SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Homepage: https://ssrpublisher.com/ssrjai/

Volume 3, Issue 6, 2026 ISSN: 3049-0413

Insider Threat Detection in Enterprise Communication Logs Using Transformer-Based NLP Models

Adiele Joshua Eze[1], Amangi-Edomo Andabi Benita[2], Brisibe Beauty Oroma[3]

Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State, Nigeria

Received: *05.05.2026* | ***Accepted:*** *25.05.2026* | ***Published:*** *01.06.2026*

***Corresponding author:** Adiele Joshua Eze

DOI: *10.5281/zenodo.20492394*

Abstract **Review Article**

Insider threats represent a critical challenge in cybersecurity, often bypassing traditional perimeter defenses due to their legitimate access privileges. This paper proposes a transformer-based natural language processing (NLP) framework for detecting insider threats through the analysis of enterprise communication logs. Leveraging contextual embeddings and behavioral features, the model identifies anomalous patterns indicative of malicious intent. Experiments conducted on the Enron email dataset demonstrate the efficacy of the proposed approach, achieving an F1-score of 0.87 in threat classification. The study also addresses ethical considerations and proposes a human-in-the-loop review mechanism for deployment.

Keywords: Insider Threat Detection, Natural Language Processing (NLP), Transformer Models, Enterprise Communication Logs, Behavioral Analysis, Cybersecurity, BERT, Anomaly Detection, Sentiment Analysis, Email Forensics.

1. Introduction

Insider threats malicious activities conducted by individuals within an organization pose significant risks to data integrity, confidentiality, and availability. Unlike external attacks, insider threats are harder to detect due to the legitimate access of the perpetrators. With the proliferation of digital communication platforms, analyzing textual data such as emails, chat logs, and reports offers a promising avenue for threat detection [4].

Natural Language Processing (NLP) has emerged as a powerful tool for extracting insights from unstructured text. Recent advances in transformer-based models like BERT and RoBERTa have enabled deeper semantic understanding, making them suitable for detecting subtle indicators of insider threats [1].

2. Related Work

The detection of insider threats has been extensively studied through a variety of approaches, including rule-based systems, statistical anomaly detection, and machine learning classifiers. Traditional methods often rely on structured data sources such as system logs, access records, and audit trails, which provide quantifiable behavioral indicators but lack semantic depth [4, 2]. For instance, Greitzer et al. [4] proposed a psychological profiling

framework that integrates structured logs with behavioral indicators to identify potential insider threats. Similarly, Eberle and Holder [2] employed graph-based anomaly detection techniques to uncover suspicious access patterns within structured datasets.

More recent efforts have explored the application of deep learning to insider threat detection. Zhang et al. [8] utilized Long Short-Term Memory (LSTM) networks to analyze email communication patterns. While their approach marked a shift toward leveraging unstructured data, it did not incorporate contextualized language representations, limiting its ability to capture nuanced linguistic cues.

Despite these advancements, the use of transformer-based Natural Language Processing (NLP) models such as BERT or GPT for insider threat detection remains underexplored. A recent study by Elbasheer and Akinfaderin [3] demonstrated the effectiveness of transformer encoders in modeling sequential user behavior using the CERT dataset, achieving state-of-the-art performance. However, their work primarily focused on structured activity sequences rather than unstructured communication content.

This paper builds upon these foundations by integrating behavioral signals with linguistic features extracted from unstructured communication data using transformer-based NLP models [1]. By doing so, it aims to bridge the gap between structured behavioral analytics and semantic understanding of insider intent.

3. Methodology

3.1 Data Collection

This study employs the publicly available Enron Email Corpus, which contains over 500,000 real-world email messages exchanged among employees of the Enron Corporation [6]. The dataset has become one of the most extensively utilized benchmark datasets in insider threat and organizational communication research because of its scale, authenticity, and diversity of communication patterns [1]. The corpus provides a realistic representation of enterprise communication behavior, thereby making it suitable for the analysis of malicious insider activities and anomalous interactions.

Prior to analysis, the dataset undergoes a rigorous preprocessing procedure to enhance data quality and analytical reliability. Duplicate emails, spam messages, automatically generated responses, corrupted entries, and irrelevant metadata are removed to ensure a clean and representative dataset. Additionally, email headers, timestamps, sender-recipient relationships, and message content are standardized to facilitate consistent feature extraction and behavioral modeling.

3.2 Preprocessing

To prepare the textual data for machine learning and natural language processing tasks, a multi-stage preprocessing pipeline is implemented. The preprocessing stages are designed to reduce textual noise, normalize linguistic structures, and enhance semantic interpretability. The pipeline includes the following components:

a. **Tokenization and Lemmatization:** Email text is segmented into tokens using standard NLP tokenization techniques. Lemmatization is subsequently applied to reduce words to their base or root forms, thereby minimizing lexical variability and improving semantic consistency across documents.

b. **Named Entity Recognition (NER):** Named Entity Recognition techniques are employed to identify and extract entities such as person names, organizations, locations, dates, and financial references. This process enables the capture of semantic and contextual relationships embedded within organizational communication patterns.

c. **Sentiment and Emotion Annotation:** Emotional and affective characteristics of emails are analyzed using the VADER sentiment analyzer [5] and the NRC Emotion Lexicon [7]. Sentiment polarity scores (positive, negative, and neutral) as well as emotional categories such as anger, fear, trust, and anticipation are extracted to support behavioral profiling and anomaly detection.

d. **Text Cleaning and Normalization:** Special characters, hyperlinks, stop words, and excessive whitespace are removed during preprocessing. Email texts are also converted to lowercase to ensure uniformity and reduce dimensional inconsistencies during feature extraction.

3.3 Feature Engineering

Feature engineering is conducted to capture the linguistic, behavioral, and contextual characteristics associated with insider threat activities. The extracted features are grouped into three major categories:

a. **Linguistic Features:** These features include sentiment polarity scores, emotional intensity indicators, urgency-related expressions, lexical diversity, and syntactic patterns derived from email content [5]. Such features assist in identifying potentially suspicious or emotionally charged communications.

b. **Behavioral Features:** Behavioral indicators are extracted from communication metadata, including email frequency, communication intervals, time-of-day activity, recipient diversity, and interaction networks [4]. These features enable the modeling of employee communication behavior and the identification of abnormal interaction patterns.

c. **Contextual Embeddings:** Deep contextual representations of email content are generated using Bidirectional Encoder Representations from Transformers (BERT) [1]. The embeddings capture semantic dependencies and contextual relationships within textual data, thereby improving the model's ability to detect subtle insider threat indicators.

3.4 Model Architecture

This study proposes a hybrid insider threat detection framework that integrates both semantic and behavioral intelligence for improved detection accuracy. The architecture combines supervised and unsupervised machine learning techniques to capture known malicious behaviors and previously unseen anomalies.

a. **BERT-Based Semantic Representation:** BERT is utilized to generate contextual embeddings from email content. The transformer-based architecture enables the extraction of deep semantic patterns and contextual relationships within organizational communications [1].

b. **Random Forest Classification:** A Random Forest classifier is employed as the primary supervised learning model for structured feature classification. The algorithm is selected because of its robustness against overfitting, ability to handle high-dimensional data, and interpretability in identifying feature importance.

c. **Isolation Forest for Anomaly Detection:** Isolation Forest is implemented as an unsupervised anomaly detection technique to identify unusual communication behaviors and outlier activities that may indicate insider threats [2]. The algorithm isolates anomalous observations based on deviations from normal behavioral patterns.

The integration of semantic embeddings with behavioral analytics provides a comprehensive detection mechanism capable of identifying both explicit and covert insider threat activities within enterprise communication systems.

3.5 Evaluation Metrics

The performance of the proposed framework is evaluated using widely accepted classification and anomaly detection metrics to ensure comprehensive assessment and model reliability. The evaluation metrics include:

a. **Precision:** Measures the proportion of correctly identified insider threat instances among all predicted threat instances.

b. **Recall:** Evaluates the model's ability to correctly identify actual insider threat cases within the dataset.

c. **F1-Score:** Provides the harmonic mean of precision and recall, ensuring balanced evaluation of classification performance.

d. **Receiver Operating Characteristic – Area Under Curve (ROC-AUC):** Assesses the discriminative capability of the model across varying classification thresholds.

e. **Confusion Matrix:** Visualizes classification outcomes by presenting true positives, true negatives, false positives, and false negatives, thereby enabling detailed error analysis and performance interpretation.

f. **Accuracy:**
Measures the overall proportion of correctly classified instances within the dataset.

4. Results and Discussion

Model	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.81	0.78	0.79	0.84
BERT + RF Hybrid	0.88	0.86	0.87	0.91
Isolation Forest	0.74	0.69	0.71	0.79

The hybrid model outperforms traditional approaches, demonstrating the value of combining linguistic and behavioral features. Case studies reveal that flagged emails often contain emotionally charged language, unusual timing, or references to sensitive topics [8].

5. Ethical Considerations

The deployment of NLP-based insider threat detection and surveillance systems within organizational environments introduces significant ethical, legal, and privacy-related concerns. Although such systems can enhance organizational security and reduce the risk of malicious insider activities, inappropriate implementation may result in employee mistrust, discrimination, and violations of fundamental privacy rights. Consequently, ethical safeguards must be integrated into every stage of system design, development, and deployment.

5.1 Privacy and Data Protection

The analysis of employee communications, particularly emails and organizational messaging data, raises substantial privacy concerns. To address these issues, all sensitive and personally identifiable information (PII) should be anonymized prior to processing and analysis. Data collection and monitoring activities must comply with established data protection regulations and organizational policies, ensuring that employees are adequately informed about monitoring practices and consent procedures where applicable. Furthermore, access to collected data should be restricted to authorized personnel, and secure storage mechanisms must be implemented to prevent unauthorized disclosure or misuse of information.

5.2 Bias and Fairness

Machine learning and NLP models are susceptible to biases embedded within training datasets, feature engineering processes, and algorithmic decision-making mechanisms. Biased datasets may disproportionately target or misclassify certain individuals or groups, thereby leading to unfair or discriminatory

outcomes. To mitigate such risks, this study emphasizes the use of balanced and representative datasets, fairness-aware model evaluation techniques, and periodic bias assessments. In addition, model outputs should be continuously monitored to identify and correct potential discriminatory behaviors during operational deployment.

5.3 Transparency and Explainability

Transparency is a critical requirement for the responsible deployment of AI-driven surveillance systems. Employees and organizational stakeholders should have a clear understanding of the objectives, limitations, and operational procedures of the monitoring framework. To enhance accountability and trust, explainable AI (XAI) techniques and human-in-the-loop review mechanisms should be incorporated into the detection process. Human oversight ensures that automated alerts and anomaly classifications are reviewed and validated before administrative or disciplinary actions are taken.

5.4 Ethical Deployment Framework

To support responsible implementation, this study proposes an ethical deployment framework that incorporates the following principles:

a. Periodic Audits: Regular technical and ethical audits should be conducted to evaluate system performance, fairness, privacy compliance, and operational transparency.

b. Employee Feedback Mechanisms: Organizations should establish communication channels through which employees can express concerns, provide feedback, and seek clarification regarding monitoring practices.

c. Accountability and Governance: Clear governance structures and accountability policies should be defined to regulate data access, decision-making authority, and incident response procedures.

d. Minimal Intrusion Principle: Monitoring activities should be limited strictly to security-related objectives, ensuring that employee privacy is not unnecessarily compromised.

By integrating these ethical safeguards, organizations can achieve a balance between effective insider threat detection and the preservation of employee rights, trust, and organizational integrity.

6. Conclusion

This study presents a novel NLP-based framework for insider threat detection using transformer models. The integration of linguistic and behavioral features significantly improves detection accuracy. Future work will explore multilingual datasets, real-time detection, and integration with enterprise security systems.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019). https://doi.org/10.48550/arXiv.1810.04805
2. Eberle, W., Holder, L.: Discovering structural anomalies in graph-based data. In: 2007 Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 393–398. IEEE (2007). https://doi.org/10.1109/ICDMW.2007.104
3. Elbasheer, M., Akinfaderin, A.: Enhancing insider threat detection using user-based sequencing and transformer encoders. arXiv preprint arXiv:2506.23446 (2025). https://doi.org/10.48550/arXiv.2506.23446
4. Greitzer, F.L., Kangas, L.J., Noonan, C.F., Brown, C.M., Ferryman, T.A.: Psychosocial modeling of insider threat risk based on behavioral and word use analysis. In: 2013 45th Annual IEEE International Carnahan Conference on

Security Technology (ICCST), pp. 42–49. IEEE (2013). https://doi.org/10.1109/CCST.2013.6679567

5. Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14), pp. 216–225. AAAI (2014).
6. Klimt, B., Yang, Y.: Introducing the Enron corpus. In: CEAS 2004 – Conference on Email and Anti-Spam (2004).
7. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Comput. Intell. 29(3), 436–465 (2013). https://doi.org/10.1111/j.1467-8640.2012.00460.x
8. Zhang, Y., Khan, M.A.R., Haq, M.A., Alshehri, M.: Insider threat detection based on NLP word embedding and machine learning. J. Inf. Secur. Appl. 63, 103030 (2021). https://doi.org/10.1016/j.jisa.2021.103030