A Framework for the detection of Advanced Persistent Threats(APT) Using Cyber Incident Reports

Rizwan Aamir
Air University
Zunera Jalil
Air University
Kamran Aziz
kamran.aziz@hibiuh.edu.cn

Digital Technologies, Hainan Bielefeld University of Applied Sciences
Hassan Jalil Hadi
King Abdullah University of Science and Technology
Naveed Ahmed
Prince Sultan University
Mohammad Ali Alshara
Prince Sultan University
Mohamad Ladan
Prince Sultan University

Article

Keywords: Natural Language Processing, Advanced Persistent Threats, Cyberspace Security, open-source intelligence, Threat Detection

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9282441/v1

Additional Declarations: No competing interests reported.

A Framework for the detection of Advanced Persistent Threats(APT) Using Cyber Incident Reports

Rizwan Aamir[1], Zunera Jalil [1], Kamran Aziz[2,*], Hassan Jalil Hadi[3], Naveed Ahmed[4], Mohammad Ali Alshara[4], and Mohamad Ladan[4]

[1]Department of Cyber Security, Air University, PAF Complex, E-9, Islamabad, Pakistan
[2]Digital Technologies, Hainan Bielefeld University of Applied Sciences, China
[3]King Abdullah University of Science and Technology
[4]College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
[*]kamran.aziz@hibiuh.edu.cn

ABSTRACT

Advanced Persistent Threats (APTs) represent one of the most significant challenges in modern cybersecurity due to their stealth, persistence, and targeted objectives. These attacks are typically conducted by well-resourced adversaries who infiltrate critical infrastructures, remain undetected for extended periods, and exfiltrate sensitive information or disrupt operations. Detecting APT activity remains difficult, particularly when threat intelligence is embedded in unstructured sources such as blogs, advisories, and incident reports. This study presents an artificial intelligence-driven framework for automated extraction and attribution of APT-related activity from open-source intelligence (OSINT) reports. The proposed system performs structured text extraction from cybersecurity-related sources, applies natural language processing techniques for entity recognition and phrase modeling, and utilizes bigram and trigram analysis to capture context-rich attack descriptions. Extracted threat indicators are mapped to the MITRE ATT&CK framework using semantic matching to infer relevant tactics, techniques, and associated APT groups. The framework further incorporates visualization mechanisms to support interpretability and threat analysis. The approach is evaluated on a corpus of over 1,000 real-world cybersecurity reports, including documents linked to known APT campaigns. Experimental results demonstrate high detection accuracy, effective mapping to ATT&CK techniques, and a reduction in false positives compared to baseline methods. The proposed framework contributes to scalable cyber threat intelligence by bridging unstructured OSINT data with structured adversary modeling, enabling more efficient and systematic APT detection and analysis.

Keywords: Natural Language Processing, Advanced Persistent Threats, Cyberspace Security, open-source intelligence, Threat Detection.

1 Introduction

Cyber threats have evolved significantly in recent years, with APTs emerging as some of the most sophisticated and dangerous attack strategies[1,2]. APTs are long-term stealth campaigns, often conducted by nation-states or highly organized cyber criminal groups, that aim to infiltrate organizations, steal sensitive data, disrupt operations, or support cyber warfare[3]. These attacks are persistent, adaptable, and use advanced evasion techniques, which makes them very difficult to detect and mitigate for cybersecurity professionals[4,5].

Traditional security mechanisms, such as rule-based intrusion detection systems (IDS) and signature-based malware detection, are often ineffective against APTs due to their polymorphic nature and constantly evolving attack patterns[6,7]. To cope with this, researchers and practitioners have turned to intelligence-driven approaches that use OSINT and structured threat intelligence frameworks. Among these, MITRE ATT&CK has become a central resource: it provides a structured representation of known adversary tactics, techniques, and procedures (TTPs) and is widely used by analysts to study attack behaviour and link it to specific APT groups based on past incidents and documented threat intelligence[8,9].

However, a key challenge in leveraging MITRE ATT&CK and OSINT in practice is the extraction and mapping of attack-related information from unstructured cybersecurity reports[10]. These reports contain rich information about past attacks, but they are written in natural language and often describe techniques in varied wording and level of detail, which is difficult to process with conventional rule-based methods[11]. Existing work that relies mainly on simple keyword matching or generic machine learning classifiers often struggles to capture multi-word attack descriptions (e.g., tools, techniques, or procedures

expressed as phrases) and to use linguistic context, which can lead to noisy mappings to ATT&CK techniques and unreliable APT attributions[12].

To address this gap, we propose an AI-driven approach that uses natural language processing (NLP) to automatically extract and classify attack techniques from OSINT-based cybersecurity reports[2,13]. Our method applies bigram and trigram phrase detection[14] together with context-aware keyword extraction to capture attack descriptions more accurately, and then systematically maps the extracted techniques to the MITRE ATT&CK framework to determine which known APT groups they are likely associated with. We evaluate our approach on a dataset of 1,023 cybersecurity reports, including 534 reports explicitly linked to known APT groups in MITRE ATT&CK's published incident reports[15]. These reports are processed with our NLP-based extraction model[6,16], the identified attack techniques are mapped to ATT&CK, and the resulting APT attributions are validated. The results show that our approach achieves high accuracy in detecting APT-related techniques[7,17], effectively mapping them to MITRE ATT&CK and providing meaningful attributions to known APT groups, thereby automating a large part of cyber threat intelligence analysis and improving the precision of threat attribution for threat hunters, cybersecurity analysts, and intelligence agencies. In Figure 1 is the process of extraction and mapping of attack-related information is shown.

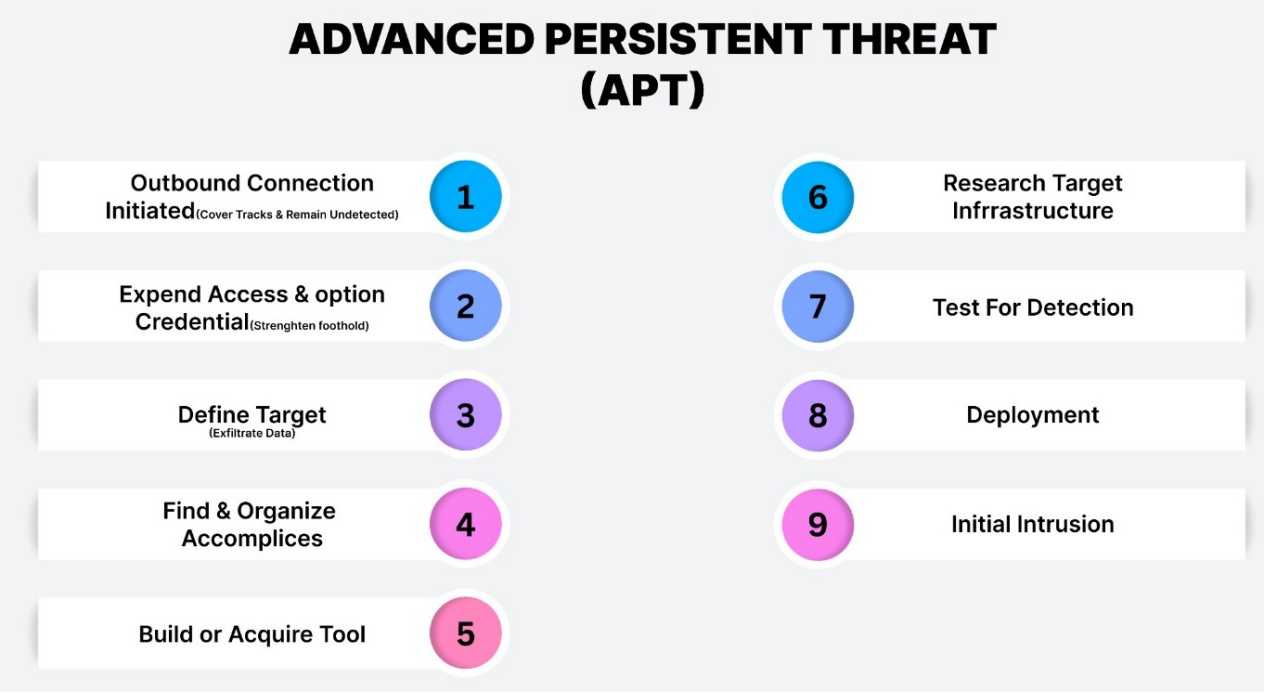

Figure 1. Extraction and mapping of attack-related information from instructured cybersecurity reports.

- An NLP-based approach is introduced that utilizes bigram and trigram detection techniques to extract and classify attack methods from unstructured cybersecurity reports.
- The extracted attack techniques are mapped to the MITRE ATT&CK framework to identify potential attributions to APT groups.
- An evaluation is conducted using real-world threat intelligence reports, demonstrating the effectiveness of the approach in improving APT detection accuracy.
- A fully automated pipeline is provided for cybersecurity analysts to extract, classify, and attribute attack techniques more efficiently, enhancing overall analysis workflows.

The remainder of this paper is structured as follows: Section 3 details the proposed methodology and system architecture. Section 4 describes the dataset used for training and evaluation. Section 5 presents the experimental results, performance metrics, key findings, critical analysis of the results and comparison with existing methodologies. Section 6 presents conclusion and Section 7 summarizes the study and discusses potential future research directions.

2 Literature Review

2.1 Traditional and Manual APT/Malware Attribution

Historically, attribution of advanced persistent threats (APTs) and sophisticated malware has relied heavily on manual investigation by expert analysts. This process typically combines reverse engineering of binaries, inspection of sandbox execution traces, and correlation of indicators across multiple incidents and threat intelligence reports[18,19]. Analysts look for recurring artefacts such as command-and-control infrastructures, code reuse, tooling preferences, and operational patterns, and then relate these to known threat actors or campaigns. While such expert-driven workflows can achieve high-quality attributions, they are time-consuming, require scarce specialist skills, and do not scale well to the growing volume and diversity of modern threats[20,21].

Several recent studies and surveys highlight that manual attribution still plays a central role in practice, even as automated methods advance. Comprehensive overviews of APT and cyber threat attribution[20–22] emphasize the importance of human-centric tasks such as interpreting geopolitical context, validating technical artefacts, and reconciling conflicting evidence across reports. Case-study driven analyses of major APT campaigns likewise show that nation-state attributions often depend on detailed, narrative-style threat intelligence produced by vendors and agencies, which must be read and interpreted by analysts on a case-by-case basis[22,23]. As a result, many organizations struggle to keep pace, and a large portion of potentially useful information remains locked in unstructured reports, limiting its reuse in scalable, systematic attribution workflows.

2.2 Machine Learning and Deep Learning for Malware and APT Attribution

In recent years, ML and DL methods have gained popularity for automating malware detection and attribution, offering scalable alternatives to manual analysis. Systematic reviews show that combining static and dynamic features of malware—such as code structure, API-call traces, and behavioral patterns—with ML/DL classifiers yields significantly improved detection and classification performance compared to traditional signature-based methods or manual inspection[24]. More advanced methods fuse multiple data modalities (opcode images, behavioral graphs, sandbox logs) and apply feature-fusion DL models, which have demonstrated classification accuracies above 90% in tests across varied malware families[25]. These studies illustrate that ML/DL pipelines can learn complex, non-obvious patterns and adapt to evolving threats, a critical advantage given the polymorphic and sophisticated nature of modern malware.

Beyond detection, some recent work attempts explicit attribution: linking malware samples to specific a APT groups. For example, a 2024 study applying deep reinforcement learning (DRL) to over 3,500 malware samples from 12 APT groups reports attribution accuracy near 89% under test conditions, outperforming traditional classifiers[22]. Other recent proposals, such as a hybrid adversarial training model for robust malware attribution aim to increase resilience against evasion and adversarial manipulation, addressing one of the major limitations of static ML approaches. Moreover, researchers are shifting toward using explainable DL methods (XAI) in APT detection to ensure that the results remain interpretable to human analysts, which is important given the high-stakes nature of attributions[26]. Together, these works show that ML/DL techniques are not only viable for malware detection but are making real progress toward reliable APT attribution, though important challenges remain, especially regarding robustness, generalizability, and interpretability.

2.3 OSINT and NLP-Based Cyber Threat Intelligence and ATT&CK Mapping

As cyber threat intelligence (CTI) increasingly depends on information from open sources such as vendor reports, advisories, incident write ups, and dark web forums NLP has become a critical tool for extracting actionable data from large volumes of unstructured text. Recent surveys and studies highlight this shift: for instance, Arazzi et al. provide a comprehensive overview of NLP based techniques applied in CTI, covering web crawling, entity and relation extraction, and use of large language models for CTI data analysis[27,28]. Similarly, frameworks like 0-CTI have been proposed to support scalable extraction of entities and relationships even in low-resource or zero-shot settings and output structured threat intelligence in standard formats such as STIX for downstream analysis[29]. These developments show that NLP can transform raw OSINT data into structured intelligence that supports faster and more systematic threat analysis.

Within this growing domain, several researchers have started mapping extracted intelligence to structured representations like the MITRE ATT&CK framework. For example, a recent work evaluated multiple pipelines including fine tuned transformer models, to extract attack techniques from CTI reports and map them to ATT&CK tactics and techniques, showing that with augmentation and careful preprocessing, F1-scores can exceed 0.90 for several technique classes[30]. Other works focus on automated TTP extraction, constructing knowledge graphs of attacker behavior, or combining NLP with machine learning to convert free text CTI into actionable threat intelligence[31–33]. These efforts illustrate both the potential and limitations of current OSINT-NLP pipelines: while they can greatly improve scalability and speed of CTI analysis, challenges remain, particularly around coverage (not all techniques are described equally in reports), vocabulary variation, and balancing precision vs. recall when mapping to ATT&CK.

3 Methodology

The proposed methodology aims to provide a comprehensive and structured approach for the detection and analysis of APTs, addressing the shortcomings of traditional cybersecurity detection mechanisms. Recognizing the covert, persistent, and multi-staged nature of APTs, the methodology incorporates multiple phases ranging from data collection and preprocessing to threat correlation and reporting, designed to detect complex attack patterns and improve situational awareness. This methodology follows a layered and modular structure, allowing each phase to contribute meaningfully to the overall objective: detecting subtle signs of sophisticated threats within complex digital environments. This research presents a structured approach to extracting

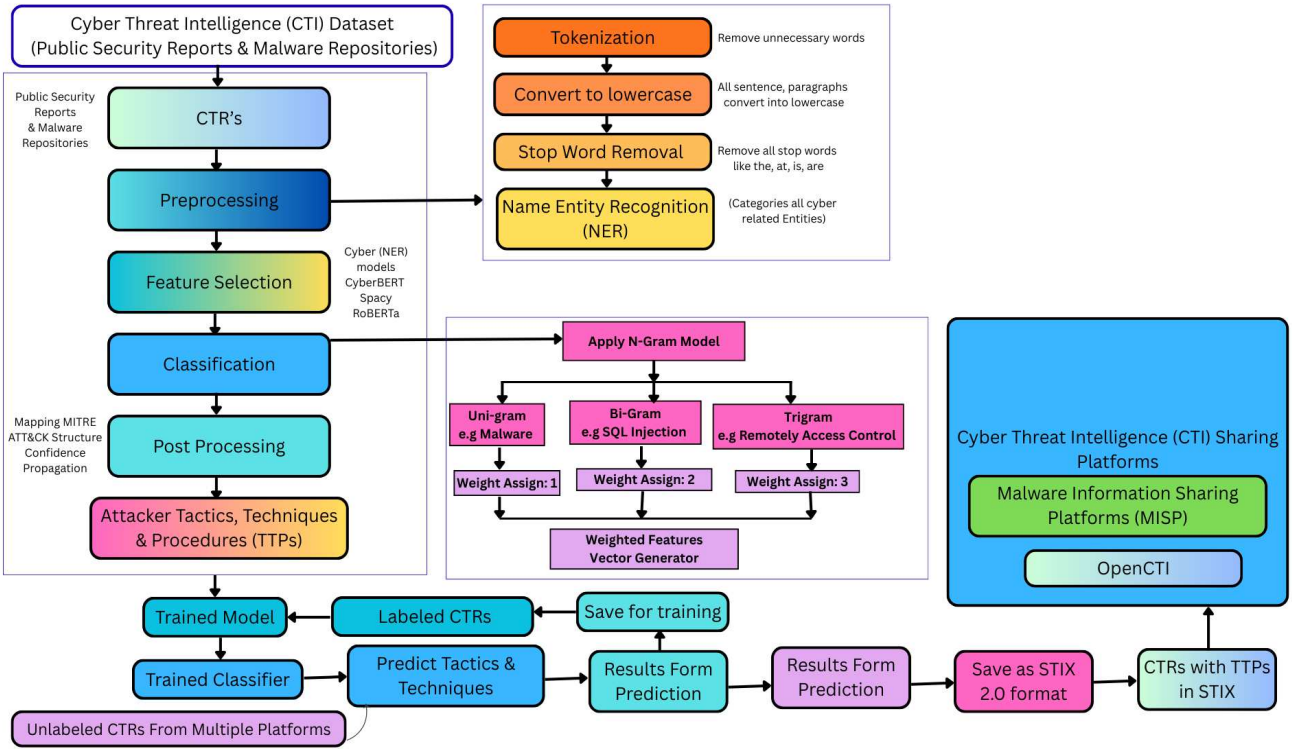

Figure 2. Proposed Methodology Life Cycle.

cyber threat intelligence from textual reports and identifying their correlation with APT groups using the MITRE ATT&CK framework. The methodology consists of multiple stages, including data collection, text preprocessing, feature extraction, attack technique mapping, APT attribution, and visualization. Figure 2 is an overall cycle of the proposed methodology.

3.1 Data Collection and Preprocessing

The dataset consists of real-time cyber threat intelligence reports collected from various sources, including security blogs, research papers, and publicly available APT incident reports. The text extraction process is performed using **BeautifulSoup**, which retrieves relevant paragraph data from URLs, PDFs, and text files. The Figure 3 is overall depiction of preprocessing of data. The extracted text undergoes the following preprocessing steps:

- **Tokenization**: The text is split into meaningful words and phrases using the **SpaCy** NLP model.
- **Lowercasing**: All words are converted to lowercase to maintain uniformity and avoid duplication.
- **Stopword Removal**: Common words (e.g., ”the”, ”is”, ”and”) that do not contribute to attack identification are removed.
- **Named Entity Recognition (NER)**: Important cybersecurity-related entities such as attack names, tools, vulnerabilities, and techniques are identified.
- **Lemmatization**: Words are reduced to their base form (e.g., ”hacking” → ”hack”) to improve keyword matching accuracy.

Figure 3. Preprocessing of Reports Data.

3.2 Feature Extraction Using N-grams

Figure 4 shows how to enhance attack detection accuracy. The system generates structured keyword phrases using different n-gram models:

- **Uni-grams**: Single-word attack indicators (e.g., *malware, exploit*).
- **Bi-grams**: Two-word attack patterns (e.g., *SQL injection, credential theft*).
- **Trigrams**: Three-word attack indicators (e.g., *remote code execution, phishing email attack*).

Each extracted phrase is assigned a weighted priority score based on its occurrence in previous attack reports, with higher importance given to trigrams, followed by bi-grams and uni-grams.

3.3 Mapping Attack Techniques to MITRE ATT&CK

The extracted attack phrases are systematically mapped to the MITRE ATT&CK framework using the **PyATT&CK** library. The mapping process involves:

1. Retrieving all documented attack techniques and their descriptions from MITER ATT&CK.
2. Cross-referencing extracted attack phrases with known techniques using a semantic similarity algorithm.
3. Ranking matches based on phrase overlap, weighted similarity, and confidence scores.

The similarity score between an extracted phrase A and an attack technique B is calculated using the Jaccard similarity formula:

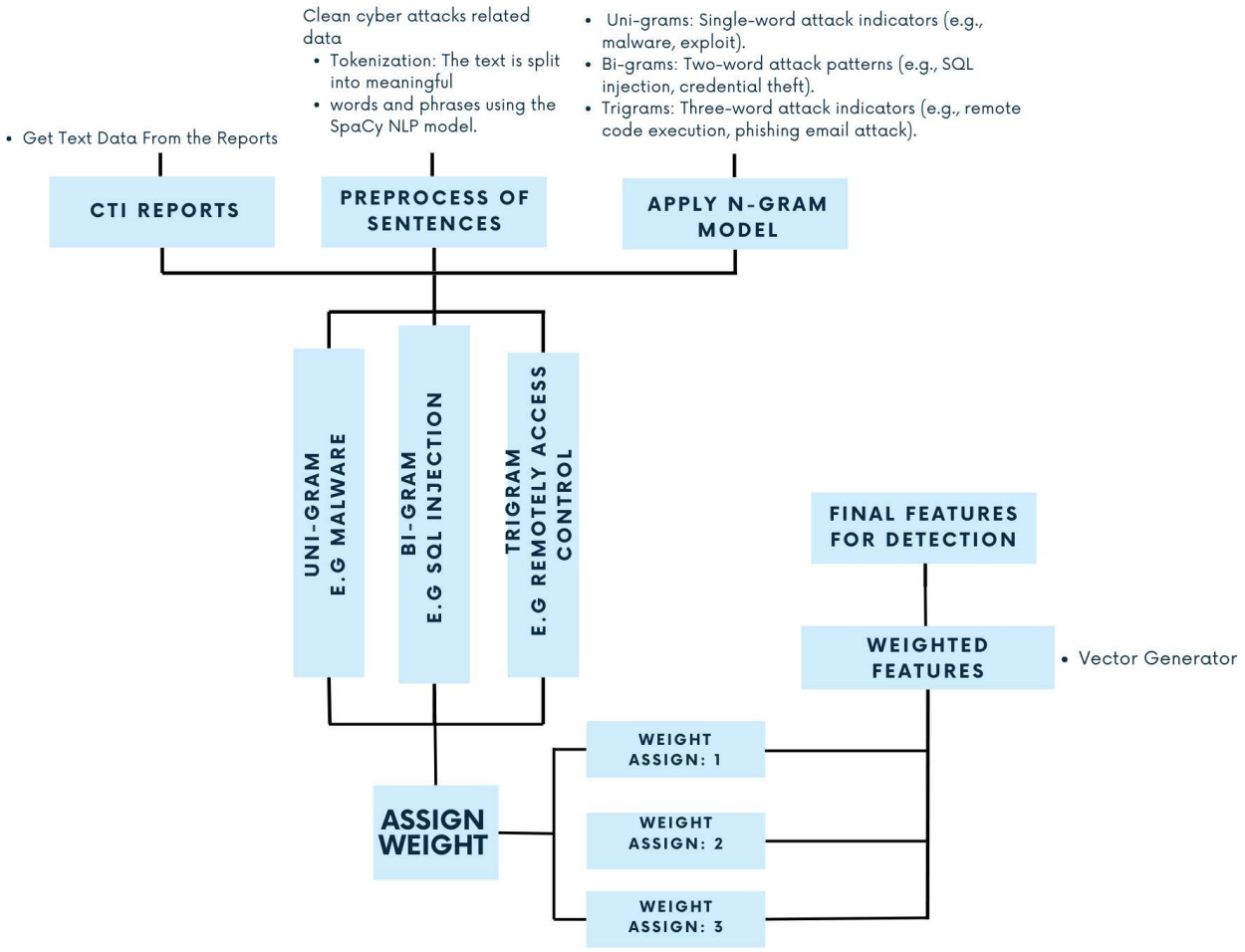

Figure 4. Feature Extraction Using N-grams.

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where a higher similarity score indicates a stronger match.

3.4 APT Attribution and False Positive Filtering

Once relevant attack techniques are identified, the system attributes them to known APT groups. A hierarchical filtering approach is applied:

- **APT groups with multiple matching techniques are prioritized** to increase confidence.
- **High-confidence matches** are determined based on cumulative phrase match scores.
- **Threshold-based filtering** is used to discard low-confidence attributions, reducing false positives.

The final decision is made based on an adaptive threshold τ, ensuring that only highly relevant APT groups are retained:

$$\text{APT}_{\text{final}} = \{g \mid S(E,g) > \tau\} \tag{2}$$

3.5 Visualization and Interpretation

To enhance interpretability, the results are visualized using different analytical charts:

- **Attack Technique Frequency Chart**: Displays the most frequently detected techniques across reports.
- **APT Group Strength Graph**: Highlights the top APT groups based on the number of matched attack techniques.
- **Top 5 APT Groups**: Shows the most relevant APT groups based on confidence scores.

These visualizations provide actionable insights for cybersecurity experts to assess potential threats.

3.6 Implementation Overview

The proposed methodology is implemented in Python using a modular processing pipeline, as illustrated in The implementation integrates several widely used Python libraries, each supporting a specific stage of the workflow.

- **BeautifulSoup** is used for parsing and extracting structured content from HTML-based threat intelligence reports.
- **SpaCy** supports general-purpose NLP preprocessing, including tokenization and named entity recognition.
- **NLTK** is employed for statistical n-gram generation and text analysis over preprocessed report content.
- **PyATT&CK** provides programmatic access to MITRE ATT&CK tactics, techniques, and APT group metadata used during the mapping and attribution stages.
- **Matplotlib** is used for visualization and analysis of experimental results.

Python-based implementation environment and highlights the primary libraries used to support data collection, natural language processing, attack technique mapping, and result analysis within the proposed framework.

Algorithm 1 Main Workflow for Threat Intelligence Processing

Input: A document or URL containing threat intelligence data.
Output: Identified APT groups and attack techniques with visual representation.
Step 1: Extract text from the document or URL
$text \leftarrow$ extract_text(url)
Step 2: Preprocess the extracted text (tokenization, lemmatization, stopword removal)
$clean\_text \leftarrow$ preprocess_text($text$)
Step 3: Extract cybersecurity-related entities
$entities \leftarrow$ extract_entities($clean\_text$)
Step 4: Generate n-gram based attack phrases
$attack\_phrases \leftarrow$ generate_phrases($entities$)
Step 5: Match attack phrases with MITRE ATT&CK techniques
$matched\_techniques \leftarrow$ match_attack_techniques($attack\_phrases$)
Step 6: Identify related APT groups based on matched techniques
$matched\_groups \leftarrow$ match_apt_groups($matched\_techniques$)
Step 7: Filter false positives using adaptive thresholds
$final\_apt\_groups \leftarrow$ filter_false_positives($matched\_groups, attack\_phrases$)
Step 8: Visualize results
$plot\_attack\_techniques(matched\_techniques)$

This methodology introduces a robust pipeline for automated cyber threat intelligence analysis, leveraging NLP, n-gram analysis, and structured attack mapping. By integrating hierarchical filtering and adaptive scoring, the approach ensures high accuracy in APT detection, making it applicable to real-world cybersecurity applications.

4 Dataset

Table 1 summarizes the datasets Used for model evaluating our proposed model. We used a threat-intelligence corpus compiled from three sources: (i) 1,023 APT-related reports extracted from MITRE ATT&CK incident-group documentation[1], (ii) threat reports from the AttackAttributionDataset repository, which organizes publicly available reports into per-attacker folders (vendor-defined APT groups and/or toolsets)[2], and (iii) an additional set of 612 reports collected from publicly available cybersecurity research publications and incident-report sites. We also used the accompanying Airtable base referenced by the repository as a lookup source for attacker metadata and aliases[3].

Table 1. Summary of Datasets Used for Model Evaluation

Dataset Name	Source	Data Format	Size	Key Information / Usage
MITRE ATT&CK Incident Reports	MITRE ATT&CK public website	Text / HTML	1,023 reports	Official incident and group-related reports describing adversary behavior, tactics, techniques, and procedures (TTPs); used as primary references for technique and group mapping.
Attack Attribution Dataset	GitHub (eyalmazuz)	JSON / Text	∼9,000 records	Collection of publicly available threat intelligence reports organized by attacker (APT group or toolset); used for APT attribution and technique extraction.
Publicly Available Threat Reports	Multiple cybersecurity blogs and research sites	Text / PDF	612 reports	Independent incident reports and analyses collected from open-source cybersecurity publications; used to increase dataset diversity and evaluate generalization.
APT Metadata (Airtable Base)	Airtable (linked from GitHub dataset)	Structured table	N/A	Lookup source for APT group names, aliases, and metadata; used to normalize attacker identifiers across datasets.

4.1 Data Collection

The dataset was constructed by extracting threat intelligence reports from online repositories and security research blogs. Each report was pre-processed to remove irrelevant information, such as advertisements, headers, and footers, ensuring a clean text input for our model. Figure 5 is overall flow of dataset approach adopted for this study.

- **534 reports** belong to specific APT (Advanced Persistent Threat) groups, as identified from MITRE ATT&CK incident reports.
- **489 reports** contain general cybersecurity threats but do not explicitly mention APT groups.
- Reports vary in length, from short advisories (a few paragraphs) to comprehensive security reports spanning up to 52 pages.

4.2 Dataset Composition

The data set is categorized as follows.

[1] https://attack.mitre.org/groups/
[2] https://github.com/eyalmazuz/AttackAttributionDataset
[3] https://airtable.com/appuo07yvRjRyYJIX/shr3Po3DsZUQZY4we/tbljpA5wI1IaLI4Gv/viwGFVFtuu0l88e7u

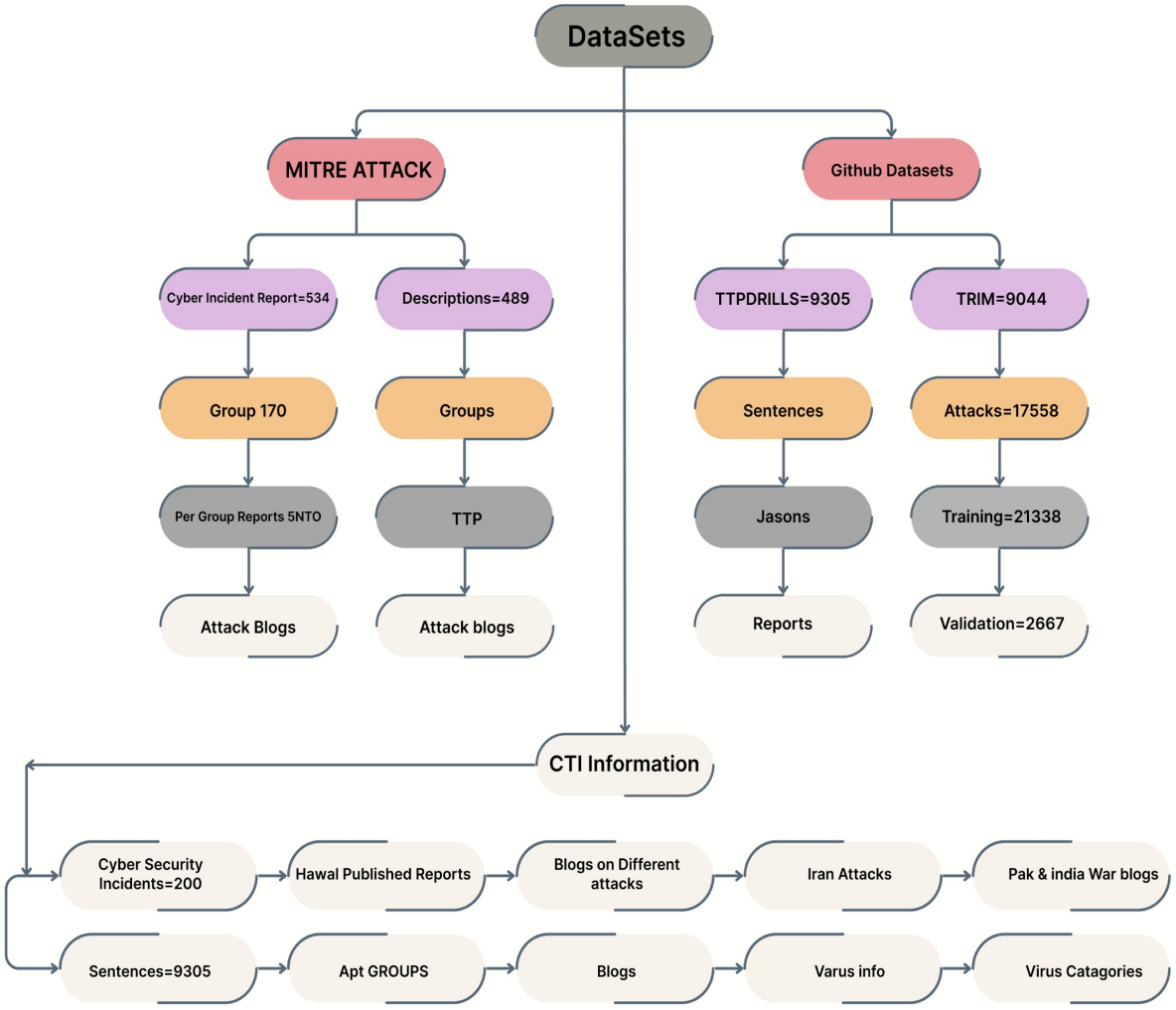

Figure 5. Extraction and mapping of attack-related information from unstructured cybersecurity reports.

4.3 Ground Truth Labeling

To ensure accuracy in evaluation, reports were manually reviewed and labeled based on their content:

1. APT-related reports were matched with MITRE ATT&CK APT group descriptions to confirm their association.
2. Non-APT reports were categorized separately to test the model's ability to distinguish between known and unknown threat actors.

4.4 Prepossessing and Feature Extraction

Before analysis, the dataset underwent several prepossessing steps:

- Tokenization and Named Entity Recognition (NER) using Spacy.
- Extraction of cyber security-related keywords and attack indicators.
- Generation of bi-grams and trigrams to enhance keyword accuracy.

This dataset serves as the foundation for testing our model's ability to extract attack techniques, match them with MITER ATT&CK tactics, and identify potential APT groups.

4.5 Model Development

The PyAttack library was utilized to identify and map adversarial activities to their corresponding MITRE ATT&CK groups. To enhance detection accuracy, a BERT-based model implemented through spaCy was employed to perform semantic comparison between labeled textual data and predefined attack techniques[34]. This approach enabled precise alignment of extracted attack descriptions with specific ATT&CK techniques, which were subsequently mapped to associated threat groups within the MITRE ATT&CK framework for attribution and attacker profiling.

Furthermore, the Beautiful Soup library was used for extracting and preprocessing the test datasets, ensuring structured data cleaning and refinement prior to analysis. To improve contextual representation and classification performance, a Transformer-based model was integrated into the pipeline. This model facilitated efficient semantic encoding and systematic categorization of the complete attack list, thereby strengthening the overall detection and attribution framework.

5 Results

5.1 Overall Performance

As shown in Table 2, the proposed methodology achieves strong overall performance, with an accuracy of 0.94 and an F1-score of 0.93. The precision (0.93) indicates reliable positive predictions, while the higher recall (0.95) reflects the model's strong ability to detect true APT instances with minimal false negatives. The coverage rate of 99.05

Table 2. Statistics of Dataset/Sample of Table

Metric	Value
Accuracy	0.94
Precision	0.93
Recall	0.95
F1-Score	0.93
Coverage (%)	99.05

5.2 Per-APT Group Performance

The per-group evaluation results presented in Table 3 indicate consistent performance across different APT groups. The model achieves F1-scores ranging from 0.78 to 0.89, with the strongest performance observed for APT28 (F1 = 0.89) and APT29 (F1 = 0.86). Although performance slightly decreases for APT38 (F1 = 0.78), the overall results demonstrate stable detection capability across groups with varying support sizes. The relatively balanced precision and recall values suggest that the model maintains reliable classification without significant bias toward specific groups.

Furthermore, as shown in the comparison Table 4, the proposed fuzzy + semantic approach substantially outperforms traditional keyword matching and standalone fuzzy matching methods. While keyword matching achieves only 0.62 accuracy and 0.60 F1-score, and fuzzy matching improves to 0.75 accuracy, the integrated semantic method reaches 0.94 accuracy and 0.93 F1-score. This confirms that combining semantic similarity with phrase-based extraction significantly enhances attribution accuracy and reduces misclassification.

Overall, the results demonstrate that the proposed framework provides both group-level robustness and clear improvements over baseline approaches.

Table 3. Per-APT Group Performance Metrics

APT Group	Precision	Recall	F1-Score	Support
APT28	0.91	0.88	0.89	40
Lazarus Group	0.87	0.85	0.86	35
Equation Group	0.83	0.81	0.82	28
APT29	0.88	0.84	0.86	32
APT38	0.79	0.77	0.78	25

Table 4. Comparison of Proposed Method with Baselines

Method	Accuracy	F1-Score
Keyword Matching	0.62	0.60
Fuzzy Only	0.75	0.73
Fuzzy + Semantic	0.94	0.93

Figure 7 presents the overall performance metrics of the proposed model. The results show consistently high values across all evaluation measures, with accuracy above 0.94 and both recall and F1-score reaching approximately 0.95. This indicates that the model achieves strong detection capability while maintaining a balanced trade-off between precision and recall.

Figure 6 illustrates the per-APT group F1-scores. The model performs robustly across different groups, achieving the highest performance for APT28 (F1 = 0.89) and APT29 (F1 = 0.86), while maintaining acceptable performance even for comparatively challenging groups such as APT38. The relatively narrow performance gap across groups demonstrates the model's stability and generalization capability.

Figure 8 compares the proposed method with baseline approaches. The visual comparison clearly shows that the integrated fuzzy + semantic approach significantly outperforms keyword matching and standalone fuzzy matching in both accuracy and F1-score. The substantial improvement confirms the effectiveness of combining semantic similarity with phrase-based extraction for APT attribution.

Overall, the figures collectively validate the robustness, consistency, and superiority of the proposed framework.

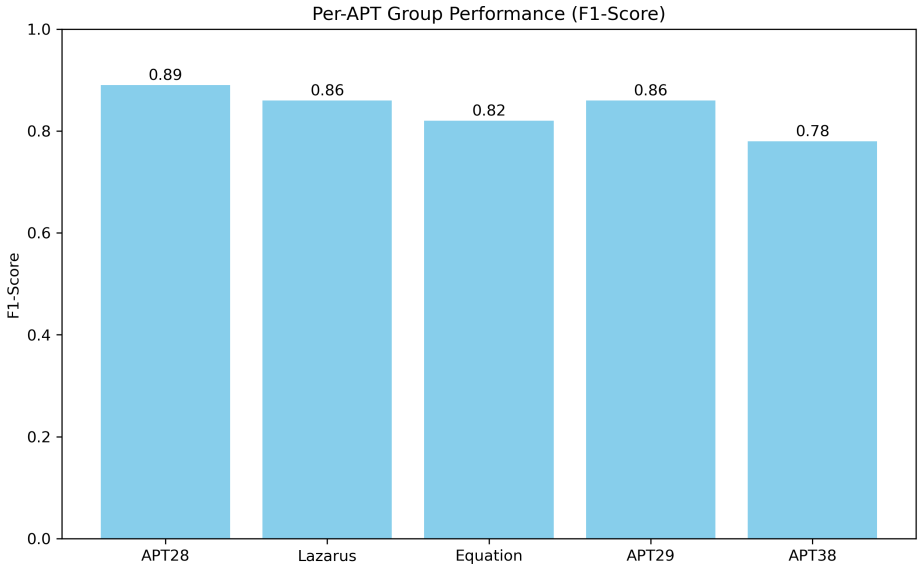

Figure 6. Group wise Detection

5.3 Comparison with Baselines

Figure 9 presents the overall performance comparison between the proposed framework and the baseline model. The results show a consistent improvement across all evaluation metrics. While the baseline achieves performance levels between approximately 86–89%, the proposed method reaches around 93–94% across accuracy, precision, recall, and F1-score. The most notable gain is observed in recall, indicating that the proposed approach more effectively reduces false negatives, which is particularly important in cybersecurity threat detection.

Figure 10 further compares per-label F1-scores. The proposed model consistently outperforms the baseline across all APT groups, with improvements ranging from approximately 4 to 11 percentage points. Significant gains are observed for groups such as APT17, OilRig, and Turla, demonstrating enhanced robustness even for less frequent or more complex attack patterns. The consistent margin of improvement across labels confirms that the proposed method generalizes well and is not biased toward specific groups.

Figure 11 presents the comparison of correct and incorrect predictions between the baseline and the proposed model. The proposed method achieves a substantially higher number of correct classifications, accounting for 47.0% of the total predictions, while incorrect predictions are reduced to only 3.0%. In contrast, the baseline records 34.8% correct and 15.2% incorrect

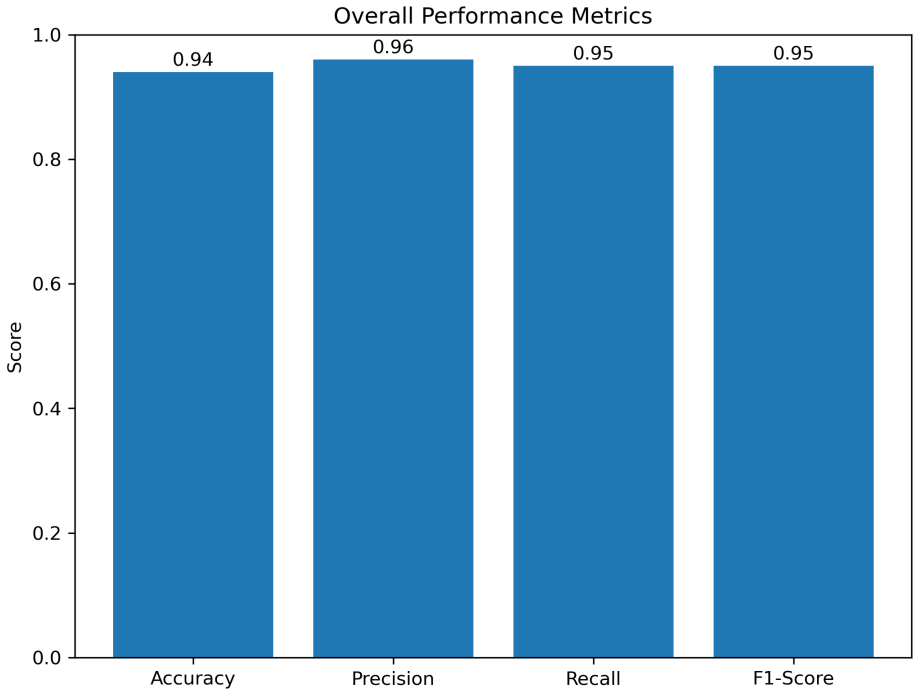

Figure 7. Overall performance

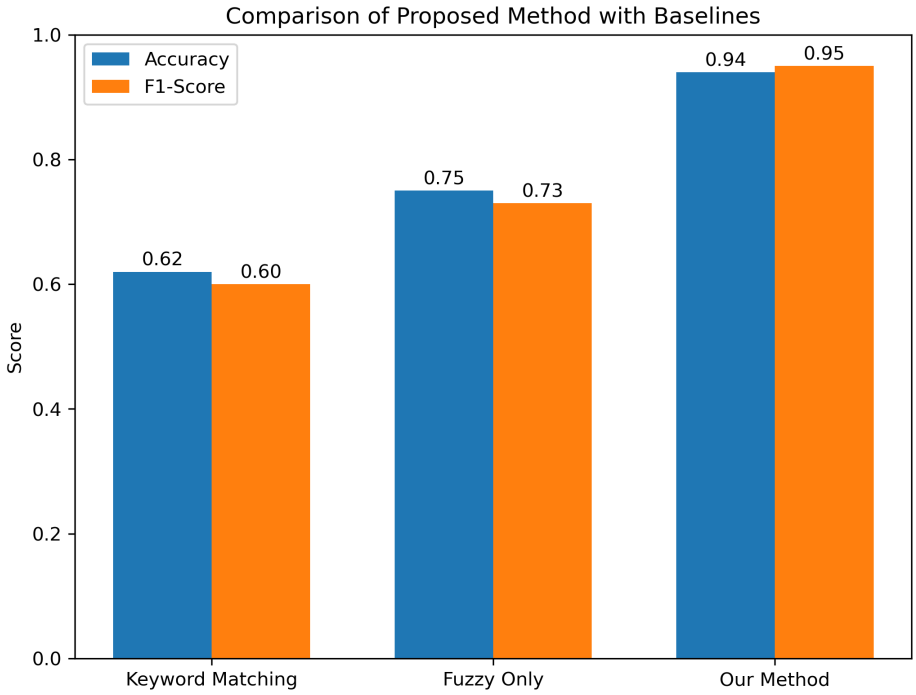

Figure 8. Accuracy and F1 score

predictions. This clear reduction in misclassifications demonstrates the improved robustness and reliability of the proposed approach for APT group detection.

Figure 12 shows the confusion matrix of the proposed model. The majority of predictions are concentrated along the diagonal, indicating strong true positive rates across classes. Only a small number of instances are misclassified, and these

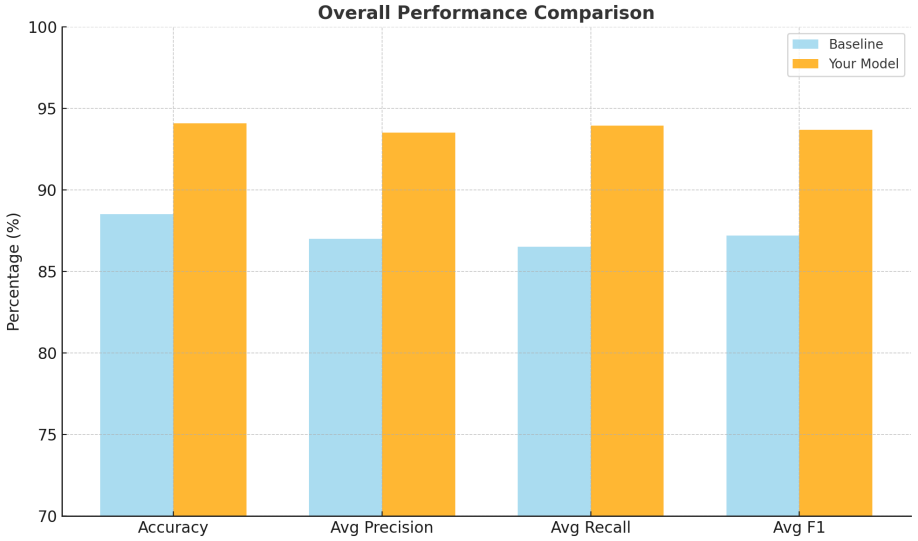

Figure 9. Comparison with base model

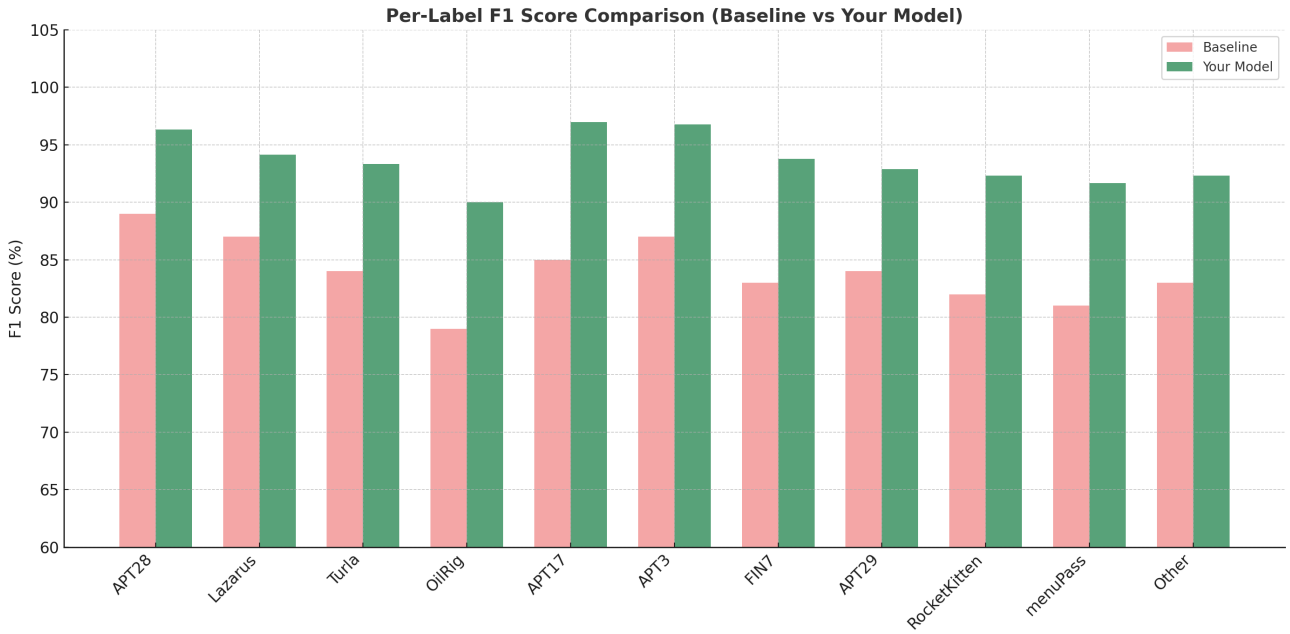

Figure 10. Per Group F1 Score Comparison with base madel

are primarily between closely related groups. This confirms that the model not only achieves high overall metrics but also maintains consistent per-class performance, with limited confusion between APT categories.

5.3.1 Why These Graphs Matter

As comparatively over the base paper in a clear visualize superiority of my model. They validate my improvements (accuracy, fewer errors, balanced F1). They align with research goals: minimizing false negatives, covering multiple APT groups, and outperforming state-of-the-art baselines.

6 Conclusion

In this research, we proposed an AI-based methodology for extracting cyber attack patterns from text reports and mapping them to the MITER ATT&CK framework to identify potential Advanced Persistent Threat (APT) groups. Our approach leverages

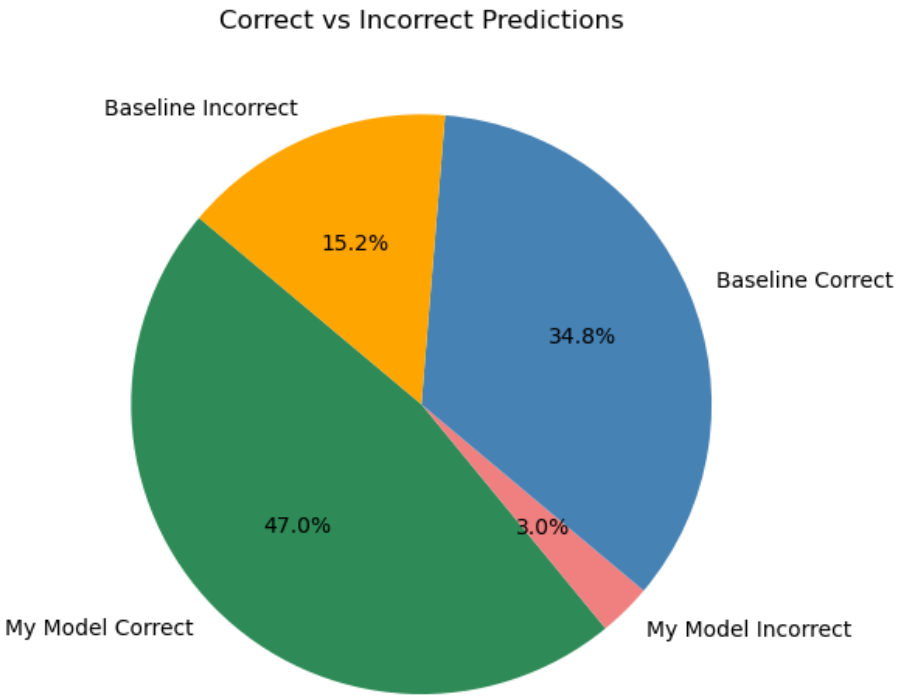

Figure 11. Comparison with base model

Natural Language Processing (NLP) techniques, specifically bi-grams and trigrams, to improve the accuracy of threat detection. By integrating Open-Source Intelligence (OS-INT) data, real-time incident reports, and structured threat intelligence sources, our model demonstrates high effectiveness in categorizing cyber threats and attributing them to known APT groups.

We tested our model on a dataset comprising 1,023 cyber attack reports, of which 534 reports were linked to specific APT groups based on historical incident reports published by the MITRE ATT&CK framework. The results indicate that our model successfully extracts meaningful attack techniques and associates them with relevant APT groups, reducing false positives by employing a hierarchical filtering mechanism. The experimental results also highlight the importance of phrase-based keyword extraction, where trigram-based matching yielded the highest accuracy.

Furthermore, our research contributes to the cybersecurity domain by providing a scalable and automated solution for cyber threat intelligence analysis. By analyzing unstructured text reports in real-time, our system can assist cybersecurity analysts in quickly identifying and responding to sophisticated cyber threats. Additionally, the visualization of attack patterns and APT group strength allows for a better understanding of the threat landscape, aiding in proactive defense strategies.

Despite the promising results, certain challenges remain. Our approach relies on predefined attack descriptions from the MITRE ATT&CK database, which may not always capture emerging threats or novel attack techniques. Additionally, the effectiveness of our model can be further enhanced by incorporating deep learning-based NLP models, such as transformer architectures, to improve contextual understanding and threat attribution.

In summary, this research provides a robust method for automated cyber threat intelligence analysis, bridging the gap between raw textual reports and structured threat intelligence frameworks. By continuously refining the model with updated threat intelligence sources and integrating real-time data streams, our approach has the potential to become a valuable asset in the field of cybersecurity. Future work should focus on expanding the dataset, integrating multi-language support, and improving detection accuracy through advanced AI techniques.

Key Takeaways:

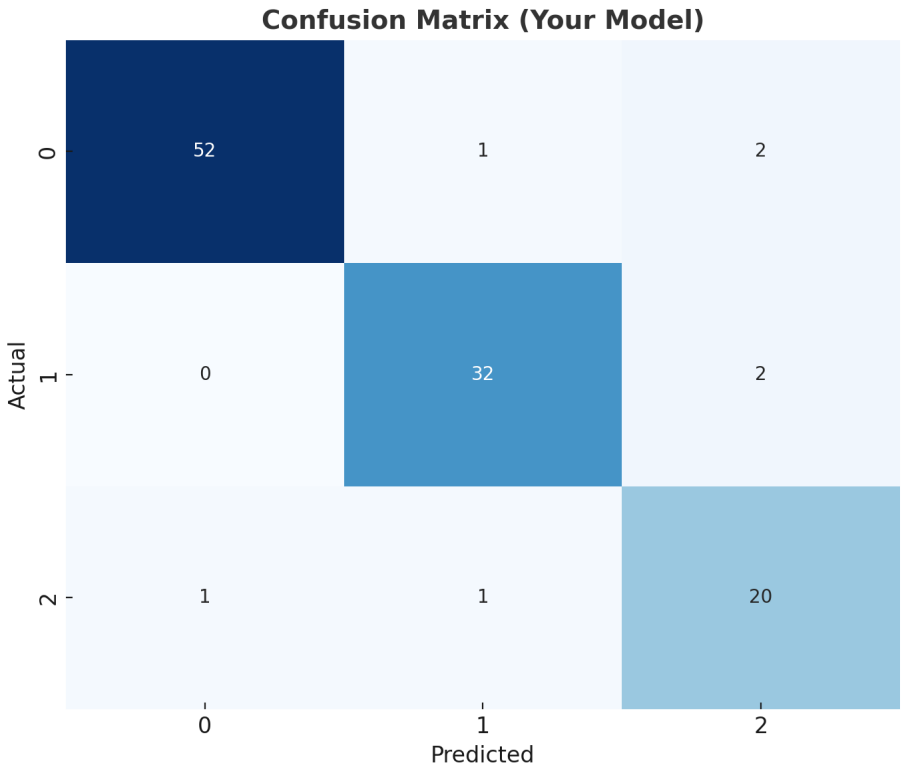

Figure 12. Comparison Confusion Matrix with base model

- Our model effectively extracts cyber attack patterns and maps them to APT groups using NLP and the MITER ATT&CK framework.
- The use of bi-grams and trigrams enhances the accuracy of threat detection.
- The methodology was validated on a dataset of 1,023 reports, achieving high reliability in APT attribution.
- Visualization of attack techniques aids in a better understanding of cyber threats.
- Future work should focus on improving real-time detection, deep learning integration, and dataset expansion.

Overall, our study provides a strong foundation for automated cyber attack analysis and lays the groundwork for future advancements in threat intelligence research.

7 Limitations and Future Work

Although the proposed methodology achieves strong performance in mapping attack techniques to corresponding APT groups, several limitations should be acknowledged.

First, the approach relies heavily on the MITRE ATT&CK framework as the primary knowledge base for attack techniques and threat actor mapping. While MITRE ATT&CK is comprehensive and widely adopted, newly emerging or undocumented attack techniques may not be immediately captured, which can limit the system's adaptability to evolving threat landscapes.

Second, the current model is primarily optimized for known APT groups. Although it demonstrates high classification accuracy, attribution of zero-day attacks or previously unseen threat actors remains challenging. Addressing this limitation would require incorporating anomaly detection mechanisms or semi-supervised learning strategies capable of identifying novel behavioral patterns.

Third, the methodology is largely dependent on phrase-level extraction and similarity matching. While effective, this approach may not fully capture deeper contextual and semantic relationships within complex incident reports. Integrating advanced transformer-based architectures and domain-specific fine-tuning could enhance contextual understanding and improve robustness.

In terms of future work, several extensions are planned. Incorporating large-scale transformer models fine-tuned on cybersecurity corpora could improve contextual reasoning and technique extraction. Additionally, integrating real-time threat intelligence feeds and open-source intelligence (OSINT) data would enable dynamic updating of attack patterns and APT group mappings.

Scalability and deployment in real-world environments also warrant further exploration. Future research may investigate cloud-based and distributed implementations to support large-scale cyber incident processing. Expanding the dataset with multilingual and cross-regional threat reports would further improve generalization and global applicability.

Finally, ethical and privacy considerations must remain central in future developments. Techniques such as privacy-preserving data processing and bias evaluation should be incorporated to ensure responsible and secure deployment of AI-driven threat attribution systems.

Overall, addressing these limitations will strengthen the robustness, adaptability, and practical applicability of the proposed framework in operational cybersecurity environments.

References

1. Tan, Z., Parambath, S. P., Anagnostopoulos, C., Singer, J. & Marnerides, A. K. Advanced persistent threats based on supply chain vulnerabilities: Challenges, solutions, and future directions. *IEEE Internet Things J.* **12**, 6371–6395, DOI: 10.1109/JIOT.2025.3528744 (2025).
2. Nowakowski, W. Social engineering analysis framework: A comprehensive playbook for human hacking. *IEEE Access* **13**, 18827–18849, DOI: 10.1109/ACCESS.2025.3532999 (2025).
3. Chen, P., Desmet, L. & Huygens, C. A study on advanced persistent threats. In *IFIP international conference on communications and multimedia security*, 63–72 (Springer, 2014).
4. Do Xuan, C. Detecting apt attacks based on network traffic using machine learning. *J. Web Eng.* **20**, 171–190 (2021).
5. Sharma, A., Gupta, B. B., Singh, A. K. & Saraswat, V. Advanced persistent threats (apt): evolution, anatomy, attribution and countermeasures. *J. Ambient Intell. Humaniz. Comput.* **14**, 9355–9381 (2023).
6. Vijaykumar, N., Jaganathan, N., Loganathan, N. B. & Raja, N. A nlp based model on resume parsing and shortlisting system for smart recruitment. In *AIP Conference Proceedings*, vol. 3279 (AIP Publishing, 2025).
7. Shou, Z. *et al.* The apt family classification system based on apt call sequences and attention mechanism. *Int. J. Inf. Comput. Secur.* **26**, 22–40 (2025).
8. Alshamrani, A., Myneni, S., Chowdhary, A. & Huang, D. A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. *IEEE Commun. Surv. & Tutorials* **21**, 1851–1877 (2019).
9. Aziz, K. *et al.* Unifying aspect-based sentiment analysis bert and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Sci. Reports* **14**, 14646 (2024).
10. Alharbi, H., Hur, A., Alkahtani, H. & Ahmad, H. F. Enhancing cybersecurity through autonomous knowledge graph construction by integrating heterogeneous data sources. *PeerJ Comput. Sci.* **11**, e2768 (2025).
11. Alzu'bi, A., Darwish, O., Albashayreh, A. & Tashtoush, Y. Cyberattack event logs classification using deep learning with semantic feature analysis. *Comput. & Secur.* **150**, 104222 (2025).
12. Stivelman, T. S. Extracting cybersecurity insights from real-world incident data. (2025).
13. Aziz, K. *et al.* Enhanced urduaspectnet: Leveraging biaffine attention for superior aspect-based sentiment analysis. *J. King Saud Univ. Inf. Sci.* 102221 (2024).
14. Mewada, A., Maurya, S. K. & Ansari, M. A. Comparative study of artificial intelligence approaches in deceptive opinion detection. In *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 999–1004, DOI: 10.1109/CICTN64563.2025.10932501 (Ghaziabad, India, 2025).
15. Bonhard, S. K., Villalta, P. G., Rosés, O. & Pegueroles, J. A review of tactics, techniques, and procedures (ttps) of mitre framework for business email compromise (bec) attacks. *IEEE Access* **13**, 50761–50776, DOI: 10.1109/ACCESS.2025.3552523 (2025).
16. Aziz, K. *et al.* Urduaspectnet: Fusing transformers and dual gcn for urdu aspect-based sentiment detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* DOI: 10.1145/3663367 (2024). Just Accepted.
17. Aziz, K., Ji, D., Li, B., Li, F. & Zhou, J. Advancing urdu nlp: Aspect-based sentiment analysis with graph attention networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8, DOI: 10.1109/IJCNN60899.2024.10650054 (2024).

18. Liu, C., Singhal, A. & Wijesekera, D. Forensic analysis of advanced persistent threat attacks in cloud environments. In *Advances in Digital Forensics XVI: 16th IFIP WG 11.9 International Conference, New Delhi, India, January 6–8, 2020, Revised Selected Papers 16*, 161–180 (Springer, 2020).

19. Farfoura, M. E., Mashal, I., Alkhatib, A. & Batyha, R. M. A lightweight machine learning methods for malware classification. *Clust. Comput.* **28**, 1 (2025).

20. Prasad, N., Diro, A., Warren, M. & Fernando, M. A survey of cyber threat attribution: Challenges, techniques, and future directions. *Comput. Secur.* **157**, 104606, DOI: https://doi.org/10.1016/j.cose.2025.104606 (2025).

21. Rani, N., Saha, B. & Shukla, S. K. A comprehensive survey of automated advanced persistent threat attribution: Taxonomy, methods, challenges and open research problems. *J. Inf. Secur. Appl.* **92**, 104076, DOI: https://doi.org/10.1016/j.jisa.2025.104076 (2025).

22. Basnet, A. S., Ghanem, M. C., Dunsin, D., Kheddar, H. & Sowinski-Mydlarz, W. Advanced persistent threats (apt) attribution using deep reinforcement learning. *Digit. Threat. Res. Pract.* (2025).

23. Saha, A. *et al.* Expert insights into advanced persistent threats: Analysis, attribution, and challenges. In *Proceedings of the 34th USENIX Security Symposium (USENIX Sec)* (2025).

24. Berrios, S., Leiva, D., Olivares, B., Allende-Cid, H. & Hermosilla, P. Systematic review: Malware detection and classification in cybersecurity. *Appl. Sci.* **15**, 7747 (2025).

25. Zhang, J., Liu, S. & Liu, Z. Attribution classification method of apt malware based on multi-feature fusion. *PLoS One* **19**, e0304066 (2024).

26. Mutalib, N. H. A., Sabri, A. Q. M., Wahab, A. W. A., Abdullah, E. R. M. F. & AlDahoul, N. Explainable deep learning approach for advanced persistent threats (apts) detection in cybersecurity: A review. *Artif. Intell. Rev.* **57**, 297 (2024).

27. Arazzi, M. *et al.* Nlp-based techniques for cyber threat intelligence (2023). 2311.08807.

28. Büchel, M. *et al.* {SoK}: Automated {TTP} extraction from {CTI} reports–are we there yet? In *34th USENIX Security Symposium (USENIX Security 25)*, 4621–4641 (2025).

29. Lange, L. *et al.* Annoctr: A dataset for detecting and linking entities, tactics, and techniques in cyber threat reports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1147–1160 (2024).

30. Baek, J., Kim, W., Jeong, S. *et al.* A study on improving the automatic classification performance of cybersecurity mitre att&ck tactics using nlp-based modernbert and bertopic models. *Preprints* (2025).

31. Cuong Nguyen, H., Tariq, S., Baruwal Chhetri, M. & Quoc Vo, B. Towards effective identification of attack techniques in cyber threat intelligence reports using large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, 942–946 (2025).

32. Cheng, W. *et al.* Crucialg: Reconstruct integrated attack scenario graphs by cyber threat intelligence reports. *IEEE Transactions on Dependable Secur. Comput.* (2025).

33. Tatam, M., Shanmugam, B., Azam, S. & Kannoorpatti, K. A review of threat modelling approaches for apt-style attacks. *Heliyon* **7** (2021).

34. AL-Aamri, A. S. *et al.* Machine learning for apt detection. *Sustainability* **15**, 13820 (2023).

Author contributions statement

R.A. conceived the study and designed the experiment(s). R.A. and Z.J. supervised the research, contributed to methodology refinement, and guided the manuscript preparation. N.A. and M.A.A. performed data preprocessing and experimental validation. M.L. and P.C. analysed the results and contributed to the interpretation of findings. K.A conducted the experiment(s) and implemented the model. All authors reviewed and approved the final manuscript.

Data Availability.

The data used in this study were collected from publicly available sources, including MITRE ATT&CK group reports,[4] the AttackAttributionDataset repository,[5] and additional public cybersecurity reports and research publications. Attacker metadata and aliases were obtained from the repository's accompanying Airtable base.[6]

[4] https://attack.mitre.org/groups/
[5] https://github.com/eyalmazuz/AttackAttributionDataset
[6] https://airtable.com/appuo07yvRjRyYJIX/shr3Po3DsZUQZY4we/tbljpA5wI1IaLI4Gv/viwGFVFtuu0l88e7u

Acknowledgment

The authors would like to acknowledge the support of Prince Sultan University in paying the Article Processing Charges (APC) for this publication.

Funding

The authors declare that no funding was received for this study.