indicator of potential risk, but situational factors are often the triggers that activate these risks, making them visible. Organisations must be vigilant in identifying both types of traits to effectively predict and prevent insider threats. This dual approach is advocated by research such as that of Greitzer et al. (2019), which emphasises the need for tools that consider behavioural indicators alongside more traditional technical surveillance to detect potential threats. For instance, an organisation could promote positive deterrence practices to reduce insider risk throughout the organisation and simultaneously use standard surveillance and data-driven detection methods (Moore et al., 2022). As one reviewer pointed out, increasing throughput of backlogged cases that require synthesis from dispositional and situational evaluations can also assist organisations to develop more nuanced strategies that protect against insider threats while respecting individual differences.

1.3 Army InT hub's challenges

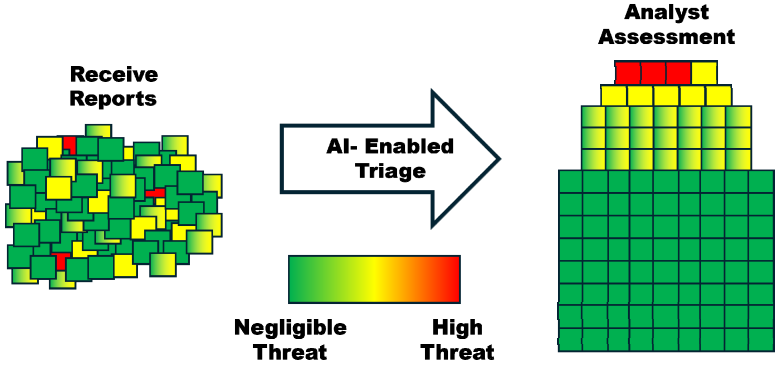
There are three challenges to the InT Hub's ability to fulfill its purpose: scope of responsibility, a queuing policy gap, and low base rates of insider threat. With an active duty end strength of nearly 450,000, the army has over 100,000 more active duty service members than the air force and the navy and nearly 300,000 more active duty members than the marines (Rogers, 2023). When including reserve personnel and DoD civilians, the difference in size between the army and the other branches of the military is even larger. Thus, the volume of insider threat reports received by the Hub is greater than all of the services, and resource limitations hinder the rate of case adjudication, which leads to the accumulation of cases in a backlog.

The second issue is that when batches of cases are assigned to individual analysts, analysts have no standard operating procedure determining which case will be investigated first. In other words, there is no policy to create an investigation queue,

given a batch of cases. As such, analysts may arbitrarily select a case for investigation, or they may use criteria such as the oldest or newest case in their batch.

Fortunately, insider threats have low base rates of occurrence. Unfortunately, this makes searching for incidents akin to searching for a 'needle in a haystack' (Lenzenweger and Shaw, 2022). Thus, the majority of cases assessed by the InT Hub are assessed to be of negligible potential threat. Altogether, these three issues create a risky environment where there is an accumulation of un-adjudicated cases. Although most of these adjudicated cases represent negligible or low potential threat, there are higher threat cases embedded in those lower threat cases. This causes those higher threat cases to have a longer waiting time in the queue before being investigated, which potentially allows them to become more problematic as time progresses. Our solution to this is a partnership between artificial intelligence (AI) and human analysts, where a classification model identifies cases as negligible, low, medium, or high potential threat so that the case investigation queue is organised in order of decreasing potential threat, thereby focusing the efforts of human analysts on the more serious cases. We illustrate this idea in Figure 1.

Figure 1 We propose a partnership between AI and human analysts to identify potentially high-threat insider threat cases in order to make effective use of the human assessor's time (see online version for colours)



1.4 Literature review: machine learning and insider threat

Due to the significant amount of research that needs to be conducted to properly track and mitigate insider threat, researchers and organisations have turned to machine learning prediction models to expedite this process (Al-Mhiqani et al., 2020; AlSlaiman et al., 2023; Alsowail and Al-Shehari, 2022; Sarhan and Altwaijry, 2022; Eftimie et al., 2020; Levy et al., 2022; Paxton-Fear et al., 2020; Symonenko et al., 2004; Meduri, 2024; Li et al., 2025; Suryotrisongko et al., 2022). Within the literature, many different modelling techniques have been used for the purpose of insider threat classification. For example, several studies have been conducted on the community emergency response team (CERT) at Carnegie Mellon University's (CMU) synthetic insider threat dataset, which contains information about employees and events related to the cyber domain such as logon/logoff, sending emails, and HTTP events (Al-Mhiqani et al., 2020; Sarhan and Altwaijry, 2022). The CERT-CMU dataset also contains a psychometric score for each

individual based on the Five Factor model for personality. It is important to note that all of the data in the CERT-CMU dataset is categorical or quantitative (Al-Mhiqani et al., 2020). Furthermore, the primary insider threats that are contained within the dataset are theft of information and privilege abuse. Several deep learning methods have been applied to this dataset, with unsupervised recurrent neural networks achieving an accuracy of 93% and supervised convolutional neural networks and convolutional neural networks achieving an accuracy of 99%. Supervised random forest achieved an accuracy of 98% and used under sampling and over sampling as a balancing measure (Al-Mhiqani et al., 2020). Overall, these models demonstrate the efficacy of machining learning models. However, for many types of insider threat, especially insider threat that does not concern information or cybersecurity, the available data is in some type of narrative text form, which cannot be traditionally processed by existing modelling methods. The solution to this problem is natural language processing (NLP).

The advantage of narrative text data is that it is easily understood and produced by humans; unfortunately, computers struggle to understand it without human assistance. This 'assistance' comes in the form of NLP is the set of methods used to analyse natural human language processing using computers (Paxton-Fear et al., 2020). Typically, data that is in the form of natural human language is written text (referred to as a corpus); however, researchers may also apply NLP to audio data. Some examples of NLP applications include, but are not limited to, translation, question answering, information retrieval, and speech recognition (Paxton-Fear et al., 2020).

The use of NLP to classify or identify insider threats is a growing field that leverages linguistic and behavioural data to detect anomalies and potentially malicious activities within an organisation. Practitioners use NLP techniques to analyse textual data, such as emails or chat logs, and can infer psychological and personality traits of users that may indicate a predisposition for risky behaviours. For instance, Eftimie et al. (2020) developed a system using NLP and personality profiles to proactively identify insider threats, utilising personality assessments based on the five-factor model and evaluating them against public datasets for feasibility. Additionally, Symonenko et al. (2004) integrated semantic analysis of insider communications with other monitoring techniques to evaluate risks, showcasing the utility of semantic NLP systems in producing conceptual representations of communication that aid in threat assessment. These applications demonstrate NLP's capability to not only understand the content but also the context of language used by potential insiders, enhancing the detection and prevention of insider threats.

In the context of countering the insider threat, information retrieval is often in the form of topic modelling because a large amount of insider threat data comes in the form of narrative text (Paxton-Fear et al., 2020). For example, police reports detailing criminal activity, human resources write ups, and counselling forms are all narrative text that are of interest for insider threat models. Using this type of data as input, topic modelling can be used to take large quantities of text data and group them into similar characteristics.

An oft-used topic modelling technique developed in 2003 is Latent Dirichlet allocation (LDA) (Blei, 2003; Axelborn and Berggren, 2023). A hierarchical Bayesian model, LDA seeks to represent text *documents* as a mixture of topics and *topics* as a mixture of words (Blei, 2003). When done well, LDA can provide the probability that a given document belongs to a particular topic, and the probability that a given topic uses a particular word. While LDA has found successful application over the past twenty years, some of the challenges the method faces are instability when applied to corpus with large

vocabulary as well as parameter interference when estimating model parameters (Egger and Yu, 2022). Furthermore, LDA is limited by its 'bag-of-words' assumption, which means that it ignores the context of the words in a corpus. In other words, LDA does not consider the order of words in text and it assumes that all topics are independent of each other. As a result, LDA is unable to differentiate whether a *bank* is a part of land that is adjacent to a river, or a financial establishment. In this regard, LDA could fail to represent a corpus accurately.

A recently developed, revolutionary alternative topic modelling technique to LDA is BERTopic (a topic modelling technique using bidirectional encoder representations from Transformers, or BERT). Importantly, BERTopic addresses the 'bag-of-words' limitation through the implementation of text embedding using BERT, which contemporary research demonstrated to produce accurate contextual vector representations for both individual words and sentences (Devlin et al., 2019; Grootendorst, 2022). This approach is distinct from previous models that typically processed text in one direction (either left to right or right to left). The model is pre-trained on a large corpus of text and then fine-tuned for specific tasks with just one additional output layer. This enables BERT to achieve state-of-the-art results on a wide range of NLP tasks, such as question answering, language inference, and others, without substantial modifications to task-specific architectures.

BERT's ability to model bidirectional context means it can understand the full context of a word based on all of its surroundings, leading to a deeper understanding of language nuances and subtleties (Devlin et al., 2019; Axelborn and Berggren, 2023). The vector representations' semantic characteristics enable the encoding of text meanings in a manner where texts with similar meanings are situated closely within the vector space. The BERTopic modelling process can be described in three steps: First, a pre-trained language model creates document embeddings that capture specific details at the document level. Next, BERTopic reduces the dimensionality of these embeddings and establishes groups of documents that share semantic connections, each group portraying a distinct topic. Finally, in contrast to similar text embedding topic modelling techniques that assume centroid-based clusters, BERTopic employs a class-based adaptation of term frequency – inverse document frequency (TF-IDF) to extract topic representations. As a result, results derived from BERTopic potentially represent documents with greater accuracy than purely lexical methods (Devlin et al., 2019).

BERT has been used in recent years to effectively extract topics and classify various types of text documents. For example, BERTopic was used for topic modelling of WhatsApp group chats, to extract coherent and interpretable themes from short, informal messages. This approach identified 12 key topics and classified the sentiment of chats as positive, negative or neutral. In this study, BERTopic outperformed traditional methods like LDA, proving more effective at analysing group chat text (Franklin et al., 2025). Another study used BERTopic to analyse over 7,000 articles from a journal. It grouped them into 35 research topics and revealed common topics, shifts in research focus, emerging topics, and potential research gaps (Baird et al., 2025). BERTopic has also been used to identify and group hate speech, particularly Islamophobic texts. In this study, BERTopic and LDA were combined with unsupervised classification techniques, and the resulting performance highlighted the differences in BERTopic and LDA (Mahmood et al., 2024). These studies show that BERT-based models like BERTopic can effectively extract meaningful topics and classify nuanced text types, even in informal or sensitive contexts.

This makes them well-suited for insider threat detection, where identifying subtle patterns, shifts in sentiment, or emerging concerns can be critical.

An additional technique that can be used with BERTtopic topic modellers is called zero-shot topic modelling. Zero-shot topic modelling is a technique designed to identify predefined topics within a large corpus of documents, leveraging the expertise of domain specialists to pinpoint expected topics (Grootendorst, 2022). This approach not only highlights these anticipated topics but also dynamically generates new topics for documents that do not align with the initial expectations, offering a flexible framework with three potential cases (Grootendorst, 2022):

- The model detects both predefined and newly clustered topics, accommodating documents that either match the predefined criteria or diverge from it.
- Only predefined topics are identified, eliminating the need for additional topic discovery.
- No predefined topics are detected, prompting the use of a standard BERTopic model to analyse the documents (Grootendorst, 2022).

The process involves labeling predefined topics, embedding these labels using an embedding model, and assessing the similarity between document embeddings and these labels using cosine similarity. Documents that meet a specific threshold are categorised under zero-shot topics, while others are processed through a regular BERTopic model, ultimately integrating both models to encompass zero-shot and non-zero-shot topics comprehensively (Grootendorst, 2022).

Including expert judgments has been illustrated to improve model performance (Tehrani et al., 2021). Examples of zero-shot topic modelling include a study that classified Reddit comments into twelve nuanced emotional and engagement categories without requiring labelled training data. Human insight played a role in defining these categories – like skepticism, concern, or technical advice – based on a manual review of sample comments, guiding the model's interpretation. The approach proved effective, revealing detailed patterns of community responses to cybersecurity discussions and demonstrating how human-informed, flexible AI can scale nuanced social analysis across large, unstructured datasets (Achuthan et al., 2025). Another study utilised BERTopic and zero-shot topic modelling for technology opportunity analysis, where a model was used to used to identify untapped or emerging solutions for a given technological challenge using patent data. Zero-shot modelling allowed for more targeted identification of technology opportunities, showing how incorporating prior knowledge can improve the relevance and precision of topic extraction. This would help in pinpointing solutions from a wide range of technological domains for given problems (Kim and Lee, 2024). These examples highlight how zero-shot topic modelling can classify complex text without labelled data, guided by human-defined categories or domain knowledge. For insider threat detection, this approach enables scalable, targeted analysis of insider threat reports, allowing organisations to surface potential risks or intent even in the absence of predefined labels.

Another modelling technique that often improves results is ensemble models. Ensemble learning is a machine learning technique that improves accuracy and robustness in predictions by combining outputs from multiple models. It addresses errors or biases in individual models by harnessing the collective insights of the ensemble. The key idea behind ensemble learning is to integrate the results of various models to produce

a more accurate prediction. By taking into account multiple viewpoints and leveraging the strengths of different models, ensemble learning enhances the overall performance of the system. This method not only increases accuracy but also offers resilience against data uncertainties. Ensemble learning has proven to be an effective tool in numerous fields, providing more reliable and robust forecasts by integrating predictions from several models (Ganaie et al., 2022).

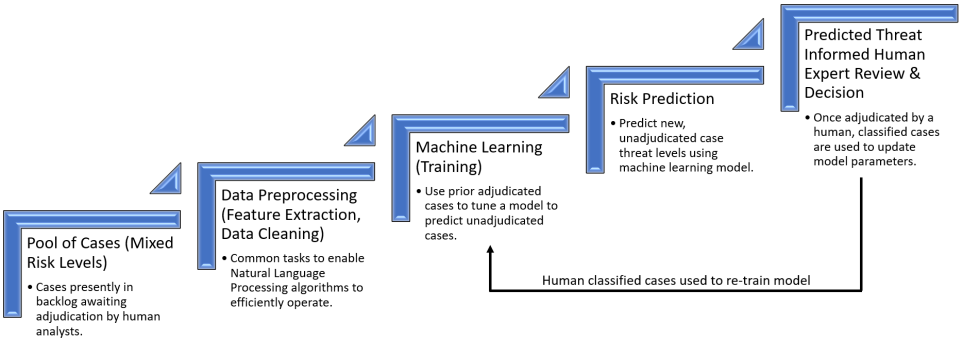
1.5 Research goal

This paper seeks to apply NLP techniques to data received from the army insider threat hub. Using a dataset of 1,306 expert-evaluated cases, we developed a classification model that enables us to assess the threat level posed by insider threat case subjects. This classification model can be used to create a threat motivated queuing system, where analysts investigate their assigned batch of cases in the order of decreasing potential threat as shown in Figure 1. In other words, the model would serve as a component of a broader decision – support system, offering analytical insights to support and enhance human judgment (Storey et al., 2024).

2 Methods

In this section, we propose a method that takes raw text and assigns a predicted threat level – derived from previously adjudicated insider threat cases. Our proposed method is shown below in Figure 2.

Figure 2 We propose a process where prior adjudicated InT cases are used to train a model to predict threat levels of unadjudicated cases (see online version for colours)



Notes: These predictions are then used to prioritise human review efforts. the adjudicated cases can then be used to strengthen the model by serving as additional data to re-run the model.

2.1 Data

The dataset for this model initially contained 1,769 adjudicated insider threat cases from the US army insider threat hub. Each of these cases was adjudicated by an expert human analyst and assigned a threat level based on specific criteria. Below is a mock-up of what one of these cases would look like using notional text:

Administrative information
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris.
Research
Law enforcement: Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem.
Human resources: Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc.
Assessment
Subject is assessed to pose a negligible potential threat to the US Army

The 'administrative information' section contained the analyst's name, case number, date of adjudication, and various information about the subject (gender, rank, incident location, etc.). Research Table 1 illustrates the features a python script was able to extract. Rows that were missing data were dropped from the dataset. 'case research' and 'army threat' made up the training and test data. 'case research' is an aggregation of all the research the investigating analyst compiled from various sources, including law enforcement, human resources, personnel security, or open sources information. All of this information is analysed to produce their assessment of the subjects potential threat to the Army, which was extracted as the 'army threat' variable.

To prepare the raw data for machine learning analysis, a python script was written to extract each of the variables described in Table 1. However, after running the script on the data, we found that there were cases that seemed to be missing a threat assessment and analyst research. Upon inspecting the original data, we found that these cases were missing many of the key variables from 1. Consequently, these cases were dropped from our data, which caused the size of our usable data to be reduced from 1,769 to 1,306 cases.

2.2 Natural language processing models: BERT, BERTopic, zero-shot topic, and ensemble models

We first applied a BERT model to train our insider threat classification model by separating our observations up into training (70%) and test (30%) sets. We chose to use BERT because of the model's ability to use bidirectional context to understand the full context of a word based on all of its surroundings, leading to a deeper understanding of language nuances and subtleties (Devlin et al., 2019). This is important when using NLP to understand and assess insider threat because the spatial relationships between words in text can help to understand actions or behaviours that were conducted repeatedly or that may be associated with specific circumstances.

Next, we applied a BERTopic model to classify the topic, or threat category, of each individual insider threat case. We chose to use BERTopic because of the model's ability to identify and represent latent topics more effectively compared to classical models like LDA, providing more nuanced insights into the underlying themes in the text data (Grootendorst, 2022). More specifically, we applied a zero-shot topic model that allowed us to specify our own topics that we thought were significant and allow BERTopic to create its own additional topics. We used the case topics as a means to help our client

understand the relative magnitude of each threat and also used the topics as a feature in our ensemble model.

Table 1 Description of variables

Variable	*Type*	*Description*
Analyst	Categorical	Numerical identifier for analyst
Case information	Text	Contains case ID, rank, incident location, etc.
Reporting threshold	Categorical	DoD insider threat management and analysis centre reporting thresholds
Case research	Narrative text	Analyst research on case subject
Army threat	Ordinal categorical	Threat ASSESSMENT (negligible, low, medium, high)

Notes: Case research contains all of the research an analyst had to create their threat assessment. Army threat was an assessment of the individuals threat to personnel, resources, or information.

Lastly, we applied ensemble model to our data. We combined the results of our BERTopic topic model with our BERT classification model using multi class logistic regression model. The data used for the multi class regression was each insider threat case's assigned topic and their predicted threat level from our chunked classification model. Those two features were used to predict a final predicted threat level. Our goal was to use the additional information from the topic classification to inform a better overall model.

2.3 Pitfalls and mitigation of unbalanced data

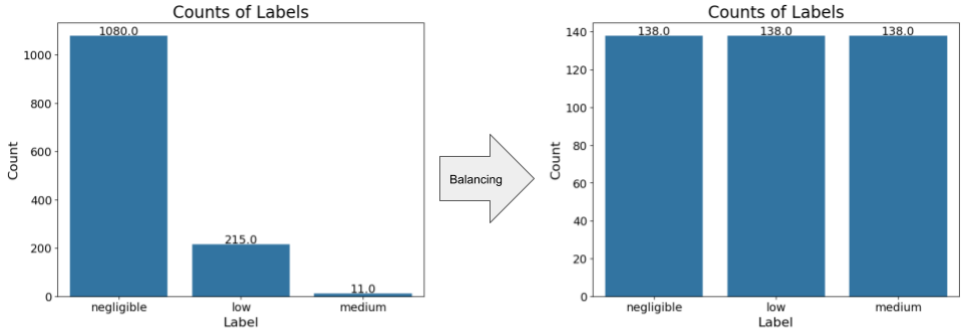
Figure 3 highlights a common challenge in applying machine learning to insider threat: unbalanced data. A majority of insider threat cases are classified as negligible threat level, whereas very few will be of medium threat level or higher. This can cause insider threat classification models to over fit to the majority class. In this case, that means that our model will tend to label all cases as negligible, which could potentially cause higher threat cases to me mislabelled as low threat cases. Two common techniques to address this issue are over-sampling and under-sampling.

Oversampling is a process that increases the size of the minority class by duplicating existing instances or generating synthetic samples, which can help improve the classifier's ability to detect the minority class (Chawla et al., 2002; Nti et al., 2024). It can result in better classifier performance as it provides more examples for the model to learn from (Oskouei and Bigham, 2016). However, over-sampling can lead to overfitting (generating accurate predictions with in-sample data, but poor predictions on out-sample data) since it makes exact copies of minority class instances or creates closely similar synthetic samples. This could make the model extremely accurate on the training data, but inaccurate on any 'new' data (Shi et al., 2023).

Under-sampling is a process that reduces the size of the majority class by removing instances, which can decrease the training time and computational cost (Jindaluang et al., 2014). It can help to balance the dataset without creating synthetic data, thus maintaining the original data's properties (Kasemtaweechok and Suwannik, 2024). However, It can lead to loss of important information if significant instances from the majority class are removed, potentially degrading the model's performance (Moghaddam and Noroozi,

2021). Also, there is a risk that under-sampling might oversimplify the problem, leading to underfitting and poor generalisation on new data.

Figure 3 Counts of the analyst threat assessment levels of 1,306 insider threat cases (see online version for colours)



Notes: in our balancing process, nelgible threat cases were undersampled, and medium threat cases were oversampled through random duplication and deletion.

For our dataset, after randomly selecting 70% of the data to serve as the training set, the negligible cases were randomly under-sampled through removal to the same number of cases as low. Similarly, the medium cases were over-sampled through duplication to the same number of cases as low (see Figure 3).

2.4 Ensemble model

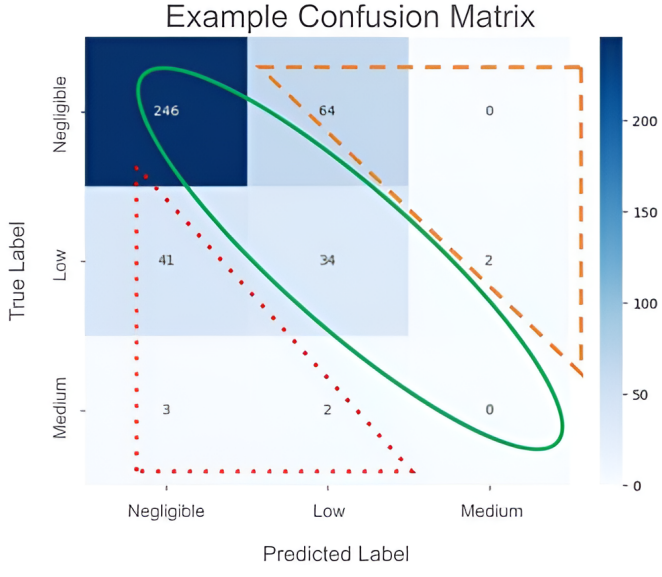
Ensemble learning is a machine learning technique that improves accuracy and robustness in predictions by combining outputs from multiple models (Han et al., 2025). It addresses errors or biases in individual models by harnessing the collective insights of the ensemble. The key idea behind ensemble learning is to integrate the results of various models to produce a more accurate prediction. By taking into account multiple viewpoints and leveraging the strengths of different models, ensemble learning enhances the overall performance of the system. This method not only increases accuracy but also offers resilience against data uncertainties. Ensemble learning has proven to be an effective tool in numerous fields, providing more reliable and robust forecasts by integrating predictions from several models (Ganaie et al., 2022).

Ensemble models have been successfully used to detect network threats. For example, a study used ensemble models to address the limitations of individual classifiers in detecting network threats, such as overfitting and inconsistent accuracy. The ensemble, which included models like Random Forest, AdaBoost, and convolutional neural networks, achieved a high overall accuracy of 97.92% with balanced precision and recall across both benign and malicious traffic. While performance gains over the best individual models were modest, the ensemble offered more stable and generalised results across varied traffic types (Ford and Berry, 2025). Another application was credit card fraud detection, where logistic regression, random forest, and AdaBoost were combined to improve classification accuracy and reduce false positives. A soft voting mechanism was used to aggregate model outputs, leveraging the strengths of each algorithm for better decision-making and adaptability to imbalanced and evolving fraud patterns

(Al-Maari et al., 2025). Thus, ensemble models are relevant for insider threat detection because they can address challenges like overfitting, class imbalance, and inconsistent accuracy – common issues in insider threat datasets. By combining multiple algorithms, ensembles provide could more robust and generalised detection across diverse and evolving threat behaviours.

We combined the results of our topic model with our classification model using multi class logistic regression. The data used for the multi class regression was each insider threat case's assigned topic and their predicted threat level from our chunked classification model. Those two features were used to predict a final predicted threat level.

Figure 4 In this example confusion matrix, the DAR can be calculated by taking the sum of cases enclosed in green and orange (accurate and over-estimations of threat) and dividing by the total number of cases (see online version for colours)



2.5 Evaluation criteria

We assess our model with a newly-developed measure we call the detection accuracy rate (DAR). This measure considers accurate predictions of threat and over-estimations of threat (low threat case predicted as medium) because in the interest of security, it is prudent to overestimate rather than underestimate threat. For example, it is better for the organisation to place greater scrutiny on a case classified as 'high,' but turns out to be 'medium,' than the reverse. Using the confusion matrix in Figure 4 as a reference, the DAR can be computed by taking the sum of the accurate predictions (outlined in solid green) and over-estimations of threat (outlined in dashed orange).

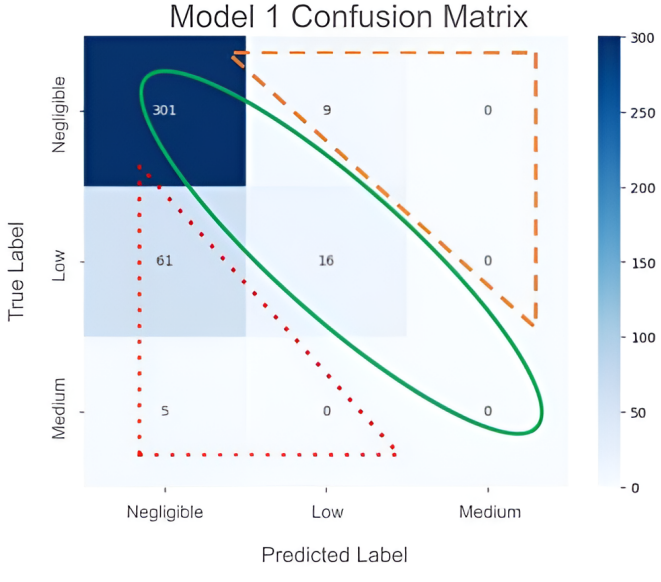
3 Results

Developing our final model results was iterative, involving three different NLP classification models, a zero shot topic model, and an ensemble model. The first three models are NLP classification models that we made sequential improvements to by adjusting the model parameters to improve the performance of each model. The fourth model is an ensemble model that combined the models of the best NLP classification model we developed (Section 3.3: Model 3: chunked model) with our zero shot topic model (Section 3.4: ensemble model) using a multi-classification logistic regression classification model. We describe each model and associated results in this section.

3.1 Model 1: baseline model

Our first model was a basic model that used various NLP methods using default parameter values assigned by package designers. This model uses the distillbert-base-uncased model and then trained and tested it on our data without special considerations for the models parameters, namely 'max_embedding_length'. The parameter 'max_embedding_length' is the maximum length of the vector representation of a corpus after tokenisation. For all of our models, an embedding length of 512 was used, so if the length of a text was greater than 512 after tokenisation, all tokens after the 512th were truncated. This means that for model 1, our model was trained on partial insider threat documents rather than complete ones. As such, model 1 does not fully 'understand' any insider threat case, which causes it to have a DAR of 83% and predict many of the low and medium cases as negligible.

Figure 5 Confusion matrix of model 1 (see online version for colours)



Notes: Model achieved a DAR of 83%. all medium and about 79% of low threat cases were predicted to be of negligible threat, indicating that this model would be effective in creating a threat motivated case queuing system.

Figure 5 shows the results of Model 1 on threat cases that were not used in the training of the model. We use a performance measurement tool for machine learning classification models called the confusion matrix to highlight results. This tool summarises the predictions of a model by comparing the actual values with the predicted values. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. This multi-class confusion matrix helps in evaluating the accuracy, precision, recall, and overall effectiveness of the classification model across all classes, providing detailed insights into where the model performs well and where it may need improvement.

When we examined model 1 predictions, we concluded that certain parts of the analyst research may carry more weight when it comes to assigning a threat level. The truncation had a profound impact on DAR - primarily due to the loss of important information.

3.2 Model 2: focused model

For model 2, we chose to naively focus our model on important parts by truncating all but the middle 512 tokens of the insider threat case data. Figure 6 illustrates that model two achieve a DAR score of 88% – higher than model 1, indicating that more medium and low threat predictions that were accurate or over estimations were made. While these results were promising, this method was still truncating potentially important information.

3.3 Model 3: chunked model

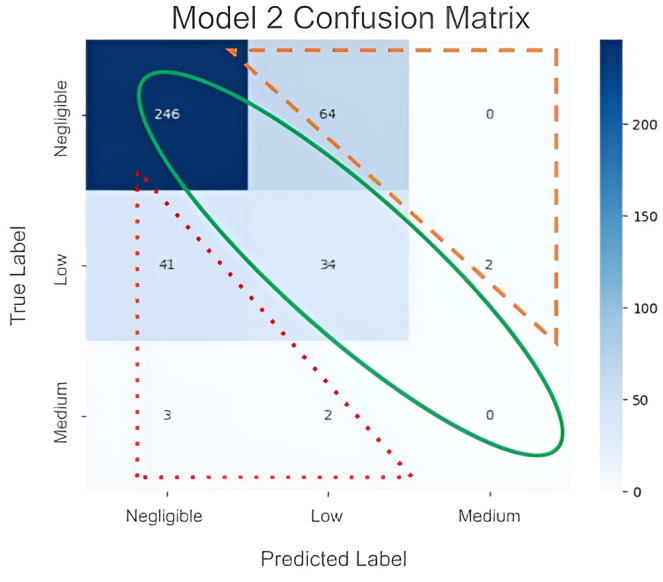
Thus for model three, we divided each insider threat case into 6 evenly sized chunks that were small enough to be 512 tokens in length. After training on chunks, the model was used for inference on chunked test data, with each chunk receiving a prediction. Insider threat cases were then reassembled and the highest threat prediction a chunk received was set as that documents overall threat prediction, This led to the results in 7 with a DAR of 96%, which indicates that our model can be used to effectively create a threat motivated queuing system.

3.4 Zero shot topic modelling

Figure 8 illustrates the results of our zero-shot topic modelling. All identified topics were threat indicators used by hub analysts to identify threat. The vast majority of documents were identified as the ‘criminal activity, arrest’ topic, but others, such as domestic abuse and sex crimes were also identified, highlighting the potential of topic models as a screening tool to determine the types of cases. This result is akin to model 1 of the classification models because no focusing or chunking was performed, and there is truncation occurring, which harms the strength of the topic representations. Nevertheless, uncovering topics within a set of cases would be useful for insider threat analysts because it enables them to focus their attention on cases with topic that are more serious or indicative of a threat. For example, there may be a strategic reason why leadership needs to triage cases to review cases of ‘topic 1’ with higher priority than other topics, regardless of threat level. Or, there may be a directive to review ‘all high and medium

threat level cases in topic 1 by a given deadline.' Topic modelling enables analysts to meet both imperatives and more.

Figure 6 Confusion matrix of model 2 ('focused' model) (see online version for colours)



Notes: Model achieves a DAR of 88%. more cases were identified as medium threat, but 53% of low cases were still predicted to be medium threat. this suggests that certain portions of insider threat cases have a stronger association with the threat level.

3.5 Model 4: ensemble model

Figure 9 is the result of combining our topic model with our classification model to create a multi class logistic regression. This model performed the worst among all of the models, with a DAR of 73%, indicating that in its current state, the ensemble model is not suitable for creating a threat motivated queuing system. This may be because the vast majority of cases were assigned to the 'criminal activity, arrest' topic, which means that that data point did not do much to differentiate the cases from each other. A potential improvement for the ensemble model is to apply the same chunking procedure that was used for our classification model. This would allow for more accurate topic identification.

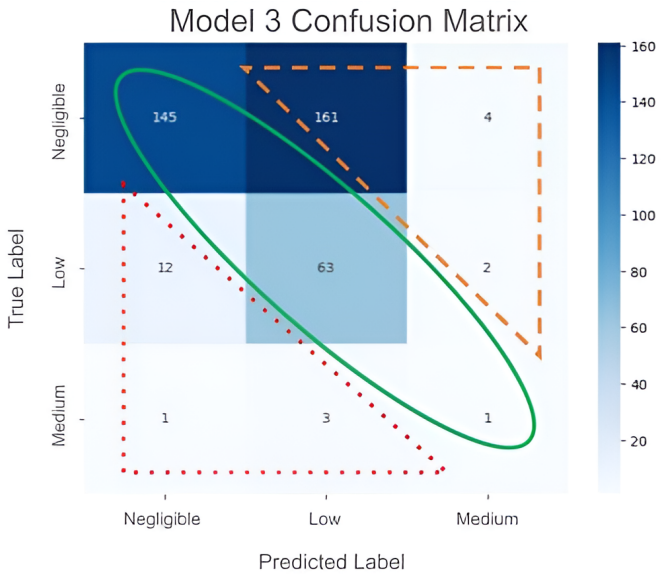
4 Limitations

We have proposed a methodology that uses state of the art machine learning models to classify insider threat cases as negligible to high threat. We demonstrated accuracy through the metric of DAR showing how AI can help to prioritise the cases that expert human analysts examine. While we believe that the partnership between AI and the human can add value to detecting and mitigating the impact of insider threats, there are some limitations to our work that fall into three main categories: data, model building, and ethical considerations. We comment on each below.

- Data. At the outset of this project, the data used to train both the topic model and the classifier had already been curated by an expert team of InT hub analysts that merged various data streams. While we have shown accurate classification on this curated data, it has not yet been shown to produce similar results on raw data streams (criminal sources, serious incident reports, personnel records, etc.) While we believe that the additional data streams will yield similar results, this work should be viewed as a stepping stone to demonstrate the utility of the NLP methodology, as it is still unproven on these raw streams. Future work aims to resolve this limitation.
- Model. NLP methods represent a research thread that is highly active with new methods developing nearly daily to improve on shortcomings. Often these threads are expanded by developments in computing capability. Tasks that were once only possible in theory have been enabled by advances in computing. NLP scholars just 10 years ago could only theorise that models (Devlin et al., 2019) built on the interaction of nouns and verbs derived from thousands of books (800 million words) and all of Wikipedia (2.5 billion words) could perform the classification tasks that we have shown in this paper. With this stated, many of these models rely on the transfer of learning from a general use to a specific use. While the authors would gladly share the more than 340 million parameter values from our transformer models to a researcher who seeks to implement the methodology on their specific use case, we would strongly encourage performing additional fine tuning on the specific use case. Further, models should be periodically re-tuned using expertly labelled data to ensure that the model is not drifting (decrease in predictive capability caused by changes in the environment). Without initial fine tuning, it is possible that features most important to a given use case are overlooked. Without periodic re-tuning, a model that performs well today may perform poorly at a later time. Stated concisely, any implementation of the models described in this paper should be accompanied by fine tuning and periodic re-tuning.
- Ethical. Any large language model must be trained on large text corpora that may contain certain biases related to race, gender, and other demographics. These biases can be reflected and potentially amplified in outputs of the model that may result in unfair or discriminatory results in the application. Special care must be taken to ensure a training corpus represents the population as a whole (Selbst et al., 2019; Barocas et al., 2023; Thomas et al., 2022; Zhang et al., 2018) – only with careful consideration can researcher achieve fair and equitable machine learning models.

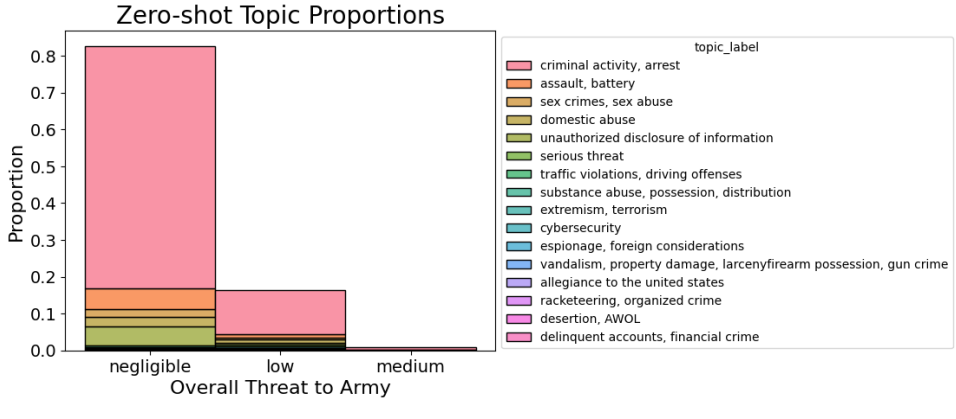
Additionally, as stated, this model is not meant to replace the analysts expert role in classifying case threat level. Rather, the goal of the classification model is to triage the priority in which analysts review cases to enable a more timely and effecitve review process. Relying solely on the classification level of the machine learning model without a human in the loop has inherent ethical considerations. We are not proposing this, as there is still the need for expert human analysis when assessing insider threat due to the intangible, human dimension. Even with the human in the loop; however, there are some ethical considerations about why the model may be classifying cases the way it is. We accept this potential risk as we see leveraging the machine learning model to give a best guess at risk level to prioritise review as more informative than random sampling based on our initial results.

Figure 7 Confusion matrix of model 3 (‘chunked’ model) (see online version for colours)



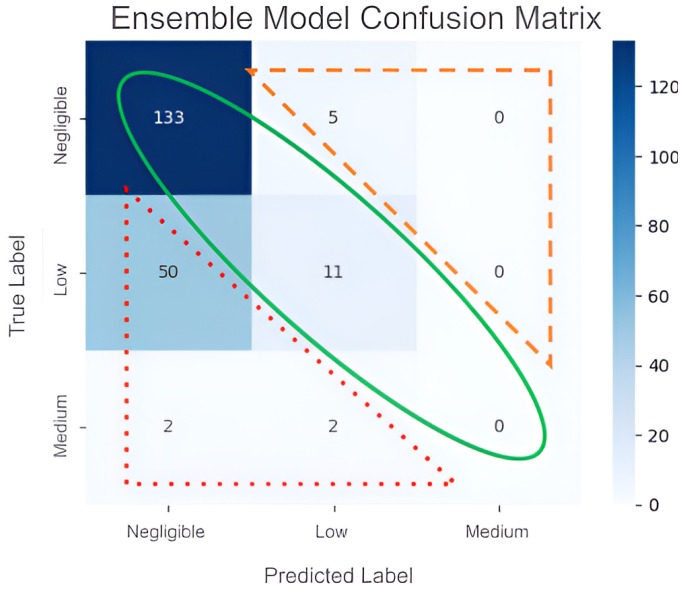
Notes: This model produced the best results (DAR = 96%), indicating the importance of capturing all relevant information related to insider threat cases when training the model.

Figure 8 Zero-shot topic proportions (see online version for colours)



Note: The vast majority of cases were labelled as ‘criminal activity, arrest’.

Figure 9 Confusion matrix of ensemble model. model achieved a DAR of 73%, indicating that in its current state, the ensemble model is not suitable for creating a threat motivated queuing system (see online version for colours)



5 Discussion

An NLP model that can classify insider threat cases as a particular threat level can help the counter-insider threat analytical team focus their efforts where it is most needed: the cases with greatest likelihood of high severity. While this model would not accelerate analysts' workflow or increase their throughput, it would increase the value of analysts' time because more of that time would be expended on potentially higher threat cases instead of cases of negligible threat. In the long run, this should enable organisations to rapidly identify and mitigate their most pressing threats, which reduces the risk of insider threat incidents. We envision analyst being able to used some type of sorting mechanism to sore their batch of cases by predicted threat level and topic.

This type of model is similar in principle to a predictive policing model, as so many of the same ethical considerations hold. The models described herein and predictive policing both involve acquisition and analysis of large amounts of data, including personal information, raising significant privacy issues. Our insider threat assessment model and a predictive policing model can perpetuate or even exacerbate existing biases in the data they use. This can lead to disproportionate targeting of specific groups. As such, we stress that this model is a tool meant to increase the productivity of analysts in threat identification, not replace the analyst. Humans are necessary for the ethical implementation of insider threat protection measures.

6 Conclusions and future research

This works serves to advance the partnership between human and machine in the areas of insider threat detection and mitigation. The following conclusions summarise the contribution of the work; we also summarise the direction of potential future work in this area.

6.1 Conclusions

- The authors provide a methodology to clean and through over- and under-sampling techniques prepare a corpus of semantic data for the application of NLP models, where expected outcomes have non-uniform distribution.
- The models described herein classify a corpus of documents by topic enabling the identification of extant categories contained within the corpus.
- The models developed through this work are trained and fine tuned to ingest analytical reports and predict the severity of an insider threat case.
- Though not described in this paper, the models created provide a mechanism to compare the distribution analytical findings by analyst, potentially highlighting biases and enabling re-training of the analytical workforce.

6.2 Future directions

- The corpus of semantic data currently analysed are reports compiled by analysts. Future work includes the development of tools to automatically ingest raw data from the same sources analysts manually 'pull' data and harmonise that data in an architecture within which analysts may run NLP and AI tools.
- Verify the efficacy of modelling outputs by testing currently trained models on data not yet evaluated by analysts and comparing the model classification of threat with the analyst's classification of threat.
- Evaluate the potential of using models similar to the insider threat models described in this work to the continuous evaluation (CE) process for security clearances.

References

Achuthan, K., Khobragade, S. and Kowalski, R. (2025) 'Public sentiment and engagement on cybersecurity: insights from Reddit discussions', *Computers in Human Behavior Reports*, March, Vol. 17, p.100573.

Al-Maari, A-A., Abdulnabi, M., Nathan, Y., Ali, A., Ali, U. and Khan, M. (2025) 'Optimized credit card fraud detection leveraging ensemble machine learning methods', *Engineering, Technology and Applied Science Research*, Vol. 15, No. 3, pp.22287–22294.

Al-Mhiqani, M.N., Ahmad, R., Abidin, Z.Z., Yassin, W., Hassan, A., Abdulkareem, K.H., Ali, N.S. and Yunos, Z. (2020) 'A review of insider threat detection: classification, machine learning techniques, datasets, open challenges, and recommendations', *Applied Sciences*, July, Vol. 10, No. 15, p.5208.

Al-Shehari, T. and Alsowail, R.A. (2021) 'An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques', *Entropy*, September, Vol. 23, No. 10, p.1258.

AlSlaiman, M., Salman, M.I., Saleh, M.M. and Wang, B. (2023) 'Enhancing false negative and positive rates for efficient insider threat detection', *Computers and Security*, March, Vol. 126, p.103066.

Alsowail, R.A. and Al-Shehari, T. (2022) 'Techniques and countermeasures for preventing insider threats', *Peer J. Computer Science*, April, Vol. 8, p.e938.

Axelborn, H. and Berggren, J. (2023) *Topic Modeling for Customer Insights*.

Baird, H.B.G., Allen, W., Gallegos, M., Ashy, C.C., Slone, H.S. and Pullen, W.M. (2025) 'Artificial intelligence driven analysis identifies anterior cruciate ligament reconstruction, hip arthroscopy and femoroacetabular impingement syndrome, and shoulder instability as the most commonly published topics in arthroscopy', *Arthroscopy, Sports Medicine, and Rehabilitation*, February, p.101108.

Barocas, S., Hardt, M. and Narayanan, A. (2023) *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press.

Blei, D.M. (2003) *Latent Dirichlet Allocation*.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and. Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, June, Vol. 16, pp.321–357.

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2019) *BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding*, May, arXiv:1810.04805 [cs].

Eftimie, S., Moinescu, R. and Racuciu, C. (2020) 'Insider threat detection using natural language processing and personality profiles', in *2020 13th International Conference on Communications* (*COMM*), Bucharest, Romania, June, pp.325–330, IEEE.

Egger, R. and Yu, J. (2022) 'A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts', *Frontiers in Sociology*, May, Vol. 7, p.886498.

Ford, J. and Berry, H.S. (2025) *Advancing Network Threat Detection through Standardized Feature Extraction and Dynamic Ensemble Learning*, Concord, NC.

Franklin, A., Emmanuella, C.M. and Bamidele, W. (2025) 'Analysing Whatsapp group chat using advanced natural language processing (NLP) techniques', *International Journal of Computer Science and Mathematical Theory*, Vol. 11, p.2025.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. and Suganthan, P.N. (2022) 'Ensemble deep learning: a review', *Engineering Applications of Artificial Intelligence*, October, Vol. 115, p.105151.

Greitzer, F.L., Purl, J., Leong, Y.M. and Sticha, P.J. (2019) 'Positioning your organization to respond to insider threats', *IEEE Engineering Management Review*, Vol. 47, No. 2, pp.75–83.

Grootendorst, M. (2022) *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*, March, arXiv:2203.05794 [cs].

Han, K., Apio, C., Song, H., Lee, B., Hu, X., Park, J., Zhe, L., Goo, T. and Park, T. (2025) 'An ensemble approach improves the prediction of the COVID-19 pandemic in South Korea', *Journal of Global Health*, March, Vol. 15, p.4079.

Harms, P.D., Marbut, A., Johnston, A.C., Lester, P. and Fezzey, T. (2022) 'Exposing the darkness within: a review of dark personality traits, models, and measures and their relationship to insider threats', *Journal of Information Security and Applications*, December, Vol. 71, p.103378.

Harris, M.N. and Teasdale, B. (2021) 'The prediction of repeated violence among individuals with serious mental disorders: situational versus dispositional factors', *Journal of Interpersonal Violence*, January, Vol. 36, Nos. 1–2, ppl.691–721.

Inayat, U., Farzan, M., Mahmood, S., Zia, M.F., Hussain, S. and Pallonetto, F. (2024) 'Insider threat mitigation: Systematic literature review', *Ain Shams Engineering Journal*, December, Vol. 15, No. 12, p.103068.

Jindaluang, W., Chouvatut, V. and Kantabutra, S. (2014) 'Under-sampling by algorithm with performance guaranteed for class-imbalance problem', in *2014 International Computer Science and Engineering Conference* (*ICSEC*), Khon Kaen, Thailand, July, pp.215–221, IEEE.

Johnston, A.C., Warkentin, M., McBride, M. and Carter, L. (2016) 'Dispositional and situational factors: influences on information security policy violations', *European Journal of Information Systems*, May, Vol. 25, No. 3, pp.231–251.

Kamatchi, K. and Uma, E. (2025) 'Insights into user behavioral-based insider threat detection: systematic review', *International Journal of Information Security*, April, Vol. 24, No. 2, p.88.

Kasemtaweechok, C. and Suwannik, W. (2024) 'Under-sampling technique for imbalanced data using minimum sum of Euclidean distance in principal component subset', *IAES International Journal of Artificial Intelligence* (*IJ-AI*), March, Vol. 13, No. 1, p.305.

Kelly, W.E. (2018) *Enemies Within Our Government*, Vol. 35, No. 2, p.2018.

Kim, J. and Lee, S. (2024) 'Technology opportunity analysis for creating innovative solutions: applying semi-supervised topic modelling on patent data', in *2024 Portland International Conference on Management of Engineering and Technology* (*PICMET*), Portland, OR, USA, August, pp.1–9, IEEE.

Lenzenweger, M.F. and Shaw, E.D. (2022) *The Critical Pathway to Insider Risk Model: Brief Overview and Future Directions*.

Levy, M., Horneman, A., Ditmore, R. and Motell, C. (2022) 'Predicting the threat: investigating insider threat psychological factors with advanced natural language processing', in *Proceedings of the 55th Hawaii International Conference on System Sciences*, January.

Li, C., Zhu, Z., He, J. and Zhang, X. (2025) *RedChronos: A Large Language Model-Based Log Analysis System for Insider Threat Detection in Enterprises*, March 2025, arXiv:2503.02702 [cs].

Mahmood, S.A., Siddiqi, R., Hameed, M. and Rafi, M. (2024) 'Content moderations without labels: unsupervised classification of hateful texts', in *2024 International Conference on IT and Industrial Technologies* (*ICIT*), December, pp.1–6, Chiniot, Pakistan, IEEE.

Meduri, K. (2024) 'Cybersecurity threats in banking: unsupervised fraud detection analysis', *International Journal of Science and Research Archive*, March, Vol. 11, No. 2, pp.915–925.

Moghaddam, S.M.J. and Noroozi, A. (2021) 'A novel imbalanced data classification approach using both under and over sampling', *Bulletin of Electrical Engineering and Informatics*, October, Vol. 10, No. 5, pp.2789–2795.

Moore, A.P., Gardner, C. and Rousseau, D.M. (2022) 'Reducing insider risk through positive deterrence', *Counter-Insider Threat Research and Practice*, Vol. 1, p.2022.

Nti, I.K., Adu, K., Nimbe, P., Nyarko-Boateng, O., Adekoya, A.F. and Appiahene, P. (2024) 'Robust and resourceful automobile insurance fraud detection with multi-stacked LSTM network and adaptive synthetic oversampling', *International Journal of Applied Decision Sciences*, Vol. 17, No. 2, pp.230–249.

Nurse, J.R.C., Buckley, O., Legg, P.A., Goldsmith, M., Creese, S., Wright, G.R.T. and Whitty, M. (2014) 'Understanding insider threat: a framework for characterising attacks', in *2014 IEEE Security and Privacy Workshops*, San Jose, CA, May 2014, pp.214–228, IEEE.

Oskouei, R.J. and Bigham, B.S. (2016) 'Over-sampling via under-sampling in strongly imbalanced data', *Int. J. Adv. Intell. Paradigms*, January, Vol. 9, No. 1, pp.58–66.

Paxton-Fear, K., Hodges, D. and Buckley, O. (2020) 'understanding insider threat attacks using natural language processing: automatically mapping organic narrative reports to existing insider threat frameworks', in Abbas Moallem (Ed.): *HCI for Cybersecurity, Privacy and Trust*, Cham, 2020. Springer International Publishing. Series Title: Lecture Notes in Computer Science, Vol. 12210, pp.619–636.

Rogers, M. (2023) *H.r.2670 – 118th Congress* (2023-2024): *National Defense authorization Act for Fiscal Year 2024*. CRS Report H.R. 2670, House Armed Services Committee, Washington, DC.

Sarhan, B.B. and Altwaijry, N. (2022) 'Insider threat detection using machine learning approach', *Applied Sciences*, December, Vol. 13, No. 1, p.259.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) 'Fairness and abstraction in sociotechnical systems', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, *FAT** '19, New York, NY, USA, Association for Computing Machinery, pp.59–68.

Shaw, E. and Sellers, L. (2015) 'Application of the critical-path method to evaluate insider risks', *Studies in Intelligence*, June, Vol. 59, No. 2.

Shi, J., Song, D., Zheng, S., Hu, Y., Chen, S. and Pei, F. (2023) 'Bidirectional sampling method for imbalanced data', in Srikanta Patnaik and Tao Shen (Eds.): *Seventh International Conference on Mechatronics and Intelligent Robotics* (*ICMIR 2023*), Kunming, China, September, p.46, SPIE.

Storey, V.C., Hevner, A.R. and Yoon, V.Y. (2024) 'The design of human-artificial intelligence systems in decision sciences: a look back and directions forward', *Decision Support Systems*, July, Vol. 182, p.114230.

Suryotrisongko, H., Ginardi, H., Ciptaningtyas, H.T., Dehqan, S. and Musashi, Y. (2022) 'Topic modeling for cyber threat intelligence (CTI)', in *2022 Seventh International Conference on Informatics and Computing* (*ICIC*), Denpasar, Bali, Indonesia, December, pp.1–7, IEEE.

Symonenko, S., Liddy, E.D., Yilmazel, O., Del Zoppo, R., Brown, E. and Downey, M. (2004) 'Semantic analysis for monitoring insider threats', in Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Hsinchun Chen, Reagan Moore, Daniel D. Zeng, and John Leavitt (Eds.): *Intelligence and Security Informatics*, Springer Berlin Heidelberg, Berlin, Heidelberg, Series Title: Lecture Notes in Computer Science, Vol. 3073, pp.492–500.

Tehrani, M.M., Mobin, M., Beauregard, Y., Rioux, M. and Kenne, J.P. (2021) 'Multi-scenario and multi-criteria approach to evaluate prediction techniques used to recognise failure patterns', *International Journal of Applied Decision Sciences*, Vol. 14, No. 4, pp.361–386.

Thomas, D., Kleinberg, S., Brown, A.W., Crow, M., Bastian, N.D., Reisweber, N., Lasater, R., Kendall, T., Shafto, P., Blaine, R., Smith, S., Ruiz, D., Morrell, C. and Clark, N. (2022) 'Machine learning modeling practices to support the principles of AI and ethics in nutrition research', *Nutritional Diabetes*, June, Vol. 12, No. 1, pp.1–10.

Whitelaw, F., Riley, J. and Elmrabit, N. (2024) 'A review of the insider threat, a practitioner perspective within the UK financial services', *IEEE Access*, Vol. 12, pp.34752–4768.

Zhang, B.H., Lemoine, B. and Mitchell, M. (2018) 'Mitigating unwanted biases with adversarial learning', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp.335–340.