Computers & Security 134 (2023) 103430

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence

Markus Bayer *, Tobias Frey, Christian Reuter

PEASEC - Science and Technology for Peace and Security, Technical University of Darmstadt, Pankratiusstraße 2, 64289 Darmstadt, Germany

ARTICLE INFO

Keywords:
Cyber threat intelligence
Few-shot learning
Transfer learning
Data augmentation
Information overload

ABSTRACT

Gathering cyber threat intelligence from open sources is becoming increasingly important for maintaining and achieving a high level of security as systems become larger and more complex. However, these open sources are often subject to information overload. It is therefore useful to apply machine learning models that condense the amount of information to what is necessary. Yet, previous studies and applications have shown that existing classifiers are not able to process information about emerging cybersecurity events, such as new malware names or novel attack contexts, due to their low generalisation capability. Therefore, we propose a system to overcome this problem by training a new classifier for each new incident. Since this requires a lot of labelled data using standard training methods, we combine three different low-data regime techniques – transfer learning, data augmentation, and few-shot learning – to train a high-quality classifier from very few labelled instances. We evaluated our approach using a novel dataset derived from the Microsoft Exchange Server data breach of 2021 which was labelled by three experts. Our findings reveal an increase in F1 score of more than 21 points compared to standard training methods and more than 18 points compared to a state-of-the-art method in few-shot learning. Furthermore, the classifier trained with this method and 32 instances is only less than 5 F1 score points worse than a classifier trained with 1800 instances.

1. Introduction

Social media are where cutting-edge and critical cyber threat information is disseminated, which is highly relevant to researchers, security providers, security operation centres, urban infrastructures, and cyber emergency response teams (CERTs), among others (Mittal et al., 2016; Rodriguez and Okamura, 2019). While there have been several research works on general cyber threat event detection (Dionísio et al., 2020; Fang et al., 2020), the aim of this work is to enable fine-grained and potentially individualized collection of cybersecurity information in open data sources such as Twitter.

A major challenge in gathering cybersecurity-related information, also called Cyber Threat Intelligence (CTI) (McMillan, 2013), which needs to be specialized, i.e. customizable, is that information in this area is very dynamic and varies greatly from past events (in terms of specific names, different attack vectors, specific attack paths, affected functions, etc.) (Chatterjee and Thekdi, 2020). As a result, supervised machine learning yields poor results because these dynamics cannot be captured in the learning process. Alternatively, new classifiers could be trained for each cyber threat event so that the new features are taken into account. However, since machine learning usually requires a large amount of data for normal training, this would result in having to label a dataset for each cyber threat event, which is unrealistic considering the effort involved and the need for fast and up-to-date information. Against this background, the concept of active learning systems takes a first step towards label reduction for supervised machine learning for cyber threat events (Riebe et al., 2021b). Active learning supports the labelling process, so that only the instances with the highest learning value need to be labelled for machine learning. However, despite this method, too much data is still needed to train a useful classifier. The endeavour sought in this work takes an even stronger stance on labelling reduction by proposing a system consisting of few-shot learning, transfer learning, and data augmentation, which are all techniques to reduce the amount of manual labelling required for a high-quality classifier. With few-shot learning, it is sufficient if the model is already trained with very few instances, as opposed to hundreds or thousands in the case of active or normal learning (Brown et al., 2020). This includes special learning techniques as well as transfer learning, where knowledge

* Corresponding author.
E-mail addresses: bayer@peasec.tu-darmstadt.de (M. Bayer), tobiasjonathan.frey@stud.tu-darmstadt.de (T. Frey), reuter@peasec.tu-darmstadt.de (C. Reuter).

https://doi.org/10.1016/j.cose.2023.103430
Received 22 July 2022; Received in revised form 2 June 2023; Accepted 11 August 2023
Available online 19 August 2023

from a previous task is transferred to the new one. Data augmentation is used to create artificial instances from the training data using label-preserving transformations (Bayer et al., 2022).

The concept of few-shot learning is extended in this work through the use of multi-level transfer learning. The different levels start with a model that has been trained on a large general dataset and thus has a basic prior knowledge. During the next steps, this model is approximated more and more to the actual task domain. In this way, it can be ensured that the model is given a basic cybersecurity reference in order to be able to counter the dynamics in the task, in addition to being familiar with the task. This is particularly relevant for urban infrastructures, which require high resilience against cyberattacks, as well as for CERTs, as they need to collect and communicate information in the most reliable and targeted way possible (Riebe et al., 2021a). The data augmentation strategy is inspired by the work of Bayer et al. (2021) and follows the example of Yoo et al. (2021) by utilizing the large generation model GPT-3 to generate new instances based on the few existing labelled ones.

Our paper includes several contributions relevant for the cybersecurity and machine learning community:

- A novel pipeline combining transfer learning, data augmentation, and few-shot learning for rapid development of effective specialized CTI classifier.
- Novel techniques of data augmentation and few-shot learning to deal with a small number of training instances.
- A new specialized CTI dataset annotated by three experts and based on the 2021 Microsoft Exchange Server data breach.

The code and dataset of this study are freely available.[1]

The remainder of the paper is structured as follows: After introducing related work on transfer learning, data augmentation, few-shot learning, and cyber threat event detection and intelligence (Section 2), we explain the concept of our method (Section 3). It is subdivided in three components which are described in detail. In Section 4 the evaluation is presented and findings are given in detail. The last section (Section 5) contains a discussion of the implications, limitations, and potentials for future research.

2. Related work

2.1. Cyber threat event detection and intelligence

Cyber threat event detection can be defined as the process of automatic scraping of the webspace and Open Source Intelligence (OSINT) to detect possible cybersecurity events (Sabottke et al., 2015; Riebe et al., 2021b; Le Sceller et al., 2017). Social media platforms like Twitter are part of OSINT and provide a great space to share and discuss cybersecurity vulnerabilities, for example. While vulnerability databases such as the National Vulnerability Database (NVD) are often of high quality and much higher credibility of vulnerability information, Twitter posts can be more up-to-date and rich (Alves et al., 2020). There are some automated systems and research that already scrape Twitter and other OSINT sources to detect cyber events. Some examples are the *CySecAlert* system from Riebe et al. (2021b) or *SONAR* from Le Sceller et al. (2017), which collect cybersecurity relevant tweets from Twitter, filter them, and present them in a manageable dashboard.

CTI on the other hand describes the process of collecting additional information after the first detection of a cyber threat event. The process helps deliver the context of the vulnerabilities found to assist CERTs and cybersecurity organizations make sound decisions and find quick solutions (Abu et al., 2018; Tounsi and Rais, 2018; Wagner et al., 2019). CTI is currently mostly accomplished by manually collecting information on different platforms (Abu et al., 2018). It relies heavily on manual tasks and is therefore labour intensive (Wagner et al., 2019). However, there are already some threat intelligence platforms, such as Facebook ThreatExchange or CrowdStrike, that are able to automatically detect, monitor, and analyze cyber threat occurrences (Tounsi and Rais, 2018). A manageable dashboard is also provided by the Cyber Threat Observatory (Kaufhold et al., 2022), which aggregates cybersecurity information from various sources, including social media, security advisories, indicators of compromise (IoCs) and CVEs. For an overview of different CTI platforms and tools, see the work by Kuehn et al. (2020).

However, these systems need too much time to adapt to a newly discovered threat that is, for example, propagated on Twitter, and cannot be extensively customised. It is possible that this has not yet been addressed because current machine learning systems are generally too rigid and cannot be easily generalized to new situations. Our work aims to solve this problem by providing a novel pipeline that allows the rapid training of new specialized CTI classifiers through significantly reduced labelling requirements. This is achieved with novel techniques in the field of transfer learning, data augmentation and few-shot learning.

Once the CTI information has been collected, there are several methods and research approaches that can be used to analyze this information and provide useful insights. TTPDrill by Husari et al. (2017), for example, extracts and constructs attack patterns from threat and blog articles. IoCMiner by Niakanlahiji et al. (2019) is able to extract IoCs from Twitter data. Similarly, GoodFATR by Caballero et al. (2023) collects IoCs from Twitter and five other sources (including Telegram and blogs). They also give a good overview of different IoC extraction works. Many of these works use or can be complemented by machine learning. While our approach aims to provide a specialised CTI collection in new cyber threat events, it could also be used to enhance the quality of these works.

2.2. Transfer learning

Transfer learning describes the process of transferring knowledge gained from training a neural network from one task to another related task (Torrey and Shavlik, 2010; Pan, 2020). This technique is now one of the standard learning methods for machine learning, especially in the field of natural language processing (NLP). It is particularly powerful for tasks where there is not enough training data or it is difficult to manually adjust the data for training. In these cases, it is possible to use a pre-trained neural network that was trained to solve a related task or with more easily accessible data. Afterwards the neural network is fine-tuned with the task-specific data to fit the wanted task. One of the most frequently used pre-trained models is BERT by Devlin et al. (2018). BERT (short for Bidirectional Encoder Representations from Transformers) is a pre-trained deep bidirectional transformer for language understanding. In essence, it is trained by predicting words in a sentence given the other words, also called masked language modelling. It has a lot of widely used descendants trained for many different tasks, such as BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), and CamemBERT (Martin et al., 2020). While BERT is already a considerably large model, nowadays far larger models, like GPT-3 from Brown et al. (2020), are trained. Compared to BERT's base model with 110 million parameters, GPT-3 has 175 billion parameters, however, GPT-3 is not publicly available and cannot be easily fine-tuned due to its size.

Transfer learning can be an important step to overcome the high labelling requirements through knowledge transfer. Unlike other work in this field, we do not just train a pre-trained model for the actual task, but propose to train a model further and further towards the actual task through several fine-tuning steps. Other work may not have addressed this because it requires multiple datasets that are more and less specific to the task at hand. Moreover, this technique can only be used to handle a small number of tasks. In our case, however, this is exactly what we want, as we need a basic model that can be easily adapted to the

[1] Code: https://github.com/PEASEC/Specialized-Cyber-Threat-Intelligence. Dataset: https://github.com/PEASEC/msexchange-server-cti-dataset/.

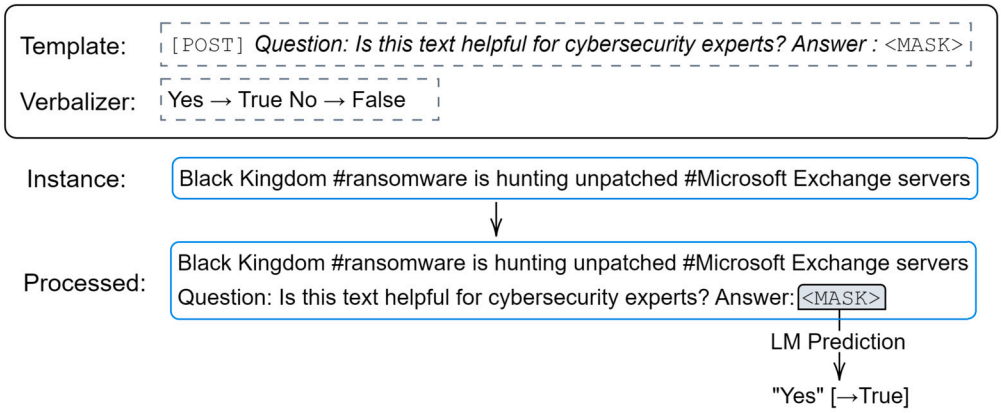

Fig. 1. Example of a template and a verbalizer and how they are applied on an instance.

different cybersecurity events and thus only very few labelled instances are needed.

2.3. Data augmentation

Data augmentation is the concept for artificially enlarging the training datasets for machine learning by transforming the existing ones. Originated and heavily used in computer vision, it is now also increasingly being explored on textual data (Bayer et al., 2022). NLP data augmentation techniques can be applied to the raw text or also on the numerical representations. Ranging from small transformations, i.e. flipping characters (Belinkov and Bisk, 2018) or inducing adversarial noise (Jiang et al., 2020), to interpolated (Sun et al., 2020) or even newly created instances (Anaby-Tavor et al., 2020), data augmentation can have great effects. Nevertheless, as Longpre et al. (2020) point out, the success of data augmentation in NLP is often not perceivable when fine-tuning large pre-trained models. A data augmentation technique needs to incorporate new linguistic patterns as otherwise the changes are too small and already captured by the pre-training phase of the model. For example, simple synonym replacement methods have not been shown to be beneficial with pre-trained models, as these synonyms are already mapped to nearly the same vector for their numerical representation (Mosolova et al., 2018). On the other hand, there are generation models that can integrate new linguistic patterns, for example, through their own training data during pre-training, as for example shown by Yoo et al. (2021) with the GPT-3 model. The challenge with using these models is to make the generations truly label preserving. This is, for example, done by Anaby-Tavor et al. (2020), Queiroz Abonizio and Barbon Junior (2020) and Bayer et al. (2021). The models are conditioned by fine-tuning on the label-induced training data (or just the class data) and are then tasked to complete a text given the label conditioned beginning (prompt). As this is oftentimes not sufficient to achieve a high label preservation, a filter mechanism is used that removes artificial instances that are unlikely to fit the class. For example, Anaby-Tavor et al. (2020) use a classifier trained on the data to predict whether the new instance can be assigned to the expected label.

In this work, we take advantage of recent research directions by combining the strategy of using GPT-3 to generate training data and then filtering out the instances that are not close enough to the respective class. In this way, we can create instances with a high degree of novelty, i.e. instances with linguistic patterns that were not previously included in the training data, but also preserve the label. If we use this for our goal of reducing the data needed for specialized CTI, we try to generate instances that have these novel linguistic patterns at best in the cybersecurity domain, but are still very close to the original data, especially since the data must be very specific.

For an overview of the data augmentation methods that could also be used in this study, we advise the reader to have a look at the survey from Bayer et al. (2022).

2.4. Few-shot learning

Few-shot learning describes the training of effective classifiers on the basis of a small number of examples. While there are several strands of research on few-shot learning (Bragg et al., 2021), in this study we focus on the use of pre-trained language models. At the latest, the large language model GPT-3 by Brown et al. (2020) paved the way for using these kinds of models, as it reaches astounding performance even without task-specific training data. However, as GPT-3 is too large for most companies and research institutes, the research field adapted smaller language models to reach similar or even better few-shot performances (Tam et al., 2021).

Pre-trained language models can be especially beneficial for few-shot settings when the instances are reformulated in a cloze-style way. Cloze tests (Taylor, 1953) are tests where some words in the text are missing and have to be completed. For few-shot learning, instances are rephrased, often into questions, so that the text contains the label (or a word that can be mapped to the label), generally within the answer to the question. The label, known (training) or not known (testing and inference), is masked out, so that the language model can fill it with the right word and a label can be inferred. Using the language model directly is more effective for few-shot learning than the classical way of training a classifier head on top of it, as there are no more randomly initialized parameters that have to be learned (Gao et al., 2021).

A pattern describes the transformation of the input instance to the cloze-like text. The verbalizer maps the predicted words for the mask to the label. An example for a pattern and a verbalizer can be seen in Fig. 1.

Gao et al. (2021) show that the choice of template and verbalizer has a major impact on the resulting performance. Since domain knowledge is often necessary for these, the authors propose a method to automatically find meaningful templates and verbalizer. For this purpose, they use a language model and the existing training instances to predict the words for the verbalizer and template. Zhang et al. (2022) take a different perspective on automatic template generation with the DART method by making the template differentiable. They use special tokens in the template that are mapped into trainable parameters. These template parameters are then optimized together with the target label. PERFECT by Mahabadi et al. (2022) leverages task-specific adapters to replace template tokens. Adapters make it possible to train only the newly added parameters, which are able to transform the hidden states, while freezing all other parameters.

Schick and Schütze (2021) propose a semi-supervised few-shot learning technique, called PET. They take several manually designed

templates and use the training data to train on each one a pre-trained language model. They take these models to generate pseudo-labels for unlabelled data. A classifier is then trained on the resulting dataset. Tam et al. (2021) adapt the PET method to not be dependent on additional training data and can even improve the performances. Contrary to the preceding PET technique, the word probabilities are computed not only for the verbalizer words, like "yes" and "no", but also for all other words. In the training, incorrect class tokens are explicitly penalized and correct tokens are encouraged. Furthermore, ADAPET (Tam et al., 2021) introduces a label conditioning step in which the model is tasked to predict other tokens in the sentence given the label.

In our pipeline, we incorporate the ADAPET technique into the proposed multi-stage fine-tuning technique and adopt it for the task of specialized CTI. Together with the novel technique of data augmentation, we can create a system that enables specialized CTI by reducing the amount of data required for high-quality classifier.

2.5. Research gap

Our study addresses several research gaps which are highly relevant for researchers as well as practitioners. Most importantly, our research paves the way for fine-grained and specialized CTI. Current research addresses CTI from a very coarse perspective, by building classifiers, like Riebe et al. (2021b), that are able to find general cybersecurity information. As a result, only a small amount of data reduction can be achieved in these information-overloaded situations. On the other hand, specialized classifiers are not designed to generalize well to new situations. Our work fills this gap by introducing a pipeline for specialized CTI, where new cyber threat events are encountered with the very fast creation of new classifiers. By addressing this fine-grained information gathering challenge, we create a novel dataset combined with a sophisticated labelling guideline for CTI. Furthermore, with our pipeline we address research gaps of machine learning low-data regimes. Our data augmentation strategy is the first to explore the generation capabilities of large language models with constraining them through filtering mechanisms. We combine the works of Yoo et al. (2021) and Bayer et al. (2021) by using GPT-3 with a human-in-the-loop filtering mechanism. We extend the few-shot learning research by proposing a multi-level fine-tuning approach. In the process, the model learns a very broad knowledge in the first levels, which in the later stages becomes more and more directed to the specific CTI task.

3. Concept

3.1. Dataset creation

The goal of dataset creation is to extract specific CTI information during a significant cyber threat event. In this work, we focus on Twitter as a data source because it provides a wide range of vulnerability information, unlike the NVD, which only provides brief information that is not as up-to-date, does not provide direct mitigation advice, and does not include exploit information (Alves et al., 2020). However, we are also aware of the disadvantages of Twitter, as the information may not be as credible, for example (more on this in the limitations in Section 5.2).

This dataset is subsequently binary-labelled according to the relevance of the information for CTI and for cybersecurity experts. We focused on the Microsoft Exchange Server data breach of 2021, where four zero-day exploits were discovered. While the first report of a vulnerability was already made in January of that year, in March various attackers were found to be exploiting the vulnerabilities and a proof of concept was released.

We used the Twitter APIv2 to gather 50,000 tweets in March that fulfil the query *"Microsoft Exchange" OR "MS Exchange" OR "CVE-2021-26855" OR "CVE-2021-26857" OR "CVE-2021-26858" OR "CVE-2021-27065"*. From this, we filtered out the tweets that were not in English,

> *New episode alert! In this Juniper Threat Labs podcast I interview @MounirHahad for his take on the MS Exchange ProxyLogon vulnerability CVE-2021-26855. https://bit.ly/3lU1347 (https://t.co/6pniZgqTju)*

> *#Microsoft Releases Patches for Older Versions of #Exchange Server https://www.zdnet.com/article/microsoft-exchange-attacks-now-microsoft-rushes-out-a-patch-for-these-unsupported-exchange-servers-too/ (https://t.co/U2jwnLMA9n)*

Fig. 2. Two examples of the labelling procedure, the upper one not relevant and the lower one relevant.

resulting in 39474 tweets. The used query is intended to replicate the process of filtering Tweets by security experts in the event of the incident. Examination of the resulting posts shows that only a subset of them are really relevant to an expert. While they contain a lot of relevant and up-to-date information such as references to patches and remedies, proof of concepts, code, IoCs and attacker names, they also contain a lot of irrelevant information aimed at the general public or, for example, only spam, podcasts and news aggregations. For these tweets, we drew a random sample of 3001 posts for labelling and resolved the links shortened by Twitter, as the full URLs could be an important indicator in the context of CTI.

For the labelling process, we have created a codebook that contains guidelines describing what content is relevant and what is not, closely following the staff of CERTs who work with this type of data. Our goal was to collect precise information that would yield maximum benefit for them. This does not include, for example, information that is primarily intended for a wider audience or information that is not current. For instance, valuable insights can be gained through details about exploits, IoCs, and the vulnerabilities themselves (including remediation, impact, solutions, etc.). Fig. 2 presents one case that is relevant and another that is not, both from the process of labelling. The relevant example contains information about patching the Microsoft software, while the example labelled as not relevant does not provide any in-depth information or technical insights that would be immediately useful to a cybersecurity professional.

The labelling of the data was performed by three cybersecurity experts guided by the codebook. The guidelines, which gave annotators clear guidance on when to mark a post as relevant or irrelevant, were iteratively updated by the annotation leader. A first draft of this was developed with the help of the CTI concept (McMillan, 2013):

> "Threat intelligence is referred to as the task of gathering evidence-based knowledge, including context, mechanisms, indicators, implications, and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard."

After an initial sifting of the tweets and again after the first labelling of 750 tweets, the process was refined by the annotation leader. The full guidelines can be found in the Appendix A.

The first round of annotation of 750 tweets was conducted by the annotation leader, who updated the guidelines after gathering several insights. He and the other two cybersecurity experts then annotated the 750 tweets again. After this round, all three experts discussed the cases they were not sure about and corrected them if necessary. Regarding the intercoder reliability the Kappa Scores were calculated (see Table 1). Subsequently, each annotator tagged 750 different examples, resulting in a total of 3001 commented Twitter posts for the complete dataset (the labels of the 750 instances of the first round were determined by majority vote).

Table 1
Intercoder reliability calculated with the Kappa Score.

Coder	Score
C1 and C2	0.8763
C2 and C3	0.7446
C1 and C3	0.8709

Table 2
Split of the dataset with count of relevant and not relevant labels in the datasets.

Split	Count	Relevant	Not relevant
Train (full)	1800	949	851
Train	32	16	16
Dev (full)	600	273	327
Dev	8	4	4
Test	601	304	297
Total	3001	1526	1475

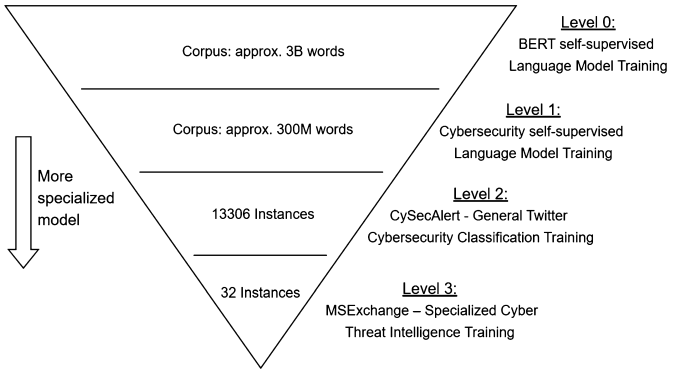

Fig. 3. Multi-level fine-tuning process that shows the model becoming more specialized as it is guided to the actual task with less data.

The dataset was then split into a full and few-shot training set and development set. The splits (train, dev) consist of 1800 and 600 instances for the full set and 32 and 32 instances for the few-shot set, respectively. The test set is the same in both cases and consists of 601 instances. For a complete overview of the dataset splits and class distribution, see Table 2.

3.2. Approach

Our system for dynamic, specialized cyber threat event detection consists of three components, all of which help to boost performance with little data. We explain the three components in detail in the following:

Multi-level fine-tuning: In light of the success of large pre-trained models such as BERT, we propose to further tune such models on several levels of domain-dependent data (see Fig. 3). The levels begin from a broader view and are narrowed down to the actual task. In our case, we first take a pre-trained BERT model (which can be seen as the 0th level of fine-tuning), train it with masked language modelling on cybersecurity data. We then tune the resulting model for classification on the CySecAlert dataset (Riebe et al., 2021b) in which Twitter posts are generally assigned to the cybersecurity domain. Finally, we train it on the few training examples of the specialized cyber threat dataset. The rationale behind this is that the model gains more and more knowledge as it is tuned to more and more fitting tasks. The 0th level is about gaining general knowledge of text. In the first level, the dataset consists of papers, blogs, web pages, and also Twitter data, from which the model gains knowledge about cybersecurity language and also how Twitter data is written in this domain. In the second level, the model should gain a general understanding of the relevance of cybersecurity information. Finally, in the third level, the model is tuned to the actual task data to which it can transfer the knowledge of the previous levels.

GPT-3 data augmentation: With data augmentation we can create new instances from existing ones, which can be particularly advantageous when the amount of data is small. We propose a data augmentation strategy based on text generation with GPT-3 (Brown et al., 2020), which is inspired by the method from Yoo et al. (2021) and Bayer et al. (2021). GPT-3 can be tasked to complete a given text, also called a prompt. We utilize this mechanism so that the generation model creates new instances based on the training data of one class. Specifically, this means that we are concatenating all instances of one class with a class specific priming token. For the class of cyber threat information we prepend every positive instance with "cybersecurity `->`". For the irrelevant class we chose "other `->`" as priming token. In both cases the priming token is also appended at the end so that the model generates the instance(s) after it. Dependent on how many remaining generation tokens the model has after the prompt, it may generate more than one instance by picking up the priming token. After the creation of the instances we perform the human-in-the-loop filtering step proposed by Bayer et al. (2021). The training examples and generated instances are mapped into an embedding space. There, the generated instances that deviate the most from the training data are discarded. The distance from which this happens is determined by an expert.

Few-shot learning: We make use of the existing ADAPET (Tam et al., 2021) few-shot learning technique and adapt it to our case. With ADAPET, in contrast to normal use, no classification head is trained on the language models. The instances are transformed to cloze-style phrases and then the language model itself is used to predict the blank word in the phrases. The predicted word is subsequently transformed with a verbalizer to one of the labels. The cloze-style phrases are automatically formed with templates. For our task we use the following template:

> "`[POST]` Question: Is this text helpful for cybersecurity experts? Answer: `<MASK>`. `[SEP]`"

The verbalizer maps the two possible words "yes" and "no" to the labels representing relevant and not relevant. As explained in Section 2.4 there also exist methods for automatically determining the pattern and verbalizer. We believe that these techniques are not necessary in our case, as we can integrate the expert knowledge regarding the task, which facilitates the learning process.

4. Evaluation

4.1. Dataset, models and evaluation settings

Following the research goal of specialized CTI for security professionals, we constructed a setting, consisting of models and datasets, representing the real conditions. For the dataset, we labelled data from the 2021 Microsoft Exchange Server data breach. The specifics of the dataset can be found in Section 3.1. The labelled dataset, including few-shot and normal-shot splits, is freely available.

In our main evaluation we have different settings regarding the dataset and models. The *baseline* and initial model of our evaluation is the bert-base-uncased model by Devlin et al. (2018). For the baseline, this model is fine-tuned on the few-shot dataset representing the standard training strategy without any few-shot or data augmentation methods. For the *best case*, on the other hand, we train the bert-base-uncased model with the full dataset of 1800 instances. This is called the best case because we consider this amount of data to be the best case in the event of a new cybersecurity attack. In addition, we also train a

model with ADAPET, as we consider this to be the current state of the art in few-shot research. In preliminary tests, we found that ADAPET performed best on the few-shot split with ALBERT (Lan et al., 2020) compared to DART and a PERFECT variant. To be consistent with our evaluation settings as opposed to the evaluation settings of ADAPET, we use the bert-base-uncased model, instead of the albert-xxlarge-v2 model by Tam et al. (2021). The evaluation settings of our procedure are divided into the three components mentioned. For the data augmentation technique we use GPT-3 (DaVinci) as text generation model, which is prompted with the specifics explained in Section 3.2. The multi-stage fine-tuning process starts with the bert-base-uncased model, which is further pre-trained on a cybersecurity dataset, which is then fine-tuned with the ADAPET few-shot method on the CySecAlert dataset. This resulting model is finally trained on the few-shot split and evaluated on the test set of the Microsoft Exchange dataset. Furthermore, in addition to the CySecAlert fine-tuning process, we also use the ADAPET few-shot method for the fine-tuning of the Microsoft Exchange Server dataset. The mentioned components are also inspected within an ablation study, showing their individual contribution to the overall pipeline.

The evaluation performance is measured in accuracy and with the F1-score. For every evaluation setting, we perform five runs to rule out random factors. The results are given with the minimum, maximum, mean, and standard deviation.

4.2. Hyperparameters

As already mentioned, we are using bert-base-uncased as base model for our experiments. The evaluations are performed on a NVIDIA A100 with 40 GB GPU memory. The training runs on the CySecAlert and Microsoft Exchange dataset are performed with 5 epochs each. Furthermore, we used a batch size of 48, 100 warmup steps with a warmup ratio of 0.06, a learning rate of 0.00001, and weight decay of 0.001. As optimization algorithm, we used the Adam algorithm. For the data augmentation technique we used the GPT-3 text-davinci-002, which has 175 billion parameters. The filtering was performed with SBERT with the all-mpnet-base-v2 model.

4.3. Evaluation

The first section of our evaluation is about the data augmentation process, as we manually inspected the instances generated by GPT-3 and compare our method to two other data augmentation techniques.

After this, the main evaluation follows where we compare our methods to a baseline, state-of-the-art and best case experiment. Finally, we inspect our method by doing ablation studies, testing how each component evaluates.

4.3.1. Data augmentation

Due to our human-in-the-loop approach, we already saw that the generated instances are of very high quality. An excerpt of the generated data is given in Table 5. For research purposes, we were also interested in the most likely original instances that the model used for generating specific instances. This is why we tried to find the training instance with the closest resemblance to the generated one. We measured the resemblance by generating sentence embeddings with SBERT (Reimers and Gurevych, 2019) and comparing them with the cosine distance. These counterparts are also given in Table 5. These examples show that the data augmentation method is capable of many different transformations. The first example demonstrates that the model sometimes replaces one or few words with synonyms (*hosting -> running*) or adds context words (*#cybersecurity*). While in the second example, one can see that the model is able to paraphrase parts of the original instance (*Another #ransomware operation known as 'Black Kingdom' is exploiting the [...] -> Black Kingdom ransomware is exploiting the [...]*), in the third example the entire instance is paraphrased (*Just as predicted, the Microsoft Exchange exploit chain #ProxyLogon now confirmed being used to install ransomware -> The ProxyLogon vulnerability in Microsoft Exchange Server is being actively exploited in the wild to install ransomware*). The fourth shown instance is an example of the method stripping away parts, while still preserving the label (*Microsoft Exchange Server Remote Code Execution CVE-2021-26855 Exploit.*). For some generated instances, like the fifth example, we were not able to find similar instances. The instances might be entirely new based on the interpolation of the given instances and the knowledge of the underlying model.

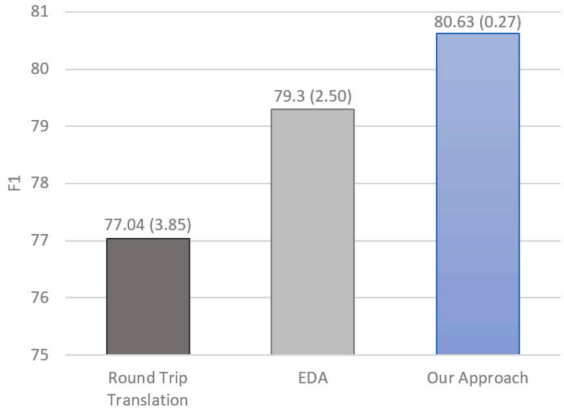

Fig. 4. Evaluation results of the data augmentation experiment. Showing the mean F1 results of 5 runs and the standard deviation in brackets.

Regarding the irrelevant class, we see that many generated instances are duplicates of the training instances, differing at most by very small changes, such as removing the hashtag in the first example (*#Microsoft-Exchange Server Attack -> Microsoft Exchange Server Attack*) or swapping the position of words in the second example (*#Technology #TechNews Microsoft [...] Authority #Cybersecurity #AiUpNow #techy -> #Technology #Cybersecurity Microsoft [...] Authority #AiUp-Now #tech*). While the third example, again, shows an instance where the content is paraphrased, the last two generated texts have no clear counterpart.

We also quantitatively evaluated our data augmentation strategy by evaluating the entire pipeline using our data augmentation method compared to two other popular methods in the field. One of these augmentation techniques was proposed by Wei and Zou (2019), called Easy Data Augmentation (EDA), and consists of several text transformations: Replacing a word with a synonym or randomly inserting a synonym as well as randomly swapping and deleting words. The other data augmentation technique is Round-trip translation (often referred to as Backtranslation (Sennrich et al., 2016)), as for example in (Fabbri et al., 2021), where the instance to be transformed is first translated into another language (German in our case) and then back into the original language.

The results of this experiment can be found in Fig. 4. Here we can see that the data augmentation proposed in this work is clearly superior for our task, achieving +1.33 and +3.59 F1 points over EDA and backtranslation, respectively.

4.3.2. Main experiments

In our main experiments, we test the whole pipeline proposed in Section 3. As a quick reminder, our method includes the multi-level fine-tuning with bert-base-uncased on cybersecurity data, the CySecAlert dataset and the actual few-shot learning task with 32 instances, as well as the GPT-3-based data augmentation technique and ADAPET for few-shot training. For a sensible comparison, we first follow the standard training procedure by fine-tuning a bert-base-uncased model with a classifier head on the few-shot training instances (baseline). Furthermore, we test a bert-base-uncased model with the ADAPET method, as it can be regarded as the state-of-the-art method for performing few-shot learning. We also perform a best case evaluation in which we train the bert-base-uncased model on the full training dataset (1800 instances) to see how a classifier would perform with enough data. A more detailed

Table 3
Detailed evaluation results of the main experiments. The values on the left show the minimum, in the middle the mean, in brackets the standard deviation, and on the right the maximum value.

Name	Model	Accuracy	F1
Best Case	BERT	84.69/ 85.36(0.07) /86.02	84.87/ 85.35(0.47) /85.81
Baseline	BERT	46.26/ 49.65(1.90) /50.58	25.06/ 58.70(18.81) /67.18
ADAPET	BERT	64.89/ 65.89(1.35) /68.05	59.30/ 62.54(4.32) /69.81
Our Approach	CySecBERT	78.54/ 79.13(0.56) /80.03	80.42/ 80.63(0.27) /81.07

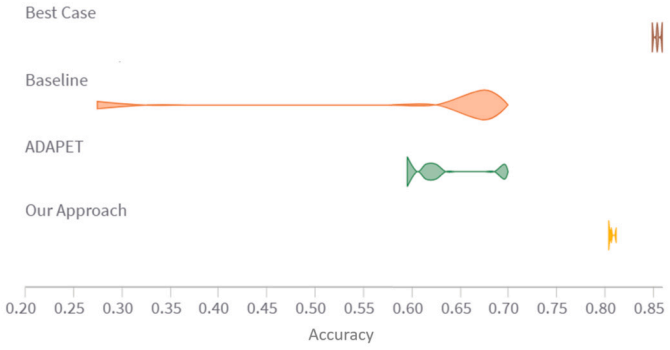

Fig. 5. Violin plots showing the accuracy differences of the main experimentation setting.

analysis of the approach itself can be found in the ablation studies in Section 4.3.3.

The results of the pipeline experiments are shown in Table 3. It is observable that the baseline is not able to learn any meaningful classification strategy with the low dataset, reaching an accuracy of about 50% and F1 score of 58.70%. ADAPET reaches a significantly higher accuracy with an additive improvement of about 15 points in accuracy and a F1 score of 62.57%. This is, nevertheless, far from a good classification quality as the best case classifier reaches a F1 score of 85.35%. With an F1 score of 80.63%, our approach proposed in this paper could even almost keep up with the best case classifier. Particularly noteworthy at this point is that the best case classifier is trained with 1800 instances, while our approach only has access to 32 instances. Furthermore, our approach improves the current state of the art with 18.09 points in F1. A look at the violin plots in Fig. 5 shows that both the best case and our approach have a very good standard deviation, which means that both are robust to random changes.

The evaluation results show that our approach is able to identify cyber threat information from which we can deduce that a new classifier can be trained for upcoming cybersecurity incidents with limited data.

4.3.3. Ablation studies

Finally, we want to give a more detailed insight into our method by showing how each component contributes to the resulting score. For this purpose, we conducted three further experiments in which we omitted one component in each case and evaluated the other two components. When multi-level fine-tuning is not used, we evaluate the BERT base model with the auxiliary data of the augmentation method and ADAPET for the learning objective. Without ADAPET, we train the cybersecurity pre-trained model on the CySecAlert dataset and the final task (with augmented data) with a classifier head. In the last experiment, the augmented data are simply omitted, while training the model in the multi-level fine-tuning process with ADAPET.

Upon examination of the results, presented in Table 4, it becomes clear that leaving out a component worsens the overall results. The highest loss is reached when the multi-level fine-tuning component is left out, showing how important it is. This behaviour could be due to the many specific cybersecurity words trained by the general language modelling of cybersecurity data (CyBERT) and to fine-tuning by a very related task that already gives the model an idea of how to distinguish between relevant and irrelevant content. Furthermore, we can clearly observe that leaving out ADAPET greatly worsens the results. When compared with the results of the main evaluation presented in Table 3, ADAPET even improves the values significantly more than compared to the baseline. This shows that ADAPET needs a strong base model to be highly beneficial. The smallest improvement is made with the augmented data. Although the data appeared to be of high quality (see Section 4.3.1), it did not significantly improve the classifier. Nevertheless, a small increase can be reached and the classifier training got more robust through the additional training data (smallest standard deviation).

Table 4
Detailed evaluation results of the ablation experiments. The values on the left show the minimum, in the middle the mean, in brackets the standard deviation, and on the right the maximum value.

Name	F1
Our Approach	80.42/ 80.63(0.27) /81.07
→ w/o Augmentation	78.48/ 80.33(1.27) /81.49
→ w/o Multi-Level Fine-Tuning	63.95/ 66.16(1.67) /67.43
→ w/o ADAPET	65.33/ 71.33(3.62) /75.08

5. Conclusion and discussion

CTI, the collection of evidence-based knowledge of cybersecurity threats, is highly relevant for identifying and remediating security incidents. Professionals, security providers, CERTs, as well as many others in the cybersecurity realm can gain important information about the incidents, such as how severe they may be, which software and systems are affected, how to be protected, and if exploits exist. The challenges lie in the information overload and the high dynamics associated with every new threat event. To counteract the flood of information or to collect certain types of information, it is necessary to train a classifier. However, a trained classifier cannot generalise to new vulnerability events due to high dynamics (new names of vulnerabilities, paths, etc.) and new requirements (focus on exploitation, mitigation, consequences, etc.). To the best of our knowledge, this is the first work to address this problem by proposing a framework that enables rapid training of new, high-performance classifiers for specialised CTI. It consists of several components that allow the end user to label only a few data instances (tested here with 32 instances) to obtain a classifier that is comparable to one trained with 1800 instances. We also constructed a dataset labelled by three cybersecurity experts showing that this method indeed overcomes the problem of information overload and addresses high dynamics by being easily adaptable to new incidents.

5.1. Practical, theoretical, and empirical contributions

Considering our findings, the study revealed (P) practical, (T) theoretical, and (E) empirical contributions:

(P) A novel pipeline for detecting specialized cyber threat information. Our work provides an approach to detecting specific cyber threat information by addressing the problem that a trained classifier

Table 5
Generated data instances and their most similar original counterparts. The instances created are displayed first and the most similar ones second. URLs are removed from the text.

Relevant	RT If you're running Microsoft Exchange Server on premises, you need to take these urgent security steps now. The zero-day exploits may have already caused a breach of your data. #infosec #cybersecurity #HAFNIUM http://..
	If you are hosting #MicrosoftExchange on premises you need to take these urgent security steps right now. The zero-day exploits may have already caused a breach of your data. #infosec #HAFNIUM http://..
	RT Black Kingdom ransomware is exploiting the Microsoft Exchange Server ProxyLogon vulnerabilities to encrypt servers. http://..
	Please take Information Security seriously. #CyberAttack can bring your reputation down. Another #ransomware operation known as 'Black Kingdom' is exploiting the Microsoft Exchange Server ProxyLogon vulnerabilities to #encrypt servers. http://..
	RT @SecureList: The ProxyLogon vulnerability in Microsoft Exchange Server is being actively exploited in the wild to install ransomware. http://..
	RT Just as predicted, the Microsoft Exchange exploit chain #ProxyLogon now confirmed being used to install ransomware #DEARCRY http://..
	RT RT @hackerfantastic: Microsoft Exchange Server Remote Code Execution CVE-2021-26855 Exploit. #BugBounty #RCE #infosec http://..
	RT Thousands of US companies have been hacked by Chinese hackers using This RCE. Microsoft Exchange Server Remote Code Execution CVE-2021-26855 Exploit. #BugBounty #RCE #infosec http://..
	RT @ryan_a_h: Microsoft just released their quarterly updates which include a patch for the Exchange zero-day. You can find more information here: http://..
	If you are hosting #MicrosoftExchange on premises you need to take these urgent security steps right now. The zero-day exploits may have already caused a breach of your data. #infosec #HAFNIUM http://..
Not Relevant	Microsoft Exchange Server Attack Escalation Prompts #Patching Panic #cybersecurity #vulnerabilities http://..
	#MicrosoftExchange Server Attack Escalation Prompts #Patching Panic #cybersecurity #vulnerabilities http://..
	#Technology #Cybersecurity Microsoft Exchange Hackers Also Breached European Banking Authority #AiUpNow #techy http://..
	#Technology #TechNews Microsoft Exchange Hackers Also Breached European Banking Authority #Cybersecurity #AiUpNow #techy via http://..
	RT Microsoft Exchange Server has been hacked – here's what you need to know http://..
	RT Here's what we know so far about the massive Microsoft Exchange hack http://..
	Microsoft Exchange Server Flaws Expose Millions of Emails to Attack http://..
	RT Here's what we know so far about the massive Microsoft Exchange hack http://..
	Protected: Microsoft Exchange Server Attacks Escalate to Government, Healthcare and Financial Institutions http://..
	The Microsoft Exchange hacks: How they started and where we are http://..

does not generalise well to new cyber security events and is able to adapt to different requirements. Furthermore, it is aligned with the circumstances of such events. These circumstances include that information has to be gathered fast in the early stages of the events and that security institutions and experts do not have the time and capacity to label many instances. Therefore, we combine few-shot learning with multi-level fine-tuning and data augmentation to produce classifiers that only need few instances to perform with high quality. For few-shot learning we utilize ADAPET by Tam et al. (2021) combined with the multi-level fine-tuning process. For data augmentation we use GPT-3 to create instances with novel linguistic patterns. Our pipeline reaches a F1-score of 80.63 on a specialized cyber threat dataset, which is 21.93 points above the score of a classical learning scheme. Other works, such as the cyber threat event detection systems of Riebe et al. (2021b) or Le Sceller et al. (2017), allow for coarse-grained information gathering. To the best of our knowledge, our system is the first to provide rapid detection of specialized cyber threat information, by needing only very few data instances to create high-quality classifiers. This way, in the event of a current cybersecurity incident, experts can quickly create a classifier tailored to their specific needs and gather important information.

Moreover, our general approach to learning with very few examples can also be used for these detection systems or other cybersecurity problems, such as in IoC extraction. This leads over to the theoretical contributions of our work.

(T) New few-shot learning technique based on multi-level fine-tuning. We propose a novel few-shot learning approach for creating classifiers of high quality with a smaller amount of training data. The idea behind this approach is to fine-tune a machine learning model in several levels where enough data is available (see Fig. 3). In our study we first further trained a BERT model on a general cybersecurity corpus. This model was then trained on a general Twitter cybersecurity relevance dataset. From this point, the model has a fundamental understanding of cybersecurity texts and is also able to distinguish cybersecurity-related content from irrelevant content. With this pre-trained knowledge, the model only needs few data instances to be able to differentiate specific cybersecurity content. As shown in this study, this new technique can also be combined with other techniques like ADAPET or data augmentation to further reduce the amount of needed training data. However, we show that this multi-stage fine-tuning approach has the greatest impact on classification quality of all techniques (+14.47 F1, see Table 4). The multi-level fine-tuning approach significantly advances research in few-shot learning, as it allows for a much higher model quality and at the same time can be combined with previous few-shot studies, such as ADAPET (Tam et al., 2021), DART (Zhang et al., 2022), or PERFECT (Mahabadi et al., 2022).

(T) New insights on data augmentation with large pre-trained language models. In our study, we also implemented a data augmentation technique that combines the works of Yoo et al. (2021) and Bayer et al. (2021). As in the former, we used the large language model GPT-3 with a prompting strategy and filtered the generated instances with a human-in-the-loop technique, as in the latter. The idea is that GPT-3 can create instances with a very high degree of novelty, resulting in some very valuable instances. However, this novelty comes with the problem of poor label preservation, as the instances may be too far away from the class. For this reason, we also introduced this filtering strategy where the original labelled data of a class is compared with the generated data and those that are too far away from the original data are discarded. The boundary is determined by an expert who examines those instances close to a predefined boundary. As shown in Section 4.3.1 and Table 5, this procedure generates instances with very different transformation patterns, including word substitution, paraphrasing, and partial

removal. It even leads to instances that are entirely novel. Furthermore, we included a quantitative evaluation in Section 4.3.1 comparing the data augmentation strategy against two of the most common NLP data augmentation strategies. It shows that our method is clearly superior for the task of specialized CTI in our pipeline.

However, in Section 4.3.3, we showed that omitting this method from the overall pipeline only slightly reduces the resulting score. This means that the model learns very little from the augmented data when multi-level fine-tuning and ADAPET are already used. Nevertheless, the evaluation results show a reduction in the standard deviation, which shows that the model has become more robust with the artificial data.

(E) A specialized CTI dataset for further research purposes. In this study we created a CTI dataset based on the 2021 Microsoft Exchange Server data breach. The dataset was constructed by three experts. The guidelines have been revised several times in an attempt to flesh out the concept of cyber threat analysis as much as possible. Along with the code and the dataset, the guidelines are available in the repository. All annotators reached a good intercoder reliability showing that the guidelines and the general annotation process was successful. Further research can benefit from this dataset as it is, to our knowledge, the first to contain a relevance coding regarding CTI in Twitter in relation to a specific cybersecurity event.

5.2. Limitations and outlook

In this work, we focus on Twitter as a data source, as it can be very up-to-date and rich, as the cybersecurity community is very active when it comes to vulnerabilities in Twitter. However, we would like to point out that using Twitter as a data source also brings some disadvantages. On Twitter, for example, anyone can share information, which can lead to a lot of speculative or even false information. A system that relies only on Twitter may not be as reliable and could easily be fooled. This is why we look forward to studies making the proposed approach more robust. In addition, we believe that other data sources should also be examined for vulnerabilities. We plan to integrate the component implemented in this work into an already developed, manageable dashboard that additionally includes a credibility component and aggregates information from many different sources.

In terms of the overall concept, we look forward to research studies testing the performance of this approach in other domains and on further cyber threat events. For example, it would be interesting to see if the same improvements can be achieved in medical or crisis domains, where data is also scarce. On a smaller scale, we also look forward to work applying our methodology to other cybersecurity events. Our pipeline was only evaluated with the MS Exchange data breach, but can be generalized to other CTI-related incidents as this was a priority in our development process. Moreover, our experiments are limited to the BERT base model. It would be interesting to see if the improvements are as high when a larger model like RoBERTa (Liu et al., 2019) is used. Likewise, one could also test other language models for the data augmentation technique. Especially interesting would be to test if open source models, like GPT-NeoX-20B (Black et al., 2022), reach a good augmentation performance.

A part of our experiments was to fine-tune the model on the CySecAlert dataset of Riebe et al. (2021b). The authors of this work propose an active learning component to achieve high classification scores with less data. With a view to future research, it might be sensible to also include active learning into the concept of our approach to further increase the classification quality. In practice, our approach would in the worst case lead to users labelling very similar examples, resulting in poor execution of data augmentation and poor classification quality, which can happen quickly when labelling such a small amount of data. Therefore, an active learning system could help to collect very different examples. Otherwise, experts can also be trained to label diverse examples.

CRediT authorship contribution statement

Markus Bayer: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft. **Tobias Frey:** Data curation, Investigation, Software, Writing – review & editing. **Christian Reuter:** Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code are available in GitHub repositories (linked in the paper).

Acknowledgements

This work has been co-funded by the German Federal Ministry of Education and Research (BMBF) in the project CYWARN (13N15407) and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1119 (CROSSING) – 236615297, as well as the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Calculations for this research were conducted on the Lichtenberg high performance computer of the TU Darmstadt.

Appendix A. Codebook

The iteratively developed codebook is shown below:

A.1. Annotation guidelines

Put yourself in the role of a cybersecurity expert who has received initial information about a Microsoft Exchange incident. Since the information is relatively new, you want to gain further insight by looking at some Twitter posts. You search Twitter with some keywords regarding Microsoft Exchange. As you notice that many posts are not insightful you start to annotate the posts in terms of their relevance to you.

A.1.1. Do not only consider the text but also the referenced websites

A.1.1.1. What is labelled as ***Relevant***?

- Timely information that might be good to know for cybersecurity experts
- Specific information about the vulnerabilities, how to perform them, how to mitigate them, what an attacker can do with exploiting the vulnerability, etc.
- Explicitly mentioned numbers, methods, proof of concepts, code, fixes
- Tweets that include IoCs (IPs, Hash values, usw.), or exploits (DearCry)
- Relevant information is more valuable than irrelevant information: If a post contains both relevant and irrelevant information, it is to be labelled as relevant
  - For example “As of March 8, over 30K servers in the US have been hit by the recent Exchange #zero-day attack, which leaves behind a web shell that allows hackers to access the server to steal data & install malware” → Relevant

A.1.1.2. What is labelled as **Not Relevant**?

- Information that is too general
- Information regarding the scope (which authorities, business branches and how many)
  - This then includes posts about how many servers have not been patched yet, for example, or the following: "Microsoft Exchange Server attacks: 'They're being hacked faster than we can count', says security company"
  - Likewise this refers to posts like "The European Banking Authority said it had been a victim of a cyberattack targeting its Microsoft Exchange Servers" → Not relevant
  - But on the other hand, a post like "RT Researchers have acquired a list of 86,000 IP addresses of MS Exchange servers infected worldwide by the mass compromises" is still relevant, as the IP list would be of interest.
- If a post only links to a page without really providing information, it should be marked as not relevant, even if the linked page contains important information. Above all, a post should be marked as not relevant if it is not exactly clear what to expect on the page. Otherwise, when, for example, the text says that it is IoC info or a security advisory, you can consider it relevant (or check the page again). Please still consider the following examples:
  - "GovInfoSecurity | Analysis: Microsoft Exchange Server Hacks https://bit.ly/2QgZb9P" → not relevant (too inaccurate)
  - "Chile's bank regulator shares IoCs after Microsoft Exchange hack https://ift.tt/3lrnfm3" → Relevant (IoCs are very interesting for security experts)
  - "Here's what we know so far about the massive Microsoft Exchange hack https://www.wxii12.com/article/here-s-what-we-know-so-far-about-the-massive-microsoft-exchange-hack/35793771?utm_campaign=snd-autopilot" → not Relevant (really no information in the text; too vague what to expect on the page)
  - "RT How the Microsoft Exchange hack could impact your organization https://tek.io/2Oieqi5 (https://t.co/un4YdQkbA3)" → Not Relevant (too inaccurate)
  - "CISA Updates Microsoft Exchange Advisory to Include China Chopper https://dlvr.it/RvhdjL" → Relevant (You know what to expect on the site; besides, CISA is a major player)
  - "At Least 10 Hacking Groups Are Exploiting Microsoft Exchange Server Flaws [PCMag] https://best.photography/articles/543893/at-least-10-hacking-groups-are-exploiting-microsoft-exchange-server-flaws/" → Not Relevant (sounds relevant at first, because possibly the 10 hacking groups are mentioned and they are possibly known. However, the link is not trustworthy and also not callable) – relevant would have been fine here too, especially if the link was not strange.
  - "It really was only a matter of time https://www.bleepingcomputer.com/news/security/new-dearcry-ransomware-is-targeting-microsoft-exchange-servers/" → Not Relevant (looks like an exciting link, but in the post there is no information about the link - you don't know what to expect) – relevant would also have been fine
- Information that seems to be spam
- Information that is highly politically motivated
- Information for the general public (non-experts)
- Information that is speculative
- "Casual news" for the general public
- General cybersecurity advises (for the general public)
- Podcasts, Interviews or personal opinions are to be marked as not relevant
- Smaller service companies that report that their systems are updated and safe
- News aggregations with several other news are not relevant
- "RT Read this week's digest to find out the latest updates in the #Microsoft Exchange vulnerabilities as well as how #hackers were able to breach 150,000 surveillance cameras from inside hospitals, jails and Tesla."

When in doubt → Not Relevant

References

Abu, M.S., Selamat, S.R., Ariffin, A., Yusof, R., 2018. Cyber threat intelligence–issue and challenges. Indones. J. Electr. Eng. Comput. Sci. 10 (1), 371–379.

Alves, F., Andongabo, A., Gashi, I., Ferreira, P.M., Bessani, A., 2020. Follow the blue bird: a study on threat data published on Twitter. In: Chen, L., Li, N., Liang, K., Schneider, S. (Eds.), Computer Security – ESORICS 2020. Springer International Publishing, Cham, pp. 217–236.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., Zwerdling, N., 2020. Do not have enough data? Deep learning to the rescue! In: Proceedings of the AAAI. arXiv:1911.03118.

Bayer, M., Kaufhold, M.A., Buchhold, B., Keller, M., Dallmeyer, J., Reuter, C., 2021. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. Int. J. Mach. Learn. Cybern. https://doi.org/10.1007/s13042-022-01553-3. arXiv:2103.14453.

Bayer, M., Kaufhold, M.A., Reuter, C., 2022. A survey on data augmentation for text classification. ACM Comput. Surv. https://doi.org/10.1145/3544558. 3544558.

Belinkov, Y., Bisk, Y., 2018. Synthetic and natural noise both break neural machine translation. In: Proceedings of ICLR.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: a pretrained language model for scientific text. arXiv:1903.10676.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U.S., Purohit, S., Reynolds, L., Tow, J., Wang, B., Weinbach, S., 2022. GPT-NeoX-20B: an open-source autoregressive language model. arXiv:2204.06745.

Bragg, J., Cohan, A., Lo, K., Beltagy, I., 2021. FLEX: unifying evaluation for few-shot NLP. arXiv:2107.07170, p. 14.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: NeurIPS. arXiv:2005.14165.

Caballero, J., Gomez, G., Matic, S., Sánchez, G., Sebastián, S., Villacañas, A., 2023. The rise of GoodFATR: a novel accuracy comparison methodology for indicator extraction tools. Future Gener. Comput. Syst. 144, 74–89. https://doi.org/10.1016/j.future.2023.02.012. arXiv:2208.00042.

Chatterjee, S., Thekdi, S., 2020. An iterative learning and inference approach to managing dynamic cyber vulnerabilities of complex systems. Reliab. Eng. Syst. Saf. 193, 106664. https://doi.org/10.1016/j.ress.2019.106664.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding (Mlm). arXiv:1810.04805.

Dionísio, N., Alves, F., Ferreira, P.M., Bessani, A., 2020. Towards end-to-end cyberthreat detection from Twitter using multi-task learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). ISSN 2161-4407, pp. 1–8.

Fabbri, A.R., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D., Mehdad, Y., 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. arXiv:2010.12836.

Fang, Y., Gao, J., Liu, Z., Huang, C., 2020. Detecting cyber threat event from Twitter using IDCNN and BiLSTM. Appl. Sci. 10 (17), 5922. https://doi.org/10.3390/app10175922. https://www.mdpi.com/2076-3417/10/17/5922.

Gao, T., Fisch, A., Chen, D., 2021. Making pre-trained language models better few-shot learners. arXiv:2012.15723.

Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., Niu, X., 2017. TTPDrill: automatic and accurate extraction of threat actions from unstructured text of CTI sources. In: Proceedings of the 33rd Annual Computer Security Applications Conference. ACM, Orlando, FL, USA, pp. 103–115.

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Zhao, T., 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 2177–2190. Online. https://www.aclweb.org/anthology/2020.acl-main.197.

Kaufhold, M.A., Basyurt, A.S., Eyilmez, K., Ag, V., Stöttinger, M., Reuter, C., Sercan, A., 2022. Cyber threat observatory: design and evaluation of an interactive dashboard for computer emergency response teams. In: ECIS 2022, p. 18.

Kuehn, P., Riebe, T., Apelt, L., Jansen, M., Reuter, C., 2020. Sharing of cyber threat intelligence between states. Sicherh. Frieden 38 (1), 22–28. https://doi.org/10.5771/0175-274X-2020-1-22.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., 2020. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv:1909.11942.

Le Sceller, Q., Karbab, E.B., Debbabi, M., Iqbal, F., 2017. Sonar: automatic detection of cyber security events over the Twitter stream. In: Proceedings of the 12th International Conference on Availability, Reliability and Security. ARES'17. Association for Computing Machinery, New York, NY, USA.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz682. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz682/5566506.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., Allen, P.G., 2019. RoBERTa: a Robustly Optimized BERT Pretraining Approach. Tech. Rep. https://github.com/pytorch/fairseq.

Longpre, S., Wang, Y., DuBois, C., 2020. How effective is task-agnostic data augmentation for pretrained transformers? In: Findings of EMNLP.

Mahabadi, R.K., Zettlemoyer, L., Henderson, J., Saeidi, M., Mathias, L., Stoyanov, V., Yazdani, M., 2022. PERFECT: prompt-free and efficient few-shot learning with language models. arXiv:2204.01172.

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É, Seddah, D., Sagot, B., 2020. CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 7203–7219. Online. https://www.aclweb.org/anthology/2020.acl-main.645.

McMillan, R., 2013. Definition: threat intelligence. https://www.gartner.com/en/documents/2487216.

Mittal, S., Das, P.K., Mulwad, V., Joshi, A., Finin, T., 2016. Cybertwitter: using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, pp. 860–867.

Mosolova, A.V., Fomin, V.V., Bondarenko, I.Y., 2018. Text augmentation for neural networks. CEUR Workshop Proc. 2268, 104–109.

Niakanlahiji, A., Safarnejad, L., Harper, R., Chu, B.T., 2019. IoCMiner: automatic extraction of indicators of compromise from Twitter. In: 2019 IEEE International Conference on Big Data (Big Data). IEEE, Los Angeles, CA, USA, pp. 4747–4754. https://ieeexplore.ieee.org/document/9006562/.

Pan, S.J., 2020. Transfer learning. Learn. 21, 1–2.

Queiroz Abonizio, H., Barbon Junior, S., 2020. Pre-trained data augmentation for text classification. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12319 LNAI. Springer Science and Business Media Deutschland GmbH, pp. 551–565. iSSN: 16113349.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. https://doi.org/10.18653/v1/d19-1410.

Riebe, T., Kaufhold, M.A., Reuter, C., 2021a. The impact of organizational structure and technology use on collaborative practices in computer emergency response teams: an empirical study. Proc. ACM Hum.-Comput. Interact. 5 (CSCW2), 1–30. https://doi.org/10.1145/3479865.

Riebe, T., Wirth, T., Bayer, M., Kühn, P., Kaufhold, M.A., Knauthe, V., Guthe, S., Reuter, C., 2021b. CySecAlert: an alert generation system for cyber security events using open source intelligence data. In: Gao, D., Li, Q., Guan, X., Liao, X. (Eds.), Information and Communications Security. In: Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 429–446.

Rodriguez, A., Okamura, K., 2019. Generating real time cyber situational awareness information through social media data mining. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), vol. 2. IEEE, pp. 502–507.

Sabottke, C., Suciu, O., Dumitras, T., 2015. Vulnerability disclosure in the age of social media: exploiting Twitter for predicting real-world exploits. In: 24th USENIX Security Symposium (USENIX Security 15). USENIX Association, Washington, D.C., pp. 1041–1056. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/sabottke.

Schick, T., Schütze, H., 2021. Exploiting cloze questions for few shot text classification and natural language inference. arXiv:2001.07676.

Sennrich, R., Haddow, B., Birch, A., 2016. Improving neural machine translation models with monolingual data. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers.

Sun, L., Xia, C., Yin, W., Liang, T., Yu, P.S., He, L., 2020. Mixup-transfomer: dynamic data augmentation for NLP tasks. iSSN: 23318422. arXiv:2010.02394.

Tam, D., Menon, R.R., Bansal, M., Srivastava, S., Raffel, C., 2021. Improving and simplifying pattern exploiting training. arXiv:2103.11955.

Taylor, W.L., 1953. "Cloze procedure": a new tool for measuring readability. Journal. Quart. 30 (4), 415–433. https://doi.org/10.1177/107769905303000401.

Torrey, L., Shavlik, J., 2010. Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global, pp. 242–264.

Tounsi, W., Rais, H., 2018. A survey on technical threat intelligence in the age of sophisticated cyber attacks. Comput. Secur. 72, 212–233. https://doi.org/10.1016/j.cose.2017.09.001.

Wagner, T.D., Mahbub, K., Palomar, E., Abdallah, A.E., 2019. Cyber threat intelligence sharing: survey and research directions. Comput. Secur. 87, 101589. https://doi.org/10.1016/j.cose.2019.101589.

Wei, J., Zou, K., 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Yoo, K.M., Park, D., Kang, J., Lee, S.W., Park, W., 2021. GPT3Mix: leveraging large-scale language models for text augmentation. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 2225–2239.

Zhang, N., Li, L., Chen, X., Deng, S., Bi, Z., Tan, C., Huang, F., Chen, H., 2022. Differentiable prompt makes pre-trained language models better few-shot learners. arXiv:2108.13161.

Markus Bayer, M.Sc. is a research assistant and doctoral student at the chair of Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science at the Technical University of Darmstadt. He applies his expertise in the CYWARN project to address challenges of cyber emergency response teams with deep learning. As an overarching goal, he seeks to address highly relevant real-world problems, such as Explainable AI and Deep Learning in low-data regimes, through focused and theoretically sound research.

Tobias Frey is a student assistant at the chair of Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science at the Technical University of Darmstadt. In addition to his experience in cyber security, he specialises in natural text processing and analysis, as well as Deep Learning.

Prof. Dr. Dr. Christian Reuter is Full Professor at Technical University of Darmstadt. His chair Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science combines computer science with peace and security research. He holds a Diplom, M.Sc. and Ph.D. in Information Systems. On the intersection of the disciplines (A) Cyber Security and Privacy, (B) Peace and Conflict Studies as well as (C) Human-Computer Interaction, he and his team specifically address (1) Peace Informatics and technical Peace Research, (2) Crisis Informatics and Information Warfare as well as (3) Usable Safety, Security and Privacy.