ABSTRACT
Due to the ever-increasing threat of cyber-attacks to critical cyber infrastructure, organizations are focusing on building their cybersecurity knowledge base. A salient list of cybersecurity knowledge is the Common Vulnerabilities and Exposures (CVE) list, which details vulnerabilities found in a wide range of software and hardware. However, these vulnerabilities often do not have a mitigation strategy to prevent an attacker from exploiting them. A well-known cybersecurity risk management framework, MITRE ATT&CK, offers mitigation techniques for many malicious tactics. Despite the tremendous benefits that both CVEs and the ATT&CK framework can provide for key cybersecurity stakeholders (e.g., analysts, educators, and managers), the two entities are currently separate. We propose a model, named the CVE Transformer (CVET), to label CVEs with one of ten MITRE ATT&CK tactics. The CVET model contains a fine-tuning and self-knowledge distillation design applied to the state-of-the-art pre-trained language model RoBERTa. Empirical results on a gold-standard dataset suggest that our proposed novelties can increase model performance in F1-score. The results of this research can allow cybersecurity stakeholders to add preliminary MITRE ATT&CK information to their collected CVEs.

CCS CONCEPTS
- Security and privacy → Software and application security
- Information systems → Information systems applications
- Data mining

KEYWORDS
CVE, MITRE ATT&CK, Cybersecurity, Transformer models, Pre-trained language models, Self-Knowledge Distillation

1 INTRODUCTION
Harmful cyber-attacks on critical cyber-infrastructure (e.g., large servers hosting confidential data) have cost on average $7.91 million per breach, leading to over 446,000,000 exposed records containing sensitive information in 2019 [22]. Thus, it is imperative to build our cybersecurity knowledge base to combat new and evolving cyber-threats. An essential component of the cybersecurity knowledge base is the Common Vulnerability and Exposures (CVE) list, overseen by the MITRE Corporation. Cybersecurity professionals commonly use CVEs to coordinate efforts to address the vulnerability. A new CVE is created whenever a security flaw is discovered in software and hardware and reported to MITRE. An example of a recent CVE is shown in Figure 1.
mitigation strategies could provide tremendous value to cybersecurity analysts and researchers.

In 2018, MITRE created the ATT&CK Matrix for Enterprise CRMF that models the tactics, techniques, and procedures (TTP) that an attacker would take when attempting to breach cyberinfrastructure [21]. The MITRE ATT&CK framework can help reliably predict and mitigate the tactics, techniques, and procedures (TTP) chain that an attacker follows after an initial breach [1]. Example tactics include “initial access,” “defense evasion,” and “exfiltration.” Each tactic in MITRE ATT&CK comes with a mitigation strategy (e.g., user training, account management, password policies, etc.) to assist cybersecurity analysts in protecting critical cyberinfrastructure, making it an excellent CRMF for our CVE labeling task. Labeling CVEs with ATT&CK tactics requires a computational model that can analyze the textual metadata available in CVE descriptions.

Despite the tremendous benefits that both CVEs and the ATT&CK framework can provide for key cybersecurity stakeholders (e.g., analysts, researchers, educators, and managers), the two entities are currently separate. With over 158,000 CVEs existing as of the beginning of 2021, it would be a non-trivial task to manually link each one to the ATT&CK framework to gather mitigation strategies for every existing CVE.

In this study, we developed a novel framework that leverages the textual features in CVEs currently linked to an ATT&CK tactic by prior research [11] to create an ATT&CK tactic label for CVEs outside our gold-standard CVE dataset. To achieve this goal, we drew upon state-of-the-art methodologies in deep learning-based text classification literature to guide the development of a novel cybersecurity model, the CVE Transformer (CVET) model. To ensure the value of our proposed approach, we rigorously evaluated CVET against benchmark models found in related CVE data mining and cybersecurity analytics literature.

The rest of this paper is as follows. First, we review related literature to CVE data machine learning, transformers for text classification, and self-knowledge distillation. Second, we identify gaps from our literature review and pose our research question for study. Third, we detail our proposed method for labeling CVEs with MITRE ATT&CK tactics. Fourth, we summarize the empirical results of our work and discussed their implications. Finally, we conclude the work with important takeaways from the paper.

2 LITERATURE REVIEW

2.1 CVE Data Machine Learning

Large undertakings have been taken recently to use CVEs to improve other cybersecurity information systems through the use of classical machine and deep learning architectures. Authors have used the convolutional neural network (CNN) to predict CVE vulnerability severity [10] and build knowledge graphs with the CVE list, the Common Weakness Enumeration (CWE) list, and the Common Attack Pattern Enumeration and Classification (CAPEC) list [29]. However, CNNs often struggle to capture long-term dependencies in textual passages [27]. Extant literature then applied the more powerful bidirectional long short-term memory (BiLSTM) model with a self-attention mechanism to predict vulnerability severity [6] and vulnerability type (e.g., boundary condition error) [7] with higher accuracy than the CNN. More recently, researchers have leveraged the pre-trained transformer model known as BERT [23] to extract information from the prominent vulnerability database ExploitDB to enhance the textual descriptions for new CVEs. Building a model that can effectively label CVEs with ATT&CK tactics using textual descriptions requires an algorithm that can effectively represent the long text sequences found in CVE descriptions. The transformer model (and its extensions) is currently the state-of-art within text classification literature and has proven to be robust against adversarial attacks [14]. We review the transformer model in-depth to gain a deeper understanding of how it can assist in our target task.

2.2 Transformers for Text Classification

Introduced in 2017, the transformer model replaces the recurrent cells found in prominent text classification deep learning models (e.g., BiLSTM, LSTM) with multi-head attention mechanisms [26]. While the original design incorporates an encoder-decoder structure (for machine translation tasks), multi-class text classification only requires the encoder stack. The encoder transformer model creates an embedding from the input, passes the embedding into the transformer block, and outputs a softmax probability score of outputs. The embedding layer is a one-hot encoding with positional encodings. The transformer block consists of a multi-head attention mechanism and feed-forward layers, which has shown to greatly improve accuracy, precision, recall, and F1-score over recurrent models in benchmark tasks [8]. Recently, transformers have been used for massive pre-training language models (PTLMs) that produce state-of-the-art results in various text classification tasks [20]. These models (e.g., BERT, GPT-2) are often trained on millions of data points and contain hundreds of millions of trainable parameters. While most researchers do not own the hardware or data to create their PTLM, these models can be carefully fine-tuned [4] and distilled [12] for improved performance in specialized tasks. Knowledge distillation is an emerging paradigm that can extract key information from the parameters of a PTLM to supplement the training of a targeted model [12].

2.3 Self-Knowledge Distillation

Knowledge distillation (KD) combines relational knowledge from a large, pre-trained model (teacher) and a prior untrained model (student) [30]. As a result, the trained student model is often more generalizable to unseen data than a model without knowledge distillation. This design allows researchers that do not have access to the computing power required to make a massive PTLM to create highly targeted state-of-the-art
One increasingly popular form of knowledge distillation is self-knowledge distillation (self-KD). Self-KD uses the same architecture for both the student and teacher, where knowledge transfer occurs within the same model [24]. This form of distillation creates a new model that often outperforms the original teacher model without requiring new data [32], due to distilling latent features from deeper to shallower sections of the network [33], improved feature importance weighting [5], or modified regularization [17]. Generally, self-KD for a natural language processing task with available target labels (e.g., a text classification task) uses the weighted sum of cross-entropy (CE) loss with the correct labels and CE loss with the soft target [9].

3 RESEARCH GAPS AND QUESTIONS

From the extant literature, we identify several research gaps: First, many tasks have been undertaken to link CVEs to vulnerabilities, CWEs, and CAPEC, but not directly to the MITRE ATT&CK framework. The closest work to attempt this task is the BRON model [11], which does not link new CVEs, but uses existing databases to create a more holistic knowledge graph. Second, the deep learning models implemented in recent literature (e.g., CNN, BiLSTM) struggle to capture long-term dependencies in text, such as the lengthy descriptions within CVE listings. These two gaps motivate our research question:

- How can we create a novel and accurate link between CVEs and ATT&CK tactics through their textual descriptions and long-term dependencies?

4 PROPOSED METHOD

Our proposed methodology is comprised of three major components: (1) Data Collection and Pre-Processing, (2) the CVE Transformer (CVET) Architecture, and (3) Benchmark Experiments. Each component is further detailed in the subsequent sections.

4.1 Data Collection and Pre-Processing

<table>
<thead>
<tr>
<th>ATT&amp;CK Tactic</th>
<th>Count of CVEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defense Evasion</td>
<td>8,482</td>
</tr>
<tr>
<td>Discovery</td>
<td>6,647</td>
</tr>
<tr>
<td>Privilege Escalation</td>
<td>5,779</td>
</tr>
<tr>
<td>Collection</td>
<td>1,748</td>
</tr>
<tr>
<td>Lateral Movement</td>
<td>715</td>
</tr>
<tr>
<td>Impact</td>
<td>594</td>
</tr>
<tr>
<td>Credential Access</td>
<td>427</td>
</tr>
<tr>
<td>Initial Access</td>
<td>309</td>
</tr>
<tr>
<td>Exfiltration</td>
<td>137</td>
</tr>
<tr>
<td>Execution</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24,863</strong></td>
</tr>
</tbody>
</table>

Table 1: Gold-Standard Dataset CVE Distribution

For our research, we use the dataset provided by the BRON knowledge graph [11]. As discussed earlier, there are currently more than 158,000 CVEs, and our gold-standard dataset only captures a fraction of them, making this linking task critical. The dataset leverages existing knowledge bases to link 24,863 CVEs into 10 of the 14 ATT&CK tactics. Many ATT&CK tactics do not require a vulnerability (e.g., “Resource Development” and “Command and Control”). Thus, we cannot link CVEs to them. Table 1 provides a distribution of how many CVEs are in each ATT&CK tactic category. About 91% of the data is within four tactic categories: defense evasion (8,452), discovery (6,647), privilege escalation (5,779), and collection (1,748).

To pre-process the CVE textual description, stop words were removed, non-alphanumeric characters were stripped. The remaining text was lower-cased, lemmatized, and padded to ensure proper lengths for all inputs. This sequence of pre-processing steps is common in deep learning-based text classification literature [15]. We used the pre-made RoBERTa tokenizer [16] to encode the data as input for our CVET model. Other metadata available in CVEs is not used, as it did not provide a benefit to model performance in preliminary testing.

4.2 CVET Architecture

4.2.1 Model Selection. We adapt a PTLM called RoBERTa [16] (a BERT-based model trained on longer sequences) for CVET due to the generalizability it has shown in text classification tasks [2]. While there are dozens of PTLMs to choose from for our target task, we chose RoBERTa due to the high performance it achieves while also allowing custom fine-tuning and self-KD designs.

4.2.2 Self-Knowledge Distillation. During fine-tuning of the CVET model, we use CVET as both a teacher and student. The student model (denoted as CVETs) is CVET at fine-tuning time step t and the teacher model (denoted as CVETT) is CVET at
All classical machine learning models were implemented using Huggingface Transformers library. The CVET model used RoBERTa-large from the Huggingface library [28], and our self-distillation and fine-tuning designs were implemented in PyTorch 1.4 [19].

All models are run with 10-fold cross-validation [3] benchmark models are evaluated with the accuracy, precision, recall, and F1-score metrics, which is the standard for multi-class text classification tasks [25]. Paired t-tests are used to identify if statistically significant differences exist between CVET and each benchmark method. Due to our unbalanced dataset, the discussion focuses on the F1-score (which is more resilient against skewed distributions than the other metrics) [13].

5.2 Results and Discussion

From Table 2, we make four key observations about the results of our experiments. First, the four classical machine learning models had the lowest F1-scores, ranging from 36.70% for Random Forest to 48.66% for SVM. Second, all deep learning models reached a higher F1-score than the classical machine learning models. The transformer model had the highest F1-score (73.61%) among the deep learning models. The transformer does not have an internal recurrent mechanism like the other models, which suggests its multi-head attention architecture helped improve the text classification performance. Third, all of the PTLMs marginally improved upon the transformer model in F1-score. Baseline RoBERTa improved over the transformer by 0.96% (from 73.61% to 74.57%). These results suggest that pre-training the transformer can improve the performances of specialized text classification tasks. Finally, our proposed CVET model outperformed all other PTLMs, deep learning models, and classical machine learning models in F1-score (76.18%). The differences were significant at p<0.05 or less in all models. Our CVET model also achieved the best accuracy (79.93%) versus all other models. These results suggest that our fine-tuning and self-KD design assisted in performance improvement in the target task.

6 CONCLUSION

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical Machine Learning</td>
<td>Random Forest</td>
<td>63.70%**</td>
<td>35.83%***</td>
<td>37.67%***</td>
<td>36.70%***</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>65.70%**</td>
<td>51.23%***</td>
<td>46.34%***</td>
<td>48.66%***</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>67.30%**</td>
<td>44.22%***</td>
<td>33.92%***</td>
<td>38.16%***</td>
</tr>
<tr>
<td></td>
<td>Logistic Regression</td>
<td>67.10%**</td>
<td>41.65%***</td>
<td>34.12%***</td>
<td>37.38%***</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>RNN</td>
<td>68.45%***</td>
<td>69.66%***</td>
<td>67.30%***</td>
<td>68.46%***</td>
</tr>
<tr>
<td></td>
<td>GRU</td>
<td>70.90%**</td>
<td>72.55%***</td>
<td>69.19%***</td>
<td>70.83%***</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>72.75%***</td>
<td>74.14%***</td>
<td>71.89%***</td>
<td>72.00%***</td>
</tr>
<tr>
<td></td>
<td>BiLSTM</td>
<td>72.55%***</td>
<td>73.71%***</td>
<td>71.71%***</td>
<td>72.70%***</td>
</tr>
<tr>
<td></td>
<td>BiLSTM with Attention</td>
<td>71.41%**</td>
<td>72.52%***</td>
<td>70.32%*</td>
<td>71.40%***</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>72.45%***</td>
<td>74.49%***</td>
<td>70.82%*</td>
<td>73.61%***</td>
</tr>
<tr>
<td>Pre-Trained Language Models</td>
<td>GPT-2</td>
<td>70.21%***</td>
<td>77.12%***</td>
<td>64.56%***</td>
<td>70.27%***</td>
</tr>
<tr>
<td></td>
<td>XLNet</td>
<td>74.12%***</td>
<td>80.12%*</td>
<td>68.56%***</td>
<td>73.88%***</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td>73.93%**</td>
<td>79.86%*</td>
<td>69.41%*</td>
<td>74.26%*</td>
</tr>
<tr>
<td></td>
<td>RoBERTa</td>
<td>74.42%*</td>
<td>81.88%</td>
<td>68.49%**</td>
<td>74.57%*</td>
</tr>
<tr>
<td>Self-Distillation</td>
<td>CVET</td>
<td>76.93%</td>
<td>81.38%</td>
<td>71.49%</td>
<td>76.18%</td>
</tr>
</tbody>
</table>

Table 2: Comparing CVET Against Benchmark Classical Machine Learning, Deep Learning, and Pre-Trained Language Models (*: p<0.05, **: p<0.01, ***: p<0.001)
In this study, we developed a novel self-distillation approach to automatically label CVEs with their associated ATT&CK tactic. The CVET model fine-tuning process included an Adam loss function with added bias correction and a self-KD design. Our model was evaluated with a series of experiments against state-of-the-art models in classical machine learning, deep learning, and pre-trained language models. Results indicated that the CVET model offers a significant benefit to labeling CVEs with MITRE ATT&CK tactics over baseline non-distillation techniques. Our architecture can greatly assist the cybersecurity community by creating an immediate link between a heavily utilized cybersecurity risk management framework and critical vulnerabilities. This can be implemented into key cybersecurity stakeholders’ workflow to associate vulnerabilities found in their scanners to the MITRE ATT&CK framework for additional information on how to combat the vulnerability.

We identify two potential future directions for related work in this domain. First, we would like to expand the connection of CVEs to other prominent CRMFs. Potential options include the CAPEC list and the National Institute of Standards and Technology (NIST) framework. Such connections can broaden the mitigation strategies provided by this work when a CVE is discovered. Second, we plan to look into other more refined textual data representation techniques (e.g., novel word embedding strategies, synonym/homonym generation, POS tagging) to improve the features of our textual inputs.

REFERENCES


