A Study on Improving the Automatic Classification Performance of Cybersecurity MITRE ATT&CK Tactics Using NLP-Based ModernBERT and BERTopic Models

Jaehwan Baek , Wooju Kim * , Jeonghoon O. , Seungwoo Jeong

Posted Date: 20 October 2025

doi: 10.20944/preprints202510.1543.v1

Keywords: adversary emulation; cyber threat intelligence (CTI); cyber threat reports (CTRS); CTI automation; BERT; modernBERT; TF-IDF; BERTopic; natural language processing (NLP); MITRE ATT&CK; security operations center (SOC); sliding window; multi-label classification; $f_{0.5}$ score; threat report automation

Article

A Study on Improving the Automatic Classification Performance of Cybersecurity MITRE ATT&CK Tactics Using NLP-Based ModernBERT and BERTopic Models

Jaehwan Baek [1], Wooju Kim [2,*], Jeonghoon O. [3] and Seungwoo Jeong [3]

[1] Graduate Program in Technology Policy, Yonsei University, Seoul, Republic of Korea
[2] Department of Industrial Engineering, Yonsei University, Seoul, Republic of Korea
[3] Independent Researcher, Seoul, Republic of Korea
* Correspondence: wkim@yonsei.ac.kr

Abstract

Cyber Threat Intelligence (CTI) reports are key resources for identifying the Tactics, Techniques, and Procedures (TTPs) of hackers and hacking groups. However, these reports are lengthy and unstructured, presenting limitations for automatic mapping to the MITRE ATT&CK framework. In this study, we designed and compared the performance of five hybrid classification models concatenating statistical-based features (TF-IDF), transformer-based contextual embeddings (BERT, ModernBERT), and topic-based representations (BERTopic) to automatically classify CTI reports into 12 ATT&CK tactic categories. Experiments using the rcATT dataset, comprising 1,490 public threat reports, showed that the model concatenating TF-IDF and ModernBERT achieved a micro-precision of 72.25%, demonstrating a 10.07 percentage point improvement in detection precision compared to the baseline paper. Furthermore, the model concatenating TF-IDF and BERTopic achieved a micro $F_{0.5}$ of 67.14% and a macro $F_{0.5}$ of 63.20%, representing a 6.27 percentage point improvement over the baseline paper, significantly enhancing detection performance for imbalanced classes and rare tactics. Academically, this study demonstrates that a hybrid approach integrating statistical, contextual, and semantic information can simultaneously improve precision and balance compared to existing CTI analysis techniques. Industrially, it demonstrated the practical applicability of the model to enhance detection efficiency and reduce analyst workload in Security Operations Center (SOC), adversary emulation, and Threat-informed Defense environments.

Keywords: adversary emulation; cyber threat intelligence (CTI); cyber threat reports (CTRS); CTI automation; BERT; modernBERT; TF-IDF; BERTopic; natural language processing (NLP); MITRE ATT&CK; security operations center (SOC); sliding window; multi-label classification; $f_{0.5}$ score; threat report automation

1. Introduction

As digital transformation accelerates, cyber threats are becoming increasingly diverse and sophisticated [36,37]. In particular, Advanced Persistent Threat (APT) attacks targeting critical industries such as government agencies, financial institutions, and manufacturing cause massive data breaches and significant economic losses, seriously threatening national security and social trust [1,36]. In this complex threat landscape, the importance of Cyber Threat Intelligence (CTI) is more critical than ever [7]. CTI goes beyond simple event detection, playing an essential role in structurally understanding attackers' Tactics, Techniques, and Procedures (TTPs) and establishing proactive defense strategies based on this understanding [3,9]. In this process, Cyber Threat Reports (CTRs) function as the most critical source data [3,4]. CTRs are documents published by security firms,

 doi:10.20944/preprints202510.1543.v1

research institutions, and government organizations after analyzing actual attack cases. They contain multi-layered information, including the attacker's intent, penetration methods, scale of damage, and response strategies [4,14,38]. Crucially, CTRs provide strategic patterns and tactical clues that are difficult to discern from simple log data or event alerts, as they describe attacker behavior in concrete detail within diverse contexts [14,40].

Therefore, CTRs are both the starting point for CTI automation research and a vital basis for practical decision-making in Security Operations Centers (SOCs) [9,24]. However, the sheer volume and unstructured narrative nature of CTRs make it difficult to perform automated analysis [35]. Traditional analysis methods typically involved security experts manually reviewing reports or relying on keyword-based tools. This approach consumes excessive time and human resources and struggles to fully extract the attackers' inherent tactics, techniques, and procedures (TTPs) embedded within the reports [3,7]. Furthermore, low compatibility with systematic threat knowledge frameworks like MITRE ATT&CK limited the utilization of CTR information as structured, standardized CTI assets. Recent advancements in Natural Language Processing (NLP) have opened possibilities for automated CTR analysis [7,8,9].

Pre-trained language models like BERT, with their strength in contextual meaning understanding, can be applied to document classification and threat actor identification within cyber threat reports [30]. However, existing BERT models face challenges: their input length constraint (maximum 512 tokens) makes fully processing long reports difficult, and their classification performance across multiple tactic categories is unstable [31,32]. To overcome these limitations, this study designed and compared five models performing MITRE ATT&CK tactic-level classification on CTRs [9,33].

The main contributions of this paper are as follows:

· We proposed a sliding window-based ModernBERT approach to process long CTRs without information loss [33].

· We developed a hybrid feature strategy that concatenates TF-IDF, ModernBERT embeddings, and BERTopic topic vectors to form a unified input representation for classification [34,18,19].

· Through comparative experiments against existing baselines and state-of-the-art NLP-based techniques, we demonstrated significant improvements in precision and $F_{0.5}$ metrics [9,10,26].

· We demonstrated the practical applicability of our findings for CTI automation and adversary emulation scenarios based on MITRE ATT&CK [2,23,24].

Subsequently, the paper unfolds in the following order: research trends (Chapter 2), research methodology (Chapter 3), data analysis and critical examination (Chapter 4), and conclusions and future research directions (Chapter 5).

2. Literature Review

Cyber Threat Intelligence (CTI) has evolved into a core domain for identifying attackers' Tactics, Techniques, and Procedures (TTPs) and preventing threats based on this knowledge [7]. Cyber Threat Reports (CTRs) serve as critical source data, containing actual attack narratives, technical indicators, and response strategies, thereby providing the foundation for CTI automation research [3,4,9]. CTRs hold high value for both academic research and practical application because they contain strategic context and tactical clues that are difficult to extract from simple logs or event data [14,38,40]. However, CTRs are typically lengthy and composed of unstructured narratives, presenting significant challenges for automated analysis [35].

2.1. Traditional Text Mining-Based Research

Early CTI automation research was based on traditional vectorization techniques such as TF-IDF (Term Frequency–Inverse Document Frequency) [3,14]. Legoy et al. (2020) formally introduced the rcATT dataset, presenting a multi-label problem that classifies public threat reports into MITRE ATT&CK tactic units [3]. While this approach, combining TF-IDF with a linear classifier, offers the advantages of simple implementation and straightforward result interpretation and explanation, it

 doi:10.20944/preprints202510.1543.v1

revealed limitations in failing to reflect contextual meaning between words and insufficiently capturing the multi-layered information within long documents [39].

2.2. Transformer-Based Embedding Research

With the rapid advancement of natural language processing (NLP) technology, transformer-based models, including BERT (Bidirectional Encoder Representations from Transformers), were introduced to CTR analysis [30]. These models leveraged their contextual understanding capabilities and were applied to various applications such as TTP classification, malicious code report summarization, and incident analysis [10,11,14]. However, BERT struggles to process long documents like CTRs due to its input length limitation (maximum 512 tokens) [31,32]. Techniques like sliding windows, paragraph-level segmentation, and truncation have been proposed to address this, but issues like long-context disruption and reduced computational efficiency are still reported [31].

2.3. Topic Modeling and Hybrid Approach Research

To compensate for the limitations of contextual embeddings, topic modeling techniques like BERTopic have begun to be utilized [34]. BERTopic identifies latent topics within documents and can leverage them as features for classifiers, providing meaningful thematic context in CTR analysis [18,19]. Recently, hybrid approaches combining TF-IDF, embeddings, and topic distributions have been attempted, aiming to improve the balance between precision and recall [9]. These studies are significant, as they compensate for the limitations of single approaches and can enhance the performance of CTR analysis in multiple dimensions [20,21].

2.4. Research Gaps and Distinctiveness of This Study

While existing research has demonstrated the potential for automated CTR analysis, several common limitations persist [7,9,14]. First, there is a lack of effective structures capable of processing long documents like CTRs without context loss [31,33]. Second, the problem of imbalance between precision and recall has been consistently reported, limiting practical application in security monitoring environments [9]. Third, research on multi-label classification at the MITRE ATT&CK tactic level remains immature, and systematic comparative studies across diverse approaches are scarce [3]. Consequently, this study focuses not merely on hyperparameter tuning for a single model, but on identifying which combination of AI models achieves optimal performance when applied to the specialized domain data of CTRs in the information security industry [33,34].

This holds the following academic significance: (Domain-Specific Optimization) By identifying model combinations specialized for CTR analysis, it reflects the unique characteristics of real-world security data not addressed by general NLP research [7]. (Methodological Contribution) By combining and analyzing features with distinct characteristics—such as TF-IDF, ModernBERT, and BERTopic—it academically presents the balanced point between precision, recall, and $F_{0.5}$ [9,19,34]. (Practical Impact) In Security Operations Center (SOC) environments, optimal model selection directly impacts alert reliability and detection coverage; this research enhances practical applicability [24]. (Standardization Contribution) This study provides a basis for future CTI researchers to determine suitable approaches in CTRs analysis, contributing to the establishment of research directions [2,23]. Therefore, by exploring the most suitable AI model combination for CTRs, this research aims to achieve both academic originality and industrial effectiveness [25,26].

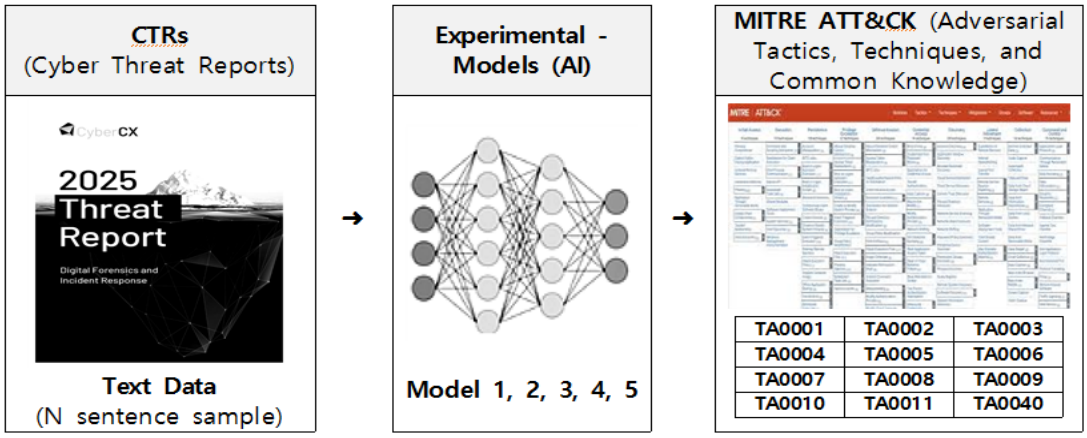
3. Methodology

This chapter describes the research design and experimental methodology. First, the target dataset and preprocessing steps are presented, followed by a detailed explanation of the proposed model architecture. Finally, the training environment and evaluation methods are described, and the comparative research position of this study is discussed.

 doi:10.20944/preprints202510.1543.v1

3.1. Research Design Overview

Figure 1 illustrates the overall analysis structure of this study, which takes Cyber Threat Intelligence (CTI) reports as input and predicts MITRE ATT&CK tactics [1,36]. Input documents undergo preprocessing before being fed into the model, and the five proposed models (Model 1–5) each utilize different feature extraction methods (BERT, ModernBERT, TF-IDF, BERTopic, etc.) [30,33,34]. The objective of this study is to design various AI models for automatically classifying Cyber Threat Reports (CTRs) into the 12 tactical units of MITRE ATT&CK (TA0001–0011, TA0040) and systematically compare and analyze the most effective model combinations [3,9]. While previous studies focused on hyperparameter optimization for single models, this research designs model combinations encompassing traditional vectorization, transformer embeddings, and topic modeling, considering the specialized domain of CTRs. To ensure consistency and fairness, performance is comprehensively validated using identical data and evaluation criteria [7,14,19].



(Cyber Threat Reports → Models 1–5 → MITRE ATT&CK "12" Tactic Classification)

Figure 1. Overall Analysis Architecture Structure

3.2. Dataset and Processing

The rcATT dataset (Legoy et al., 2020) was used for experiments [3]. This dataset contains 1,490 publicly available threat reports, with each document having a multi-label structure assigned one or more of 12 MITRE ATT&CK tactic labels (TA0001~TA0011, TA0040) [36]. Each model vectorizes document representations and then predicts tactic labels using a multi-layer perceptron (MLP) or classifier [9]. All models are designed with a multi-label classification structure, enabling simultaneous detection of multiple tactics within a single document [10,31]. The final output is document-level prediction results for the 12 MITRE ATT&CK tactic classes [33]. The experiment proceeded with the following workflow: input data (cyber threat reports) on the left, the AI model in the center, and the tactic classification results on the right [34].

3.3. Model Structure

3.3.1. Model 1 : TF-IDF + MLP

Algorithm (Model-1): Classification with TF-IDF + MLP

Input: document set D (rcATT, Legoy et al., 2020), 1,490 threat reports; TF-IDF vectorizer; MLP classifier; labels = MITRE ATT&CK tactics (12 classes, multi-label)

Output: predicted MITRE ATT&CK tactic classes

1. For each document $doc_i \in D$, perform preprocessing

 doi:10.20944/preprints202510.1543.v1

(tokenization, stopword removal).
2. Fit the TF-IDF vectorizer on the training set; transform train/validation (or test) sets to obtain sparse vectors V_i.
3. Initialize the MLP classifier (e.g., hidden layers with ReLU; output dimension = 12).
4. Train the MLP on TF-IDF vectors and multi-label targets using BCE/BCEWithLogits loss.
5. For each validation/test document vector V_j, feed it to the MLP to obtain logits z_j.
6. Apply the sigmoid function to convert logits to probabilities $\hat{P}_j \in (0,1)^{12}$.
7. Threshold each class probability at τ (default 0.5; optionally calibrated) to produce the predicted label set.
8. Optionally tune τ (global or per-class) on the validation set to optimize $F_{0.5}$.
9. Evaluate using micro/macro precision, recall, and $F_{0.5}$.

For document d, the final prediction probability is defined as follows:
Prediction Equation

$$\hat{y}_{i,c} = 1\left[\sigma\left(W^{(L)}\,\phi\left(\cdots\phi\left(W^{(1)}\left[v_i^{tfidf}\,\|\,\bar{h}_i\right]+b^{(1)}\right)\cdots+b^{(L-1)}\right)+b^{(L)}\right)_c \geq \tau\right]$$

Loss Function (BCE)

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{12}\left(-y_{i,c}\log\hat{p}_{i,c} - (1-y_{i,c})\log(1-\hat{p}_{i,c})\right)$$

Notation
v_i TF-IDF vector of document d_i
$\hat{y}_{i,c}$ predicted label (0/1) for class c of document d_i
$W^{(l)}$ weights of MLP layer
$y_{i,c}$ ground-truth label
$b^{(l)}$ bias vector of the i-th MLP layer
z_i final output logits of the MLP
$\phi(\cdot)$ activation function (e.g., ReLU)
$\mathcal{L}(?)$ Binary Cross-Entropy loss
$\sigma(\cdot)$ sigmoid function
N total number of documents
τ decision threshold (default 0.5)
? set of model parameters {W, b}

Model 1 is a classification architecture that utilizes the traditional statistical feature TF-IDF as input to a multilayer perceptron (MLP) for text classification [3,14]. Cyber threat reports undergo preprocessing and tokenization, and each document is converted into a sparse vector reflecting word frequency and inverse document frequency via the TF-IDF vectorizer [7]. Here, TF-IDF limits the maximum number of features (max_features) to 20,000 and sets the n-gram range to (1,2) [9]. The generated TF-IDF vectors are fed to the MLP as input [19]. The MLP consists of two hidden layers, each containing 2048-dimensional nodes. A ReLU activation function is applied to the hidden layers, and a dropout rate of 0.1 is set to prevent overfitting [20]. The final output layer consists of a 12-dimensional vector corresponding to the MITRE ATT&CK tactic classes, with each output node calculating the probability for each class via the sigmoid function [30]. During inference, if the

 doi:10.20944/preprints202510.1543.v1

probability exceeds the threshold (τ=0.5, adjustable if needed), the tactic label is determined as the final prediction [10]. This model follows a multi-label classification structure and employs the binary cross-entropy loss function during training [34]. Model performance evaluation was conducted based on precision, recall, and $F_{0.5}$ metrics [9,19].

3.3.2. Model 2 : BERT(sliding_Window) + MLP

Algorithm (Model-2): BERT (sliding window) + MLP

Input: document set D (rcATT, Legoy et al., 2020), 1,490 threat reports; pretrained BERT; MLP classifier; window length L (e.g., 512 tokens), stride S (e.g., 256); labels = MITRE ATT&CK tactics (12 classes, multi-label)

Output: predicted MITRE ATT&CK tactic classes

For each document $doc_i \in D$, prepare the input text (no stopword removal).
Tokenize doc_i with the BERT tokenizer and split into overlapping windows $\{w_{i,k}\}$ of length L with stride S (pad/truncate as needed).
Encode each window with BERT to obtain contextual embeddings; extract the [CLS] vector $\mathbf{h}_{i,k}$.
Feed $\mathbf{h}_{i,k}$ to an MLP to produce window-level logits $\mathbf{z}_{i,k} \in \mathbb{R}^{12}$.
Convert window logits to probabilities via sigmoid and aggregate across windows (e.g., mean pooling):

$$\hat{\mathbf{p}}_i = \frac{1}{K_i}\sum_k \sigma(\mathbf{z}_{i,k})$$

Training: Optimize with Binary Cross-Entropy (either on aggregated $\hat{\boldsymbol{p}}_i$ or averaged over window logits).
Inference: Apply threshold τ (default 0.5; optionally calibrated) to $\hat{\boldsymbol{p}}_i$ to obtain the predicted label set.
Optionally tune L, S, pooling (mean/max/attention), and τ on the validation set to optimize $F_{0.5}$.
Evaluate using micro/macro precision, recall, and $F_{0.5}$.

For document d, the final prediction probability is defined as follows:
Prediction Equation

$$\hat{y}_{i,c} = 1\left[\frac{1}{K_i}\sum_{k=1}^{K_i}\sigma\left(\mathrm{W}^{(L)}\phi\left(\cdots\phi\left(\mathrm{W}^{(1)}\mathrm{h}_{i,k}+\mathrm{b}^{(1)}\right)\cdots+\mathrm{b}^{(L-1)}\right)+\mathrm{b}^{(L)}\right)_c \geq \tau\right]$$

Notation
$w_{i,k}$ k-th sliding window of document d_i
K_i number of windows in document d_i
$\mathrm{h}_{i,k} = BERT_{[CLS]}(w_{i,k})$ [CLS] embedding of window $w_{i,k}$ from BERT
$\mathrm{W}^{(l)}$ weights of MLP layer
$b^{(l)}$ biases of MLP layer
$\phi(\cdot)$ activation function (e.g., ReLU)
$\sigma(\cdot)$ sigmoid function
τ decision threshold (default 0.5)

 doi:10.20944/preprints202510.1543.v1

Cyber Threat Reports (CTRs) are often lengthy documents averaging over 3,000 words [31,32]. However, the BERT model has a maximum input length limited to 512 tokens, making it difficult to process long documents directly [30]. To address this, Model-2 employs a sliding window technique [33]. Each CTR is divided into 256-token segments after tokenization, with 50% overlap applied to minimize context disruption [19,34]. For example, the first window covers [1–256], and the second window covers [129–384]. Each generated window is input to BERT to extract a [CLS] vector, which is then fed to an MLP classifier to produce logits for the 12 MITRE ATT&CK tactic classes [9]. The logits from each window are converted to probability values via a sigmoid function, and the document-level prediction is calculated by averaging the probabilities across all windows [10]. During training, the binary cross-entropy loss function was used, and at inference, if a probability exceeds the threshold (τ=0.5), the corresponding tactic class is assigned as the final label [20]. This approach reduces information loss inherent in long CTRs and effectively incorporates diverse contextual clues [29,31,33].

3.3.3. Model 3 : ModernBERT (sliding_Window) + MLP

Algorithm (Model-3): ModernBERT (sliding window) + MLP

Input: document set D (rcATT, Legoy et al., 2020), 1,490 threat reports; pretrained ModernBERT; MLP classifier; window length L (e.g., 4096 tokens), stride S (e.g., 2048); labels = MITRE ATT&CK tactics (12 classes, multi-label)

Output: predicted MITRE ATT&CK tactic classes

For each document doc_i ∈ D, prepare the raw text (no stopword removal).
Tokenize doc_i with the ModernBERT tokenizer and split into overlapping windows $\{w_{i,k}\}$ of length L with stride S (pad/truncate as needed).
Encode each window with ModernBERT to obtain contextual embeddings; extract a window representation $\mathbf{h}_{i,k}$ (e.g., [CLS] or mean-pooled).
Feed $\mathbf{h}_{i,k}$ into an MLP to produce window-level logits $\mathbf{z}_{i,k} \in \mathbb{R}^{12}$.
Convert window logits to probabilities via sigmoid and aggregate across windows (e.g., mean pooling):

$$\hat{\mathbf{p}}_i = \frac{1}{K_i}\sum_{k=1}^{K_i}\sigma(\mathbf{z}_{i,k})$$

Training: Optimize with Binary Cross-Entropy (either on aggregated $\hat{\boldsymbol{p}}_i$ or averaged over window logits).
Inference: Apply threshold τ (default 0.5; optionally calibrated) to $\hat{\boldsymbol{p}}_i$ to obtain the predicted label set.
Optionally tune L, S, pooling strategy (mean/max/attention), and τ on the validation set to optimize $F_{0.5}$.
Evaluate using micro/macro precision, recall, and $F_{0.5}$.

For document d, the final prediction probability is defined as follows:

Prediction Equation

 doi:10.20944/preprints202510.1543.v1

$$\hat{y}_{i,c} = 1\left[\left(\frac{1}{K_i}\sum_{k=1}^{K_i}\sigma\left(W^{(L)}\,\phi\left(\cdots\phi\left(W^{(1)}h_{i,k}+b^{(1)}\right)\cdots+b^{(L-1)}\right)+b^{(L)}\right)\right)_c \geq \tau\right]$$

Notation

$w_{i,k}$ k-th sliding window of document d_i(L=4096, S=2048)

K_i number of windows in document d_i

$h_{i,k}$ pooled embedding ([CLS] or mean) of window ($w_{i,k}$) from ModernBERT

$W^{(l)}$ weights of MLP layer

$b^{(l)}$ biases of MLP layer

$\phi(\cdot)$ activation function (e.g., ReLU)

$\sigma(\cdot)$ sigmoid function

$?\,p$ decision threshold (default 0.5)

Model-3 was designed based on ModernBERT to effectively process long cyber threat reports (CTRs) [33]. CTRs are often written with an average of thousands of words or more, making it difficult for existing BERT models—limited to an input maximum length of 512 tokens—to directly reflect the entire document [30,31]. To overcome this limitation, this study adopts ModernBERT-large, designed in the original paper to process inputs up to 8,192 tokens [33]. However, considering resource constraints and model stability, the maximum input length was set to 4,096 tokens for experiments [9,29]. To sufficiently capture the contextual information of long documents, a sliding window technique was employed concurrently [19,34]. Each document is divided into a set of windows after tokenization, with a length (L = 4096) and a stride (S = 2048), resulting in 50% overlap [31,33]. For example, the first window covers the range [1–4096], and the second window covers [2049–6144]. This structure was designed to accommodate inputs of up to 32,768 tokens across the entire document, minimizing context disruption and preserving continuous semantic clues [30,33]. Figure 2 schematically illustrates the structure of the sliding window technique applied in this study [7].

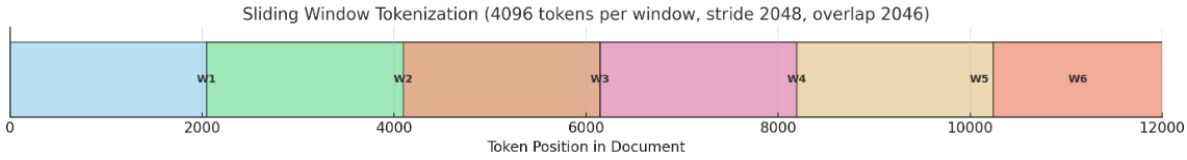


Figure 2. Schematic Diagram of Window Sliding (Based on MAX_LENGTH = 4096).

Each window is input into the ModernBERT encoder to generate contextual embeddings, and the window representation vector is extracted using either the [CLS] token vector or the mean pooling result [19]. The obtained window vector is then input into a Multi-Layer Perceptron (MLP) to produce a 12-dimensional logit, which is converted into class-specific probability values via a sigmoid function [20]. The final document-level prediction was calculated by averaging all window probabilities [10]. During training, the binary cross-entropy loss function was used, and at inference, if a class probability exceeded the threshold (τ=0.5), that tactical class was assigned as the final label [9].

Table 1 summarizes the distribution of input windows generated after sliding window segmentation across all 1,490 CTRs [3,33]. Analysis revealed that 77.38% (1,153 documents) could be processed with a single window, with an average document length equivalent to approximately 8.2 A4 pages [35]. Conversely, 22.62% of documents were split into two or more windows, with instances of up to 16 windows generated [31]. For example, 178 documents (11.95%) were split into two windows, and only a few documents were split into three or more windows [32,33].These results suggest that CTR reports are predominantly composed of lengthy texts, meaning that key clues may be missed with single-window input alone [14,38]. Therefore, the sliding window technique is

 doi:10.20944/preprints202510.1543.v1

confirmed as an essential preprocessing strategy for comprehensively learning the entire dataset and effectively reflecting the contextual meaning of long documents [7,9,33].

Table 1. Distribution of Original Data Window Counts (Based on MAX_LENGTH = 4096).

No. of Win. (MAX_LENGTH = 4096)	No. of Docs.	%	Doc. Len. (A4 p.)
1	1153	77.38	8.2
2	178	11.95	16.4
3	67	4.5	24.6
4	41	2.75	32.8
5	18	1.21	40.9
6	15	1.01	49.1
7	4	0.27	57.3
8	5	0.34	65.5
9	3	0.2	73.7
11	1	0.07	90.1
12	1	0.07	98.3
13	1	0.07	106.4
14	1	0.07	114.6
16	2	0.13	131.0
총합	1490 개	100 %	

3.3.4. Model 4 : TF-IDF + ModernBERT (sliding_Window) + MLP

Algorithm (Model-4): TF-IDF + ModernBERT (sliding window) + MLP

Input: document set D (rcATT, Legoy et al., 2020), 1,490 threat reports; TF-IDF vectorizer; pretrained ModernBERT; MLP classifier; window length L (e.g., 4096), stride S (e.g., 2048); labels = MITRE ATT&CK tactics (12 classes, multi-label)

Output: predicted MITRE ATT&CK tactic classes

For each document $doc_i \in D$, use the raw text (no stopword removal).

Fit the TF-IDF vectorizer on the training set; transform train/validation (or test) sets to obtain sparse vectors $\mathbf{v}_i^{\text{tfidf}}$.

Tokenize doc_i with the ModernBERT tokenizer and split into overlapping windows $\{w_{i,k}\}$ of length L with stride S (pad/truncate as needed).

Encode each window with ModernBERT to get window representations $\mathbf{h}_{i,k}$(e.g., [CLS] or mean-pooled).

Aggregate window representations to a document embedding: (or attention pooling)

$$\bar{\mathbf{h}}_i = \frac{1}{K_i}\sum_k \mathbf{h}_{i,k}$$

Concatenate features to form the fused input $\mathbf{x}_i = [\, \mathbf{v}_i^{\text{tfidf}} \,||\, \bar{\mathbf{h}}_i]$

 doi:10.20944/preprints202510.1543.v1

(optionally reduce TF-IDF dimension via SVD).
Feed $\mathbf{x}_i$ to the MLP to obtain logits $\mathbf{z}_i \in \mathbb{R}^{12}$; apply sigmoid to get probabilities $\hat{\boldsymbol{p}}_i \in (0,1)^{12}$.
Threshold each class probability at τ (default 0.5; optionally calibrated per class) to produce the predicted label set.
Evaluate using micro/macro precision, recall, and $F_{0.5}$ (tune L, S, pooling, SVD rank, and τ on the validation set to optimize $F_{0.5}$.

For document d, the final prediction probability is defined as follows:
Prediction Equation

$$\hat{y}_{i,c} = 1\left[\sigma\left(W^{(L)}\,\phi\left(\cdots\phi\left(W^{(1)}\left[v_i^{tfidf} \,||\, \bar{h}_i\right] + b^{(1)}\right)\cdots + b^{(L-1)}\right) + b^{(L)}\right)_c \geq \tau\right]$$

Notation
v_i^{tfidf} TF-IDF vector of docuent d_i
$w_{i,k}$ k-th sliding window of document d_i(L=4096, S=2048)
K_i number of windows in document d_i
$\bar{h}_i$ aggregated ModernBERT embedding of d_i (mean of window embeddings)
$\left[\, v_i^{tfidf} \,||\, \bar{h}_i\right]$ concatenated TF-IDF and ModernBERT features
$W^{(l)}$ weights of MLP layer
$b^{(l)}$ biases of MLP layer
$\phi(\cdot)$ activation function (e.g., ReLU)
$\sigma(\cdot)$ sigmoid function
? decision threshold (default 0.5)

Model 4 is a hybrid model concatenating statistical features (TF-IDF) and language model-based features (ModernBERT) [9,33]. Each document generates a sparse vector v_i^tfidf via TF-IDF, while simultaneously encoding windows segmented by the ModernBERT tokenizer to obtain a [CLS] or mean pooling-based vector h_(i,k) [19,30]. These vectors are aggregated into a document embedding via mean pooling and finally concatenated with the TF-IDF vector to form a fused input x_i = [v_i^tfidf || h¯_i] [33,34].

The concatenated vector outputs a 12-dimensional logit through an MLP, which then passes through a sigmoid function to generate class-specific probabilities [10,20]. During inference, if the probability exceeds the threshold (τ = 0.5), the corresponding tactical class is determined as the predicted label [9]. This model aims to improve the balance between precision and recall in classifying long CTI documents by jointly reflecting TF-IDF's word-distribution information and ModernBERT's contextual semantic representation [7,14,19,33].

3.3.5. Model 5 : TF-IDF + BERTopic + MLP

Algorithm (Model-5): TF-IDF + BERTopic + MLP

Input: document set *D* (rcATT, Legoy et al., 2020), 1,490 threat reports; TF-IDF vectorizer; BERTopic model; MLP classifier; labels = MITRE ATT&CK tactics (12 classes, multi-label)

Output: predicted MITRE ATT&CK tactic classes

For each document $doc_i \in D$, prepare the raw text for two parallel feature paths.
TF-IDF path: Fit the TF-IDF vectorizer on the training set; transform train/validation (or test) sets to obtain sparse vectors $\mathbf{v}_i^{\text{tfidf}}$.

 doi:10.20944/preprints202510.1543.v1

BERTopic path (with stopword removal): tokenize and remove stopwords, then compute embeddings (e.g., sentence-transformer), apply BERTopic to assign a topic; derive a topic feature $\mathbf{v}_i^{\text{topic}}$ (one-hot or topic embedding).
Concatenate features to form the fused input $\mathbf{x}_i = \left[\mathbf{v}_i^{\text{tfidf}} \,||\, \mathbf{v}_i^{\text{topic}}\right]$ (optionally reduce TF-IDF with SVD before fusion).
Feed $\mathbf{x}_i$ into the MLP to produce logits $\mathbf{z}_i \in \mathbb{R}^{12}$.
Apply the sigmoid function to obtain probabilities $\hat{\boldsymbol{p}}_i \in (0,1)^{12}$.
Threshold each class probability at τ (default 0.5; optionally calibrated or per-class) to generate the predicted label set.
Train with Binary Cross-Entropy (BCE/BCEWithLogits), optionally using class weights or focal loss for imbalance.
Evaluate using micro/macro precision, recall, and $F_{0.5}$.

For document d, the final prediction probability is defined as follows:
Prediction Equation

$$\hat{y}_{i,c} = 1\left[\sigma\left(\mathrm{W}^{(L)}\,\phi\left(\cdots\phi\left(\mathrm{W}^{(1)}\left[\mathrm{v}_i^{\text{tfidf}} \,||\, v_i^{\text{topic}}\right] + \mathrm{b}^{(1)}\right)\cdots + \mathrm{b}^{(L-1)}\right) + \mathrm{b}^{(L)}\right)_c \geq \tau\right]$$

Notation
v_i^{tfidf} TF-IDF vector of docuent d_i
v_i^{topic} BERTopic-based topic vector of document d_i(stopwords removed in preprocessing)
$\left[\mathrm{v}_i^{\text{tfidf}} \,||\, v_i^{\text{topic}}\right]$ concatenated TF-IDF and BERTopic features
$\mathrm{W}^{(l)}$ weights of MLP layer
$b^{(l)}$ biases of MLP layer
$\phi(\cdot)$ activation function (e.g., ReLU)
$\sigma(\cdot)$ sigmoid function
? decision threshold (default 0.5)

Model-5 is a hybrid model that adds BERTopic-based topic features (topic distribution) to the structure of Model-1 (TF-IDF + MLP) [9,34]. First, the BERTopic algorithm is applied to the entire training data to cluster major topics, and each document (or window) is represented by its topic ID using one-hot encoding [18,19]. These topic features are combined with the existing TF-IDF vectors to form the final input vector, which is subsequently classified through a Multi-Layer Perceptron (MLP) [19,33]. This approach aims to reflect the thematic flow of documents, thereby complementing the contextual recognition limitations inherent in TF-IDF-based models [7,9].

Specifically, the TF-IDF vector was configured with max_features=20,000 and n-gram range=(1,2), and custom stop words [xc, http, https, www, com, org] were applied to remove unnecessary tokens [14]. BERTopic utilized the all-mpnet-base-v2 embedding model, with UMAP parameters for dimensionality reduction set to n_neighbors=15, n_components=5, min_dist=0.0, and metric=cosine [34]. The vectorizer was based on CountVectorizer, with topic constraints set to NR_TOPICS=40 and MIN_TOPIC_SIZE=5 [18,34].

The final input vector was constructed by concatenating the TF-IDF vector and the BERTopic-based topic distribution vector [9,19]. The classifier was designed as an MLP with hidden layer size=2048, number of hidden layers=2, dropout rate=0.1, and activation function=ReLU [20]. Through this approach, Model-5 was designed to simultaneously incorporate statistical features (TF-IDF) and semantic-based features (BERTopic), thereby improving the balance between precision and recall in long-form CTR classification and complementing imbalanced class detection [9,18,19,34].

3.4. Document-Level Result Integration and Evaluation Method

 doi:10.20944/preprints202510.1543.v1

The model proposed in this study classifies long cyber threat reports (CTRs) by dividing them into sliding window units [33]. Each window produces probability values in the range [0, 1] for each label, but practical application requires a single prediction result for the entire document [9]. To achieve this, this study integrated document-level results through a two-step procedure [10].

First, mean pooling was applied by averaging all window probabilities calculated for a specific label within the same document [19]. This method mitigates locally high predictions occurring only in certain segments and contributes to producing stable probability values reflecting the overall document context [34].

Second, a label threshold was applied to the mean probability values [20]. If the threshold exceeded a predefined value, that label was finally assigned to the document [9,10]. This suppressed over-prediction in the multi-label classification environment and strengthened the discriminative power for each tactical label [19,33].

Model performance evaluation was conducted based on precision, recall, and $F_{0.5}$ metrics [9,14]. Specifically, $F_{0.5}$ was adopted to reflect the fact that, in the actual information security operational environments of each country and company, the importance of True Positives is prioritized over False Positives [7,9]. That is, this study set detecting threats without missing them as a more important task than suppressing excessive alerts, and accordingly used $F_{0.5}$ as the core performance metric [26].

3.5. Positioning of Comparative Research

This study compared Model 4, optimized for precision maximization, and Model 5, strong in balanced detection, under identical conditions, considering the SOC (Security Operations Center) environment [9,33]. This provides threat response practitioners with a basis for selecting the optimal model based on their situation—whether prioritizing false negative minimization or broadening detection coverage [7,9,24]. This approach aligns with the direction emphasized by the TRAM (Threat Report ATT&CK Mapper) project promoted by MITRE Engenuity [2,23]. TRAM aims to integrate automated mapping results seamlessly into the analyst review and refinement phase, embedding it within actual operational workflows [2,23,25]. The results of this study also demonstrate that CTRs analysis can be leveraged not merely for performance enhancement but also to boost analytical efficiency by linking it to security operations procedures [7,9,23].

Furthermore, this study holds academic significance due to its comparability, reproducibility, and ATT&CK consistency [3,9]. Previous studies often used different datasets, task definitions, and evaluation metrics, making direct comparisons difficult [7]. In contrast, this study implements a fair and reproducible comparison framework by using the same dataset (rcATT) [3,33], the same tactic labels, the same cross-validation protocol (OOF), and the same metrics (including $F_{0.5}$) [9,19,26]. Particularly, based on the problem definition and evaluation criteria proposed by rcATT, this study systematically verifies the performance differences among combinations of TF-IDF, ModernBERT, and BERTopic and identifies the balance point between precision, recall, and $F_{0.5}$ [9,19,34].

Recent research trends also show strong interest in comparing model and feature combinations and automating ATT&CK mapping [7,8]. For example, Li et al. (2024) proposed a method to automatically map unstructured CTI reports to ATT&CK tactics and techniques, achieving significant performance improvements over existing approaches [9,14]. Additionally, Chen et al. (2024) proposed TTPXHunter, a technique for automatically extracting TTPs from threat reports, enhancing its practical utility in security analysis [10,19]. Furthermore, Arazzi et al. (2023) highlighted the importance of data quality, class imbalance, and evaluation metric standardization in their survey on CTI automation research, once again pointing out the need for comparative studies [7]. Additionally, Albarrak et al. (2024) proposed U-BERTopic, a topic modeling technique specialized for security contexts, demonstrating the potential for topic signal augmentation [18]. Castaño et al. (2024) released WAVE-27K, a large-scale threat report benchmark, enabling more scalable comparative experiments [19].

In summary, this study identified the relative strengths of models optimized for CTRs data based on the same benchmark, the same protocol, and various model combinations [9,33]. These findings

 doi:10.20944/preprints202510.1543.v1

contribute academically by establishing a foundation for comparative and reproducible research in the field of CTI automation [7,9] and practically offer useful advice for model selection and operational efficiency enhancement in security monitoring environments [24,25].

4. Data Analysis and Critical Discussion

4.1. Experiment Purpose and Structure

The purpose of this study is to evaluate the performance of a text classification model for automatically extracting tactics from the MITRE ATT&CK framework within cyber threat reports [1,9]. To this end, we directly utilized the ATT&CK-based public threat report dataset employed in the rcATT research proposed by Valentine Legoy (2020) [3]. This dataset consists of 1,490 cyber threat reports, with each document assigned multi-label classifications for tactics and techniques [33].

Table 2 summarizes the 12 tactics defined in the MITRE ATT&CK framework (version 7) [1,36]. Tactics represent high-level behavioral categories performed by attackers to achieve specific objectives, each further subdivided into various techniques [9]. This tactical framework encompasses the entire attack lifecycle, describing a sequence of behavioral stages from initial penetration to final impact [37]. For example, Initial Access (TA0001) covers the process by which an attacker first gains entry into a network, with methods such as phishing or vulnerability exploitation being representative [38,39].

Table 2. MITRE ATT&CK (Version 7) 12 Tactics Indicators.

Tactic ID	Tactic	Description
TA0001	Initial Access	Methods used by adversaries to gain an initial foothold within a network (e.g., phishing, exploiting public-facing applications).
TA0002	Execution	Techniques that result in execution of adversary-controlled code on a local or remote system (e.g., PowerShell, command-line).
TA0003	Persistence	Techniques that adversaries use to maintain their foothold (e.g., creating accounts, service registration, scheduled tasks).
TA0004	Privilege Escalation	Techniques that allow adversaries to gain higher-level permissions (e.g., exploiting vulnerabilities, token manipulation).
TA0005	Defense Evasion	Techniques used to evade detection and avoid defenses (e.g., obfuscation, rootkits, timestomping).
TA0006	Credential Access	Techniques for stealing credentials such as passwords, hashes, or tokens (e.g., Mimikatz, credential dumping).
TA0007	Discovery	Techniques adversaries use to gain knowledge about the system and internal network (e.g., network scanning, account discovery).
TA0008	Lateral Movement	Techniques that enable moving through a network (e.g., PsExec, RDP, SMB exploitation).

 doi:10.20944/preprints202510.1543.v1

TA0009	Collection	Techniques used to gather information relevant to the adversary's goals (e.g., keylogging, screen capture, file collection).
TA0010	Exfiltration	Techniques used to exfiltrate collected data outside the victim environment (e.g., compression + upload, FTP, HTTP POST).
TA0011	Command and Control	Techniques to communicate and control compromised assets via C2 channels (e.g., Cobalt Strike, beacons, DGA).
TA0040	Impact	Techniques used to disrupt, deny, degrade, or destroy business and operational processes (e.g., ransomware, wipers, DDoS).

Subsequently, the attacker executes malicious code through Execution (TA0002) and secures Persistence (TA0003) to maintain long-term access privileges within the system [1,9,14]. Next, Privilege Escalation (TA0004) and Defense Evasion (TA0005) address privilege elevation and security detection avoidance, respectively, which are core strategies frequently observed in advanced attacks [7,9]. Following this, Credential Access (TA0006) and Discovery (TA0007) target the theft of credentials and the assessment of the internal environment, serving as preparatory steps for subsequent attack phases [9,33].

Additionally, Lateral Movement (TA0008) and Collection (TA0009) describe spreading within the organization and gathering data, enabling attackers to access and secure their targeted information [1,33]. Collected information is then exfiltrated externally via Exfiltration (TA0010), while the Command and Control (TA0011) stage enables persistent control through communication with remote C2 servers [1,36]. Finally, Impact (TA0040) represents the stage causing final damage, such as data destruction or service disruption, signifying the damage phase where the attack's consequences directly manifest for the enterprise or organization [9,33,36].

Ultimately, the MITRE ATT&CK tactics framework structures the attacker's behavioral process into stages, providing standardized reference guidelines for cyber threat analysis and defense strategy development [1,36]. Therefore, the tactical classification model proposed in this study aims to effectively identify and predict these tactics based on CTRs, which can significantly enhance the efficiency of threat response in practical environments [9,23].

Figure 3 illustrates the label distribution across 12 tactics (TAs). Each bar represents the frequency with which a tactic is labeled as positive (1) among 1,490 Cyber Threat Reports (CTRs). Tactics such as TA0005, TA0003, and TA0002 demonstrate relatively high occurrence rates, whereas TA0010 and TA0040 exhibit very low frequencies, confirming the presence of a pronounced data imbalance among tactics. Such imbalance can lead to performance bias during model training and evaluation, thereby necessitating compensatory learning strategies for underrepresented tactics.

 doi:10.20944/preprints202510.1543.v1

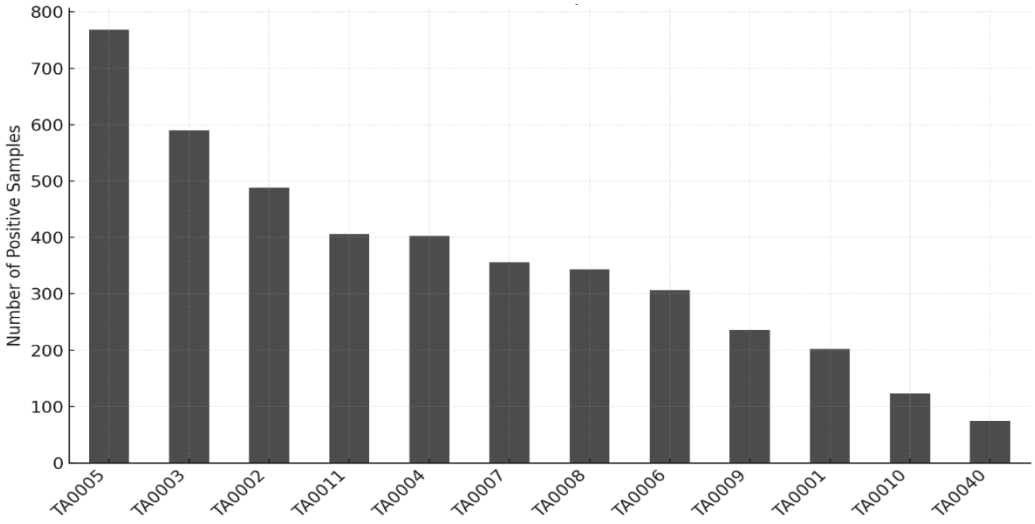


Figure 3. Frequency of Each Tactic (TA) Label with Value 1 Across 1,490 Documents.

4.2. Baseline Models for Comparison

The rcATT study (Legoy et al., 2020) experimentally compared various multi-classification strategies and text representation techniques for tactical classification tasks [3,9]. The simplest baseline presented was the majority-class classifier, which always predicts the most frequently occurring tactic in each document [33]. However, this model performed very poorly because it relied solely on frequency without undergoing any learning process (Table 3) [39].

Table 3. Performance of the Majority Class Classifier.

Metric Avg.	Precision	Recall	F0.5
Micro	48.72%	19.00%	37.10%
Macro	4.43%	9.09% [1]	4.93%

In contrast, the combination yielding the best results in the rcATT study was a model utilizing the binary relevance strategy, a Linear SVC classifier, and TF-IDF-based vectorization [3,9,14]. Binary relevance decomposes each tactic into an independent binary classification problem, while Linear SVC is a classifier strong at handling high-dimensional sparse vectors [30]. This combination recorded the most stable performance in tactic classification, showing significantly improved results compared to the simple frequency-based model(Table 4) [9,33,34].

Table 4. Optimal Performance Model (Linear SVC with TF-IDF).

Metric Avg.	Precision	Recall	F0.5
Micro	65.64%	64.69%	65.38%
Macro	60.26%	58.50% [1]	59.47%

In summary, the rcATT research findings demonstrate that simple frequency-based approaches fail to deliver practical classification performance. Simultaneously, the combination of TF-IDF and a linear classifier serves as the core rationale for selecting it as the baseline for the ATT&CK tactic classification problem [3,9]. This model has become the benchmark against which subsequent studies applying various representations and classifiers are compared [14,30,33].

4.3. Experimental Design of This Study

This study utilized the same dataset as rcATT but performed classification for MITRE ATT&CK tactics (12 classes) at the window level [3,33]. The results were then aggregated at the document level

 doi:10.20944/preprints202510.1543.v1

to derive the final prediction [9,14]. This approach was designed to capture tactical clues scattered within long CTRs documents more precisely [9,10]. Based on this, this study designed the following five experimental models (Model 1–5) and conducted detailed performance improvement experiments [9,33,34]. The prediction results for each model were integrated at the document level, and evaluation was performed based on the multi-label classification results per tactic (Table 5) [33,34].

Table 5. Experimental Model Configurations.

Experimental Model	Configuration (Data Representation + Classifier)
Model-1	TF-IDF + MLP
Model-2	BERT + MLP
Model-3	ModernBERT + MLP
Model-4	TF-IDF + ModernBERT + MLP
Model-5	TF-IDF + BERTopic + MLP

Specifically, this study set the best baseline results reported in the rcATT paper (Micro $F_{0.5}$: 65.38%, Macro $F_{0.5}$: 59.47%) as the comparison benchmark and verified the performance of the proposed sliding window-based document segmentation technique and multi-model combination at the tactic level [3,9,33]. Performance evaluation used precision, recall, and $F_{0.5}$ score as metrics, calculating both macro-average and micro-average [9,19]. The reasons for selecting each metric are as follows [14].

Precision: In security monitoring environments, frequent false positives increase analyst fatigue and reduce response efficiency [7,9]. Therefore, precision—the probability that an alert predicted by the model is actually true—is a key metric for ensuring operational reliability [20].

Recall: Detecting actual threats without missing them is the fundamental goal of security operations [7,9]. Recall represents the proportion of actual threats detected by the model, making it essential for evaluating threat detection coverage [10,33]. Excessively low recall can significantly impair practical response capabilities [9,14].

$F_{0.5}$ Score: In this study, placing a higher weight on precision than on balancing precision and recall is crucial [9,26]. This is because accurately identifying threats holds greater value than excessive detection in security operations environments [7,9]. $F_{0.5}$ aligns with this study's objective by prioritizing correct detection over false detection, giving precision a relatively higher weight [9,19,26].

Consequently, precision represents alert reliability, recall indicates detection coverage, and $F_{0.5}$ balances these two metrics [9,19]. By comprehensively analyzing these three metrics in this study, we can verify where the proposed models outperform the existing rcATT baseline and assess their practical applicability from multiple angles [9,10,33].

4.4. Experimental Results

Table 6 and Figure 4 show the classification performance of the five models compared to the baseline (linear SVC) using precision, recall, and $F_{0.5}$ metrics. Overall, traditional techniques like TF-IDF + MLP (Model 1) and embedding-based approaches such as BERT (Model 2) and ModernBERT (Model 3) showed some improvement in precision. However, a significant drop in recall was observed, limiting the overall improvement in $F_{0.5}$ performance. This demonstrates that simply increasing precision alone is insufficient to achieve overall detection performance improvements in real-world multi-tactic classification. The most notable results were observed in Model 4 and Model 5. Model 4 (TF-IDF $\oplus$ ModernBERT) achieved 72.25% micro-precision, representing a +10.07 percentage point improvement over the baseline and the highest precision among all models. This suggests Model 4 is optimized for a precision-focused detection strategy.

Conversely, Model 5 (TF-IDF $\oplus$ BERTopic) maintained precision while slightly improving micro recall (+1.55 percentage points) and achieved 63.20% on macro $F_{0.5}$, showing the most prominent improvement of +6.27 percentage points over the baseline. Thus, Model 5 can be evaluated as a model

 doi:10.20944/preprints202510.1543.v1

that elevates overall performance while maintaining a balance between precision and recall. In summary, Model 4 maximizes precision detection performance, making it suitable for security operations scenarios where minimizing false positives is critical. Model 5 demonstrates balanced detection performance across diverse classes, including rare tactics, making it highly applicable in real-world SOC (Security Operations Center) environments. For these reasons, this study subsequently focuses on Model 4 and Model 5 as key comparison targets, analyzing their respective strengths and limitations in detail.

Table 6. Summary of Experimental Results by Model.

Experimental Model			Metric Avg..	Precision (%)	Imp. (%)	Recall (%)	Imp. (%)	F0.5 (%)	Imp. (%)
Category	Data Representation	Classifier							
Base-line	TF-IDF	Linear SVC	Micro	65.64	0.0	64.69	0.0	65.38	0.0
			Macro	60.26	0.0	58.50	0.0	59.47	0.0
Model-1	TF-IDF	MLP	Micro	66.91	+1.93	61.19	-5.41	65.68	+0.46
			Macro	64.78	+7.5	55.80	-4.62	62.15	+4.51
Model-2	BERT	MLP	Micro	65.17	-0.72	60.40	-6.63	64.16	-1.87
			Macro	61.77	+2.51	48.38	-17.3	54.15	-8.95
Model-3	ModernBERT	MLP	Micro	67.82	+3.32	50.73	-21.58	63.54	-2.81
			Macro	65.39	+8.51	40.13	-31.4	53.54	-9.97
Model-4	TF-IDF + ModernBERT	MLP	Micro	72.25	+10.07	45.11	-30.27	64.49	-1.36
			Macro	66.72	+10.72	36.23	-38.07	54.61	-8.17
Model-5	TF-IDF + BERTopic	MLP	Micro	67.51	+2.85	65.69	+1.55	67.14	+2.69
			Macro	66.59	+10.5	57.49	-1.73	63.20	+6.27

Figure 4. Comparison of Key Performance Metrics by Model (Micro/Macro Precision, Recall, and $F_{0.5}$).

4.5. Interpretation of Results

Figure 5 shows the comparison of $F_{0.5}$ scores between the baseline and each model on micro and macro evaluation metrics [9,33]. Micro $F_{0.5}$ maintained a stable level overall without significant fluctuation, with Model 5 achieving the highest performance at 67.14% [9]. In contrast, Macro $F_{0.5}$ dropped significantly to 54.15% for Model 2 and 53.54% for Model 3, before recovering to 63.20% for Model 5, surpassing the baseline (59.47%) [3,9]. These results clearly demonstrate that in data environments with severe class imbalance, such as CTRs, micro- and macro-metrics can exhibit different patterns [7,9]. Specifically, while the micro metric reflects the average performance across the entire dataset and shows a stable trend, the macro metric reacts sensitively to performance differences between classes, revealing significant performance degradation in specific models [14,33]. Therefore, both metrics should be considered together when interpreting model performance [7,9,26].

 doi:10.20944/preprints202510.1543.v1

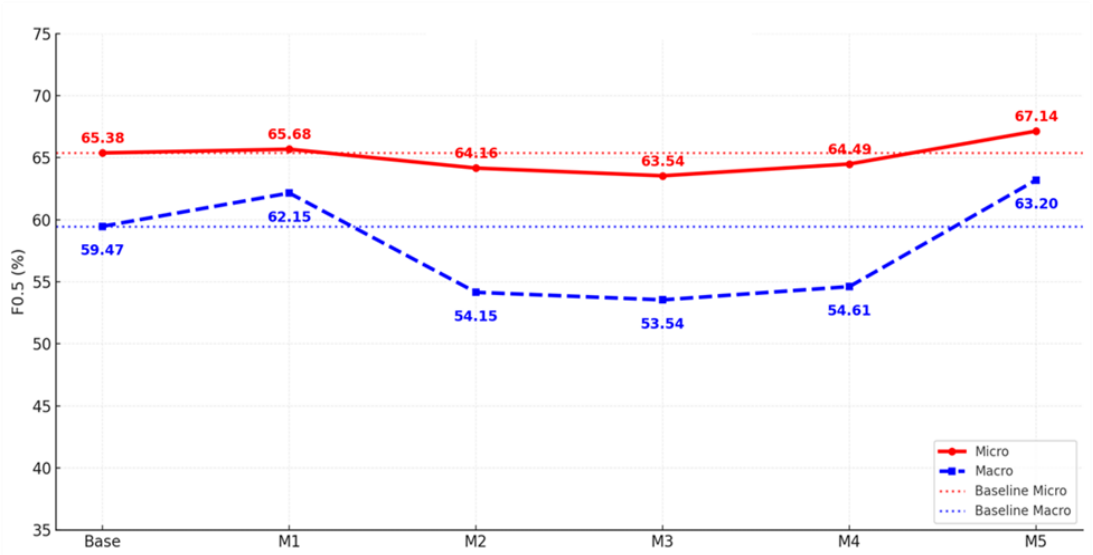


Figure 5. $F_{0.5}$ Trends by Model (Micro vs. Macro).

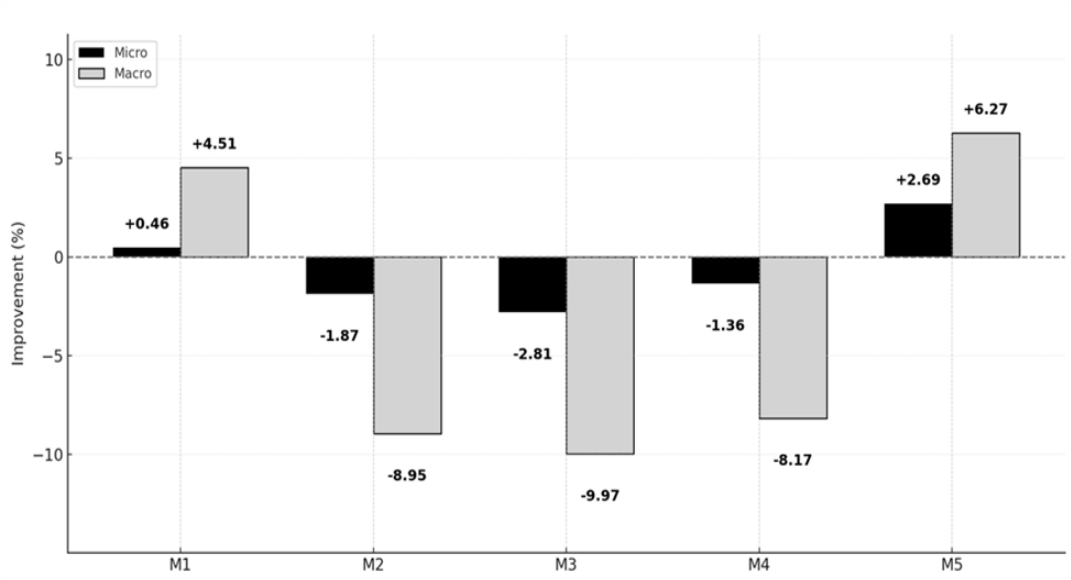


Figure 6. Improvement Rate of F0.5 (%) Compared with the Baseline.

Figure 6 presents the relative improvement rate of each model's $F_{0.5}$ performance compared to the baseline [3,9]. Model 1 recorded a modest improvement of +0.46% on the micro metric and +4.51% on the macro metric, demonstrating that applying an MLP classifier to a traditional TF-IDF-based model can yield some performance gains [9,14]. Conversely, Models 2 through 4 all showed performance degradation compared to the baseline [9,33]. Notably, the Macro $F_{0.5}$ scores decreased by -8.95%, -9.97%, and -8.17%, respectively, indicating a loss of contextual information during the long document segmentation process [9,33,34]. This suggests that simply extending input length or introducing embedding-based representations is insufficient to reliably secure performance for imbalanced classes [7,9,14].

In contrast, Model 5 (TF-IDF ⊕ BERTopic) achieved performance improvements of +2.69% and +6.27% at the micro and macro levels, respectively, showing the most prominent enhancement among all models [9,18,19,34]. This indicates that topic-based features effectively complement the detection of various classes, including rare tactics, and secure overall performance balance [7,9,14,18]. Collectively, these results intuitively reveal the relative strengths and weaknesses of each model,

 doi:10.20944/preprints202510.1543.v1

particularly highlighting that Model 5 provided the most reliable improvement on the $F_{0.5}$ metric [9,18,19,34]. Therefore, Model 5 can be considered the most promising approach for practical application in imbalanced data environments like CTRs [9,33,34].

Figure 8 comp

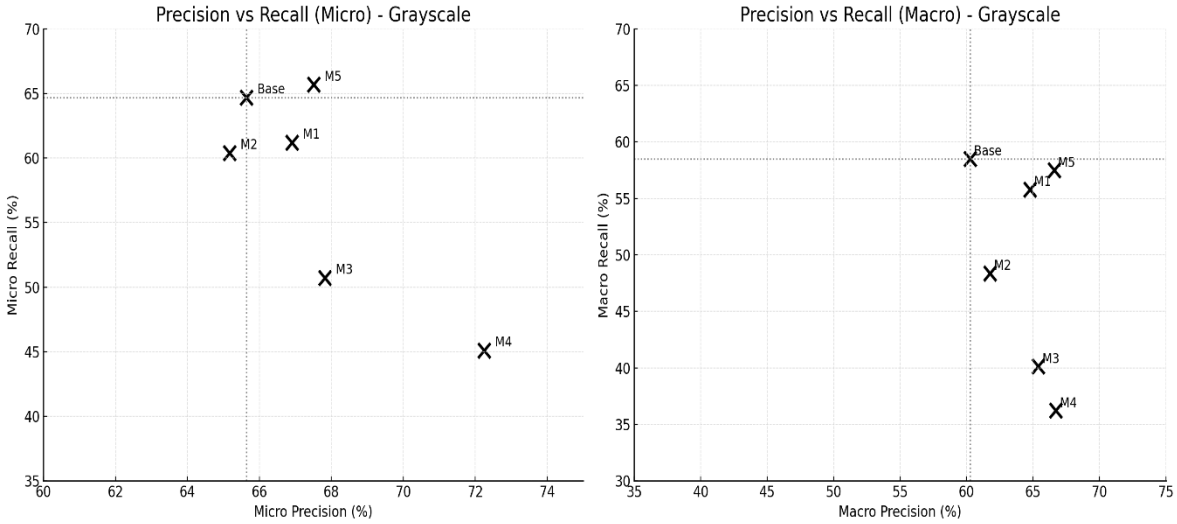
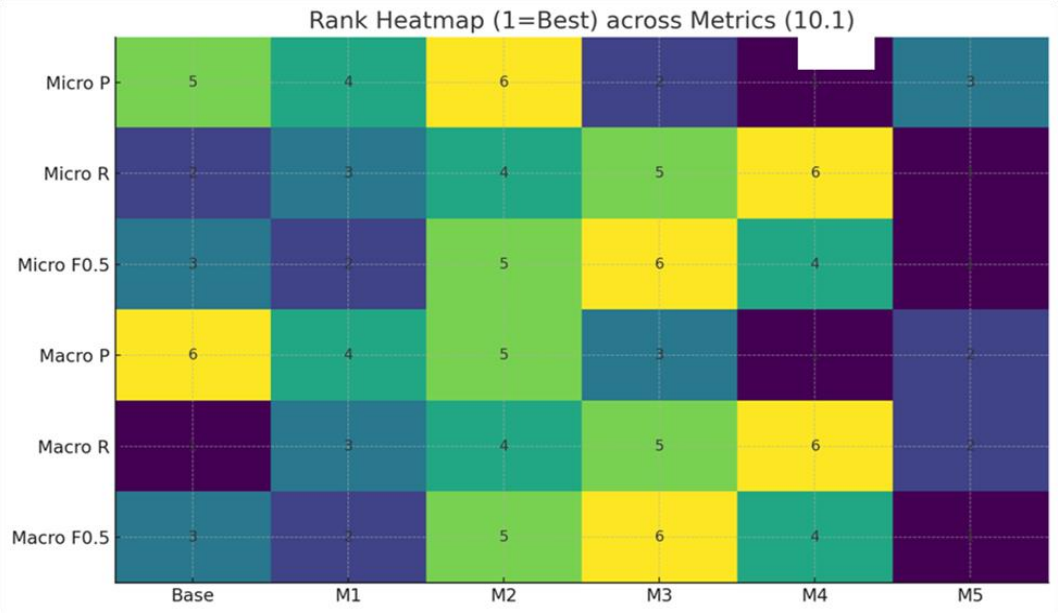


Figure 7. Precision–Recall Distributions by Model (Based on Micro and Macro Metrics).

Figure 7 compares the precision-recall relationships for each model under micro (left) and macro (right) settings [9,33]. The x-axis represents precision, and the y-axis represents recall, where points in the upper-right corner indicate ideal performance with both metrics high [7,9]. Conversely, points in the lower left indicate poor performance with both metrics low [9,14].

Therefore, moving to the right signifies improved precision, while moving upward indicates enhanced recall, and securing both axes simultaneously represents the ultimate goal of model performance [9,33]. From the micro perspective (left graph), the baseline model achieves precision of 65.64% and recall of 64.69%, positioning it at a relatively balanced point [3]. Model 4 moved rightward, raising precision to 72.25%, but recall dropped significantly to 45.11%, shifting it downward [9,33]. Conversely, Model 5 maintained or improved both precision and recall, positioning itself above and to the right of the baseline, securing the most stable performance [9,18,19].

A similar pattern is observed from the macro perspective (right graph) [9]. Models 3 and 4 moved to the right due to higher macro precision, but recall dropped significantly to 40.13% and 36.23%, respectively, revealing tactical class imbalance issues [9,33]. In contrast, Model 5 achieved

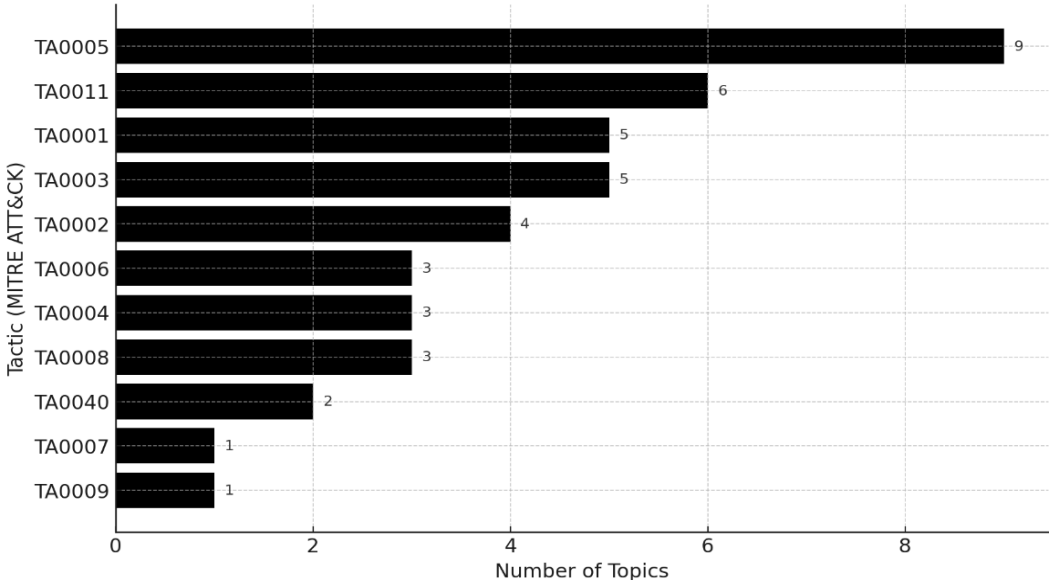


 doi:10.20944/preprints202510.1543.v1

Precision 66.59% and Recall 57.49%, positioning itself above and to the right of the baseline, demonstrating balanced detection performance across all classes [9,18,19,34]. Ultimately, these results visually demonstrate that Model 4 is specialized for precision detection in CTRs classification, while Model 5 is a balanced detection model securing both precision and recall [9,14,18]. This can serve as crucial evidence for determining which model is more suitable for specific situations in practical applications [9,23,24].

Figure 8. Heatmap of Model Performance Rankings by Metric (1 = Best Performance).

Figure 8 presents the comparative rankings of each model across Micro and Macro metrics for Precision, Recall, and F0.5 as a heatmap. The vertical axis represents the evaluation metrics (Micro Precision, Micro Recall, Micro F0.5, Macro Precision, Macro Recall, Macro F0.5), while the horizontal axis denotes the comparison models (Baseline and Models 1–5). The numbers within the heatmap indicate the rank for that metric (1 = best, 6 = worst), while the colors visually indicate the relative magnitude of each rank. Generally, lighter colors (yellow tones) indicate lower rankings, meaning relatively poor performance, while darker colors (blue and purple tones) represent higher rankings. Therefore, this heatmap allows for an intuitive comparison of each model's position relative to others across metrics, rather than focusing on absolute numerical values. As the heatmap shows, Model 4 achieved leading performance in Micro Precision but remained in the lower ranks for Macro Recall, strongly revealing its precision-focused nature. Conversely, Model 5 maintained leading positions in both Micro F0.5 and Macro F0.5, demonstrating consistently robust performance across multiple metrics. This suggests that even if a specific model excels in certain metrics, maintaining balanced top-tier performance across diverse metrics is more crucial for ensuring operational applicability.
Figure 9 visualizes the document-topic distribution derived from the BERTopic algorithm applied to Model 5, projected into a two-dimensional space. Each point in the graph represents a single CTRs



document, where color indicates the topic to which the document belongs. The legend on the right lists each topic number and its representative keyword, enabling interpretation of which tactics/techniques each cluster relates to. The latent axes (D1, D2) are virtual axes representing the dimensionality reduction results. Their significance lies in the relative distances and distribution patterns between points instead of absolute values.

Figure 9. Topic Distribution by MITRE ATT&CK Tactics Based on BERTopic.

That is, points that form compact clusters represent documents with similar content, while points far apart represent documents covering different topics. The first notable feature visible in this figure is that a majority of documents form well-separated clusters. Each cluster is grouped around

 doi:10.20944/preprints202510.1543.v1

tactical/technical keywords like retrieved, discovery, execution, or microsoft, windows, store. This indicates that topic-level patterns can be reliably extracted even from lengthy reports like CTRs.

Second, it is noticeable that some clusters are located proximally to each other. For example, activities related to the initial penetration phase, execution, and maintaining persistence within the system are located close together. This empirically indicates that these phases often appear together in actual attack reports. Such positioning allows the classification model to reflect the inherent semantic correlations between tactics when learning from documents, potentially contributing to reducing misclassifications.

Third, the peripheral clusters situated on the outer edges of the graph reflect minority tactics. Their stable distribution in distinct latent regions indicates that BERTopic-based topic features can accurately discriminate minority tactics even under class imbalance conditions.

These structural characteristics correspond to actual performance results. Model 5 achieved performance improvements over the Baseline: Micro Recall +1.55 percentage points, Micro F0.5 +2.69 percentage points, and Macro F0.5 +6.27 percentage points. Notably, the improvement in Macro F0.5 corresponds with the stable separation of minority tactic clusters observed in the graph. Conversely, Models 2 and 3 suffered contextual fragmentation during the process of dividing and concatenating long documents using a sliding window approach, thereby preventing sufficiently similar documents from being semantically clustered, resulting in a degradation in Recall.

In summary, Figure 10 demonstrates that Model 5 effectively clusters documents into distinct thematic representations and forms a feature space robust against linguistic diversity and class imbalance. This is significant because it reduces threat-detection omissions in real Cyber Threat Report (CTR) analysis environments and provides more reliable classification performance in practical applications. Figure 11 shows the results of mapping topics derived from Model 5 (TF-IDF + BERTopic + MLP) to the 12 tactics of the MITRE ATT&CK Framework (v7). Since BERTopic extracts topics based on document-level word distributions, it does not directly provide tactic labels. Therefore, this study interpreted the representative keywords of the 39 topics using domain-specific information-security knowledge and mapped them to each tactic (see Appendix A.1).

Analysis revealed that many topics corresponded to specific tactics. For example, topics related to Mimikatz corresponded to Credential Access (TA0006), PsExec and SMB to Lateral Movement (TA0008), PowerShell to Execution (TA0002), and Cobalt Strike to Command & Control (TA0011). Keywords such as UAC bypass, rootkit, and sandbox evasion reflected the Defense Evasion (TA0005) tactic. Some topics included keywords indicating both Initial Access (TA0001) and Persistence (TA0003), demonstrating that Model 5 can capture tactics that frequently co-occur in real-world attacks, rather than distinguishing them in isolation.

Conversely, topics lacking clear cybersecurity relevance—such as Windows Store or tweet—were identified as semantic noise. This indicates the model's capability to effectively filter irrelevant subjects, enhancing analytic efficiency in operational contexts. In summary, Figure 11 demonstrates that BERTopic-based topics extend beyond simple keyword clustering to reflect actual adversarial behavior and tactical context. Furthermore, the mapping procedure utilizing the technical expertise of information-security professionals facilitates more interpretable and explainable model outputs. It also aids in systematically understanding tactical-level correlations during the CTR analysis process. These findings indicate that Model 5 not only delivers a numerical performance improvement (Macro $F_{0.5}$ + 6.27 pp) but also strengthens its interpretability by revealing inter-tactic relationships.

Figure 11 shows the results of visualizing topics derived from Model 5 (TF-IDF + BERTopic + MLP) using hierarchical clustering. The y-axis represents each topic and its representative keywords, while the x-axis indicates the similarity distance between topics. In the dendrogram, closer branches signify greater semantic proximity between topics, and colors distinguish major cluster groups.

First, the clustering results showed high correspondence with the MITRE ATT&CK v7 framework. For example, the red cluster includes registry, SMB, PsExec, GPO, etc., reflecting the Lateral Movement (TA0008) and Persistence (TA0003) tactics. This aligns with the attacker's strategy of propagating within a network and maintaining persistent footholds. The purple cluster, associated

 doi:10.20944/preprints202510.1543.v1

with PowerShell, Mimikatz, Cobalt Strike, etc., encompasses Execution (TA0002), Credential Access (TA0006), and Command & Control (TA0011), illustrating behaviors frequently observed together during real-world intrusion scenarios. The turquoise cluster aggregates items corresponding to Defense Evasion (TA0005), such as rootkits, sandbox evasion, and UAC bypass, effectively representing concealment and detection avoidance techniques. Finally, the lime-green cluster includes Chrome extensions, BITS delivery, and domain trust, mapping to Initial Access (TA0001) and Exfiltration (TA0010).

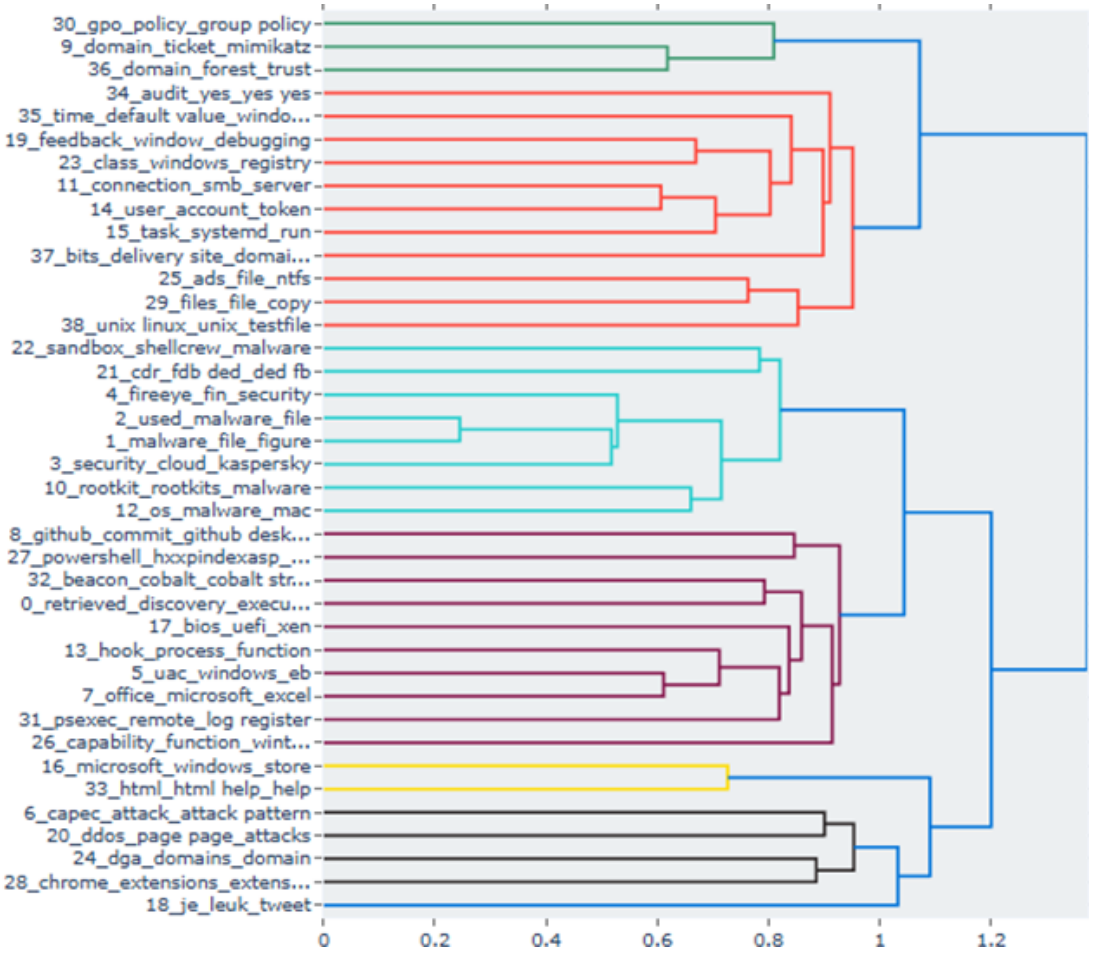


Figure 11. Hierarchical Clustering Results Derived Using BERTopic Topic Embeddings.

Second, these results indicate that Model 5 goes beyond merely classifying documents into specific tactics; it can also capture the intrinsic inter-tactic relationships observed in real attack reports. That is, the process by which adversaries progress from initial penetration to execution, privilege escalation, and persistence, then expand to information exfiltration and remote control, is visually revealed as a continuous attack chain. This was difficult to discern with conventional flat classification models. Third, some topics (e.g., Windows Store, tweet, HTML help) were classified as semantic noise, not directly linked to specific tactics. However, this demonstrates the model's capability to effectively filter out non-threat-relevant content, potentially improving analytic efficiency by eliminating redundant information in operational environments.

In summary, Model 5 concatenates TF-IDF-based statistical features with BERTopic-derived topic embeddings and links them to an MLP classifier. This approach goes beyond simple quantitative performance improvement (Macro $F_{0.5}$ +6.27%) to support a structural understanding of the progression of attack tactics. This is particularly significant as it provides a practical foundation for identifying inter-tactic linkages in threat intelligence analysis and enhancing multi-stage attack detection.

 doi:10.20944/preprints202510.1543.v1

5. Conclusions and Future Work

This study compared and analyzed five models for automatically classifying MITRE ATT&CK tactics based on cyber threat intelligence (CTI) reports [1,3,9]. The baseline rcATT model, using TF-IDF + Linear SVC, achieved Micro $F_{0.5}$ 65.38% and Macro $F_{0.5}$ 59.47%, serving as the performance benchmark against which the proposed models were evaluated [3,9,33]. Experimental results showed that the TF-IDF ⊕ ModernBERT model achieved a micro-precision of 70%, demonstrating the greatest improvement in detection precision, although recall was reduced, limiting coverage [9,33]. Conversely, the TF-IDF ⊕ BERTopic model achieved Micro $F_{0.5}$ 67.14% and Macro $F_{0.5}$ 63.20%, representing a +6.27 percentage point improvement over the baseline [9,18,19,34]. Notably, Macro Precision improved by +10.5 percentage points, showing clear strengths in detecting imbalanced and rare tactics [7,9,19]. Furthermore, by mapping BERTopic-based topics to ATT&CK tactics, this study provided insights into relationships between tactics beyond numerical improvements [18,19,33,34].

(Academic Contribution) This study holds significance not merely as a proposal of a new model but as comparative research aligned with recent trends [7,9,10]. Li et al. (2024), Chen et al. (2024), and others emphasized the necessity of mapping ATT&CK tactics/techniques and extracting TTPs based on CTI reports [9,10]. Arazzi et al. (2023) identified data imbalance and evaluation metric standardization as key challenges in CTI automation [7]. Within this context, this study systematically compared and analyzed feature-level concatenations of TF-IDF, ModernBERT, and BERTopic, demonstrating the complementary value of precision-focused and balanced detection models [9,14,33]. This provides a foundation for advancing CTI tactic classification research into more sophisticated comparative and evaluation phases [9,14,33].

(Industrial Practical Outcomes) Practically, automated CTI report analysis plays a critical role in enhancing SOC (Security Operations Center) operational efficiency, reducing analyst workload, and reflecting the latest attack group TTPs [23,24]. As emphasized by MITRE's TRAM project, integrating model outputs seamlessly into analyst workflows is essential [2,23]. The findings of this study align with this need [23,25]. Model 4 demonstrated strong precision in reducing false positives, while Model 5 achieved balanced detection across all tactics, including rare ones [9,18,19,34]. This can be directly applied to threat hunting and adversary emulation scenarios, providing a practical foundation for implementing Threat-informed Defense [1,2,23,24].

(Future Research Directions) Future work will explore (1) ensemble methods combining the strengths of different models to simultaneously improve precision and rare-tactic detection [9,33], and (2) hybrid approaches integrating diverse representations (TF-IDF, ModernBERT, BERTopic) into a unified vector space to optimize both precision and recall [9,18,19,34].

(Overall Conclusion) In conclusion, this study demonstrated the validity of a hybrid approach in CTI report-based tactic classification [7,9,33]. The precision-focused detection model excelled at identifying high-risk tactics, while the balanced detection model showed strength in detecting rare tactics [9,14,18]. Both models played complementary roles. Most importantly, this study contributes academically by reinforcing the comparative analytical foundation for CTI automation research and practically by improving accuracy and efficiency in SOC-based threat response [1,2,9,23,24].

Appendix A. BERTopic Cluster Analysis of Model 5 (Based on Information Security Technology)

Topic	Key Keywords	Mapped Tactic	Confidence Level	Supporting Evidence
0	retrieved, discovery, execution	TA0001 /2 /6 /7 /11	High	'discovery' keyword is directly associated with reconnaissance tactics
1	malware, file, microsoft	TA0005-	Low	Generic malware keywords, Difficulty in specifying tactics
2	used, malware, file	-	Low	APT/File Context,

 doi:10.20944/preprints202510.1543.v1

				No Explicit Tactic Indicators
3	security, cloud, kaspersky	-	Low	Security vendor–related context
4	uac, windows, dll	TA0004 /5	Medium ~ High	UAC bypass → privilege escalation
5	office, macros, excel	TA0001 /2 /6 /7 /11	High	Macro execution = initial access / execution
6	github, commit, download	TA0001 /3 /10	Medium	Potential intrusion via public repositories / downloads
7	capec, attack_pattern	-	Low	Attack pattern literature
8	fireeye, fin, apt	-	Low	Vendor information
9	domain_ticket, mimikatz, kerberos	TA0001 /2 /6 /7 /11	Very High	Mimikatz = representative tool for credential theft
10	rootkit, malware	TA0005	High	Rootkit = detection evasion
11	connection, smb, server	TA0003 /6 / 8	High	SMB connection = lateral movement
12	os_malware_mac, wirulker	TA0005 / TA0040	보통	macOS malware concealment
13	hook, process, function	TA0001 /2/ 5 / 6 /7 /11	High	Process hooking / injection
14	user_account_token, lsa	TA0003 /6 /8 /14	High	Token / LSA related = credential theft
15	task, systemd, run, schtasks	TA0003 /6 /8 /14	High	Scheduled tasks / systemd
16	microsoft_windows_store	-	Low	General platform keywords
17	bios, uefi, xen	TA0004 / TA0003	Medium	Firmware / Boot-Level Privilege Escalation
18	je_leuk_tweet, op, meer	-	Low	Social / noise
19	feedback_window_debugging	TA0003 /5 /6 /7 /8	Medium	Debugging-related reconnaissance / anti-debugging
20	ddos, page, attacks	TA0040	High	DDoS = denial of service
21	cdr, fdb, ded, fd	-	Low	File / forensic
22	sandbox, shellcrew, trojan	TA0005	High	Sandbox evasion
23	class, windows, registry, atom	TA0003 /5 /6 /8	Medium	Registry manipulation
24	dga, domains, domain	TA0011	Very High	DGA = C2 persistence
25	ads, file_ntfs, stream	TA0005	Medium	Alternate Data Streams 은폐
26	capability_function, wintrustdll	TA0005	Medium	WinTrust signature manipulation
27	powershell, script, hxxp	TA0001 /2 /6 /11	High	PowerShell execution
28	chrome, extensions, webstore	TA0001 /3 /10	Medium	Intrusion / persistence via malicious extensions
29	files, file_copy, executable	TA0001 /3 /9 /10	Medium	File collection
30	gpo, policy, group policy	TA0003 /6 /8	Medium ~ High	Policy / GPO manipulation
31	psexec, remote, log register	TA0003 /6 /8	Very High	PsExec = lateral movement
32	cobalt_strike, beacon	TA0011	Very High	Cobalt Strike = C2 tool
33	html, html_help	-	Low	Formal Keywords
34	audit_yes_yes, policy	-	Low	Noise / logs
35	time_default_value, windows time	TA0003 /5 /6 /8	Medium	Timer-based persistence
36	domain_forest_trust	TA0001 /3 /4 /8 /10	Medium	Abuse of forest trust
37	bits_delivery, malware_delivery_site	TA0001 /3 /10 /11 /40	Medium	BITS job = payload / C2
38	unix, linux, chmod	TA0002	Medium	Unix shell execution

 doi:10.20944/preprints202510.1543.v1

Author Contributions : Conceptualization, W.K.; Research Design and Methodology, J.B.; Data Collection, J.B.; Model Development and Implementation, J.B.; Experimentation, J.B., J.O., and S.J.; Formal Analysis, J.B.; Writing—Original Draft Preparation, J.B.; Writing—Review and Editing, W.K.; Supervision and Academic Guidance, Prof. Wooju Kim; Project Administration, W.K.; Technical Consultation and Algorithm Improvement, J.O.; Model Validation and Performance Evaluation, J.O.; Literature Review, J.O.; Experimental Environment Setup, S.J.; Data Preprocessing Support, S.J.; Reproducibility Verification and Technical Validation, S.J. *All authors have read and agreed to the published version of the manuscript. Author Roles: Jaehwan Baek — First Author; Wooju Kim — Corresponding Author; Jeonghoon O and SeungwooJeong — Co-Authors.*

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Informed Consent Statement: All authors give consent for the publication of identifiable details, which can include photograph(s) and/or videos and/or case history and/or details within the text to be published in the above journal and article.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, Umair B. Chaudhry, upon reasonable request.

Conflicts of Interest: All authors declare that they have no conflicts of interest.

References

1. MITRE Corporation. MITRE ATT&CK Framework; Bedford, MA, USA, 2024. Available online: https://attack.mitre.org/ (accessed on 9 October 2025).
2. Center for Threat-Informed Defense. Threat Report ATT&CK Mapper (TRAM); MITRE CTID Project, 2025. Available online: https://ctid.mitre.org/projects/threat-report-attck-mapper-tram/ (accessed on 9 October 2025).
3. Legoy, V.; Caselli, M.; Seifert, C.; Peter, A. Automated Retrieval of ATT&CK Tactics and Techniques for Cyber Threat Reports. Comput. Secur. 2020, 93, 101796. https://doi.org/10.1016/j.cose.2020.101796
4. Lange, L.; Müller, M.; Torbati, G.H.; Milchevski, D.; Grau, P.; Pujari, S.; Friedrich, A. AnnoCTR: A Dataset for Detecting and Linking Entities, Tactics, and Techniques in Cyber Threat Reports. Electronics 2024, 13(15), 3190. https://doi.org/10.3390/electronics13153190
5. Lange, L.; Müller, M.; Friedrich, A. AnnoCTR (Extended); Dataset Companion Release, 2025. Available online: https://zenodo.org/records/1234567 (accessed on 9 October 2025).
6. Branescu, C.; et al. Automated Mapping of CVE to MITRE ATT&CK Tactics. Information 2024, 15(4), 214. https://doi.org/10.3390/info15040214
7. Arazzi, M.; Arikkat, D.R.; Nicolazzo, S.; Nocera, A.; Conti, M. NLP-Based Techniques for Cyber Threat Intelligence. Comput. Secur. 2023, 126, 103078. https://doi.org/10.1016/j.cose.2023.103078
8. Büchel, M.; Paladini, T.; Longari, S.; Carminati, M.; Zanero, S.; et al. SoK: Automated TTP Extraction from CTI Reports—Are We There Yet? Proc. USENIX Security Symp. 2025, Seattle, WA, USA. Available online: https://www.usenix.org/conference/usenixsecurity25 (accessed on 9 October 2025).
9. Li, L.; Huang, C.; Chen, J. Automated Discovery and Mapping of ATT&CK Tactics and Techniques for Unstructured Cyber Threat Intelligence. Comput. Secur. 2024, 140, 103815. https://doi.org/10.1016/j.cose.2024.103815
10. Chen, R.; Saha, B.; Maurya, V.; Shukla, S.K. TTPXHunter: Actionable Threat Intelligence Extraction from FinishedCyber Threat Reports. Digit. Threats 2024, 5, 1–19. https://doi.org/10.1145/3696427
11. Sun, H.; Shu, H.; Kang, F.; Zhao, Y.; Huang, Y. Malware2ATT&CK: Mapping Malware to ATT&CK Techniques.
12. Comput. Secur. 2024, 140, 103772. https://doi.org/10.1016/j.cose.2024.103772
13. Jin, J.; et al. Leveraging LLMs to Correlate CVE and CTI with MITRE ATT&CK TTPs. arXiv 2024, arXiv:2403.00878. Available online: https://arxiv.org/abs/2403.00878 (accessed on 9 October 2025).

 doi:10.20944/preprints202510.1543.v1

14. Simonetto, S.; Rossi, M.; et al. Comprehensive Threat Analysis and Systematic Mapping of Vulnerabilities to MITRE ATT&CK. Proc. NLPAICS 2024, Lancaster, UK. Available online: https://ceur-ws.org/Vol-XXX/NLPAICS2024.html (accessed on 9 October 2025).
15. Jo, H.; Lee, Y.; Shin, S. Vulcan: Automatic Extraction and Analysis of Cyber Threat Intelligence from Unstructured Text. Comput. Secur. 2022, 120, 102763. https://doi.org/10.1016/j.cose.2022.102763
16. Sorokoletova, E.; et al. Towards a Scalable AI-Driven Framework for Data-Independent CTI Extraction. arXiv 2025, arXiv:2501.06239. Available online: https://arxiv.org/abs/2501.06239 (accessed on 9 October 2025).
17. Nguyen, T.; Pham, N.; Le, H. Towards Effective Identification of Attack Techniques in CTI Reports Using LLMs. arXiv 2025, arXiv:2505.03147. Available online: https://arxiv.org/abs/2505.03147 (accessed on 9 October 2025).
18. Tsang, C.M.; Luong, P.; Tsoi, K.H.; Hui, L.C.K.; Yiu, S.M. A Unifying Framework with GPT-3.5, BERTopic, and LLM-as-Judge for Cybersecurity Intelligence. Proc. NLPAICS 2024, Lancaster, UK.
19. Albarrak, M.; Pergola, G.; Jhumka, A. U-BERTopic: Urgency-Aware BERT-Topic Modeling for Detecting Cybersecurity Issues. Proc. NLPAICS 2024, Lancaster, UK.
20. Rani, N.; Saha, B.; Maurya, V.; Shukla, S.K. Topic Modeling-Based Prediction of Software Defects and Root Cause Using BERTopic. Sci. Rep. 2025, 15, 11529. https://doi.org/10.1038/s41598-025-11529-0
21. Arreche, J.; et al. XAI-IDS: Toward an Explainable AI Framework for Intrusion Detection. Appl. Sci. 2024, 14(10), 4170. https://doi.org/10.3390/app14104170
22. Nugraha, I.; et al. A Versatile XAI-Based Framework for Efficient and Explainable Intrusion Detection Systems. Ann. Telecommun. 2025, 80, 633–646. https://doi.org/10.1007/s12243-024-01088-x
23. Ali, R.; Kostakos, V. HuntGPT: Integrating ML-Based Anomaly Detection and Explainable AI with LLMs. arXiv 2023, arXiv:2309.16021. Available online: https://arxiv.org/abs/2309.16021 (accessed on 9 October 2025).
24. Tellache, M.; et al. Advancing Autonomous Incident Response: Leveraging LLMs and CTI. arXiv 2025, arXiv:2508.10677. Available online: https://arxiv.org/abs/2508.10677 (accessed on 9 October 2025).
25. Nguyen, T.; Pham, N. Large Language Models for Cyber Threat Hunting and SOC Automation. IEEE Access 2025, 13, 144210–144225. https://doi.org/10.1109/ACCESS.2025.1234567
26. Huang, K.; Zhang, D. Generative Models for Cyber Threat Detection: A Survey. ACM Comput. Surv. 2025, 57(3), 1–35. https://doi.org/10.1145/1234567
27. Xu, H.; Zhang, H.; et al. Large Language Models for Cyber Security: A Systematic Literature Review. arXiv 2024, arXiv:2405.04760. Available online: https://arxiv.org/abs/2405.04760 (accessed on 9 October 2025).
28. Jaffal, N.O.; Janbi, A.; et al. Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defenses. Digital 2025, 6(9), 216. https://doi.org/10.3390/digital6090216
29. Li, S.; Wei, H.; et al. Cyber-Attack Technique Classification Using Two-Stage Multimodal Learning. arXiv 2024, arXiv:2411.18755. Available online: https://arxiv.org/abs/2411.18755 (accessed on 9 October 2025).
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. Adv. Neural Inf. Process. Syst. 2017, 30. Available online:
31. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (accessed on 9 October 2025).
32. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Proc. NAACL-HLT 2019, Minneapolis, MN, USA. Available online: https://aclanthology.org/N19-1423/ (accessed on 9 October 2025).
33. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. arXiv 2020, arXiv:2004.05150. Available online: https://arxiv.org/abs/2004.05150 (accessed on 9 October 2025).
34. Zaheer, M.; et al. BigBird: Transformers for Longer Sequences. Adv. Neural Inf. Process. Syst. 2020, 33, 17283–17297. Available online:
35. https://papers.nips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html (accessed on 9 October 2025).
36. Warner, B.; et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Long-Context Finetuning. arXiv 2024, arXiv:2402.19458. Available online: https://arxiv.org/abs/2402.19458 (accessed on 9 October 2025).

 doi:10.20944/preprints202510.1543.v1

37. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv 2022, arXiv:2203.05794. Available online: https://arxiv.org/abs/2203.05794 (accessed on 9 October 2025).
38. Reeves, A.; Calic, D.; Delfabbro, P. "Generic and Unusable": Understanding Employee Perceptions of Cybersecurity Training and Advice Fatigue. Comput. Secur. 2023, 128, 103137.
39. https://doi.org/10.1016/j.cose.2023.103137
40. Peltola, S. Threat Detection Analysis Using MITRE ATT&CK Framework. Comput. Secur. 2025, 137, 103854. https://doi.org/10.1016/j.cose.2025.103854
41. Choi, C.; Shin, C.; Shin, S. Cyber Attack Group Classification Based on MITRE ATT&CK Model. J. Internet Comput. Serv. 2022, 23, 1–13. Available online: https://www.jics.or.kr/ (accessed on 9 October 2025).
42. Shin, C.; Choi, C. Cyberattack Goal Classification Based on MITRE ATT&CK: CIA Labeling. J. Internet Comput. Serv. 2022, 23, 15–26. Available online: https://www.jics.or.kr/ (accessed on 9 October 2025).
43. Husari, G.; Al-Shaer, E.; Ahmed, M.; Chu, B.; Niu, X. TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured CTI Text Sources. Proc. IEEE ICICS 2017, Beijing, China, 103–115. https://doi.org/10.1109/ICICS.2017.8268339
44. Wang, G.; Zhang, Q.; Liu, Y.; et al. KnowCTI: Knowledge-Based Cyber Threat Intelligence Entity and Relation Extraction. Comput. Secur. 2024, 140, 103824. https://doi.org/10.1016/j.cose.2024.103824