Translating Cybersecurity Descriptions into Interpretable MITRE Tactics using Transfer Learning

Reza Fayyazi, Yiyan Steve Wufeng, Pradumna Gautam, Shanchieh Jay Yang

Introduction

- Intrusion logs and threat intelligence reports have been developed to assist security analysts
- Description in these logs and reports, however, can be cryptic and not easy to interpret. Thus:
  - We ask: Given a description of cyberattack techniques, how to interpret the intended effects (MITRE Tactics [1]?)
  - E.g.1: Initialization scripts can be used to perform administrative functions, which may often execute other programs or send information to an internal logging server.
  - E.g.2. Custom Outlook forms can be created that will execute code when a specifically crafted email is sent.

Privilege Escalation? Persistence? Both?

Related Works

PATRL (Pseudo-Active Transfer Learning) [2]
- A semi-supervised process leveraging ULMFiT [3] to determine the attack stage of IDS alert signatures

BERT [4]
- A Transfer Learning technique to uncover the semantic information conveyed in a sentence

ExBERT [5]
- A framework that applies Transfer Learning to BERT to predict exploitability
- Word embedding for the pre-trained and fine-tuned BERT with cybersecurity words

SecBERT [6]
- A BERT model trained on cybersecurity texts

SecureBERT [7]
- A language model based on RoBERTa [8] that is trained on cybersecurity texts

Methodology

- Multi-Label Classification for the total of 14 MITRE Tactics
- Total of 4500+ Descriptions with their corresponding tactic(s)

Results

Table 1. Results for running the three BERT models with 30 epochs using 5-fold cross-validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>BERT</th>
<th>SecBERT</th>
<th>SecureBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Training Loss</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Avg. Test Loss</td>
<td>0.15</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>Abs. Accuracy</td>
<td>0.63</td>
<td>0.59</td>
<td>0.65</td>
</tr>
<tr>
<td>Micro Avg. F1 Score</td>
<td>0.75</td>
<td>0.72</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 2. Results for per-tactic F1 score for the three models to measure the differences in values for single-label and multi-label descriptions.

Observations & Future Works

Based on the Results:
- The 0.76 Micro F1 score in SecureBERT is promising in capturing semantic features of cybersecurity descriptions and dealing with multi-label data.
- The models could reasonably capture overlapping MITRE tactic descriptions

Future Works:
- How to 1) better reflect the model’s performance, 2) treat limited labeled data, 3) leverage label semantics, and 4) use a novel NSP-tuning approach to predict the intended consequences.

References