PEGASUS와 RoBERTa 모델을 활용한 사이버 공격 캠페인 분석
Cyber threat campaign analyze based on PEGASUS and RoBERTa model

최창희 · 이인섭 · 신찬호 · 이성호
Changhee Choi · Insup Lee · Chanho Shin · Sungho Lee

국방과학연구소
(changhee84@add.re.kr)

ABSTRACT

The success and popularity of ChatGPT has spawned a number of derivatives. In the cybersecurity domain, BERT-based models have also started to emerge, which are based on ChatGPT. In this paper, we introduce how these models can be used to analyze cyber attack campaign. In particular, we analyze attacks by APT29, Lazarus and other nation-state cyber attack campaigns using PEGASUS and BERT-based models. This study shows the potential of using models such as PEGASUS and RoBERTa model to analyze cyber attack campaigns in the future.

Key Words : RoBERTa, BERT, PEGASUS, ChatGPT, cybersecurity, cyber attack campaign

1. 서론

사이버 보안 문제는 계속해서 증가하고 있으며, 특히 중요한 국가시설이나, 무기체계를 노리는 사이버 공격의 심각성이 커지고 있다. 과거 행해진 고도의 사이버 공격에 대해서 사이버 보안 업체나 국가 기관이 침해 분석 보고서를 작성하는데, 이를 읽고 분석하여 자료화하는 것은 많은 시간이 소요된다.

본 논문에서는 사이버 위협 분석 보고서를 자동으로 분석하여 MITRE ATT&CK[1]의 공격 기술로 요약하는 기술을 제안한다. 과거에는 TF-IDF와 같이 오래된 기법을 적용하였다[2]. 본 논문에서는 PEGASUS[3]기반의 요약 모델을 사용하여 사이버 보안 문장 및 ATT&CK의 설명문을 요약하였다. RoBERTa[4] 증류 버전 모델을 사용하여 요약된 문서를 각각 벡터로 변환하였으며, 코사인 유사도를 사용하여 최종적으로 공격 캠페인에 해당하는 공격 기술을 식별하였다. 실험 결과 전문가가 수동으로 분석한 공격 기술의 30%의 공격 기술을 자동으로 식별하는 것을 알 수 있었다.

2. 배경 지식

2.1 MITRE ATT&CK® 공격 기술

미국의 비영리 연구단체인 MITRE에서는 사이버 위협을 공격자 관점에서 분석한 ATT&CK 모델을 연구해 왔다. 2023년 5월에 버전 13을 공개하였는데, 14개의 전술, 196개의 공격 기술을 모델링하였다. 이는 사이버 보안 분야에서 광범위하게 활용되고 있다. 본 논문에서는 보고서에 공격 기술이 라벨링된 시기를 고려하여 버전 10을 사용하였다.

2.2 사이버 보안 문서 요약 모델

Zhang 연구진은 2019년 입력 문서로부터 중요 단어를 제거하거나 마스킹하는 방식으로 요약을 수행하는 PEGASUS 기법을 제안하였다. 이들은 마스킹된 문장들이 중요 내용일수록 출력이 요약에 가까워지도록 설계하여 학습을 수행하였다. 본 논문에서는 사이버 보안 분야의 데이터를 학습한 파생 모델[5]을 이용하여 사이버 위협 분석 보고서를 요약하였다.

2.3 문서 임베딩 모델

Liu 연구진이 2019년에 제안한 RoBERTa 방법은 기존 BERT에 비해 많은 데이터, 더 긴 시퀀스, 더 큰 배치사이즈로 학습을 수행하고 동적마스킹 방법을 이용하여 성능을 향상시켰다. 본 논문에서는 RoBERTa의 증류(Distillation) 버전을 사용하여 요약된 문장을 768 차원의 벡터로 변환시켰다[6].

3. 제안하는 방법

그림 1은 사이버 문서를 문장으로 분해하고, 요약한 후 벡터로 변환하여 유사도를 측정하는 과정을 나타낸 것이다.

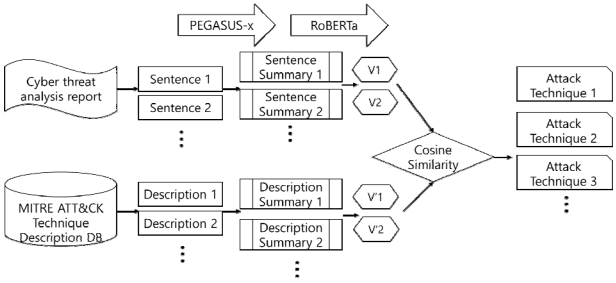

그림 1. 제안하는 방법의 알고리즘 절차

3.1 사이버 보안 문서 추출

제안하는 방법에 대한 검증을 수행하기 위해 본 논문에서는 Cyber Criminal Campaign Collections[7] 보고서 데이터셋을 사용하였다. 총 1,211건의 보고서를 텍스트로 변환하였고, python의 nltk 모듈을 사용하여 문장으로 나누어 주었다. 길이가 짧은 경우에는 문장의 형태를 갖추지 못한 경우가 많아서, 20글자보다 적은 문장은 제외하였다. 나누어진 문장은 2.2절에서 소개한 사이버 보안 문서 요약 모델을 이용하여 짧게 요약을 수행하였다.

MITRE ATT&CK 버전 10에는 188개의 기술이 존재하는데, 기술에 대한 상세한 설명이 기재되어 있다. 위의 보고서 문장과 마찬가지의 방법으로 요약을 수행하였다. 표 1은 문장 요약을 수행한 예제이다. 기술 이름의 경우, 구체적인 공격 행위에 관해서 서술되어 있지 않고, 원문의 경우 서술이 896자로 길어서 전부 읽고 파악하기에는 어려움이 따른다. 요약문은 기술 이름에 간단한 목적까지 작성되어 있어 사용자의 이해를 도와준다.

Table 1. 문장 요약 예제

기술 이름	T1059.003 : Windows Command Shell
원문	Adversaries may abuse the Windows command shell for execution. The Windows command shell (cmd) is the primary command prompt on Windows systems. The windows ...(896자)
요약문	Use the Windows command prompt to execute commands.

3.2 문장 유사도 계산 및 공격 기술 식별

문장의 유사도를 계산하기 위해서는 우선 문장을 벡터 형태로 변환해야 한다. 본 논문에서는 2.3절에서 소개한 RoBERTa 방법의 증류 버전을 이용하여 문장을 벡터로 변환하였다. 그 후 식 1과 같이 두 벡터 A와 B의 코사인 유사도를 이용하여 문장 유사도를 산출하였다. 가장 유사도가 MITRE ATT&CK 공격 기술 설명 요약을 채택하였으며, 이때 0.5 이하의 값을 가지는 것은 유사하지 않다고 판단하여 제외하였다.

$$Similarity = \frac{A \cdot B}{\|A\| \, \|B\|} \qquad \text{Eq. 1}$$

4. 실험 결과

전처리가 완료된 723개의 사이버 위협 분석 보고서에 대해 시험이 수행되었고, 231개의 공격 그룹 혹은 캠페인에 대해서 분석을 실시하였다. 각 보고서는 보안 전문가의 수동 분석을 통해 정답 공격 기술이 라벨링 되었다. 제안하는 방법으로 보고서에 대해 공격 기술을 식별한 결과 평균 30%의 정확도로 자동 식별하는 것을 알 수 있었다. 그림 3는 주요 공격 그룹의 정확도를 나타낸 것이다. 최대 67%의 확률로 공격 식별을 하는 것을 알 수 있었다.

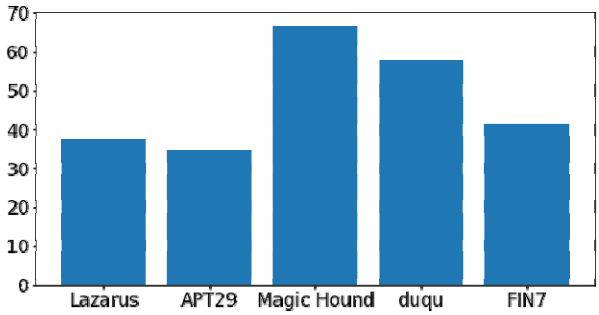

Fig. 2. 주요 그룹의 공격 기술 식별 정확도

5. 결론

본 논문에서는 최신 언어 모델인 PEGASUS와 RoBERTa를 이용하여 사이버 위협 분석 보고서를 MITRE ATT&CK의 공격 기술로 요약하는 방법에 제시하였다. 실험 결과 평균적으로 약 30%의 확률로 자동으로 공격 기술을 요약하는 것을 알 수 있었다. 향후 정확도를 높이기 위하여 사이버 공격 관련 문서를 모델에 전이 학습하는 방법을 연구할 계획이다.

References.

[1] MITRE ATT&CK, https://attack.mitre.org

[2] 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).

[3] J. Zhang, Y Zhao, M. Saleh and P Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization”, International Conference on Machine Learning, pp. 11328-11339, 2020.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, and D. Chen, “Roberta: A robustly optimized bert pretraining appoach,” arXiv preprint arXiv:1907.11692(2019).

[5] starcatmeow, “https://huggingface.co/starcatmeow/autotrain-cybersecurity-summarization-pegasus-x-book-43369110299”, accessed at May 8, 2023

[6] huggingface, “https://huggingface.co/sentence-transformers/all-distilroberta-v1”, accessed at May 8, 2023

[7] APT & CyberCriminal Campaign Collections, https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections