An NLP-Based Framework Leveraging Email and Multimodal User Data

Neda Baghalizadeh-Moghadam[a], Frédéric Cuppens[b] and Nora Boulahia-Cuppens[c]
Polytechnique Montreal, Canada

Keywords: Insider Threat Detection, Natural Language Processing (NLP), Feature Engineering, Isolation Forest, One-Class SVM.

Abstract: Traditional approaches for insider threat detection rely on analyzing activity logs to detect abnormal user activities. In this paper, we investigate how the exchange of messages between users could also contribute to detecting insider threats. This work presents an NLP-driven anomaly detection framework that incorporates feature engineering and prompt engineering across multimodal user activities, such as emails, HTTP requests, file access, and logon events. This study employs Named Entity Recognition (NER), Sentiment Analysis, and Prompt Engineering, to extract semantic, contextual, and behavioral insights that enhance anomaly detection. These enriched representations are processed by an Isolation Forest and One-Class Support Vector Machine (One-Class SVM) for the unsupervised detection of deviations from normal user behavior. Unlike most previous works that focus solely on user log activity datasets, our method incorporates both user log activity and email communication data for insider threat detection. Experimental results on the CERT r4.2 dataset demonstrate that the proposed multimodal approach improves anomaly detection with high accuracy, greater precision, and reduced false alarm rates. Hence, our framework offers greater explainability and scalability in addressing sophisticated insider threats.

1 INTRODUCTION

Insider threats pose a very serious threat to the security environment of organizations in modern times, as threats due to misuse of access by insiders, willingly or otherwise, are spreading. These threats are hard to detect because whatever resource the insider will misuse, their access is usually legitimate; therefore, detection of aberrant behavior is challenging with traditional security mechanisms like firewalls and external threat monitoring (Cappelli et al., 2012). This has attracted interest in the use of high-value machine learning methods that are able to find trace signals of user behavior (Borky and Bradley, 2018).

The current paper presents a novel robust framework for the detection of abnormal user behavior from different data sources, which may include, among others, HTTP requests, file access logs, logon events, and email communications. The paper aims to couple large-scale, NLP-driven feature extraction with machine learning models such as DistilBERT, ALBERT, RoBERTa, BERT, and GPT-4 for deep semantic understanding. Preprocessing specific to email data is done: prompt engineering, Named Entity Recognition, and Sentiment Analysis are used to capture the fine-grained patterns of user intent and emotional tone.

Then, these segregated features are provided as inputs to unsupervised anomaly detection using methods such as Isolation Forest and one-class SVM (OCSVM), that find deviations in normal user behavior patterns (Aldrich and Jain, 2013; Awad and Khanna, 2015). To evaluate the efficacy of the framework in effectively classifying behavioral anomalies typical of insider threats, different performance metrics such as accuracy, precision, recall, F1 score, False Positive Rate (FPR), and True Negative Rate (TNR) are used (Manning et al., 2008; Larose and Larose, 2015). We summarize the key contributions of this paper as the following:

- **NLP-Driven Anomaly Detection Framework:** This framework combines advanced NLP-driven feature engineering with unsupervised learning models, specifically Isolation Forest and One-Class SVM. These models effectively detect behavioral anomalies in user log activity by identifying deviations from statistical norms, providing

[a] https://orcid.org/0009-0004-7361-5396
[b] https://orcid.org/0000-0003-1124-2200
[c] https://orcid.org/0000-0001-8792-0413

Baghalizadeh-Moghadam, N., Cuppens, F. and Boulahia-Cuppens, N.
An NLP-Based Framework Leveraging Email and Multimodal User Data.
DOI: 10.5220/0013524000003979
In *Proceedings of the 22nd International Conference on Security and Cryptography (SECRYPT 2025)*, pages 168-178
ISBN: 978-989-758-760-3; ISSN: 2184-7711

a robust solution for insider threat scenarios.

- **Advanced NLP Feature Engineering:** We employ cutting-edge NLP techniques such as Named Entity Recognition (NER), sentiment analysis, and prompt engineering to extract semantic and emotional insights from communications. These features enhance the detection of nuanced insider behaviors that traditional methods might miss.
- **Empirical Validation and Performance Evaluation:** The efficacy of the framework is verified using a benchmark CERT 4.2 insider threat detection dataset, while performance metrics such as accuracy, precision, recall, and F1 score show its capability of detecting complex threat patterns with low false positives.

The remainder of this paper is organized as follows. Section 2 provides an overview of the background and related work. In Section 3, we describe the data set and the pre-processing steps required for our analysis. Section 4 details our proposed insider threat detection framework, outlining key components such as feature extraction and anomaly detection techniques. We present the experimental results along with their interpretations in Section 5, followed by a discussion of key findings and implications in Section 6. Finally, we conclude the paper in Section 7, summarizing our contributions and discussing future research directions.

2 BACKGROUND AND RELATED WORK

Insider threats continue to be some of the most critical challenges to organizational security, as they deal with actors with legitimate access to sensitive systems and data. These can be malicious in nature, such as data theft or sabotage, while others are unintentional, such as accidental leakage of sensitive information. These threats mostly evade traditional security measures like firewalls and intrusion detection systems since they leverage legitimate access paths (Cappelli et al., 2012). In order to solve these challenges, solutions must provide context that can reveal smaller behavior deviations that may indicate potential risks.

Recent advances stress adaptive methods that would continuously monitor active users by leveraging dynamic behavioral contexts to detect anomalies over time. This paper discusses a multilayered approach using ML, NLP, with capabilities that extend the conventional detection mechanisms. The proposed framework thus combines statistical anomaly detection with relational and semantic analysis to effectively identify insider threats in diverse organizational contexts.

2.1 Email Content-Based Detection on CERT Dataset

One approach to insider threat detection focuses on creating psychological profiles of employees based on sentiment analysis of email content. Using the CERT r4.2 dataset, Jiang et al. (Jiang et al., 2018) analyze the content of emails and browse history to build such profiles. By monitoring sentiment trends, it becomes possible to identify early signs of risky behaviors. While this method demonstrates potential use cases for insider threat prediction, comprehensive detection metrics are not provided.

Another line of research emphasizes anomaly detection through email content analysis. Garba et al. (Garba et al., 2021) provide a clustering-based technique, applied to the CERT r6.2 dataset, that involves preprocessing emails through tokenization and stopword removal, followed by the application of K-means clustering with Principal Component Analysis (PCA). This method achieves a detection rate of 89%, showcasing its effectiveness in identifying anomalous behaviors.

Mittal et al. (Mittal and Khurana, 2022; Mittal et al., 2023) present an approach based on Linear Discriminant Analysis (LDA) to reduce text length and Sequential Minimal Optimization (SMO) to understand the polarity of emails (i.e. identify emails with the highest weight in negative words). This approach outperforms traditional ML detection methods, but its ranking of emails by negativity only captures a limited understanding of emails and can miss more nuanced meanings of text.

2.2 NLP: A Paradigm Shift in Insider Threat Detection

Transformer-based NLP models such as BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and GPT-4 (OpenAI, 2023) have revolutionized insider threat detection by enabling deeper insights into subtle variations in communication patterns. These models go beyond simple keyword analysis, capturing context, semantics, and emotional undertones, essential elements for identifying nuanced anomalies.

This paper leverages the strengths of transformer-based models: BERT for its bidirectional context (Devlin et al., 2019), ALBERT for its memory efficiency

(Lan et al., 2019), RoBERTa for its robust pretraining (Liu et al., 2019), DistilBERT for its real-time suitability (Sanh et al., 2019), and GPT-4 for its capability of analyzing long-term behavioral trends (OpenAI, 2023). Combining the powers of these models empowers us to introduce one such framework that bridges the semantic analysis with anomaly detection, catering to both the depth and scalability of such an integration a perfect paradigm shift to arm the Enterprise to detect nuanced threats with precision and efficiency.

2.3 Feature and Prompt Engineering in Insider Threat Detection

Feature engineering and prompt engineering are important in uncovering the relevant patterns in structured and unstructured data when it comes to insider threat detection.

Feature engineering transforms structured activity logs, like login records, file access events, and HTTP requests, into measurable features that highlight deviations from normal behavior. Techniques such as Named Entity Recognition also extract key entities, while sentiment analysis extracts emotions and intent indicators from communication data (Grishman, 1997). These structured representations enable anomaly detection models to identify suspicious behavioral trends effectively.

Prompt engineering, on the other hand, amplifies unstructured textual analysis through the use of pretrained NLP models such as GPT-4. In place of feature extraction rules manually defined, task-oriented prompts in extracting urgency, authority, emotional tone, and security-sensitive phrases from user communications (Reynolds and McDonell, 2021). These prompts lead the NLP model to find subtle contextual patterns that could reveal insider threats.

2.4 Extracting Unsupervised Anomaly Detection in an NLP-Driven Framework

Integrating NLP-driven features with unsupervised anomaly detection models such as Isolation Forest and One-Class SVM (OCSVM) introduces proactive and adaptive dimensions to insider threat detection. Although in certain scenarios NLP provides an understanding of user intent and sentiment, unsupervised models point out deviations from established norms of behavior, a comprehensive framework that learns from every incoming data stream on a continuous basis. This is especially useful in settings where the volume of labeled data is small or simply does not exist. Coupled with our framework are insights from NLP, such as NER and sentiment scores, that empower anomaly detection models to dynamically adapt to emerging threat patterns without predefined rules (Aldrich and Jain, 2013; Awad and Khanna, 2015). Proactive detection is the ability of the framework to pick up early warning signals of abnormal behavior and deviation from set baselines before they build up into an insider threat. Consequently, this framework will be empowered to enable organizations to make highly dynamic discoveries of emerging patterns continuously through data streams in order to enable the detection of emerging potential threats in their nascent stage for effective intervention. The shift from reactive to proactive detection ascertains that an organization can mitigate risks and address anomalies before massive damage is caused.

3 DATASET

This research uses the CERT4.2 dataset, a benchmark dataset for insider threat detection research. The dataset is structured into specialized fields that capture various activity types across users, thus enabling comprehensive analysis of user behavior. These range from key areas of user interaction, such as email communications, log-on/log-off activities, file access, HTTP requests, and device connections/disconnections. The key fields are:

- **Emails:** Contains fields like *id_email*, *user*, *to*, *from*, *size*, *attachments*, and *content*, allowing for the analysis of communication patterns and potential data leakage. **It is important to note that the email content in this dataset does not correspond to real email text but consists of a list of keywords extracted from the original communication.**
- **Log-on/Log-off Activities:** Tracks session details with fields such as *id_logon*, *pc_logon*, and *activity*, supporting detection of abnormal access times.
- **File Access:** Captures file interactions through fields like *id*, *pc*, *filename*, and *content_file*, helping to monitor unauthorized access and data handling.
- **HTTP Requests:** Fields such as *id_http*, *pc_http*, *url*, and *content_http* facilitate tracking of web access patterns, useful for detecting risky external communications.
- **Device Activities:** Includes fields like *id*, *date*, *user*, *pc*, and *activity* (e.g., *Connect*, *Disconnect*),

enabling the detection of unauthorized device usage or unusual connection patterns.

As illustrated in Figure 1, these fields comprehensively capture user activities across multiple domains, making the dataset suitable for multi-faceted insider threat detection.

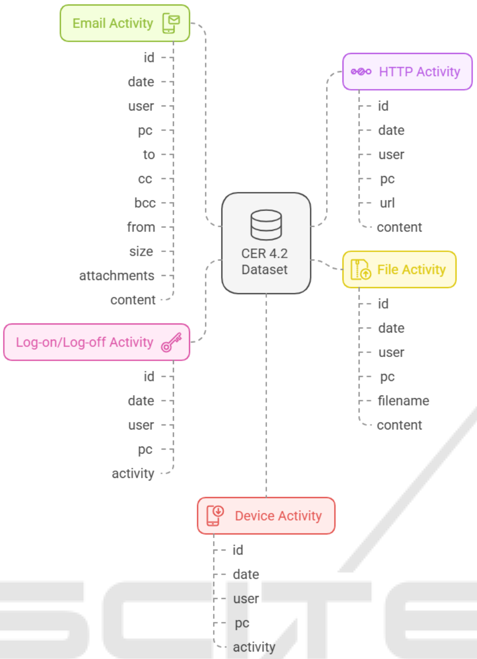


Figure 1: CERT4.2 Dataset Structure and Activity Types.

4 METHODOLOGY

4.1 NLP-Driven Unsupervised Anomaly Detection Framework

Our approach integrates NLP-driven features with unsupervised anomaly detection models, such as Isolation Forest and One-Class SVM, to construct an adaptive insider threat detection framework. The proposed framework is proactive; it monitors user activities continuously and learns from them to detect deviations from normal behavior before they escalate into security incidents. This is particularly useful in scenarios where labeled data is limited or absent. Our framework automatically adapts to newly emerging threat patterns without any predefined anomaly detection rules by utilizing the knowledge extracted from NLP, including named entities and sentiment scores. This is obtained in (Gamachchi et al., 2018).

The proposed framework and the related work represented by the graph-based anomaly detection system proposed by Gamachchi et al. in (Gamachchi et al., 2018) represent efforts toward leveraging structured relationships in user activity data for insider threat detection. For instance, the DTITD framework proposed by Wang and El Saddik (Wang and El Saddik, 2023) relies on self-attention mechanisms to enhance anomaly detection accuracy. Contrary to these, our framework tries to integrate NLP-derived features such as NER and sentiment analysis with unsupervised statistical models.

Feature engineering is indispensable in this framework for converting the raw user activity logs into representations that will aid in the detection of anomalies. We extract statistical patterns, contextual insights, and behavioral indicators such as NER and sentiment analysis to enable unsupervised models to detect deviations that might be difficult to catch in unprocessed data. Motivated by methods in Aggarwal and Yu (Aggarwal and Yu, 2015), we preprocess data to reflect both statistical patterns and contextual insights with a view to capturing effective representation for anomaly detection.

4.2 Holistic Daily User Activity Representation

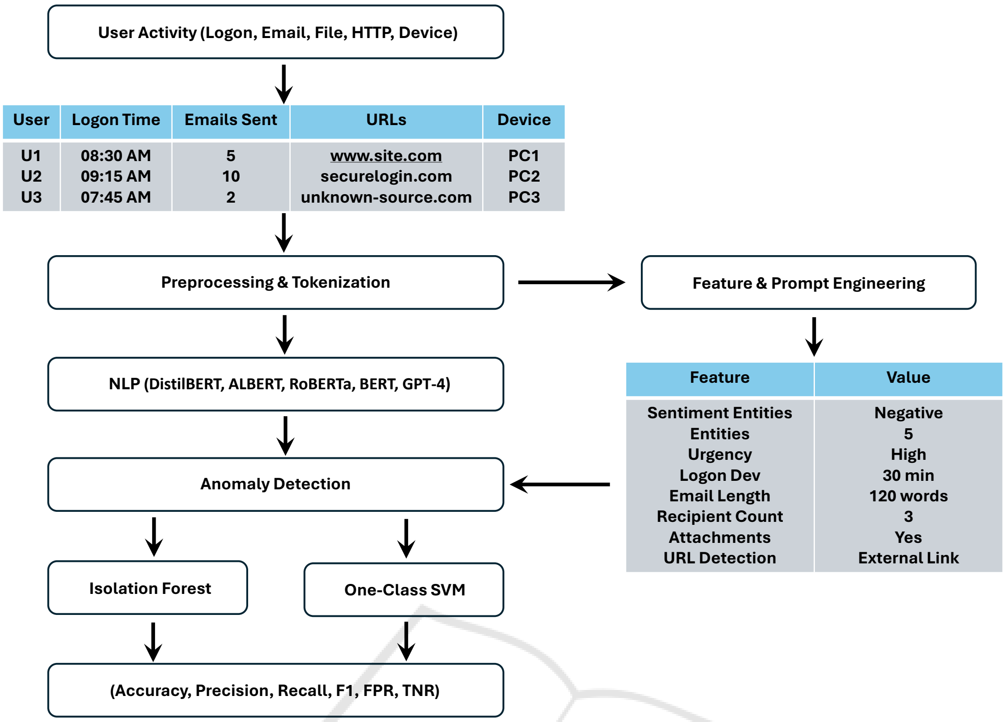
This work adopts a formatted and consistent approach to representing the daily activities of a user. Each row in the dataset corresponds to one user's activities for a single day. Various domains of activity, such as sending and receiving emails, file access, HTTP requests, and events related to logging on and off, are condensed into a single-row format. This representation captures the daily digital footprint of users, ensuring the dataset reflects the diversity and complexity of user behavior while providing a temporal perspective crucial for insider threat detection.

Each row consolidates key metadata and content from multiple sources, including email content and metadata (e.g., sender, recipients, attachments), file operations (e.g., read, write, delete), accessed URLs, and device activity logs. This unified representation encapsulates a user's multimodal activities within a day, offering a comprehensive view of their behavior. Such a structure enables natural language processing (NLP) models to extract semantic and contextual insights, facilitating the identification of deviations from normal patterns that may indicate anomalies or potential threats. In the following sections, we will delve into the role of NLP models in processing this structured data for anomaly detection.

4.2.1 Anomaly Detection Workflow

The anomaly detection framework incorporates various stages, starting with data sources like HTTP requests, file access, logon events, and email activity, followed by preprocessing, NLP models (e.g., BERT, RoBERTa), feature engineering, and the application of anomaly detection techniques, such as Isolation Forest and One-Class SVM (OCSVM). Each stage in the framework plays a crucial role in continuously monitoring and analyzing user behavior, allowing for the timely identification of anomalies. The workflow is illustrated in Figure 2.

1. **BERT - Bidirectional Encoder Representations from Transformers**
   - **Purpose:** To generate contextual embeddings that capture the relationships between different activities within a user's daily behavioral profile.
   - **Working:** BERT tokenizes each row in the dataset, representing a user's daily activities (including email metadata, file access, HTTP requests, and device activity). By analyzing this multimodal data, BERT uses its bidirectional attention mechanism to encode the contextual relationships among these activities, creating a unified representation of user behavior.
   - **Impact:** BERT effectively identifies subtle variations or anomalies in the daily activity patterns of users, such as unusual sequences of activities or deviations from typical behavior. This ability makes it a powerful tool for detecting insider threats and other anomalous behaviors (Devlin et al., 2019).
2. **RoBERTa (Robustly Optimized BERT Pretraining Approach)**
   - **Purpose:** To enhance contextual embeddings through optimized pretraining, enabling a deeper understanding of the interconnections among activities in a user's daily behavior.
   - **Working:** RoBERTa processes each row of user daily activity data, capturing nuanced details across long and complex activity sequences, such as interactions between email communication and file access or patterns of HTTP requests. Its advanced pretraining techniques, like dynamic masking, enhance its ability to detect subtle behavioral patterns.
   - **Impact:** By focusing on the relationships between different activities within a day, RoBERTa identifies irregularities in behavioral patterns that may indicate potential threats or deviations from the norm (Liu et al., 2019).
3. **DistilBERT**
   - **Purpose:** To achieve computational efficiency while retaining the ability to extract meaningful patterns from daily user activity data.
   - **Working:** DistilBERT tokenizes and processes rows of user daily activities, efficiently generating embeddings that summarize complex behavioral patterns across multimodal inputs. Its lightweight architecture allows real-time analysis of large-scale datasets.
   - **Impact:** DistilBERT supports scalable anomaly detection by rapidly identifying deviations in user behavior without compromising the quality of insights, making it ideal for handling high-dimensional datasets (Sanh et al., 2019).
4. **ALBERT (A Lite BERT)**
   - **Purpose:** To optimize memory usage and computational efficiency for processing large-scale datasets of user daily activities.
   - **Working:** ALBERT processes rows of daily user activity data using parameter-sharing and factorized embeddings to minimize computational overhead while preserving the quality of extracted representations.
   - **Impact:** ALBERT delivers robust performance in detecting anomalies in user behavior, even in resource-constrained environments, by leveraging its efficiency in handling large datasets (Lan et al., 2019).
5. **GPT-4**
   - **Purpose:** To perform advanced analysis and prompt engineering, extracting critical insights from user daily activities, such as identifying urgency, intent, or emotional undertones.
   - **Working:** GPT-4 analyzes patterns within the daily activity data, identifying key sequences or phrases that deviate from normal behavior. Its generative capabilities also provide detailed contextual expansion for better interpretability of flagged anomalies.
   - **Impact:** By analyzing the temporal and semantic aspects of daily user activities, GPT-4 enhances the detection of high-risk behaviors, such as emotionally charged communications or suspicious patterns of activity, supporting proactive anomaly detection (OpenAI, 2023).


User Activity (Logon, Email, File, HTTP, Device)
Preprocessing & Tokenization
Feature & Prompt Engineering
NLP (DistilBERT, ALBERT, RoBERTa, BERT, GPT-4)
Anomaly Detection
Isolation Forest
One-Class SVM
(Accuracy, Precision, Recall, F1, FPR, TNR)

User	Logon Time	Emails Sent	URLs	Device
U1	08:30 AM	5	www.site.com	PC1
U2	09:15 AM	10	securelogin.com	PC2
U3	07:45 AM	2	unknown-source.com	PC3

Feature	Value
Sentiment Entities	Negative
Entities	5
Urgency	High
Logon Dev	30 min
Email Length	120 words
Recipient Count	3
Attachments	Yes
URL Detection	External Link

Figure 2: Anomaly detection workflow illustrating the integration of NLP models, feature engineering, and detection methods.

4.3 Feature Engineering of User Activity Data

Feature engineering involves deriving meaningful quantitative and qualitative attributes from various user activity logs, such as email communications, HTTP requests, file access records, or logon/logoff events. The process transforms raw data into structured features that reflect behavioral patterns along with context. Feature engineering in the code is based on the following key aspects:

4.3.1 Feature Engineering Purpose

The intuition behind feature engineering involves converting raw user activity data into structured characteristics representative of explicit and implicit aspects of behavior in various domains. In such a feature-based approach, the model can:

- Identify statistical anomalies regarding user behavior within various activity logs, such as email size, number of recipients, frequency of file access, and URLs of HTTP requests.
- Identify contextual anomalies, such as tone and sentiment, or those with highlighted keywords, for example, urgent or emotional communication (Carta et al., 2020).
- Employ time-of-day and day-of-week characteristics to indicate behavior outside of typical work hours or on unusual days.
- Knowledge-based on the input domain, such as anomalies in file access or suspicious HTTP queries on an integrated platform that flags potentially anomalous behaviors deviating from established norms (Carta et al., 2020).

The structured approach strengthens the interpretability of the anomaly detection framework and hence allows effective discrimination between normal and suspicious behaviors.

For a summary of the feature engineering techniques used in this work, refer to Table 1.

4.4 Prompt Engineering

Prompt engineering in this context entails keyword, phrase, and pattern identification that conveys urgency, authority, sensitivity, or emotional tone. The semantic indicators herein contextualize the framework by allowing a deeper intent understanding of user actions across multiple domains, such as email, HTTP requests, file access, and logon/logoff data. The framework computes urgency, tone, and sensitivity metrics, bringing out suspicious activities while incorporating these scores as informative input features for anomaly detection. This structured approach

Table 1: Summary of Feature Engineering Techniques.

Category	**Details**
URL Detection	Flags external links in emails, HTTP queries, or file metadata to detect potential phishing or external communication.
Email Length	Measures the number of words in an email or text content, identifying unusually short or excessively long messages as potential anomalies.
Recipient Count	Calculates the total number of recipients in `To`, `CC`, and `BCC` fields for email communication to evaluate communication breadth.
File Access Frequency	Tracks the number of file accesses by a user in a specific period to identify unusual patterns.
Login Duration	Computes the time between logon and logoff events, flagging unusually short or long sessions as suspicious.
Day/Hour of Activity	Captures the day of the week and hour of activity, highlighting actions outside of normal working hours or unusual days.
Has Attachment	Flags whether an email or file activity includes attachments, which may indicate sensitive or malicious content.

helps enhance the interpretability of the proposed anomaly detection framework, whereby normal and suspicious behaviors can be effectively discriminated. Such features extracted by NLP-based prompt engineering are then combined with statistical features, such as the size of the email, count of recipients, or access frequency of files, and temporal attributes, such as out-of-hours activity, to obtain a complete feature matrix. The resulting feature matrix feeds into anomaly detection models.

Prompt engineering enables semantic insights to identify high-risk communications and activities related to unauthorized data sharing, emotionally charged messages, or unusual patterns in HTTP queries. The enrichment of the statistical and relational analyses forms a comprehensive understanding of user behavior and allows the enhancement of the detection of anomalous activities (Manning et al., 2008; Alsowail, 2021). As summarized in Table 2, the key components of prompt engineering involve identifying specific indicators and extracting relevant features for anomaly detection.

4.5 Named Entity Recognition (NER)

Named Entity Recognition (NER) allows the framework to extract and categorize named entities, such as people, organizations, locations, and sensitive terms, from user communications across a variety of domains: email contents, HTTP requests, and file metadata. This feature underlines the frequency or unusual patterns of named entities that may be indicative of suspicious activity. For example, unduly high mentions of individuals, projects, or terms of secrecy could point to an insider threat. By identifying such entities for scrutiny, NER further improves the detection of communications outside established norms and thereby advances the anomaly detection framework (Lample et al., 2016; Neumann, 2012).

4.5.1 Sentiment Analysis

Sentiment analysis determines the emotional tone of user communications by classifying messages as positive, neutral, or negative. This shall be applied to text-based activity logs, including e-mail content and HTTP queries, to catch behavior signals such as frustration, urgency, and satisfaction. A rapid increase in negative sentiment may indicate dissatisfaction or frustration, possibly tied to risky behaviors. At the same time, unusually positive sentiment in critical or sensitive contexts would indicate manipulation or deceit. These sentiment scores are integrated into the anomaly detection framework to inject a behavior dimension that allows the models to detect subtle emotional cues that might signal anomalous or high-risk activity (Alsowail, 2021).

The following code snippet illustrates our approach to extracting features from email content, focusing on Named Entity Recognition (NER) and sentiment analysis. NER captures entity-specific information such as individuals and organizations, while sentiment analysis gauges the emotional tone, aiding in behavioral insights crucial for anomaly detection. For a detailed look at the code used to extract features from user communications, including Named Entity

Table 2: Summary of Prompt Engineering Features for Anomaly Detection.

Prompt Engineering	
Keyword Detection	Identifies sensitive terms like ”urgent,” ”password,” and ”credentials” to detect high-risk communication.
Urgency Metrics	Extracts urgency-related keywords such as ”asap,” ”immediately,” and ”urgent” to assess the priority of communication.
Sensitivity Metrics	Detects critical terms like ”confidential,” ”classified,” and ”proprietary” to identify potentially sensitive exchanges.
Tone Analysis	Captures tone indicators such as formal (e.g., ”please,” ”sincerely”) or urgent (e.g., ”deadline,” ”immediate”) patterns.
Sentiment Analysis	Assigns a sentiment score (positive, negative, neutral) to communications to detect emotionally charged messages.
Suspicious Indicators	Flags phrases like ”verify,” ”confirm,” or ”account breach” to identify security risks or unusual activity.
Named Entity Recognition	Extracts entities (e.g., names, organizations) from text, highlighting abnormal frequency or content patterns.

Recognition (NER) and sentiment analysis, please refer to the appendix.

4.6 Anomaly Detection Using Isolation Forest and One-Class SVM (OCSVM)

The approach presented in this paper is based on ensemble classification, which combines the Isolation Forest and One-Class SVM (OCSVM) models for detecting insider threats. Insider threats in complex organizational ecosystems often evade detection and pose significant security challenges, frequently hiding in plain sight (Cappelli et al., 2012). Traditional detection methods fail to capture subtle yet critical distinctions between normal and potentially malicious insider behavior (Colwill, 2009).

This paper proposes an extended, layered anomaly detection framework: leveraging ensemble classification techniques to dynamically analyze user behavior from multiple perspectives. These range from considering individual actions across diverse sources such as file access logs and login sessions down to HTTP requests and email communication (Miller, 2020). By combining the advantages of Isolation Forest and OCSVM, both normal and abnormal behaviors are given improved detection rates in this work, thus developing a robust solution towards insider threat detection.

4.6.1 Hybrid Anomaly Detection with Ensemble Classification

We use the ensemble approach by combining the Isolation Forest and OCSVM methods on the context-aware features extracted from user logs, emails, HTTP requests, and file access events (Miller, 2020). We have identified semantic and statistical patterns rather than purely data-driven features for enhancing the threat detection accuracy of our approach.

- **Structured Features:** Include statistical deviations such as logon time variations, email recipient count, file access frequency, and unusual HTTP request patterns.
- **NLP-Derived Features:** Named entities, changes in sentiment, urgency detection, and security-sensitive phrases extracted via transformer-based analysis.

These enhanced features are then fed into the classifier ensembler, which, for better performance of anomaly detection.

4.6.2 Isolation Forest: Detecting Statistical Anomalies

Isolation Forest is fit for detecting anomalies via recursive partitioning of feature space (Liu et al., 2008). This approach efficiently isolates the rare event showing:

- Unusual login times, which can be evidenced by access during nonworking hours.
- Sudden file modification spikes or email activities that hint at potential data exfiltration.
- Changes in session lengths or activity bursts as indicative of abnormal user behavior.

4.6.3 One-Class SVM: Capturing Behavioral Deviations

One-Class SVM bridges the gap by modeling a flexible boundary around normal behavior and, consequently, labeling temporal deviations over time

(Schölkopf et al., 2001). Besides, it can serve in various cases like the detection of the following events:

- Gradual changes in file access.
- Changes in the style of communications, such as urgency indicators in emails.
- HTTP access anomaly trends: repetition in visits to security-sensitive URLs (Colwill, 2009).

5 RESULTS

The performance of the proposed NLP-driven anomaly detection framework was evaluated on the CERT4.2 dataset using metrics such as **Accuracy**, **Precision**, **Recall**, **F1 Score**, **False Positive Rate (FPR)**, and **True Negative Rate (TNR)**. These metrics offer a comprehensive evaluation of the models, measuring both their ability to correctly identify anomalies (**Recall**) and their reliability in minimizing false alarms (**Precision** and **FPR**). The key findings, summarized in Table 3, are discussed below:

- **RoBERTa:** RoBERTa achieves a strong **Accuracy** of 95.98% and a high **F1 Score** of 92.40%. It also exhibits a high **Precision** of 91.82% and a low **FPR** of 2.43%, indicating a strong balance between detecting anomalies and minimizing false positives. With a **Recall** of 93.13% and a **TNR** of 97.57%, RoBERTa is particularly suitable for applications where both precision and recall are important, such as detecting anomalies with a low tolerance for false negatives.
- **DistilBERT:** With an outstanding **Recall** of 93.86%, DistilBERT is unparalleled in its ability to identify nearly all anomalies in the dataset. Its **Precision** of 91.85% and **Accuracy** of 97.77% indicate strong overall performance, although its slightly higher **FPR** of 2.46% compared to ALBERT suggests a marginally increased rate of false positives. This makes DistilBERT suitable for scenarios where the cost of missing an anomaly is significantly higher than the cost of investigating false positives.
- **BERT:** BERT achieves a solid **Accuracy** of 96.21% and a **Precision** of 89.53%, indicating its reliability in minimizing false positives. Its **Recall** of 92.13% and **F1 Score** of 90.81% suggest balanced performance, though it is slightly lower across some metrics compared to ALBERT and RoBERTa. BERT's low **FPR** of 1.32% and high **TNR** of 98.68% further emphasize its strength in minimizing false alarms.
- **GPT-4 :** GPT-4 achieves an **Accuracy** of 94.15%, with a balanced **Precision** of 88.58% and a **Recall** of 89.11%. However, its **F1 Score** of 88.84%, **FPR** of 2.18%, and **TNR** of 97.82% suggest it is less effective than ALBERT, RoBERTa, and DistilBERT in minimizing false positives while maintaining high detection rates.
- **DBN OCSVM:** The prior model, DBN OCSVM, achieves an **Accuracy** of 87.79% and a **Recall** of 81.04%. However, it lacks values for **Precision**, **F1 Score**, and **TNR**, highlighting its limited capability in comparison to the proposed NLP-driven models.
- **PCA OCSVM:** PCA OCSVM exhibits the lowest **Accuracy** of 79.66% and a **Recall** of 77.20%, with a high **FPR** of 20.33%. This underscores its relatively poor performance in anomaly detection compared to more advanced models.

6 DISCUSSION

The results of this study have validated the huge potential of the proposed NLP-driven framework in insider threat detection. Indeed, the application of state-of-the-art NLP models such as ALBERT and DistilBERT allowed for the accurate detection of subtle behavioral anomalies. Its lightweight architecture means that ALBERT is computationally efficient, particularly for large-scale datasets, yet it achieved the highest precision of 92.82% and the lowest false positive rate of 1.02%. DistilBERT demonstrated the highest F1 score (92.85%) and recall (93.86%), making it particularly suitable for scenarios where identifying the majority of anomalies is critical.

RoBERTa also fared impressively, with an F1 score of 92.40%, balancing high precision at 91.82% with recall at 93.13%. Though slightly weaker in performance compared to DistilBERT on the F1 score, its robust metrics do mean that this will be suitable for detecting a wide range of anomalies with minimal false alarms.

The combination of Isolation Forest and One-Class SVM proved effective in handling anomalies of diverse patterns. While the isolation forest was efficient in finding the isolated outliers, such as anomalies in usual login times or sudden spikes in file access activities, the One-Class SVM identified gradual boundary anomalies, thus spotting subtle deviations in behavior. This complementary approach helped the framework to handle obvious and not-so-obvious threats with aplomb.

Advanced feature engineering techniques and prompt-based analysis considerably improved the

Table 3: Comparison of detection results between proposed models and prior work. 'NA' indicates that the value is not recorded. **Bold** indicates the best value obtained for each metric.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**FPR (%)**	**TNR (%)**
DistilBERT	97.77	91.85	**93.86**	**92.85**	2.46	97.54
ALBERT	**98.98**	**92.82**	89.00	90.88	**1.02**	**98.98**
RoBERTa	95.98	91.82	93.13	92.40	2.43	97.57
BERT	96.21	89.53	92.13	90.81	1.32	98.68
GPT-4	94.15	88.58	89.11	88.84	2.18	97.82
DBN OCSVM (Lin et al., 2017)	87.79	NA	81.04	NA	12.18	NA
PCA OCSVM (Lin et al., 2017)	79.66	NA	77.20	NA	20.33	NA

performance of the framework. Features extracted from NER and sentiment analysis provided contextually deeper features by capturing the entities, emotional tone, and intent in user communications. These features allowed the detection of anomalous behaviors that may be difficult to identify using traditional methods. While large models such as GPT-4 led to furthering contextual understanding even by finding latent intents like urgency and emotional distress from texts, together with prompt engineering, these techniques were contributing to holistic understanding in user activities.

Indeed, the proposed framework outperformed traditional approaches, such as DBN OCSVM and PCA OCSVM, by a large margin on all metrics with a significant decrease in false positives. This underlines the importance of integrating state-of-the-art NLP-driven insights with unsupervised anomaly detection techniques to develop more robust and accurate detection systems.

While these results are promising, there are some limitations of the framework that need to be considered in future work: the initial fine-tuning of the NLP model using labeled data limits the generalization for domains with poor training data. Second, while the framework has proved good in offline detection, it remains to be seen if this framework will prove suitable for real-time applications. Adaptive learning methods can improve the ability of the framework to evolve with emerging threat patterns in dynamic organizational environments.

7 CONCLUSION

This paper proposes a new framework that combines NLP-driven feature engineering with ensemble anomaly detection methods, namely Isolation Forest and One-Class SVM, for the effective detection of insider threats. By embedding both structured statistical and unstructured semantic features from various user activities such as email communication, file access, and HTTP requests, the framework enhances the detection of behavioral anomalies that could signal potential threats.

It has been evaluated extensively on the CERT4.2 dataset, and the results are that the framework demonstrated remarkable performance; ALBERT and DistilBERT generated the highest precision and recall, while RoBERTa maintained a balanced performance. These results confirmed the capability of the proposed framework in finding sensitive changes of user behavior while suppressing false positives, an important factor for practical deployment within real-world environments.

In the future, efforts will go into making the framework continuous with real-time monitoring to cater for improvements in evolving threats. Further scalability research in more organizational contexts through integrating data sources from other areas, such as collaboration tools and social media, extends the generalisability of this framework. It lays a foothold in making systems of Insider Threat Detection resilient and scalable using the merged strength of NLP techniques with unsupervised learning models.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to Mitacs, National Bank, Desjardins, Mondata, and Qohash for their invaluable support and contributions to this research.

REFERENCES

Aggarwal, C. C. and Yu, P. S. (2015). *Outlier Analysis*. Springer.

Aldrich, C. and Jain, R. (2013). Anomaly detection in systems using data mining techniques. *Journal of Computer Science*, 9(5):501–512.

Alsowail, M. (2021). A framework for insider threat detection in organizations. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(5):641–648.

Awad, M. and Khanna, R. (2015). *Machine Learning for Big Data: Hands-On for Developers and Technical Professionals*. Apress.

Borky, J. M. and Bradley, T. H. (2018). *Effective Model-Based Systems Engineering*. Springer.

Cappelli, D. M., Moore, A. P., and Trzeciak, R. F. (2012). *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud)*. Addison-Wesley Professional.

Carta, S. et al. (2020). Local feature selection for anomaly detection in user activity logs. *Journal of Machine Learning Research*, 21:1059–1087.

Colwill, C. (2009). Insider threats in the cyber security context. *Cybersecurity Review*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gamachchi, A., Lakshmanan, V., and Mathur, A. (2018). Graph-based anomaly detection in user activity data for insider threat detection. *Journal of Cybersecurity and Privacy*, 5(2):205–218.

Garba, M., Bello, F., and Lawal, S. (2021). Email anomaly detection using clustering techniques: A case study on cert insider threat datasets. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(5):200–212.

Grishman, R. (1997). Information extraction: Techniques and challenges. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, 1299:10–27.

Jiang, W., Li, Y., and Chen, P. (2018). Prediction of insider threats using psychological profiling based on email content analysis in cert dataset. *Journal of Cybersecurity Research*, 10(3):45–57.

Lample, G., Conneau, A., Ranzato, M., and Denoyer, L. (2016). Neural machine translation with attention mechanism. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2336–2345.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Larose, D. T. and Larose, C. D. (2015). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.

Lin, X., Zhang, W., and Wang, J. (2017). A study on principal component analysis for anomaly detection. In *2017 International Conference on Green Informatics*, pages 345–350. IEEE.

Liu, F., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. *Data Mining and Knowledge Discovery*, 17(3):411–421.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Miller, D. (2020). Anomaly detection for insider threat detection. *Journal of Cybersecurity*.

Mittal, P. and Khurana, N. (2022). Proposed insider threat detection framework using email sentiment analysis and machine learning techniques. *International Journal of Cybersecurity and Digital Forensics*, 15(4):78–92.

Mittal, P., Khurana, N., and Sharma, R. (2023). Prediction and detection of insider threats using lda and sentiment polarity analysis. *Journal of Information Security and Applications*, 36(1):14–27.

Neumann, P. (2012). *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes*. Addison-Wesley Professional.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond few-shot learning. *arXiv preprint arXiv:2102.07350*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Support vector machine for novelty detection. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 582–588. MIT Press.

Wang, R. and El Saddik, A. (2023). Dtitd: Deep transformer-based insider threat detection framework. *IEEE Transactions on Information Forensics and Security*, 18:123–135.