基於大型語言模型的資料擴增技術在網路威脅情資的攻擊手法分類研究

陳羿琪 陳怡安 李熹琳 黃意婷
電機工程系 電機工程系 電機工程系 電機工程系
國立臺灣科技大學 國立臺灣科技大學 國立臺灣科技大學 國立臺灣科技大學
M11107506@mail.ntust.edu.tw M11207513@mail.ntust.edu.tw lsl410778003@mail.ntust.edu.tw ythuang@mail.ntust.edu.tw

摘要

隨著資安事件頻傳，網路威脅情資報告提供最新的威脅與攻擊事件相關資訊，協助系統防禦新型網路攻擊的威脅。大多數情資報告都是基於自然語言編寫的非結構化文章，目前有許多相關研究探討如何自動對應文章內容於 MITRE ATT&CK 框架，以便於提供資訊安全分析人員閱讀與理解，進一步制定緩解與偵測策略。然而，目前大部分的方法，依賴監督式學習，當某些攻擊手法僅有少量資料被分析與發現，無法作爲有效分類；另外，在處理情資報告時，並非所有的報告內容都與攻擊手法相關。爲了解決這個問題，我們採用大型語言模型作爲資料擴增手法之一，並利用 Transformer Encoder 架構作爲分類器骨幹，以建立攻擊手法相關度二元分類器與攻擊手法分類器。實驗結果呈現，不論在我們收集 MITRE ATT&CK 知識庫中的流程範例與 MITRE TRAM 專案中的情資報告的實驗資料集，我們將所提出的方法於兩個資料集能夠更準確地分類攻擊手法 82.66% 和 76.86% 的 F1 分數，明顯高於其他方法。

關鍵字：網路威脅情資分析、攻擊手法分類、大型語言模型、MITRE ATT&CK

Abstract

As cyberattacks become more frequent and sophisticated, cyber threat intelligence (CTI) reports play a crucial role in providing up-to-date information on emerging threats and campaigns, enabling systems to defend against novel attacks. However, most open-source CTI reports are written in unstructured natural language, making automated analysis challenging. Recent research has shown that mapping CTI reports to the MITRE ATT&CK framework can assist security analysts in more effectively interpreting the reports and developing targeted mitigation and detection strategies. However, despite these advancements, existing methods frequently rely on supervised learning, which faces challenges when certain Techniques have limited data, making TTPs classification more difficult. Furthermore, not all content in CTI reports is relevant to specific Techniques. To address these challenges, we employ large language models for data augmentation and utilize a Transformer Encoder architecture as the backbone to develop two classifiers: a binary classifier for Technique relevance and another for Technique classification. We evaluate our method on two datasets from MITRE ATT&CK website and the MITRE TRAM project. Our Technique classifier achieved F1-scores of 82.66% and 76.86% on the respective test datasets, demonstrating a significant improvement over existing methods.

Keywords: Cyber Threat Intelligence, Technique Classification, Large Language Model, MITRE ATT&CK

1 前言

近年來，資安事件頻傳，從針對大型企業的勒索軟體，到竊取個人開發者電腦資訊，網路攻擊事件層出不窮。隨著攻擊變得更加普遍和複雜，即時檢測潛在的網路攻擊，才得以發展緩解策略。因此，具備即時性且高品質的網路威脅情資至關重要。網路威脅情資主要用於提供已知威脅與攻擊事件的資訊，其中包含入侵指標 (Indicators of Compromise，IoCs)，以及其攻擊相關描述等等資訊。一般來說，網路威脅情資可以分成四種類型：策略型、戰術型、營運型與技術型。技術類型的情資泛指有結構化的具體情資資訊，以資料交換爲目的，並交由機器設備來進行自動化讀取的有結構化情資，如威脅情資的結構化語言 (STIX) 與威脅情資的傳輸機制 (TAXII)。而其他類型的情資源自於非結構化的文字描述，如資安公司公布的報告、資安相關部落格等，也提供最新的威脅資訊，如惡意程式具體行爲、攻擊組織採用的手法等，並且呈現整個攻擊活動的流程，可以提供網路威脅的脈絡。

MITRE ATT&CK 框架 [1] 描述惡意程式或攻擊組織採用的攻擊的生命週期，被視爲開源的網路威脅情資資料來源之一，近年來廣受學業界採用。它是基於現實世界觀察的攻擊活動，發展攻擊者採用的攻擊策略 (Tactic)、攻擊手法 (Technique) 與其流程 (Procedure) 所建立的開源知識庫，又稱 TTPs。換句話說，攻擊策略指的是攻擊者執行步驟的意圖，攻擊手法爲實現攻擊的方法，流程用以描述攻擊者具體採取的惡意行爲。舉例來說，在 Windows 系統中的啓動資料夾安裝加密檔案，可以達到自動啓動該檔案的目的。這個流程範例可以對應到 MITRE ATT&CK 框架攻擊手法 —— 「T1547.001 自動啓動或自動登入：啓動資料夾」，以達到「持續潛伏」和「權限提升」的意圖。

MITRE ATT&CK 框架中所提及的攻擊手法可以從非結構化的網路威脅情資中獲得。這些網路威脅情資是由資安分析從業人員，透過觀察第一手的最新惡意程式，或是分析已知攻擊事件，所獲得的眞實案例流程。然而，必須熟

悉 MITRE ATT&CK 框架與網路攻擊事件的實際手法，才能準確辨識與流程相關的攻擊手法。目前已有相關研究 [2]–[7] 專注於從情資報告中，偵測對應 MITRE ATT&CK 框架的攻擊手法 (本文將攻擊手法與 TTP 作爲同義詞使用)。爲了從情資報告中，自動辨識文字描述所對應的攻擊手法，目前仍有兩個研究問題尚未被解決：(1) 目前研究使用的公開資料集，大多爲 MITRE ATT&CK 框架所提供的流程範例，然而大多數的攻擊手法的流程範例數量並不平均，其中只有 51% 的攻擊手法有超過 30 個以上的流程範例可供參考，不利於監督式學習相關研究發展。(2) 當給定一篇網路威脅情資報告，其中有部分文字描述是與攻擊手法無關，如 *Here's an overview of their research.*，不利於直接做攻擊手法分類。

本研究針對以上研究問題，提出 (1) 利用大型語言模型進行資料擴增，以針對每個與攻擊相關的文字描述，進行攻擊手法分類器的深度學習訓練；並 (2) 建立攻擊手法相關度二元分類器，以區分情資報告中，給定的文字描述是否與攻擊手法相關。因此，當給定網路威脅情資報告，先針對每段文字描述，判斷其是否與攻擊手法相關，再藉由大型語言模型的輔助所建立的模型，進行攻擊手法的分類。

2 文獻探討

目前多家防毒軟體供應商皆聘請專業資安分析師撰寫 CTI 報告，例如：Fortinet、Trend Micro、Kaspersky，這些報告是將主要的情資報告來源，以自然語言繕寫而成。然而，如何有效整合、分析從各種情資報告來源收集的大量文字資料，需要投入大量的時間與人力資源，爲了更有效率地從情資報告中獲取攸關駭客如何入侵系統的重要資訊，目前有多項研究致力於開發自動化從報告中找到關鍵語句的工具，將報告中含有駭客入侵手法語意的文字，對應到 MITRE ATT&CK 框架中的攻擊手法，以整合爲結構化的資訊。

TTPDrill [2] 先使用正則表達式將入侵指標替換成相應的字串後，找出可能具有攻擊手法相關句子的「主詞」、「動詞」及「受詞」，將這些由「主詞」、「動詞」及「受詞」組合作爲可能代表攻擊手法的句子，接著使用 BM25 TF-IDF，比較已知入侵手法敘述類型的相似度，以判斷這些句子是否描述了此類型的攻擊行爲。

rcATT [3] 利用 TF-IDF 計算每個文字在文章中出現的頻率作爲分類特徵，並刪除在 NLTK 中所有的停用詞後，使用多種機器學習模型對文章中可能有描述的所有攻擊策略及攻擊手法進行分類，其中線性支援向量機模型在實驗中取得最好的分類結果。

AttacKG [4] 從情資報告中自動提取網路攻擊相關的資訊，並將這些攻擊行爲整合成「攻擊行爲圖」，用於識別所採用的攻擊技術，最後與由攻擊手法中描述生成的攻擊行爲圖進行比較，以此辨識此情資報告中出現的攻擊手法。

MITRE 提供自動化標記攻擊手法的工具 TRAM [5]，使用了來自 150 篇情資報告、共 4,070 個經標記攻擊手法的句子做爲資料集，利用預訓練的 SciBERT [8] 模型將這些句子由自然語言轉換爲 Embedding 後，將含有攻擊手法意涵的句子，分類爲可能的攻擊手法。

LADDER [6] 利用預訓練的 RoBERTa [9] 等大型語言模型偵測句子中的惡意程式、工具等命名實體 (named entity) 及其關係，再找出描述攻擊行爲的句子，最後透過語意相似度與攻擊手法的描述進行比較，以此辨識含有攻擊行爲意義的句子。

而 aCTION [7] 使用大型語言模型對網路威脅情資報告的內容進行語意整合，首先因爲大型語言模型輸入的字數有限，使用提示詞 (Prompt) 對文章段落進行總結，再利用總結出來的內容，讓大型語言模型判斷其中的實體及關係，最後再利用大型語言模型總結出含有攻擊行爲的句子，最後將含有攻擊行爲的句子與 MITRE ATT&CK 所提供的攻擊手法句子進行比較，來分類句子屬於哪個攻擊手法。然而在這個研究中，因大型語言模型對輸入字數有限，在分類攻擊手法時，無法涵蓋 MITRE 所提供的所有攻擊手法類別。

3 方法

3.1 問題定義

當給定網路威脅情報 $d = \{s_1, s_2, ..., s_m\}$，其中 s 爲文章中 m 個文字描述，如句子。我們發展攻擊手法相關度二元分類器，判斷給定的句子是否有包含攻擊手法的描述 $s = \{s_1, ..., s_n\}$，接者我們發展攻擊手法分類器來判斷每個文字描述各自對應到 MITRE ATT&CK 框架中哪一個攻擊手法 MITRE Technique (TTP) $y = \{y_1, ..., y_n\}$，其中包含兩個元件：利用大型語言模型以提供少量資料做資料擴增，與發展以句子爲主的攻擊手法分類器。資料擴增是給定 s_i，產生與該段文字描述相關描述 $\{s_{i1}, ..., s_{ik}\}$ 以作爲 s_i 的 k 筆擴充訓練資料；在本研究當中，攻擊手法分類器爲多類別分類問題，意即每個與攻擊手法相關的文字描述 s_i，分類爲多個攻擊手法之一 y_i。

本研究的系統流程圖如1所示，給定一篇網路威脅情資報告，透過前處理將整篇報告切成獨立的句子作爲文字描述的基本單位。首先攻擊手法相關度二元分類器會判斷這些句子是否爲攻擊手法，並將預測爲攻擊手法的句子挑選出來，使用攻擊手法分類器預測其所屬的攻擊手法。建立攻擊手法分類器包含兩個階段：資料擴增與建立分類器。首先針對具有攻擊手法標籤的訓練資料集進行資料擴增，爲原本的訓練集加入更大量、更多元的資料，接著使用經過資料擴增的資料集訓練攻擊手法分類器，預測測試集中的句子所屬的攻擊手法。最後，系統自動化提取整篇報告所涵蓋的攻擊手法。

3.2 斷句

當給定純文字的網路威脅情資報告，我們使用 NLTK [10] 作爲斷句的工具。由於將情資報告轉換成文字檔的過程中，有可能會產生影響斷句的資料格式，例如：(1) 句子結尾處沒有「.」、(2) 含有全形文字、(3) 單字拼寫錯誤等問題、(4) 含有目錄及頁碼等與分類攻擊手法不相關的資訊，使得 NLTK 工具造成斷句的誤判。爲了避免這樣的情況，我們會在斷句之前，會對情資報告做 (1) 在常用的句子開頭 (例如：The、This、That、There 等等) 前加上換行符號、句號「.」及空格，來符合 NLTK 斷句規則；(2) 將所有全形文字統一爲半形文字；(3) 使用 pyspelling [11] 進行拼字檢查；(4) 手動移除目錄、頁碼等與分類攻擊手法不相關的資訊等處理。

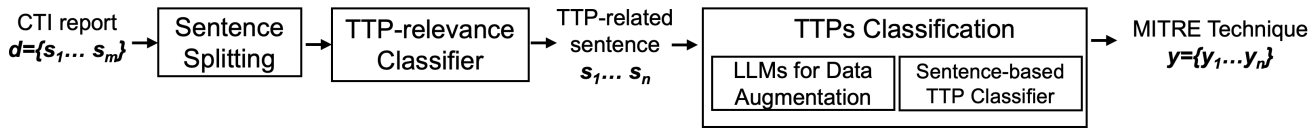

圖 1: 系統流程圖

3.3 攻擊手法相關度二元分類器

我們使用基於 Transformer Encoder 的模型 [12] 作爲本研究的分類器基礎骨幹，將文字輸入透過多頭注意力機制 (Multi-Head Attention Mechanism) 取得考慮上下文的動態嵌入向量，使每個文字嵌入向量都保有上下文涵意的資訊。將給定的句子 **s** 中的每個字 $\{w_1, w_2, ..., w_n\}$，利用 BERT [13] 預訓練模型取得動態嵌入向量，藉由 BERT 本身龐大的訓練資料和模型深度，可使文字得到更好的嵌入向量。接著利用 Transformer Encoder 學習 MITRE Technique 或 relevance 相關之知識，我們提取 [CLS] token 來當作分類特徵，用一層的線性層來作爲分類器，此訓練模型是屬於二元分類任務，用於判斷句子是否描述攻擊手法。

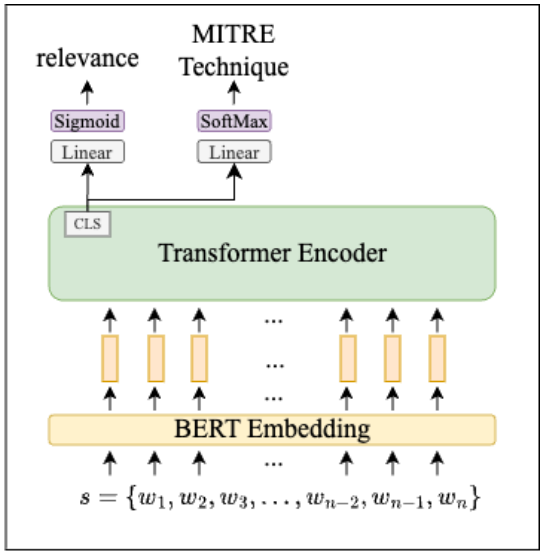

圖 2: 分類器骨幹模型架構圖，分別用於攻擊手法相關度二元分類器與攻擊手法分類器

3.4 攻擊手法分類器

1) 大型語言模型於資料擴增

在目前的攻擊手法分類任務中，部分攻擊手法類型存在其所包含的標籤句子資料量不足的問題。爲了克服這項挑戰，本章節探討如何使用大型語言模型 (Large language model，LLM)，針對用於訓練模型的資料集進行資料擴增。本章節設計適用於大型語言模型資料擴增的提示詞 (prompt)。第一部份要求針對輸入模型的句子進行重新撰寫，生成語義相似、文字內容不同的新標記資料，目的是爲了增加訓練資料集的多樣性，提升模型在攻擊手法分類任務上的表現，如圖 3所示；第二部分爲驗證語言模型所生成的資料是否符合其攻擊手法標籤意涵，如圖 4。

> You are an expert at MITRE ATT&CK framework.
> The following sentence is a procedure example of {label}.
> {sentence}
> Please generate {num_sentences} paraphrased sentences directly without any additional comments or confirmations.

圖 3: 用於生成擴增資料的提示詞

當給定某一個攻擊手法 {label} 與其相關描述 {sentence}，我們設計提示詞，要求大型語言模型換句話說，做爲資料擴增的數據使用。爲了使大型語言模型了解自己的任務，參考 [14]，一開始透過角色扮演使得爲大型語言模型設定專家角色，賦予其專業與背景知識 (You are an expert at MITRE ATT&CK framework)。接下來，提供任務介紹，包含 MITRE ATT&CK 框架中定義的攻擊手法的 ID 與其名稱 {label}，以及任務需求 {sentence} 所對應的攻擊手法。下一步爲指派任務，要求語言模型改寫 {num_sentences} 數量的文字。這段文字透過祈使句直接說明任務，要求大型語言模型針對方才提供的句子，直接進行特定次數的重寫，以生成特定次數的新標籤樣本。需要注意的是，在文字的最後加上不須額外生成其他資訊，是爲了避免大型語言模型在回應的階段生成擴增資料以外的回應用語。

> You are an expert at MITRE ATT&CK framework.
> Please determine if each of the following sentences matches the technique {label}.
> Respond 'yes' if it matches and 'no' if it does not.
> Here are the sentences: {sentences}

圖 4: 用於驗證擴增資料的提示詞

爲了確保生成的資料 {sentences} 是符合給定的攻擊手法描述 {label}，同樣地，我們利用大型語言模型驗證，以確保生成的資料符合擴增需求。在此提示詞當中，同樣透過角色扮演與任務指派的方式，要求大型語言模型回答這些句子是否屬於此攻擊手法，以此強調驗證方才所生成的標籤資料是否符合其攻擊手法的任務。

由於並不是每個文字描述都適合用於資料擴增，我們從訓練集中剔除以下內容：指令、API 函式、字數過短 (字數小於 3 個字)，或是不足以單獨表達任何攻擊手法資訊的句子，例如由一個字所構成的資料 *download*，接著將經過內容過濾的資料集整理爲一份獨立於訓練集外的新資料集，稱爲「擴增基礎」，接著使用大型語言模型對擴增基礎資料集中的每一筆資料進行重新撰寫以生成數筆資料，每生成一次，就使用驗證提示詞驗證一次方才所生成的擴增資料，確認內容是否符合對應的攻擊手法標籤。如

圖 5所示，大型語言模型依照生成提示詞，根據資料的攻擊手法標籤針對句子進行重寫，除了進行同義詞替換，例如將 *added junk bytes* 替換爲 *incorporated irrelevant bytes*，以及更改語句結構，例如將句尾的 *over HTTP* 改寫爲 *In its HTTP command and control activity* 並調換至句首，由於大型語言模型具有基礎電腦科學知識，因此能將 *C2* 改寫爲 *command and control*，也因爲大型語言模型能夠參考標籤內容，故能爲整個句子加上有關攻擊手法意涵的總結，如 *as a form of obfuscation*。

Procedure Example (T1001.001 Data Obfuscation: Junk Data)	
Original	SUNBURST added junk bytes to its C2 over HTTP.
Rephrased	In its HTTP command and control activity, SUNBURST incorporated irrelevant bytes as a form of obfuscation.
TRAM (T1041 Exfiltration Over C2 Channel)	
Original	which will then send the results to the C2 server
Rephrased	After executing this procedure, the outcomes will be sent to the command and control server.

圖 5: 擴增資料範例

總結來說，驗證提示詞用於指示大型語言模型檢查透過生成提示詞所生成的句子，是否符合該句子所屬的攻擊手法標籤。如果不是，則重新生成。最後，我們將擴增資料加入原本的訓練資料集中，訓練一個以句子爲基礎的攻擊手法分類器。

2) 句子爲基礎的攻擊手法分類

相同於攻擊手法相關度二元分類器使用 Transformer Encoder 架構，將每個句子都附有一個攻擊手法標籤的資料集作爲輸入，讓模型學習句子的特徵，進行攻擊手法的分類，這個分類任務的類型是屬於多分類、單標籤的任務，也就是我們所訓練的模型將可以用於將一個具有攻擊手法意涵的句子，分類至其所描述的攻擊手法。

4 實驗

在本章節，我們欲回答以下研究問題：

- RQ1: 本研究所提出的系統架構在攻擊手法分類上的準確性如何？
- RQ2: 當給定一篇網路威脅情資，本研究提出的系統如何幫助攻擊手法上的分類？

4.1 資料集

1) 攻擊手法實驗資料集

本實驗使用了兩種資料集驗證攻擊手法的分類結果，分別爲 MITRE 流程範例資料集、TRAM (single) 資料集。其中 MITRE 流程範例資料集的資料來源爲 MITRE ATT&CK 框架第 14 版的流程範例 [15]，並且限定爲 Windows 平台的資訊，此資料集採納 30 筆以上的標籤資料作爲研究材料，包含 11,054 個流程範例數量，涵蓋 83 種不同的攻擊手法標籤，並依照各佔 8：1：1 的比例，切分爲訓練集、驗證集及測試集，以此確保每個攻擊手法標籤下的資料在訓練集、驗證集及測試集中的比例都是相同的。TRAM 資料集源自於 MITRE Engenuity 所發表的 TRAM (Threat Report ATT&CK Mapper) 專案 [5]，其中的資料集 single_label.json 符合本研究任務定義，意即給定單一文字描述，分類某攻擊手法。此資料集共有 4,930 個文字描述，對應 50 個常見於報告的攻擊手法標籤。然而我們發現，在該資料集中 (1) 相同文字描述包含多種攻擊手法、(2) 源自於不同來源的相同文字描述，與 (3) 僅有大小寫不同的文字描述，我們將移除以上資料筆數，剩下 4,754 個文字描述於最後資料集中。最後進行五次的交叉驗證，依照 8:1:1 的方式切分訓練集、驗證集及測試集，實驗數據驗證於測試集，回報平均後的整體表現。

2) 攻擊手法相關度實驗資料集

攻擊手法相關度資料需要包含與攻擊手法相關和不相關的句子，在 MITRE TRAM (Threat Report ATT&CK Mapper) 專案中，提供 multi_label.json 資料集，其中包含與攻擊手法無關的句子。此資料集用於訓練攻擊手法相關度二元分類器，以 1/0 表示文字描述是否有對應的攻擊手法。該資料以 8:1:1 的比例切割成訓練集、驗證集與測試集。其中訓練集有 10,836 筆資料無對應攻擊手法，3,178 筆資料有對應攻擊手法；驗證集有 1355 筆資料無對應攻擊手法，397 筆資料有對應攻擊手法；測試集有 1,354 筆資料無對應攻擊手法，398 筆資料有對應攻擊手法；我們所建立的二元分類器，精確率及召回率分別達到 73.48% 和 79.98%，可用於偵測威脅情資報告中的句子，是否與攻擊手法有所關聯。

4.2 評估方法

我們使用精確率 (Precision)、召回率 (Recall) 和 F1 分數評估模型在每個攻擊手法分類任務中的表現。考慮資料筆數不均，採用 Macro-Averaged 使得每個類別具有相同的權重，再取算術平均數去計算每個類別的實驗結果。

4.3 比較對象

比較了本研究與現今四種最新研究在不同資料集上的攻擊手法分類表現，包含 TTPDrill [2]、AttacKG [4]、LADDER [6]、TRAM [5]；與兩種其他資料擴增的方法：WordNet [16] 與過取樣 (oversampling) [17]。其中 WordNet 是指透過將句子中特定數量的詞語更換同義詞進行資料擴增的方法，我們依照給定句子中每個文字的詞性來選擇 WordNet 中相同詞性且相同文字起始的同義詞作爲句子中可替換的字，例如：*use* 爲動詞，則選擇以 *use* 作爲開頭且詞性爲動詞的同義詞集形成可替換字典，最後按照建立的可替換字典對句子中的可替換文字進行隨機替換來生成新句子，並確保每個生成句子都與原句不同；過取樣是指在資料擴增上考量資料不平衡問題，針對部分類別下標記資料較少的生成較多的資料、按照各類別資料筆數決定生成數量的方法。

4.4 實作細節

所有實驗都運行在相同實驗設備上，並使用以下的設備、超參數或環境進行攻擊手法預測任務，以確保實驗環境的一致性與穩定性。實驗所使用的硬體設備包括 Intel i7-13700K 處理器，顯示卡爲搭載 24GB 記憶體的 NVIDIA Geforce RTX 4090。在超參數設定方面，每次訓練皆採用 10 個 Epoch，學習率設爲 1e-5，優化器選用 Adam，攻擊手法相關度二元分類器採用 Binary Cross-Entropy 損失函數，攻擊手法分類器則使用 Cross-Entropy 損失函數。在 Transformer Encoder 的參數設定中，隱藏層維度設爲 768，頭數爲 8，每個頭的維度爲 64，線性層維度爲 512，批

表 1: 模型在不同資料集上的表現

Dataset	MITRE 攻擊流程範例			TRAM		
Model	Precision	Recall	F1	Precision	Recall	F1
AttcaKG	2.03%	5.25%	2.13%	2.82%	6.48%	3.77%
TTPDrill	15.77%	34.57%	19.15%	12.95%	36.27%	18.79%
LADDER	59.64%	41.42%	43.71%	58.94%	39.24%	40.38%
TRAM(SciBERT)	83.98%	81.82%	81.47%	77.57%	**76.76%**	75.98%
Transformer	82.48%	78.34%	79.15%	78.32%	75.80%	75.65%
Transformer + Oversampling	82.11%	79.93%	79.81%	74.39%	76.21%	75.30%
Transformer + WordNet	84.25%	78.07%	79.17%	78.14%	76.33%	75.83%
Transformer + LLM (Our)	**85.07%**	**82.58%**	**82.66%**	**79.16%**	75.87%	**76.86%**

次大小爲 16。我們選取在驗證階段中損失率達到最低的 Epoch 對輸入進行預測。

進行資料擴增前，我們先將訓練資料集中不適合用於資料擴增的標籤資料過濾，整理爲「擴增基礎資料集」。而在資料擴增階段，我們使用的模型爲 gpt-4o-mini，其中生成提示詞的 temperature 設爲 0.8，驗證提示詞的 temperature 設爲 0.7。我們使用這兩種提示詞，針對擴增基礎資料集中每一筆句子生成 5 筆新資料，再對這 5 筆資料進行驗證，每個句子共經過兩輪的生成與驗證，因此一個句子共可用於生成 10 筆擴增資料。其中，MITRE 流程範例資料集共有 88,050 筆擴增資料，TRAM 資料集則有 31,926 筆擴增資料。在流程範例資料集當中，考量每筆資料的第一個字都是與駭客組織 (group) 和軟體 (software) 相關，因此我們針對 88,050 筆擴增資料，將其主詞取代爲 name。這是因爲資安威脅的環境變化快速，若只是針對駭客組織、特定的駭客活動名稱或是惡意程式進行攻擊手法分類，會使得模型過度依賴這些變化性高的資料進行訓練，透過主詞替換的方式就可以避免模型學習過多資安專有名詞的特徵。其他未具體列出的超參數及設定均採用預設值。在 MITRE 流程範例資料集上的 Transformer + LLM 的平均訓練時間爲 48 分鐘 32 秒，而在 TRAM 資料集上的 Transformer + LLM 的平均訓練時間爲 23 分鐘，在兩個資料集的平均測試時間爲 5 秒。

4.5 RQ1: 攻擊手法分類準確性

表 1顯示各研究在攻擊手法的分類結果。總體來說，本研究所提出的方法，不論在 MITRE 和 TRAM 資料集當中，在精確率、召回率和 F1 分數都獲得最佳的結果。相較於目前相關研究，以 TRAM 系統爲目前最頂尖的研究成果之一 (state-of-the-arts)，從 MITRE 資料集當中，當給定單一文字描述的情況下，透過資料擴增的方法，提供訓練模型更多不同變化的語句，可以加強分類結果；而在情資文本 TRAM 資料集，可看出擴增資料可以提高精確率，但卻略微損失召回率。部分原因有可能是大型語言模型提供的擴增資料的文字描述造成負面的影響，如幻覺 (Hallucination) 的發生 [18]，造成模型分類有誤。

在比較不同資料擴增技術在攻擊手法分類上的表現時，我們以使用大型語言模型進行資料擴增後進行訓練的 Transformer Encoder 爲基準，比較未使用擴增資料訓練的 Transformer Encoder 基礎模型，與其他擴增方法如 WordNet、過採樣的方式，以了解基於大型語言模型在資料擴增技術上的優勢，並將結果列於 1。

相對於基於 WordNet 的擴增方法以及過取樣方法，基於大型語言模型的資料擴增方法展現出更明顯的優勢。WordNet 擴增方法主要根據詞彙關聯性來生成新的資料樣本，但這種方法往往侷限於詞彙層面的變化，生成的樣本可能缺乏足夠的上下文多樣性和語義豐富性，而我們使用大型語言模型實作的資料擴增方法，能夠根據上下文生成更具多樣性和語義豐富的樣本，因此在複雜的攻擊手法分類任務上表現更好。過取樣方法主要集中於擴增數量，將所有資料筆數擴增至最大類別的筆數，相較之下，我們採用的資料擴增方法則是藉由使用大型語言模型，對每筆資料生成至多 10 個不同的樣本，避免使用過取樣方法時，需要將所有資料擴增至最大類別筆數，不可避免地對少數資料進行多次重複生成的問題，故我們的方法不僅能爲僅有少數標籤資料的類別生成多元化的資料，也能避免資料重複的問題，以確保每種類別的擴增資料的多樣性。

4.6 RQ2: 案例分析

本章節將介紹在現實生活中，實際的情資報告將如何運用我們的模型找到文章中對應的攻擊手法。我們選擇了一段情資報告 [19] 作爲範例，以此來詳細介紹模型的用法及表現。

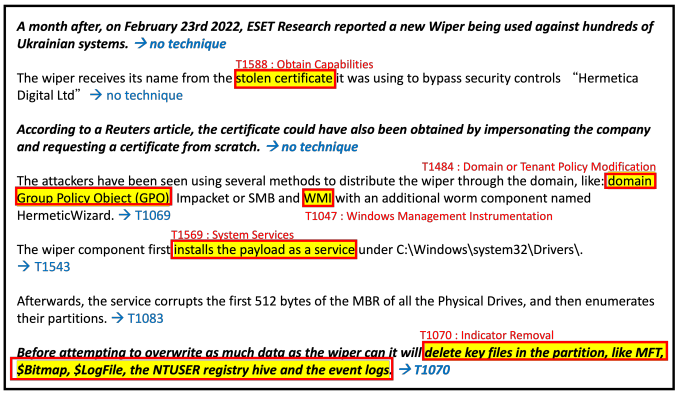

圖 6: 實際情資報告結果。藍色字代表預測結果，紅色字代表正確結果。

在報告中擷取一段有寫出攻擊手法 ID 的段落，對其進行資料清洗。爲了驗證模型的準確性，將文章中寫出的攻擊手法 ID 作爲該句的正確答案，並將攻擊手法 ID 從文字中剔除。接著，將經過處理的段落斷句後，放入攻擊手法分類器中幫助句子找到其對應的攻擊手法。此處選擇使用實驗結果 F1 分數表現最佳的「使用 MITRE 流程範例資料集所訓練的 Transformer + LLM」作爲我們的攻擊手法分類器。這段情資報告包含七個句子，其中三句是沒有攻擊手法的句子、三句是與 1 個攻擊手法相關、一句與 2 個攻擊手法相關。圖6爲此段落的預測攻擊手法與實際攻擊手法

的比較，紅色框起段落爲提及攻擊手法的部分，紅字爲其對應的攻擊手法，藍字爲預測出的攻擊手法。其中粗斜體爲預測成功的句子。

參考圖6，第一句、第三句及第七句有預測正確，我們的模型有能力分辨出無相關攻擊手法的句子，且在句子與1 個攻擊手法相關時能夠正確預測。

但依然會有預測錯誤的狀況發生。第二句相關的攻擊手法爲 T1588 Obtain Capabilities，這個攻擊手法沒有被預測出來的原因是訓練的資料集並未涵蓋這個攻擊手法，導致模型無法預測出該攻擊手法。第四句與兩個攻擊手法有關，分別爲 T1484 Domain or Tenant Policy Modification 及 T1047 Windows Management Instrumentation，預測出 T1069 Permission Groups Discovery 的原因可能爲句中提到的分發方法表明，攻擊者很可能利用收集到的群組和用户信息來設計具體的攻擊，因此誤判 T1069。第五句模型預測的攻擊手法是 T1543 Create or Modify System Process，而實際的攻擊手法則是 T1569 System Services，由於這兩個攻擊手法的常用字非常相似，因此導致了模型誤判。

此外，第六句原本並無標記攻擊手法，但其實有提到 *enumerate files and directories*，與 T1083 File and Directory Discovery 的手法相似，造成誤判。

5 結論

本研究專注於分析網路威脅情資報告，基於 MITRE ATT&CK 框架，提供文章內容惡意活動對應到 ATT&CK 框架中記錄的攻擊手法。藉此資安團隊不僅能夠更準確地識別攻擊者的策略，還能根據這些知識來擬訂防禦策略，使得網路安全不僅限於被動防禦，更能預測攻擊者行爲，從而更有效地防範未來的威脅。

本研究應用大型語言模型以進行資料擴增，然而大型語言模型可能會產生所謂的「幻覺」，即生成與眞實情況不符或錯誤的資料，可能導致生成的資料在某些情境下偏離實際情況，影響模型的預測準確性。因此即使我們採用驗證策略，檢查其生成資料，仍有生成不正確的標籤資料的風險，影響模型正確性。未來可以嘗試使用 RAG (Retrieval Augmented Generation，擷取增強生成) [20] 擷取網路威脅情資報告中的相關資訊來優化大型語言模型，讓大型語言模型生成更加準確或完整的描述。

本研究假設一個文字描述僅對應一個攻擊手法，爲單一標籤分類任務。然而，攻擊手法分類實際上可能涉及多標籤和多分類問題。未來的研究將會探索如何擴展我們的方法以處理這些更具挑戰性的多標籤和多分類問題，例如藉由改進模型架構或採用新的演算法，來提升分類的準確性和實用性。此外，由於大型語言模型在自然語言處理中的應用廣泛，我們可以透過設計特定的提示詞，運用其語意理解能力與基礎的特定領域知識，進行資料擴增，亦可以將其應用於情資報告中的命名實體識別 (Named Entity Recognition，NER) [21]，藉由自動化提取情資報告中的命名實體與實體間的關係，建立知識圖，從而預測關鍵的威脅情資，協助進一步的攻擊手法分類。

誌謝

本研究感謝邵冠崴與郭英仁的研究上的討論與技術支援，並感謝國科會 113-2634-F-001 -002 -MBK、112-2222-E-011 -011 -MY2 部分經費支持。

參考文獻

[1] MITRE Corporation, “Mitre att&ck™: Adversarial tactics, techniques, and common knowledge,” https://attack.mitre.org/, 2024.

[2] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu, “Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources,” in *Proceedings of the 33rd annual computer security applications conference*, 2017, pp. 103–115.

[3] V. Legoy, M. Caselli, C. Seifert, and A. Peter, “Automated retrieval of att&ck tactics and techniques for cyber threat reports,” *arXiv preprint arXiv:2004.14322*, 2020.

[4] Z. Li, J. Zeng, Y. Chen, and Z. Liang, “Attackg: Constructing technique knowledge graph from cyber threat intelligence reports,” in *European Symposium on Research in Computer Security*. Springer, 2022, pp. 589–609.

[5] MITRE Engenuity, “Our tram large language model automates ttp identification in cti reports,” https://medium.com/mitre-engenuity/our-tram-large-language-model-automates-ttp-identification-in-cti-reports-5bc0a30d4567, 2023.

[6] M. T. Alam, D. Bhusal, Y. Park, and N. Rastogi, “Looking beyond iocs: Automatically extracting attack patterns from external cti,” in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 92–108.

[7] G. Siracusano, D. Sanvito, R. Gonzalez, M. Srinivasan, S. Kamatchi, W. Takahashi, M. Kawakita, T. Kakumaru, and R. Bifulco, “Time for action: Automated analysis of cyber threat intelligence in the wild,” *arXiv preprint arXiv:2307.10214*, 2023.

[8] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.

[9] Y. Liu, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[10] E. L. Bird, Steven and E. Klein, “Natural language processing with python. o’reilly media inc.” https://github.com/nltk/nltk.

[11] facelessuser, “pyspelling,” https://github.com/facelessuser/pyspelling.

[12] A. Vaswani, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

[13] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong, “Better zero-shot reasoning with role-play prompting,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 4099–4113.

[15] MITRE, “Mitre att&ck v14,” https://attack.mitre.org/versions/v14/, accessed: 2023-10-31.

[16] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to wordnet: An on-line lexical database,” *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[17] R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine learning with oversampling and undersampling techniques: overview study and experimental results,” in *2020 11th international conference on information and communication systems (ICICS)*. IEEE, 2020, pp. 243–248.

[18] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *arXiv preprint arXiv:2311.05232*, 2023.

[19] Fernando Martinez, “Analysis on recent wiper attacks: examples and how wiper malware works,” https://cybersecurity.att.com/blogs/labs-research/analysis-on-recent-wiper-attacks-examples-and-how-they-wiper-malware-works, 2022.

[20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[21] Y. Hu, F. Zou, J. Han, X. Sun, and Y. Wang, “Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model,” *Computers & Security*, vol. 145, p. 103999, 2024.